ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

PROPUESTA METODOLÓGICA DE ANALÍTICA DE DATOS PARA ESTUDIO Y ANÁLISIS DE TRÁFICO EN REDES DE TELECOMUNICACIONES

TESIS DE GRADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN CONECTIVIDAD Y REDES DE TELECOMUNICACIONES

MARIO ANDRÉS ORBE ORDÓÑEZ
mario.orbe@hotmail.com

DIRECTOR: LUIS FELIPE URQUIZA
luis.urquiza@epn.edu.ec

CODIRECTOR: MARTHA CECILIA PAREDES
cecilia.paredes@epn.edu.ec

Quito, octubre 2017

AVAL

Certificamos que el presente trabajo fue desarrollado por Mario Andrés Orbe Ordóñez, bajo nuestra supervisión.

LUIS FELIPE URQUIZA
DIRECTOR DEL TRABAJO DE TITULACIÓN

MARTHA CECILIA PAREDES
CODIRECTOR DEL TRABAJO DE TITULACIÓN

DECLARACIÓN DE AUTORÍA

Yo, Mario Andrés Orbe Ordóñez, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

MARIO ANDRÉS ORBE ORDÓÑEZ

DEDICATORIA

A mi madre Marina, a quien admiro y es mi mejor ejemplo a seguir.

A mi sobrino Lucas, quien trajo hace poco a este mundo un espíritu alegre y gentil.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El presente trabajo de titulación realiza una propuesta metodológica de analítica de datos para el estudio y análisis de tráfico en redes de telecomunicaciones. Dicha propuesta resulta necesaria para el sector de telecomunicaciones puesto que, si bien existen diferentes documentos de referencia de carácter general que sirven de guía para realizar proyectos de analítica de datos, ninguno de ellos ha sido formulado como una metodología que considera las características, retos, desafíos y ventajas que las empresas de telecomunicaciones deben contemplar para llevar a cabo este tipo de proyectos. La propuesta metodológica realizada resulta también oportuna dados los cambios profundos que está atravesando la sociedad y la industria debido a una transformación económica que está siendo impulsada por lo que se conoce como la cuarta revolución industrial.

Este trabajo realiza una exposición de las circunstancias que motivan la elaboración de la metodología propuesta para posteriormente describir las condiciones que moldean su diseño y características, comparar documentos de referencia generalmente utilizados en proyectos de analítica de datos, presentar el proceso que permitió estructurar y desarrollar la propuesta metodológica, y mostrar su aplicación práctica en un caso de estudio.

La propuesta metodológica resultante se ha consolidado íntegramente en el Anexo 2 de este documento con el fin de facilitar su revisión y aplicación.

PALABRAS CLAVE: ANALÍTICA, METODOLOGÍA, TELECOMUNICACIONES, TRÁFICO, CRISP-DM, ASUM-DM

ABSTRACT

The present work makes a proposal of a data analytics methodology for the study and analysis of traffic in telecommunications networks. This proposal is necessary for the telecommunications sector since, although there are different general reference documents that serve as a guide for carrying out data analytics projects, none of them has been formulated as a methodology that considers the characteristics, challenges and advantages that telecommunications companies must contemplate to carry out this type of projects. The methodological proposal made is also timely given the profound changes that are going through society and industry due to an economic transformation that is being driven by what is known as the fourth industrial revolution.

This paper sets out the circumstances that motivate the elaboration of the proposed methodology to later describe the conditions that shape its design and characteristics, compare reference documents generally used in data analysis projects, expose the process that allowed the structuring and development of this methodology, and show its practical application in a case study.

The resulting methodology has been fully consolidated in Annex 2 of this document to facilitate its revision and implementation.

KEYWORDS: ANALYTICS, METHODOLOGY, TELECOMMUNICATIONS, TRAFFIC, CRISP-DM, ASUM-DM

1. INTRODUCCIÓN

La tercera revolución industrial, ligada a la revolución digital, ha permitido el desarrollo de lo que se conoce como la sociedad de la información. Esto es una sociedad cuya economía está fuertemente apalancada en actividades que están relacionadas con información, como por ejemplo investigación, educación, comunicación y computación [1]. La tercera revolución industrial inició en los años 60 con la mejora de los transistores y posteriormente se concretó gracias al perfeccionamiento de las computadoras y del Internet. A pesar de que esta tercera revolución industrial aún está en curso, el Foro Económico Mundial ya está hablando de una cuarta revolución industrial que es impulsada por tecnologías que están en auge y entre las cuales se mencionan a "Big Data" y a la "Analítica de Datos" ("*Data Analytics*") [2]. Esta nueva revolución permitirá a la humanidad evolucionar a una sociedad del conocimiento que no solo utilice información para dinamizar su economía, sino que sea capaz de extraer conocimiento de dicha información para utilizarlo como eje de sus actividades económicas [3]. Este proceso de transformación ha sido analizado por la Unión Europea bajo el concepto de "Industria 4.0" (*Industry 4.0 - Digitalization for productivity and growth*) [4] o por la CEPAL (Comisión Económica para América Latina y el Caribe) bajo el concepto de "Internet de la producción" [5].

La cuarta revolución industrial, como las anteriores, traerá consigo cambios profundos que crearán oportunidades y amenazas. Aquellas instituciones, empresas y profesiones que logren adaptarse a estos cambios podrán afrontar el futuro con éxito, mientras que aquellas que no logren adaptarse estarán condenadas a desaparecer o a perder su posición relevante en el mercado. En uno de estos dos casos se encontrarán las empresas de telecomunicaciones, que gracias a la tercera revolución industrial se volvieron empresas de gran importancia en la economía moderna, pero que corren el riesgo de perder liderazgo debido a que sus servicios de conectividad serían percibidos a futuro como un bien primario, necesario pero carente de valor agregado [6]. El liderazgo perdido por las empresas tradicionales de telecomunicaciones sería arrebatado por empresas tecnológicas de alto valor agregado, que están mejor adaptadas para aprovechar la cuarta revolución industrial y entre las cuales se puede mencionar a empresas OTT (*Over The Top*) como Netflix, Spotify, Uber, AirBnB o Google.

Las empresas de telecomunicaciones, y en especial las operadoras celulares, han decidido adaptar y transformar sus modelos de negocio para que se adecuen a los tiempos venideros con el fin de mantenerse como empresas líderes en nuestras

economías. La nueva visión de estas empresas incluye la explotación de los datos que cursan por sus redes aprovechando las bondades de Big Data y de Analítica de Datos [7]. Esta intención y esfuerzo están acompañados por el desarrollo de conferencias, estándares, mejores prácticas, guías y metodologías por parte de instituciones como GSMA (*Global System for Mobile communications Association*) [8], IBM (*International Business Machines*) [9], ITU (*International Telecommunication Union*) [10] y TM-Forum (*TeleManagement Forum*) [11].

ITU, por ejemplo, en diciembre 2015, publicó la recomendación Y.3600 "Big data - Requisitos y capacidades basados en la computación en la nube" [12] donde provee una definición para "Big Data" de la siguiente manera: "*Un paradigma para permitir la recolección, almacenamiento, gestión, análisis y visualización, potencialmente bajo restricciones de tiempo real, de conjuntos de datos extensos con características heterogéneas*". El análisis de datos al que hace mención ITU en la citada definición se realiza mediante lo que se conoce como "Analítica de datos", lo cual según el diccionario Oxford es "*el análisis computacional sistemático de datos o estadísticas*".

Posteriormente, en julio de 2016, ITU publicó el suplemento 40 de la Serie Y de recomendaciones, denominado "*Big data standardization roadmap*" [10], donde provee la hoja de ruta de estandarización para el área Big Data del sector de las telecomunicaciones. Este documento expone, entre otras cosas, el trabajo de estandarización relacionado con Big Data que están realizando varias organizaciones internacionales como por ejemplo ITU, ISO/IEC (*International Organization for Standardization / International Electrotechnical Commission*), OASIS (*Organization for the Advancement of Structured Information Standards*) y W3C (*World Wide Web Consortium*). En dicho suplemento se menciona a la analítica de datos motivada por redes ("*network-driven data analytics*") como un área de estandarización de Big Data y refiere a las mejores prácticas para Big Data provistas por TM-Forum en su guía GB979, denominada "*Guide book for big data analytics*" [11]. En la versión 2.0.2 de este último documento se indica que "el valor de Big Data reside en los resultados del análisis [de los datos] y en las acciones y predicciones que derivan de dichos resultados".

De esta forma, Big Data hace referencia a la tecnología que permite realizar una analítica de grandes volúmenes de datos, y la analítica corresponde al proceso de análisis computacional sistemático realizado a dichos datos con el fin de generar valor.

De acuerdo al TM-Forum, el documento GB979 es una guía para proveedores de servicios de comunicaciones que proporciona los componentes principales requeridos para implementar casos de uso de analítica Big Data en la vida real. No obstante, existen

metodologías maduras, difundidas y detalladas que también pueden guiar proyectos de analítica de datos en el sector de las telecomunicaciones y que pueden servir como una mejor guía que el documento GB979, debido a su mayor madurez, versatilidad y aplicación en multitud de entornos. Ejemplos de estas metodologías son CRISP-DM (*Cross Industry Standard Process for Data Mining*) [13] y ASUM-DM de IBM (*Analytics Solutions Unified Method for Data Mining*) [14], las cuales, si bien no fueron diseñadas exclusivamente para el sector de telecomunicaciones, tienen un ámbito de aplicación general, por lo que se puede encontrar su uso en casos que van desde la prevención de fraudes bancarios [15] hasta la predicción de "*churn*" (cantidad de clientes que en un período de tiempo abandona a su proveedor de servicios) de clientes en la industria de telecomunicaciones móviles [16].

Realizar una comparación de estas metodologías con la guía de TM-Forum, así como adaptar, extender y complementar una de ellas de forma que pueda emplearse exitosamente en el sector de telecomunicaciones será un aporte valioso para aquellas empresas, instituciones y profesionales que busquen desarrollar de forma ágil y acertada proyectos de analítica de datos en dicho sector. Este tipo de trabajo es necesario para permitir una adaptación oportuna a los cambios que vive nuestra sociedad.

1.1. Pregunta de investigación

El presente trabajo de titulación busca responder a la siguiente pregunta: ¿Qué adaptaciones son necesarias en las metodologías más difundidas y maduras de analítica de datos para que puedan ser aplicadas al estudio y análisis de tráfico de redes de telecomunicaciones?

1.2. Objetivo General

Desarrollar una metodología de analítica de datos que incluya métodos, técnicas y herramientas aplicables al estudio y análisis de datos que cursen redes de telecomunicaciones, usando como base metodologías maduras existentes.

1.3. Objetivos Específicos

- Exponer la importancia y necesidad de disponer de una metodología de analítica de datos para el estudio y análisis de datos que cursen sobre las redes de telecomunicaciones.
- Determinar el estado del arte de metodologías, métodos, técnicas y herramientas aplicables al estudio y análisis de datos que cursen redes de telecomunicaciones.

- Adaptar una metodología base de forma que pueda aplicarse a proyectos de analítica de datos de redes de telecomunicaciones.
- Demostrar la practicidad, aplicabilidad y ejecutividad de la metodología propuesta mediante la aplicación a un caso real a manera de ejemplo.

1.4. Alcance

Este trabajo de titulación hará una propuesta metodológica con las siguientes características:

- Metodología de analítica de datos de carácter general pero enfocado al estudio y análisis de los datos que cursan por las redes de telecomunicaciones, independientemente del tipo de red en el que sea aplicado.
- La metodología propuesta empleará como base metodologías existentes (CRISP-DM y ASUM-DM, por ejemplo) y las adaptará realizando las reestructuraciones, modificaciones y extensiones que sean necesarias para que la propuesta metodológica se adapte a las necesidades, requerimientos y condiciones que las redes de telecomunicaciones impongan.
- Se abarcará únicamente el tema de analítica de datos, sin incurrir en el estudio de las áreas técnicas relacionadas con el ecosistema Big Data. Esta limitación en el alcance se debe a que, de acuerdo con la ITU, Big Data es un ecosistema complejo que abarca una gran variedad de áreas técnicas y que principalmente provee de herramientas de procesamiento de datos [10]. Por otro lado, “analítica de datos” hace referencia a un proceso sistemático computacional y “metodología” hace referencia a un conjunto de métodos usados en una investigación científica. Si bien la analítica de datos está en el centro del ecosistema Big Data [17], esta hace referencia a un tema muy distinto y particular que no debe ser confundido con las herramientas provistas por Big Data.

El trabajo de titulación expondrá un caso de uso de la metodología elaborada, así como una evaluación de los resultados, utilizando:

- Una muestra real de paquetes de una red de telecomunicaciones, que por razones de privacidad y seguridad de la información deberá ser de libre utilización y exposición. En Internet existen fuentes que proveen estas muestras de forma libre y segura para fines de investigación.
- Se realizará un proceso de analítica de datos utilizando la muestra de paquetes obtenida. Se definirá un objetivo que se tendrá que alcanzar utilizando la propuesta metodológica. El caso de uso no necesariamente abarcará todas las

características de la propuesta metodológica, sino aquellas que puedan evaluarse considerando el objetivo que se habrá de plantear.

- La evaluación de los resultados se realizará comparando el objetivo planteado con los resultados obtenidos y detallando el grado de sencillez o dificultad que se tuvo para alcanzar los resultados usando la propuesta metodológica.

1.5. Estructura del documento

El presente documento está organizado en 4 capítulos.

El primer capítulo corresponde a la introducción que se acaba de realizar y ofrece un panorama de la situación actual de los desafíos a nivel de negocio que enfrenta el sector de las telecomunicaciones y explica la motivación para generar una propuesta metodológica de analítica de datos para el estudio y análisis de tráfico de redes de telecomunicaciones.

El segundo capítulo corresponde al marco teórico utilizado a lo largo del desarrollo de la propuesta metodológica objeto del trabajo. En este capítulo se ha recopilado y expuesto los componentes más relevantes que sirven para entender el entorno y los requerimientos que afectan el diseño metodológico propuesto en este trabajo, así como una breve exposición de las referencias bibliográficas pertinentes.

El tercer capítulo contiene una descripción del proceso y del desarrollo que se siguió para diseñar, estructurar y formular la propuesta que se entrega a través de este documento. Parte integral de este capítulo es la prueba y evaluación que se realizó a la metodología generada en este trabajo.

El cuarto capítulo corresponde a las conclusiones y discusiones realizadas como finalización del documento.

El documento está acompañado por el registro bibliográfico y dos anexos. El primer anexo corresponde a una tabla comparativa de las características correspondientes a varios documentos de referencia que pueden guiar procesos de analítica de datos. El segundo anexo corresponde a la propuesta metodológica resultante de este trabajo, la cual incluye un modelo de referencia y una guía de usuario de las actividades que integran la metodología propuesta.

2. MARCO TEÓRICO

El marco teórico de este trabajo está conformado por seis componentes que ayudan a entender el entorno y los requerimientos que afectan el diseño metodológico del entregable de este escrito. Estos componentes son: definiciones de términos clave, tecnologías relacionadas con el objeto de este trabajo, relación de estas tecnologías con las telecomunicaciones, características relevantes del sector de telecomunicaciones, interacción requerida con una red de telecomunicaciones para el estudio y análisis de su tráfico, y un análisis comparativo de varias guías referenciales que serán utilizadas como base para la propuesta metodológica que se entregará a través de este trabajo.

2.1. Definiciones

El primer punto que es necesario exponer corresponde a las definiciones que la Real Academia Española (RAE), el Diccionario Oxford (Oxford), la Ley Orgánica de Telecomunicaciones del Ecuador (LOT) y algunas fuentes bibliográficas proporcionan para los términos clave que constan en el objeto de este proyecto:

- **Metodología**: Ciencia del método; conjunto de métodos que se siguen en una investigación científica o en una exposición doctrinal. (RAE)
- **Analítica de datos**: Proceso de análisis computacional sistemático de dichos datos. (Oxford) - Proceso de análisis de datos multidisciplinario que combina matemáticas, estadísticas, técnicas y modelos para extraer conocimiento valioso a partir de dichos datos. [18] [19] (definición disponible también en Wikipedia)
- **Método**: Procedimiento que se sigue en las ciencias para hallar la verdad y enseñarla. (RAE)
- **Técnica**: Conjunto de procedimientos y recursos de que se sirve una ciencia o un arte. (RAE)
    - **Procedimiento**: Método de ejecutar algunas cosas. (RAE)
- **Herramienta**: Instrumento, por lo común de hierro o acero, con que trabajan los artesanos. (RAE)
    - **Instrumento**: Objeto fabricado, relativamente sencillo, con el que se puede realizar una actividad; Cosa o persona de que alguien se sirve para hacer algo o conseguir un fin. (RAE)
- **Estudio**: Esfuerzo que pone el entendimiento aplicándose a conocer algo; Trabajo empleado en aprender y cultivar una ciencia o arte; Aplicación, maña, habilidad con que se hace algo. (RAE)
- **Análisis**: Distinción y separación de las partes de algo para conocer su composición. (RAE)

- **Telecomunicaciones**: Se entiende por telecomunicaciones toda transmisión, emisión o recepción de signos, señales, textos, vídeo, imágenes, sonidos o informaciones de cualquier naturaleza, por sistemas alámbricos, ópticos o inalámbricos, inventados o por inventarse. La presente definición no tiene carácter taxativo, en consecuencia, quedarán incluidos en la misma, cualquier medio, modalidad o tipo de transmisión derivada de la innovación tecnológica. (LOT [20])
- **Red de telecomunicaciones**: Conjunto de medios (transmisión y conmutación), tecnologías (procesado, multiplexación, modulaciones), protocolos y facilidades en general, necesarios para el intercambio de información entre los usuarios de la red. [21]

2.2. Tecnologías relacionadas

El segundo punto que es necesario exponer corresponde a la naturaleza, propósito y funcionalidad de las tecnologías Big Data y Analítica de Datos que están relacionadas con el objeto de este trabajo. Existe gran variedad de fuentes de información que estudian a estas tecnologías, sin embargo, este trabajo se enfocará en la visión que tiene ITU al respecto, por ser uno de los organismos de estandarización más importantes en el sector de las telecomunicaciones, y en el concepto de "*cadena de valor de los datos*" que ha desarrollado la Comisión Europea en relación con la economía digital.

2.2.1. La visión de Big Data de la ITU

La visión que tiene ITU de Big Data para el sector de las telecomunicaciones se recopila en el suplemento 40 de la Serie Y de recomendaciones, denominado "Big data standardization roadmap" [10], y empieza por definir a Big Data como "*Un paradigma para permitir la recolección, almacenamiento, gestión, análisis y visualización, potencialmente bajo restricciones de tiempo real, de conjuntos de datos extensos con características heterogéneas*". De acuerdo con la Real Academia Española, un paradigma es una "*teoría o conjunto de teorías cuyo núcleo central se acepta sin cuestionar y que suministra la base y modelo para resolver problemas y avanzar en el conocimiento*".

ITU indica que los conjuntos de datos con los que se puede trabajar en la actualidad están comenzando a tener volúmenes tan grandes, a generarse a velocidades tan rápidas y a tener estructuras tan complejas que los métodos y herramientas tradicionales de procesamiento de esos datos se han vuelto inadecuados. Big Data es un paradigma en desarrollo que busca resolver el problema mencionado permitiendo

hacer una analítica eficiente, en tiempos tolerables, de los conjuntos de datos descritos en este párrafo.

Al hablar de Big Data, sería necesario pensar en todo un ecosistema destinado a procesar datos caracterizados por lo que comúnmente se conoce como las 4 Vs de Big Data: Volumen, Variedad, Velocidad y Veracidad. ITU incluye en esta lista una quinta V que hace referencia al "Valor" de los datos. Este ecosistema abarcaría varias áreas técnicas relacionadas, como por ejemplo "*computación en la nube*", "*Internet de las cosas*", "*seguridad y privacidad*", "*redes definidas por software*" e "*inspección profunda de paquetes*". Por otro lado, ITU lista algunas áreas potenciales de estandarización que son de su interés, y entre las cuales se menciona: "*casos de uso*", "*experiencia de red personalizada*", "*anonimización y desidentificación de datos personales*", "*estándares y guías para abordar asuntos relacionados con las implicaciones legales de usar big data en el sector de las telecomunicaciones*", y "*analítica de datos motivada por redes (network-driven analytics)*" que es el tema de interés del presente trabajo.

El modelo conceptual del ecosistema Big Data que presenta ITU se muestra en la Figura 2.1. Si bien no corresponde al alcance de este trabajo explicar a profundidad este modelo conceptual, es importante exhibirlo puesto que la analítica de datos, materia de este trabajo, pese a no pertenecer propiamente a un bloque específico del modelo, es una parte central de dicho ecosistema, tal como se explicará en la siguiente sección.

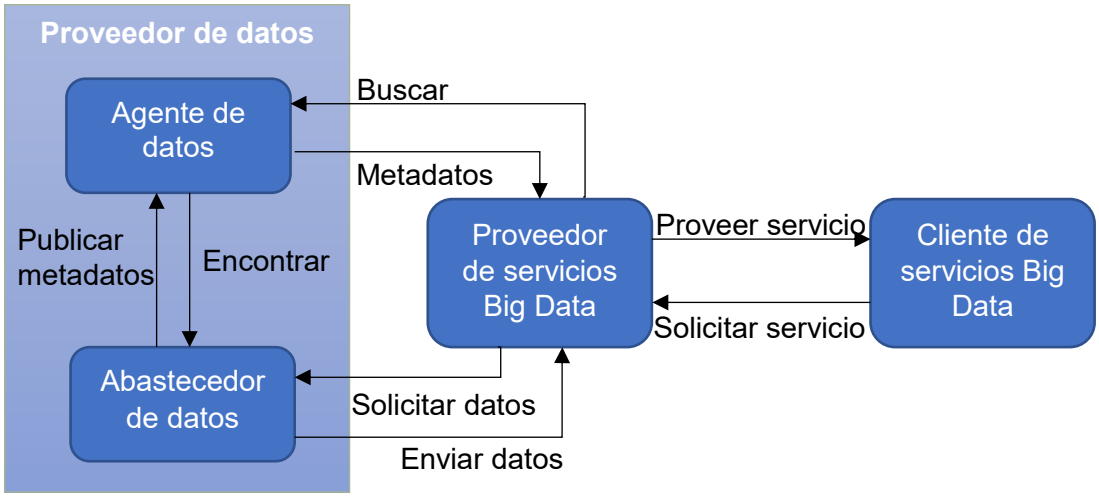

Figura 2.1. Modelo conceptual Big Data de ITU

El modelo conceptual representado en la Figura 2.1 está estructurado de acuerdo al modelo de negocios ontológico denominado E3-Value [22] y tiene tres roles principales: el proveedor de datos, el proveedor de servicios Big Data y el cliente de servicios Big Data. El proveedor de datos se divide a su vez en dos sub-roles que son el abastecedor

de datos y el agente de datos. Dentro del marco teórico que corresponde al presente trabajo de titulación, es importante mencionar que, en el modelo, los datos están separados de los servicios Big Data, y que el rol "Proveedor de servicios Big Data" soporta las prestaciones e infraestructura necesarias para realizar la analítica de estos datos, incluyendo como parte de sus actividades la provisión de herramientas de análisis y visualización de los mismos.

2.2.2. La analítica de datos en la cadena de valor de los datos

El suplemento 40 de la serie Y de recomendaciones de ITU no hace referencia a la tecnología de analítica de datos más allá de lo ya mencionado en la sección anterior. Para poder exponer sobre este tema, se ha decidido recurrir al concepto de "Cadena de Valor de los Datos" que delineó la Comisión Europea [17] como parte de una estrategia que busca mejorar la posición de liderazgo a nivel mundial que ostentaría Europa en la economía digital. Esta comisión, denominada "DG Connect", propone una estrategia que deberá:

a) Nutrir un ecosistema coherente europeo de datos que agrupe diversos grupos de interés de índole académico, económico e industrial;
b) Estimular la investigación y la innovación alrededor de los datos; y,
c) Establecer una serie de acciones que mejoren las condiciones para la extracción de valor a partir de dichos datos.

El libro "*New Horizons for a Data-Driven Economy, a roadmap for usage and exploitation of Big Data in Europe*" [23], que recoge el concepto de "cadena de valor de los datos" de la Comisión Europea, explica que la implementación del ecosistema Big Data en Europa deberá afrontar varios retos y que dichos retos han sido agrupados en dimensiones clave de acuerdo a lo ilustrado en la Figura 2.2. Esta figura muestra que los retos en cuestión tienen características legales, sociales, económicas, tecnológicas, y de aplicación, y que en el centro de estas dimensiones se encuentra lo que se ha denominado como "Datos y Habilidades".

De esta forma se habla de un ecosistema centrado en los datos y en la extracción de valor que éstos contendrían. La disponibilidad y el acceso a dichos datos se vuelve por tanto un tema fundamental, así como la necesidad de disponer de personal diestro capaz de trabajar en la extracción de valor. Por personal diestro se hace referencia a ingenieros y científicos de datos, con pericia en analítica, estadística, inteligencia artificial, minería de datos y gestión de datos. Adicionalmente se menciona que los expertos técnicos deberán combinarse con expertos de negocio que comprendan los

datos procesados y que conozcan cómo aplicar sus habilidades en la creación de valor para sus organizaciones.

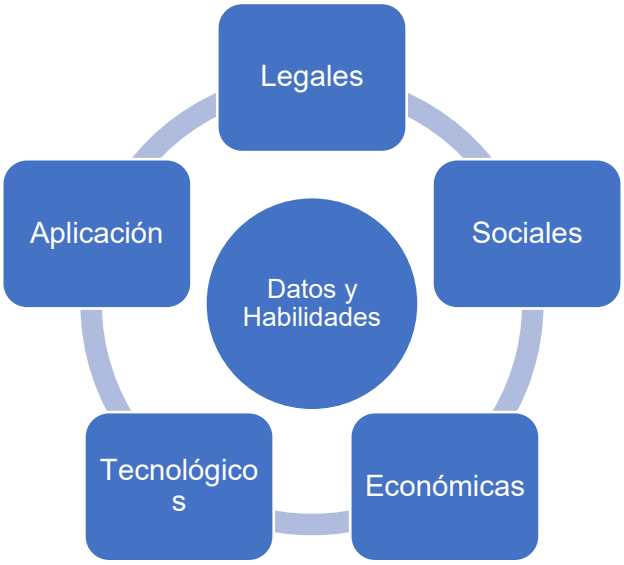

Figura 2.2: Dimensiones del valor de un ecosistema Big Data

El libro citado también ofrece un diagrama de la cadena de valor de Big Data que se presenta en la Figura 2.3 y un diagrama de los niveles macro, meso y micro del ecosistema Big Data que se presenta en la Figura 2.4.

Según los autores, la cadena de valor de los datos es una serie de pasos necesarios para generar valor y hacer revelaciones importantes a partir de dichos datos, y sería vista por la Comisión Europea como el centro de la futura economía del conocimiento. Estos pasos parten de la adquisición de datos, pasando por el análisis de los mismos, para luego asegurar su calidad, almacenarlos y utilizarlos.

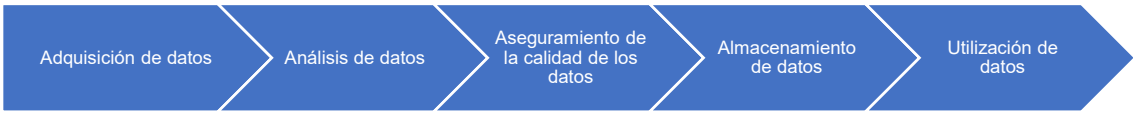

Figura 2.3. Cadena de valor de los datos

En este punto es necesario aclarar que el componente análisis de datos de la cadena de valor hace referencia a la tecnología analítica de datos, ya que, de acuerdo a la definición provista en la sección 2.1, la analítica de datos no es más que un proceso de análisis de datos multidisciplinario que combina matemáticas, estadísticas, técnicas y modelos para extraer conocimiento valioso a partir de dichos datos.

Esta cadena de valor se la encuentra en el centro de los distintos niveles del ecosistema Big Data representados en la Figura 2.4. Esta figura permite observar que dicho

ecosistema involucra tres niveles. El nivel macro abarca a organismos de amplio impacto como por ejemplo gobierno, entes reguladores, inversionistas y a la industria en general. El nivel meso abarca a organismos de menor impacto, pero de amplio alcance, como proveedores minoristas o los usuarios en general. El nivel micro, por su lado, abarca a la cadena de valor de los datos descrita previamente y a los actores que interactúan directamente con dicha cadena, los cuales son, por ejemplo, los proveedores mayoristas y los canales de distribución.

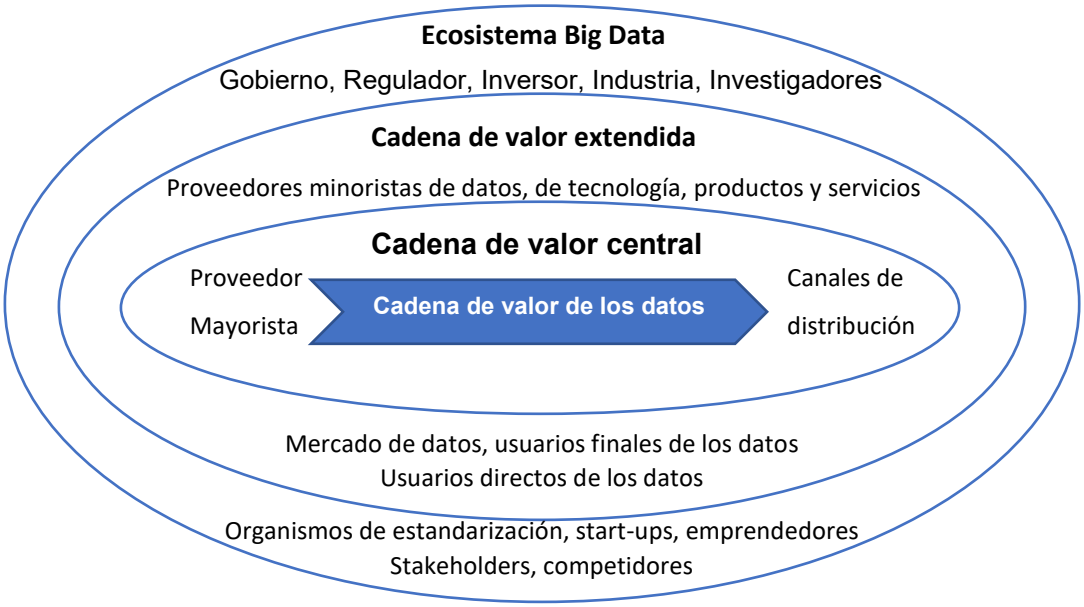

Figura 2.4. Niveles macro, meso y micro del ecosistema Big Data

La serie de pasos que conforman el proceso de análisis de datos y que abarcan diferentes disciplinas en la analítica de datos son tratados en la sección 2.6 de este trabajo de titulación.

2.2.3. Síntesis

A manera de síntesis, se puede decir que Big Data es un ecosistema complejo y amplio que abarca una gran variedad de áreas técnicas, que busca solucionar las limitaciones que las tecnologías tradicionales tienen para procesar datos cuyas características cumplen con las 5 Vs de Big Data (Volumen, Variedad, Velocidad, Veracidad y Valor), y que contiene la infraestructura, las prestaciones y las herramientas adecuadas que permiten llevar a cabo dicho procesamiento. En el centro del ecosistema se encuentra la analítica de datos, materia de este trabajo, que es un proceso de análisis multidisciplinario, sistemático y computacional, que forma parte de la cadena de valor de los datos, y que se encarga de extraer de los datos analizados valor y conocimiento.

2.3. Relación entre Telecomunicaciones, Big Data y Analítica de datos.

El tercer punto que es necesario exponer es la relación que existe entre las telecomunicaciones y las tecnologías “Big Data” y “Analítica de datos” que se acaban de explicar. Para comprender esta relación se debe considerar los siguientes factores.

2.3.1. Crecimiento de tráfico “explosivo”.

El sector de telecomunicaciones está experimentando un crecimiento “explosivo” del volumen de datos que transporta. Existen tres causas que explican este comportamiento:

- La penetración de los servicios de telecomunicaciones en la población avanza constantemente, e incluso algunos servicios de algunos países tienen una cobertura que supera al 100% de su población. No está por demás mencionar que la población mundial crece sin cesar y que se espera que aumente de 7 mil millones de personas en el año 2011 a 9 mil millones de personas en el año 2044, de acuerdo con la Oficina de Censos de Estados Unidos de América.
- Las fuentes de tráfico que se conectan a las redes de telecomunicaciones son cada vez más numerosas y variadas. Al inicio, las redes de telecomunicaciones eran utilizadas exclusivamente por personas, pero ahora son usadas también por diferentes tipos de dispositivos que se comunican entre sí. Este tipo de tecnologías se están desarrollando bajo el nombre de “Internet de las cosas” (IoT por sus siglas en inglés) y es uno de los impulsores de la cuarta revolución industrial [2].
- Los servicios de comunicación están en constante evolución y desarrollo, generando nuevas tecnologías que demandan cada vez mayor tráfico por parte de cada usuario que los utiliza. Ejemplos de estos servicios son la computación en la nube, la realidad virtual, la inteligencia artificial y el video en “*ultra alta resolución*” [2].

De esta forma, como cada vez hay más fuentes de tráfico conectadas a las redes de telecomunicaciones, y cada fuente tiene una tendencia a generar más tráfico que antes, el resultado es un crecimiento exponencial “explosivo” de tráfico que debe ser soportado por las redes de telecomunicaciones.

CISCO, mediante su reporte “The Zettabyte Era-Trends and Analysis” [24] y su herramienta web “VNI Forecast Highlights Tool” [25], estima que en el año 2021 el tráfico

anual IP que transportarán las redes de telecomunicaciones de todo el mundo llegará a tener una magnitud de varios zettabytes (1 zettabyte es 10^{21} bytes).

2.3.2. El valor potencial de los datos.

Los datos y la información empiezan a tener un valor muy importante en nuestra sociedad y economía, pues permiten obtener conocimiento que puede ser explotado de muchas formas. En una entrevista realizada en enero del año 2016 por el portal digital prnoticias [26], Elisa Martín Garijo, directora de Innovación y Tecnología de IBM España, indicó que "*la información se está convirtiendo en un nuevo recurso natural"*. Este concepto resulta ser muy importante, pues dota a los datos de un valor intrínseco que podría ser explotado. Tanto es así que la Comisión Europea, como parte del trabajo que realizó para delinear la estrategia expuesta en la sección 2.2.2 de este documento, emitió en el año 2014 al parlamento europeo, al consejo europeo, al comité económico y social europeo, y al comité de las regiones de Europa, un comunicado denominado "*Hacia una economía próspera impulsada por los datos*" en el cual se hace un llamado a que se tomen las acciones necesarias para guiar a la economía y a la sociedad hacia un modelo centrado en los datos y en el conocimiento [27].

2.3.3. La motivación de las empresas de telecomunicaciones

Tal como se explicó en la introducción, nuestra sociedad está iniciando procesos de cambio que son impulsados por la cuarta revolución industrial. Estos cambios ponen en riesgo el liderazgo económico que lograron alcanzar las empresas de telecomunicaciones gracias a la tercera revolución industrial, y con el fin de mantener dicho liderazgo, la industria de telecomunicaciones está desarrollando nuevos modelos de negocio que le permitan incrementar el valor que sus servicios ofrecen a la sociedad.

Las empresas de telecomunicaciones transportan por sus redes cada segundo una cantidad impresionante de datos que tiene un valor económico potencial y que podría ser transformado en una nueva fuente de ingresos. Estas empresas, por lo tanto, además de tener el poder financiero y logístico necesario para tomar las acciones requeridas para explotar esos datos [28], tienen también una fuerte motivación para hacerlo. El principal cuestionamiento que afrontan las empresas de telecomunicaciones es ¿cómo explotar estos datos?

2.3.4. La relación que surge entre las tecnologías descritas y las telecomunicaciones.

De acuerdo con lo que se expuso en la Sección 2.2.1 se destaca que Big Data proporcionará las herramientas e infraestructura necesarias para explotar los grandes volúmenes de datos que cursan por las redes de telecomunicaciones y la analítica de

datos proveerá los métodos necesarios para extraer valor de dichos datos. Sin embargo, para que este propósito pueda realizarse, se requiere adicionalmente modelos, guías o metodologías que coordinen y dirijan las acciones y relaciones requeridas por estas tecnologías hacia objetivos específicos. Los documentos que cumplen con este propósito serán tratados en la sección 2.6.

De esta forma se logra cubrir los medios necesarios que permiten explotar los datos que cursan por las redes de telecomunicaciones. Sin embargo, también es necesario analizar las características del sector de las telecomunicaciones que condicionarán el desarrollo de proyectos de analítica de datos.

2.4. Características del sector de telecomunicaciones relevantes para la analítica de datos.

El sector de telecomunicaciones se somete a características particulares que no suelen estar presentes en otros sectores. Algunas de estas características son una fuerte presencia regulatoria, una evolución desenfrenada de la tecnología, altos costos que afrontan las empresas de telecomunicaciones, un dinamismo y evolución constante de la industria, y una gran cantidad y variedad de actores que compiten en el mercado [28]. Este entorno genera desafíos y ventajas que no suelen existir dentro de la analítica de datos tradicional y que se exponen a continuación.

2.4.1. Convergencia, neutralidad de la red y privacidad de los datos.

La industria de telecomunicaciones está fuertemente regulada y debe someterse a reglas muy estrictas que varían de país en país. Algunos de los desafíos que se deben afrontar en esta industria para poder desarrollar proyectos de analítica de datos sobre el tráfico de telecomunicaciones están ligados a los conceptos de convergencia, neutralidad de tecnología y de red, y privacidad de los datos.

Convergencia [29] se puede definir como un proceso mediante el cual el usuario deja de estar obligado a acceder a diferentes servicios mediante dispositivos separados y redes independientes, y pasa a poder acceder a cualquier tipo de servicio mediante un mismo dispositivo y una única red de telecomunicaciones. Este concepto responde a una evolución eficiente de la infraestructura y de las tecnologías de telecomunicaciones.

Neutralidad [30] de tecnología y de red, es un concepto similar al de la convergencia, pero está relacionado con los derechos de los usuarios, y prohíbe que las empresas de telecomunicaciones restrinjan, limiten o interfieran en las comunicaciones de sus usuarios, permitiéndoles acceder a cualquier servicio que ellos deseen mediante cualquier dispositivo de su elección.

Finalmente, el concepto de privacidad [31], que está asociado a los conceptos de intimidad y secreto de las comunicaciones, se refiere a que los datos personales generados y transmitidos por cada usuario le pertenecen únicamente a dicho usuario y no pueden ser accedidos y explotados por terceros sin su consentimiento.

Estos tres conceptos por lo general deben ser respetados por cualquier empresa de telecomunicaciones. En Ecuador, por ejemplo, la Ley Orgánica de Telecomunicaciones [20] tiene como parte de sus objetivos: a) el promover la convergencia de redes, servicios y equipos, b) promover la neutralidad tecnológica y la neutralidad de red, y c) garantizar que los derechos de las personas sean respetados.

Para cumplir esto objetivos, la mencionada ley estipula:

- En su Artículo 4, que la provisión de servicios públicos de telecomunicaciones responderá a principios de no discriminación, privacidad y acceso universal.
- En su Artículo 12, que el Estado impulsará el establecimiento y explotación de redes y la prestación de servicios de telecomunicaciones que promuevan la convergencia de servicios. La Agencia de Regulación y Control de las Telecomunicaciones emitirá reglamentos y normas que permitan la prestación de diversos servicios sobre una misma red e impulsen de manera efectiva la convergencia de servicios y favorezcan el desarrollo tecnológico del país, bajo el principio de neutralidad tecnológica.
- En su Artículo 22, que los abonados, clientes y usuarios tendrán derecho al secreto e inviolabilidad del contenido de sus comunicaciones, y a la privacidad y protección de sus datos personales.
- En su Artículo 24, que los prestadores de servicios de telecomunicaciones tienen el deber de garantizar el secreto e inviolabilidad de las comunicaciones cursadas a través de las redes y servicios de telecomunicaciones, adoptar las medidas para la protección de los datos personales de sus usuarios y abonados, y adoptar las medidas para garantizar la seguridad de las redes.
- En su artículo 35, que las redes se operarán bajo el principio de regularidad, convergencia y neutralidad tecnológica.
- En su artículo 76, que los prestadores de servicios deberán adoptar las medidas técnicas y de gestión adecuadas para garantizar el secreto de las comunicaciones y de la información transmitida por sus redes.
- En su artículo 78, que para la plena vigencia del derecho a la intimidad, establecido en el artículo 66, numeral 20, de la Constitución de la República del Ecuador, los prestadores de servicios de telecomunicaciones deberán garantizar

la protección de los datos de carácter personal, y que la información suministrada por los clientes, abonados o usuarios no será utilizada para fines comerciales ni de publicidad, ni para cualquier otro fin, salvo que se cuente con el consentimiento previo y autorización expresa de cada cliente, abonado o usuario.

- Artículo 82, que los prestadores de servicios no podrán usar datos personales, información del uso del servicio, información de tráfico o el patrón de consumo de sus abonados, clientes o usuarios para la promoción comercial de servicios o productos, a menos que el abonado o usuario al que se refieran los datos o tal información, haya dado su consentimiento previo y expreso.

De igual manera, el Manual Regulatorio de la Asociación Interamericana de Empresas de Telecomunicaciones ASIET [32], indica que sus integrantes creen:

- Que la Gobernanza de Internet debe reflejar valores inalienables como la libertad de expresión, la libertad de asociación, la privacidad, accesibilidad, libertad de información y de acceso a la misma, y de utilizar Internet para el desarrollo económico.
- Que el proveedor de servicios de telecomunicaciones es responsable de asegurar el tráfico pero no su contenido, y debe proteger la intimidad y privacidad de los usuarios.
- Que un enfoque de neutralidad tecnológica promueve el despliegue de nuevas tecnologías, incentivando la innovación y los beneficios de la convergencia.
- Que el marco regulatorio debe ser neutral, independientemente de la modalidad o red desde la que se proveen los servicios.
- Que se debe avanzar hacia la neutralidad de plataformas y la neutralidad end-to-end.

De esta manera, estos tres conceptos (i.e., convergencia, neutralidad y privacidad) generan un desafío para las empresas de telecomunicaciones que quieran explotar los datos que cursan por sus redes, debido a tres razones:

- La cantidad y variedad de tipos de comunicación que cursarían por una red podría llegar a ser impredecible, compleja e imposible de controlar.
- Al momento de interactuar con los datos que cursan por la red, se deberá tener especial cuidado de no violar el principio de privacidad al que están obligadas las empresas de telecomunicaciones.
- La ley y regulación aplicable podría cambiar en el tiempo o de acuerdo con el país donde se encuentre la empresa de telecomunicaciones que está analizando

los datos, pero también podría ser necesario considerar las leyes, políticas y regulaciones de los lugares de procedencia y destino de los datos analizados.

2.4.2. Áreas de dominio de las empresas de telecomunicaciones.

Las empresas de telecomunicaciones suelen ser en su mayoría empresas de gran tamaño y poder económico. Esto se debe a que los costos de lanzamiento, desarrollo, operación y mantenimiento de su infraestructura resulta ser muy elevado y requiere de un músculo financiero fuerte [28]. Como consecuencia de esta característica del sector, resulta que estas compañías han desarrollado a lo largo de los años estructuras organizacionales y funcionales robustas que ponen énfasis en el desarrollo exitoso de proyectos, en la gestión eficiente de su infraestructura, en la seguridad que brindan a su información y en la calidad de los servicios prestados. Estos temas constituyen por tanto áreas de dominio de las empresas de telecomunicaciones.

Para cada una de estas áreas existen diferentes documentos que recopilan las mejores prácticas de la industria, marcos de referencia y estándares internacionales. Las empresas de telecomunicaciones se apoyan en estos documentos para elaborar las políticas y procesos internos que guiarán el desarrollo de las diferentes áreas de dominio mencionadas. A continuación, se cita algunos de los documentos más relevantes que se manejan en el sector de telecomunicaciones:

- A nivel de gestión de proyectos, en etapa de implementación, podemos citar a las mejores prácticas recopiladas en PMBOK (*Project Management Book Of Knowledge*) del instituto PMI (*Project Management Institute*), a las mejores prácticas recopiladas en PRINCE2 (*PRojects IN Controlled Environments*) de la compañía Axelos, y al estándar ISO 21500 "*Guidance on project management*" de la Organización Internacional de Normalización ISO.
- A nivel de gestión de proyectos, en etapa de formulación y evaluación ex-post, podemos citar a las mejores prácticas recopiladas en MoP (*Management of Portfolios*) de la compañía Axelos y al estándar ISO 2505 (*Project, programme and portfolio management — Guidance on governance*) de la Organización Internacional de Normalización ISO.
- A nivel de gestión, seguridad y calidad de servicios sobre infraestructura de tecnologías de la información, podemos citar a las mejores prácticas recopiladas en ITIL (*Information Technology Infrastructure Library*) de la empresa Axelos, a las mejores prácticas y marco de referencia para procesos de negocio eTOM de la empresa TM-Forum y a los estándares de la Organización Internacional de Normalización: ISO 20000 (gestión de servicios de tecnología de la información),

ISO 270001 (técnicas y sistemas de gestión de seguridad en tecnología de la información), e ISO 9001 (Sistema de Gestión de Calidad).

2.4.3. Síntesis

A manera de síntesis, se puede decir que el sector de telecomunicaciones presenta retos y desafíos puntuales que no suelen observarse en otros sectores de la economía, al mismo tiempo que cuenta con áreas de dominio altamente desarrolladas que buscan brindar una ventaja competitiva a las empresas que operan en dicho sector. Estos retos, desafíos y ventajas deben ser consideradas al momento de planificar y ejecutar proyectos de analítica de datos dentro de este tipo de empresas.

2.5. Interacción requerida con una red de telecomunicaciones para el estudio y análisis de sus datos.

Un aspecto crucial que es necesario considerar dentro del marco teórico es la forma en que se debe interactuar con una red de telecomunicaciones y con los datos que ésta transporta con el fin de poder estudiarlos y analizarlos. Si bien existe una gran cantidad y variedad de fuentes de información que abordan esta problemática desde distintos ángulos, se ha decidido investigar la información que proveen las siguientes fuentes:

- Center for Applied Internet Data Analysis (CAIDA – http://www.caida.org). CAIDA es un esfuerzo de colaboración entre organizaciones pertenecientes a sectores comerciales, gubernamentales y de investigación que estudia aspectos teóricos y prácticos relacionados con Internet con el fin de hacer revelaciones macroscópicas de su infraestructura, comportamiento, utilización y evolución. CAIDA fomenta un entorno de colaboración donde los datos pueden ser adquiridos, analizados y compartidos de forma adecuada.
- “Internet End-to-end Performance Monitoring”, grupo perteneciente al laboratorio SLAC que es gestionado por Stanford University en nombre del Departamento de Energía de Estados Unidos de Norte América (IEPM - http://www-iepm.slac.stanford.edu). IEPM es un grupo cuyo propósito es monitorear el desempeño de la comunicación por Internet que existe entre distintos laboratorios de aceleradores de partículas y las universidades que colaboran con dichos laboratorios.
- Distribución Kali-Linux de la empresa Offensive Security, especializada en entrenamiento, certificación y servicios de seguridad digital. (Kali - https://www.kali.org). Kali-Linux ofrece una distribución basada en Linux con una gran cantidad de programas que permiten auditar, analizar, espiar, estresar, atacar y penetrar redes de telecomunicaciones IP.

Las fuentes citadas se han escogido por tratarse de proyectos especializados de gran envergadura, con un respaldo organizacional importante, que han existido por varios años y que concentran información actualizada de interés para el desarrollo del tema que nos compete en este trabajo. Estas fuentes no solo se dedican a analizar, monitorear e interactuar con redes de telecomunicaciones, sino que recopilan y publican información actualizada de tipo académica, técnica y comercial que permitieron listar las interacciones que podrían requerirse con una red de telecomunicaciones para efecto de estudiar y analizar su tráfico, e identificar métodos, técnicas y herramientas que faciliten dichas interacciones. La lista de actividades necesarias para dicho efecto es la siguiente:

- Generar y/o capturar tráfico.
- Generar y recolectar datos estadísticos.
- Identificar, clasificar y/o filtrar el tráfico.
- Anonimizar y/o censurar el contenido del tráfico.
- Analizar, visualizar el tráfico capturado o las estadísticas generadas.
- Modelar el tráfico capturado o las estadísticas generadas.
- Transformar, reorganizar, estructurar o pre-procesar el contenido del tráfico.
- Complementar o completar el tráfico o las estadísticas recolectadas.
- Correlacionar varias fuentes de tráfico.

Algunas de las actividades indicadas, al estar específicamente relacionadas con el estudio del tráfico de una red de telecomunicaciones, no forman parte de la literatura tradicional de Analítica de Datos o de Big Data y requieren ser tratadas de forma especial dentro del presente trabajo de titulación. Estas actividades tienen un nivel de especialización tecnológica importante por lo que a continuación se realizará un resumen general de los métodos, técnicas y herramientas que pueden emplearse para llevar a cabo dichas actividades, referenciado documentos con información más detallada.

2.5.1. Métodos, técnicas y herramientas para generar y/o recolectar información proveniente de una red de telecomunicaciones.

El proceso de recolección de información de una red de telecomunicaciones para llevar a cabo procesos de analítica de datos se puede efectuar, en primer lugar, en distintos puntos físicos de la red [33] [34].

- El proceso puede realizarse en los puntos orígenes y destinos de la red, es decir en los dispositivos de los usuarios, en los servidores o en los equipos de telecomunicaciones a los que se conectan los usuarios para acceder a ciertos

servicios. En este caso se estaría haciendo un monitoreo en el extremo o borde de la red.

- El proceso también podría realizarse en puntos intermedios o internos de la red, como por ejemplo en medios de transmisión o en equipos de telecomunicaciones como *routers* o *switchs*.
- Adicionalmente, están empezando a emerger soluciones de redes virtualizadas por lo que, en estos casos, el proceso mencionado podría ejecutarse en un entorno virtualizado. Este tipo de entornos pueden encontrarse, por ejemplo, en soluciones desplegadas en la nube, donde un grupo de servidores virtualizados en distintos lugares del mundo se interconectan por una red virtual que no corresponde a la topología física que interconecta los servidores físicos [35].

Los procesos de recolección de información, independientemente de que se realicen en el borde o en el interior de una red, pueden basarse en contadores o en tráfico. En el primer caso se estaría generando y recolectando datos estadísticos mientras que en el segundo caso se estaría generando y/o capturando tráfico real proveniente de la red [34] [36] [37].

- Para la generación y recolección de datos estadísticos, existen diferentes herramientas que pueden emplearse, dependiendo de su disponibilidad. Estas herramientas corresponden a aquellos protocolos y paquetes de software que permiten generar los datos estadísticos deseados. Protocolos comunes para estas herramientas son SNMP (*Simple Network Management Protocol*), RMON (*Remote Monitoring*) o Netflow. Estas herramientas generan contadores en los dispositivos, servidores o equipos de telecomunicaciones monitoreados y los actualizan en base a mediciones que luego son recuperadas para análisis. Estos contadores pueden ser, por ejemplo, cantidad de paquetes recibidos, cantidad de paquetes enviados, cantidad de errores detectados, etc.
- Para la generación y/o recolección de tráfico, se puede utilizar métodos activos, pasivos o una combinación de ambos.
    - Los métodos activos consisten en generar tráfico en la red de telecomunicaciones. La herramienta más conocida, básica y utilizada para este propósito es el protocolo ICMP junto el programa PING, disponibles en la mayoría de sistemas operativos y equipos de telecomunicaciones. Sin embargo, existen herramientas más especializadas como, por ejemplo, sondas activas capaces de generar y simular llamadas telefónicas celulares. El dispositivo Mobile Robot R400 de la empresa iQsim [38] es una opción disponible en el mercado que puede emplearse para pruebas de calidad de

servicio, de roaming y de aseguramiento de ingresos, entre otras cosas. Estas herramientas pueden generar y recolectar datos estadísticos y el tráfico en sí.

- Los métodos pasivos consisten en capturar tráfico de la red de telecomunicaciones, sin que haya generado tráfico adicional como parte de dichos métodos. El proceso de captura podría realizarse gracias a que existen medios de transmisión compartidos, como, por ejemplo, medios inalámbricos o redes basadas en hubs. En estos casos, la técnica consiste en ubicar herramientas apropiadas (por ejemplo, computadoras con software de captura como Wireshark) en lugares adecuados, de forma que reciban las señales de telecomunicaciones y las capturen. Sin embargo, las redes de telecomunicaciones más comunes son aquellas donde se utilizan medios de transmisión guiados no compartidos y en las cuales se deben seguir los métodos y usar las herramientas que se detallan más adelante.
- Los métodos combinados usan de forma conjunta métodos activos y métodos pasivos. Un ejemplo puede ser utilizar métodos pasivos para capturar tráfico en la red proveniente de usuarios reales, y en caso de que no existan usuarios reales generando tráfico, aplicar métodos activos que generen el tráfico necesario para que sea capturado con los métodos pasivos en uso.

Los métodos de captura de tráfico en medios guiados no compartidos pueden realizarse en los equipos que están involucrados en las comunicaciones, como computadoras, servidores o equipos de telecomunicaciones, o pueden realizarse interceptando el tráfico en los medios y equipos de transmisión/red [37].

- Un método que permite capturar tráfico en equipos es la utilización de interfaces internas virtuales provistas por los sistemas operativos de dichos equipos y el uso de software capaz de capturar tráfico a través de dichas interfaces. Se puede configurar el software que se desea monitorear y/o el sistema operativo para que curse tráfico por dichas interfaces virtuales. Posteriormente, en sistemas tipo Unix se puede emplear el software denominado dumpcap para capturar el tráfico deseado, mientras que en sistemas Windows se puede emplear el software rawcap.
- Otro método que permite capturar tráfico en equipos es la utilización del modo promiscuo en las tarjetas de red junto con software de captura como Wireshark o tcpdump.
- Por otro lado, para capturar tráfico en medios de transmisión guiados no compartidos existen los siguientes métodos:

- Se puede configurar opciones de duplicación de tráfico en equipos de red como switchs y ruteadores. Estas opciones instruyen a los equipos de red a que generen copias del tráfico deseado y a que lo envíen a través de una interfaz específica, donde se puede conectar un equipo que se encargará de capturar el tráfico duplicado, generalmente usando interfaces de red en modo promiscuo. Estos métodos suelen conocerse bajo la denominación SPAN (Switched Port Analyzer), Port Mirroring o Port Monitoring.
- También se puede insertar en el medio de transmisión dispositivos que permitan capturar el tráfico sin interferir en las comunicaciones, mediante técnicas conocidas como Machine in the Middle. Estos dispositivos pueden ser Hubs (que se encargarán de enviar el tráfico por todas las interfaces existentes), Network Taps (que permitirán "leer" el tráfico que cursa por el medio de transmisión interceptado), o equipos de captura especializados.
- Otra opción más avanzada y riesgosa es la de utilizar técnicas lógicas de intercepción. Estas técnicas son, por ejemplo, ataques denominados *Man in the Middle* o *MAC Flooding*, las cuales son artilugios que consiguen que equipos de red como switchs, ruteadores, servidores o computadores tengan un comportamiento distinto al normal provocando que envíen su tráfico a un destino definido por el interceptor, donde se captura la información.

2.5.2. Métodos, técnicas y herramientas para identificar, clasificar y/o filtrar tráfico proveniente de una red de telecomunicaciones.

Una vez que se ha obtenido el tráfico, las mediciones o estadísticas deseadas de la red de telecomunicaciones analizada, puede ser necesario identificar, clasificar y/o filtrar los datos obtenidos. De acuerdo a la empresa Sandvine, [39] [40], la clasificación del tráfico de una red de telecomunicaciones abarca y requiere tanto su identificación, como su categorización, su medición y la extracción de su información.

La identificación del tráfico consiste en determinar el tipo de tráfico observado. De acuerdo a Sandvine, se puede utilizar tres técnicas para realizar esta identificación:

- Firmas (*signatures*): consiste en detectar patrones de comportamiento de un flujo de datos y compararlo con firmas de comportamiento conocidas. Si el comportamiento detectado corresponde a una firma conocida, entonces se determina que el tráfico correspondiente al flujo observado proviene del servicio conocido que produjo la firma equivalente. Una firma de tráfico correspondiente a una llamada de voz puede ser un flujo constante de tráfico de baja velocidad, mientras que una firma de tráfico correspondiente a una transmisión de video

puede ser un flujo constante de tráfico de alta velocidad y una firma de tráfico correspondiente a navegación de páginas de internet puede ser la detección de ráfagas de tráfico espaciadas en el tiempo.

- Rastreadores (*trackers*): consiste en hacer un seguimiento del tráfico de control que gestiona un flujo de datos. Los rastreadores deben conocer el comportamiento básico de control de los servicios que rastrea, y cuando observa que el tráfico de control corresponde al de un servicio conocido, entonces se determina que el tráfico gestionado corresponde a dicho servicio. Un rastreador podría detectar una llamada celular al observar una secuencia específica de intercambio de mensajes entre un controlador de radio base (RNC), las bases de datos de visitantes y de base (VLR y HLR).
- Analizadores (*analyzers*): Los analizadores son similares a los rastreadores, pero tienen un nivel de conocimiento profundo de los protocolos y flujos que corresponden a distintos servicios. Los analizadores son capaces, no solamente de detectar patrones de comportamiento en los flujos de datos y en el tráfico de control correspondiente, sino que también son capaces de extraer información e incluso interactuar o interferir con los servicios.

El resultado de la identificación de tráfico debe revelar información importante sobre aspectos que son relativos al tráfico observado. Ejemplos de esta información son Protocolo, Aplicación, Servicio, Proveedor y Red de origen/destino.

Una vez identificado el tráfico, se puede proceder a categorizarlos. Ejemplos de categorías son: Servicio de almacenamiento, Juegos en línea, Servicios de comunicación y Tráfico de navegación.

El proceso de identificación y clasificación puede aprovecharse para realizar tomar mediciones y métricas del tráfico observado. Se puede así conocer el volumen de tráfico, la velocidad de los flujos, la calidad del servicio experimentado, la cantidad de usuarios que usan un mismo servicio, entre otras cosas.

Finalmente, se puede aprovechar el proceso para extraer información como atributos y características propias del servicio detectado y tomar ciertas acciones. Se puede, por ejemplo, detectar los códecs y la resolución empleada para un servicio de streaming de video, definir eventos que disparen alertas en caso de que se detecte problemas de calidad en el servicio, asociar el flujo con un usuario específico de la red y alimentar una base de datos que sea utilizada por las áreas de inteligencia del negocio para contactarse con el usuario e investigar más afondo los problemas de servicio que experimentó.

El proceso de clasificación de tráfico, de acuerdo con Sandvine, debe superar una serie de retos que se listan a continuación:

- Los servicios pueden intentar ocultar su información mediante técnicas de encapsulamiento, cifrado, ofuscación, empleo de *proxies*, compresión y empleo de túneles de comunicación.
- El proceso puede generar falsos positivos y falsos negativos en la identificación y clasificación del tráfico.
- El tráfico puede cursarse a través de múltiples flujos y sesiones, por lo que el proceso debe conocer a fondo el diseño y funcionamiento de los protocolos y aplicaciones involucrados y puede requerir recursos de procesamiento importantes.
- El tráfico puede cursarse por distintos recursos físicos de red, por lo que la extracción de datos de la red puede ser tan compleja como la misma red.

Finalmente, en cuanto a herramientas se refiere, la tecnología más apropiada para llevar a cabo procesos de identificación y clasificación de tráfico es la denominada *Deep Packet Inspection* (DPI) o Inspección Profunda de Paquetes, y sus variaciones *Network Intrusion Detection* y *Network Intrusion Prevention*. Estos sistemas están diseñados para procesar en gran detalle el tráfico proveniente de una red y pueden emplear las técnicas aquí descritas para así determinar la naturaleza de los flujos de datos analizados. Ejemplos de paquetes de software DPI que pueden emplearse de forma libre son Bro [41] y Snort [42].

2.5.3. Métodos, técnicas y herramientas para anonimizar y/o censurar el contenido del tráfico.

La necesidad de anonimizar y/o censurar el contenido del tráfico se explicó en la Sección 2.4.1 y consiste en evitar que información privada e íntima de los usuarios sea expuesta a terceros, así como prevenir que información que revele la infraestructura de una red sea revelada [43]. Esta información no solo se encuentra en el *payload* de cada paquete con datos generados por los usuarios, sino también en las cabeceras de cada paquete, aunque estas no hayan sido generadas directamente por los usuarios. Las direcciones IP, por ejemplo, que se encuentran en las cabeceras, también son información sensible pues permiten rastrear a los usuarios y por tanto violar su derecho a la privacidad e intimidad.

El proceso de anonimización y/o censura de tráfico puede realizarse mediante la eliminación o transformación de la información que se requiere proteger. La eliminación de información es un método sencillo, pero puede remover datos indispensables para

realizar un análisis acertado del tráfico. La transformación de información, en cambio, es un método complejo pues la modificación de datos debería realizarse de una forma congruente en el tiempo y con respecto a la naturaleza del tráfico procesado, con el fin de no alterar el resultado del análisis de dicho tráfico.

Para el caso de información no generada por los usuarios y que generalmente se encuentra en las cabeceras de los paquetes procesados, el método más utilizado es la transformación. Se puede, por ejemplo, modificar las direcciones MAC, las direcciones IP, los *time-stamps* de los paquetes, los contadores y los números de protocolos y puertos [44]. Para el caso de información generada por los usuarios y que generalmente se encuentra en los *payloads* de los paquetes procesados, el método más utilizado es la eliminación.

Herramientas que pueden servir para anonimizar tráfico son: Crypto-Pan [45], Anontool [43], FLAIM [46] y tcpdpriv [47].

2.5.4. Herramientas para analizar, visualizar y modelar información proveniente de una red de telecomunicaciones.

El análisis, visualización y modelado de información proveniente de una red de telecomunicaciones puede realizarse mediante los métodos, técnicas y herramientas tradicionales descritas en libros de minería de datos [48] [49] [50] [51] [52], los cuales se han aplicado exitosamente al análisis de tráfico de redes de telecomunicaciones [53] [54]. Sin embargo, herramientas de análisis como Wireshark [55] y de visualización como QlikView [56] o Google Charts [57], pueden resultar de gran utilidad para analizar, visualizar y modelar información proveniente de redes de telecomunicaciones, haciéndolas más fáciles e intuitivas de entender. La sección 3.4.4 muestra cómo el uso de estas herramientas visuales puede ayudar a visualizar y modelar información proveniente de una red de telecomunicaciones.

Como parte del diseño de la propuesta metodológica que se realizará en el presente trabajo de titulación se deberá asegurar que las actividades que conforman la metodología aborden cada uno de las interacciones citadas en esta sección. La identificación de las actividades que conformarán la metodología se realizará a partir de los documentos referenciales que se tratan a continuación.

2.6. Referencias para desarrollo de proyectos de analítica de datos.

La analítica de datos es un campo que lleva varias décadas desarrollándose y aplicándose en distintos ámbitos, que van desde la detección de fraudes bancarios [15]

hasta la predicción de *churn* (cantidad de clientes que en un período de tiempo abandona a su proveedor de servicios) de redes móviles de telecomunicaciones [16]. Durante este tiempo, varias empresas, organizaciones y personas han contribuido a su perfeccionamiento generando y publicando documentos que pueden servir de guía para quienes busquen desarrollar proyectos de analítica de datos.

Múltiples fuentes que tratan el tema de analítica de datos, tanto en el mundo académico como en el mundo empresarial, suelen hacer referencia a los siguientes documentos: CRISP-DM, SEMMA y KDD Process [58] [59]. Adicionalmente, existen trabajos recientes de actores importantes en el sector industrial como IBM y TM-Forum, que prometen convertirse en aportes valiosos para el campo de la analítica de datos. Estos trabajos son conocidos bajo los nombres ASUM-DM, SMAM y GB979.

A continuación, se presenta una descripción de cada uno de estos documentos para luego realizar un análisis de los mismos y de su nivel de aplicación en el estudio y análisis de tráfico en redes de telecomunicaciones.

2.6.1. KDD Process

El proceso KDD hace referencia al proceso de descubrimiento de conocimiento en bases de datos y fue publicado en 1996 mediante un artículo en la revista de la Asociación Americana de Inteligencia Artificial (AAAI por sus siglas en inglés) [60].

La Figura 2.5 representa los principales pasos descritos para el proceso KDD, los cuales conforman un proceso iterativo que parte de la comprensión del objetivo del proceso, pasando por diferentes interacciones con los datos analizados, hasta llegar a la evaluación de los resultados y la actuación sobre dichos resultados.

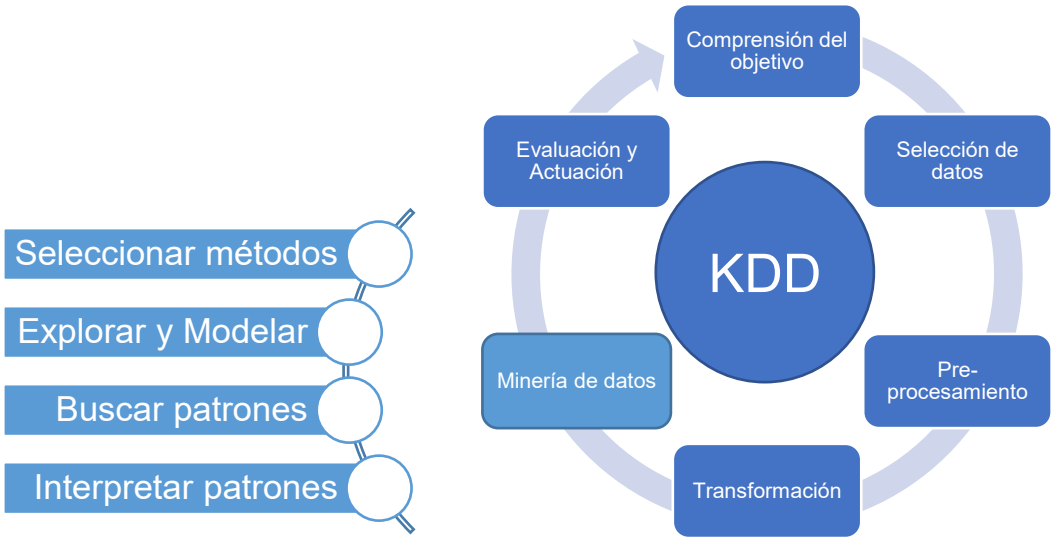

Figura 2.5. Descripción del proceso KDD

Uno de los pasos que conforman un proceso KDD es aquel denominado “Minería de Datos”, el cual a su vez está conformado por una serie de actividades que buscan modelar los datos y encontrar patrones en los conjuntos de datos analizados.

El artículo publicado en 1996 en la revista AAAI ofrece una perspectiva de los campos de la “minería de datos” y del “descubrimiento de conocimiento en bases de datos” (KDD por sus siglas en inglés), explicando la relación que existe entre estos, así como con otros campos asociados. El artículo expone las razones por las cuales estos campos son necesarios dando ejemplos de su aplicación en el mundo real y técnicas específicas que son utilizadas en la minería de datos.

2.6.2. SEMMA

El proceso SEMMA corresponde a una iniciativa de la empresa SAS, fue publicada en el año 1997 en el *SAS User Group International Conference* SUGI 22 e implementada en su paquete de software SAS Enterprise Miner. El nombre del proceso referencia a los cinco pasos que conforman el proceso de acuerdo a su nombre en el idioma inglés: Sample, Explore, Modify, Model, Assess [51].

SEMMA, de acuerdo a la empresa SAS, no es una metodología sino un proceso central para realizar minería de datos. Se trata de una organización lógica de las herramientas funcionales que el software SAS Enterprise Miner utiliza para realizar las tareas centrales de minería de datos.

Los pasos referidos por el acrónimo SEMMA son, en español, el muestreo, la exploración, la modificación y el modelado de los datos analizados, y la evaluación de los resultados obtenidos tras un proceso de minería de datos. El documento ofrece una breve explicación de cada paso sin ahondar en cada tema.

La Figura 2.6 representa los principales pasos del proceso SEMMA.

Figura 2.6. Descripción del proceso SEMMA

2.6.3. CRISP-DM

CRISP-DM es un acrónimo para CRoss Industry Standard Process for Data Mining y corresponde a un documento que describe una metodología para minería de datos, junto con un modelo de referencia, una guía de usuario, una lista de reportes y un apéndice con información relacionada adicional. Este documento fue elaborado por el consorcio CRISP-DM, formado por las empresas SPSS Inc., NCR Systems Engineering Compenhagen, DaimlerChrysler AG y OHRA Verzekeringen en Bank Groep B.V. junto con la Comisión Europea a través del proyecto ESPRIT (SPSS Inc. sería luego adquirido por IBM). La primera publicación de esta metodología fue en el año 1996 y, a pesar de que se trabajó en una segunda versión, no se volvió a actualizar el documento CRISP-DM [61].

La metodología CRISP-DM está descrita a través de un modelo de proceso jerárquico, con cuatro niveles de abstracción, que contiene una serie de tareas que deberían ser ejecutadas en un proyecto de minería de datos. La Figura 2.7 muestra un esquema conceptual de los cuatro niveles de abstracción y sus componentes.

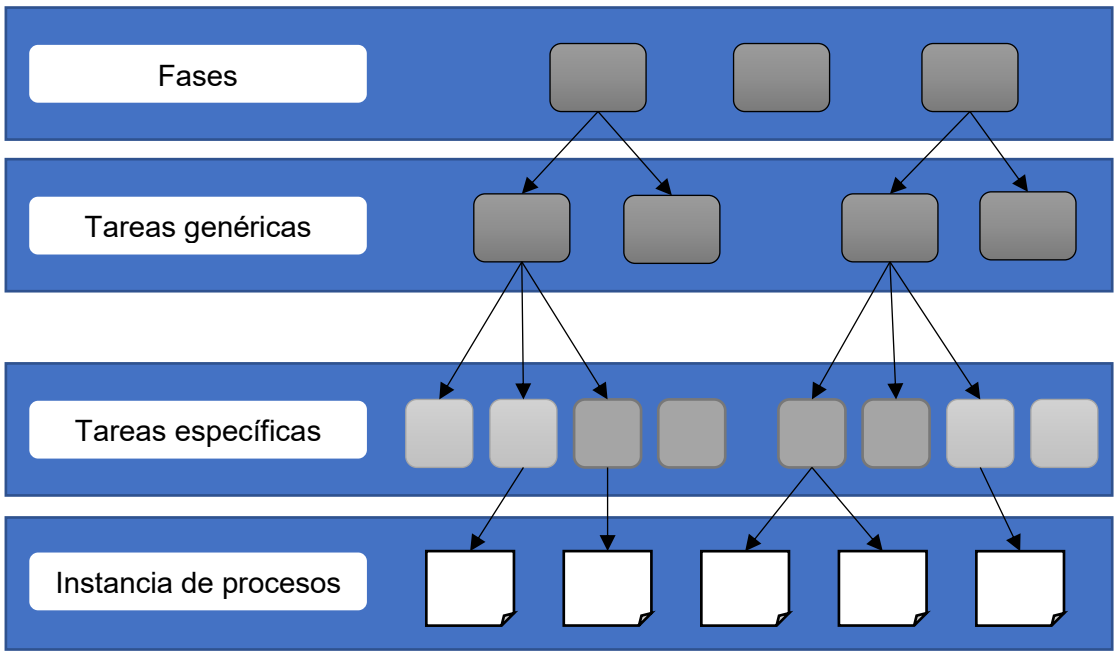

Figura 2.7. Niveles de abstracción y componentes de modelo CRISP-DM

El primer nivel de abstracción corresponde a las fases que organizan un proceso de minería de datos. Cada fase contiene varias tareas genéricas de segundo nivel, las cuales fueron diseñadas con el fin de ser aplicables a cualquier caso práctico de minería de datos. Cada tarea genérica abarcará en el tercer nivel tareas especializadas que fueron diseñadas para ser utilizadas en casos específicos de minería de datos. El cuarto y último nivel consiste en documentos o registros que detallan el trabajo que

efectivamente se habría realizado en cada tarea a lo largo de la ejecución de un proceso de minería de datos.

Tanto las fases como las tareas genéricas conforman el modelo de referencia CRISP-DM, que podrá emplearse en cualquier proyecto de minería de datos, mientras que las tareas específicas y los registros deberán seleccionarse y ejecutarse de acuerdo a las necesidades de un proceso específico de minería de datos, el cual es denominado Proceso CRISP-DM.

El modelo de referencia CRISP-DM se destaca de los documentos descritos previamente por cuanto a cada fase de su proceso le acompaña una detallada lista de tareas genéricas, las cuales a su vez contendrán tareas específicas y registros. Esta lista detallada se conoce como la guía de usuario.

La Figura 2.8 describe el proceso iterativo que vincula las distintas fases del modelo de referencia CRISP-DM.

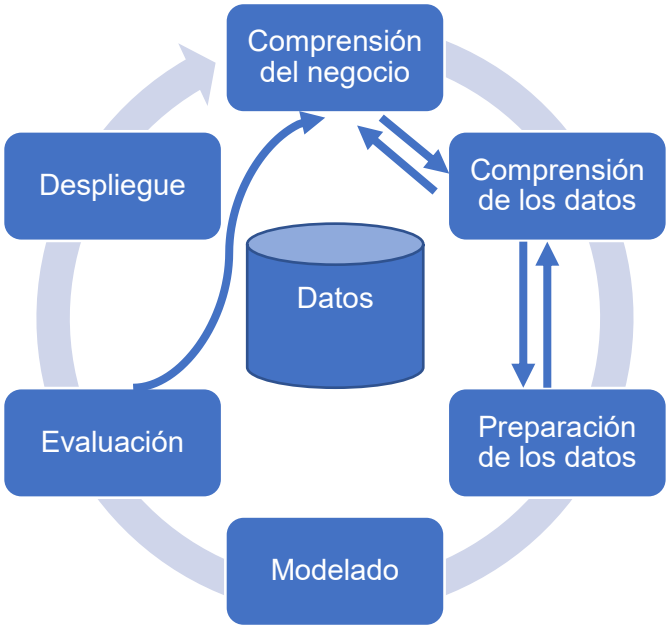

Figura 2.8. Descripción del proceso iterativo del modelo CRISP-DM

La Figura 2.9, por otra parte, muestra las distintas fases del proceso (cuadros superiores) junto con sus tareas genéricas correspondientes (texto en negritas) y ejemplos de los respectivos documentos o registros de salida (texto en itálica).

Un hecho a recalcar de la metodología CRISP-DM es que toma en consideración tareas que no pertenecen a la ciencia analítica, pero que corresponden a procesos de gestión de proyectos y a actividades relacionadas con el despliegue de soluciones o servicios, y de la operación y mantenimiento de los mismos.

Comprensión del negocio	Comprensión de los datos	Preparación de los datos	Modelado	Evaluación	Despliegue
Determinar objetivos del negocio *Criterios de éxito* **Evaluar estado** *Inventario de recursos* **Determinar objetivos de la analítica** *Criterios de éxito* **Producir el plan de proyecto** *Evaluación inicial de herramientas y técnicas*	**Recolectar datos iniciales** *Reporte* **Describir datos** *Reporte* **Explorar datos** *Reporte* **Verificar calidad de los datos** *Reporte*	**Seleccionar datos** *Razones de inclusión o exclusión* **Limpieza de datos** *Reporte* **Construcción de datos** *Nuevos Atributos* **Integrar datos** *Datos incluidos* **Formatear datos** *Datos formateados* *Conjunto de datos con su descripción*	**Seleccionar técnicas de modelado** *Supuestos empleados* **Generar diseño de pruebas** *Diseño de pruebas* **Construir modelo** *Configuración de parámetros* **Evaluar modelo** *Configuración revisada de parámetros*	**Evaluar resultados** *Modelos aprobados* **Revisar proceso** *Proceso revisado* **Determinar siguientes pasos** *Lista de posibles acciones*	**Plan de despliegue** *Plan* **Plan de monitoreo y mantenimiento** *Plan* **Producir reporte final** *Reporte y presentación final* **Revisar proyecto** *Lecciones aprendidas*

Figura 2.9. Fases y tareas genéricas correspondientes al modelo CRISP-DM

2.6.4. ASUM-DM

ASUM-DM es un acrónimo de Analytics Solutions Unified Method for Data Mining y corresponde a una iniciativa de la empresa IBM para extender y refinar CRISP-DM. ASUM-DM fue publicada en el año 2015 y puede ser obtenida libremente a través del portal de IBM previa solicitud de acceso. IBM provee esta metodología vía un instalador que genera un documento web con un mapa del sitio, una descripción de los procesos que conforman ASUM-DM [13].

ASUM-DM está definido por IBM como un proceso iterativo para la implementación de proyectos de analítica predictiva de minería de datos basado en una combinación de la metodología ASUM de IBM con una metodología CRISP-DM extendida y refinada por IBM. La Figura 2.10 muestra un esquema conceptual de ASUM-DM.

En lo que respecta a la metodología original ASUM, esta es una metodología genérica de IBM usada por la compañía para trabajar junto con sus clientes en la implementación de soluciones o servicios de analítica a través de las herramientas que ésta comercializa. ASUM tiene un enfoque que relaciona fuertemente la gestión de proyectos con el despliegue e implementación de soluciones y servicios. En la variación

denominada ASUM-DM, la metodología CRISP-DM se encuentra sumergida en el modelo ASUM, dando como resultado una combinación de fases y tareas relacionadas con analítica de datos, fases y tareas relacionadas con la gestión de proyectos, tareas relacionadas con la implementación y despliegue de soluciones, y tareas relacionadas con la operación y optimización de dichas soluciones.

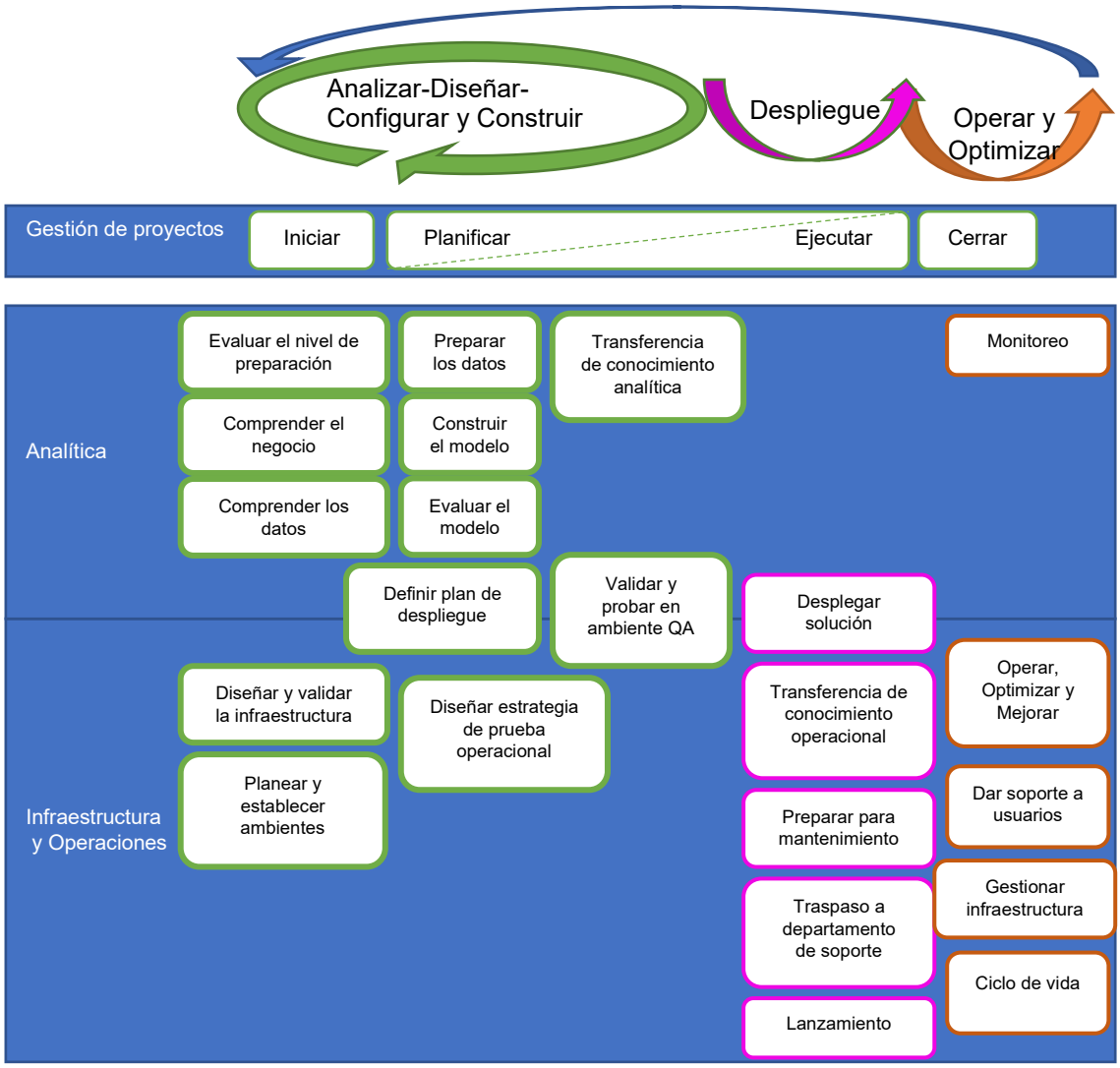

Figura 2.10. Descripción del modelo ASUM-DM.

Al igual que CRISP-DM, la metodología ASUM-DM está organizada en varios niveles. En el primer nivel, se distinguen los procesos que corresponden por un lado a las actividades inherentes a la gestión de proyectos, por otro a las actividades inherentes a la minería de datos y por otro a las actividades inherentes a la infraestructura subyacente y su operación. El proceso correspondiente a gestión de proyectos se divide en las cuatro fases “Inicio”, “Planificación”, “Ejecución” y “Cierre”. Los otros dos procesos se dividen en tres fases que son “Analizar, Diseñar, Configurar y Construir”, “Desplegar” y “Operar y Optimizar”. Al igual que CRISP-DM, cada fase contiene tareas genéricas

llamadas “Actividades”, las cuales a su vez contienen tareas específicas llamadas simplemente “Tareas”. Algunas Tareas están acompañadas de la descripción de los pasos que se deben seguir para ejecutarlas, pero en el caso de ASUM-DM no se especifica qué documentos se deben generar al finalizar la ejecución del proceso completo, como lo hace CRISP-DM, sino que define actividades en las cuales se debe generar dicha documentación.

Dado el fuerte enfoque que tiene ASUM-DM en la gestión de proyectos, la metodología también presenta una lista de roles que se deberá considerar a lo largo del proceso y especifica cuales roles intervienen en cada Actividad o Tarea.

2.6.5. SMAM

SMAM es un acrónimo de Standard Methodology for Analytical Models y corresponde a un trabajo de Olav Laudy, un colaborador de IBM, que publicó en 2015 esta propuesta metodológica en su página web personal [62].

SMAM es una metodología estándar para modelos analíticos que muestra ocho fases que se deben considerar al momento de crear modelos analíticos. Esta metodología fue considerada como referencia para el presente trabajo debido a que fue elaborada recientemente por un profesional de analítica de datos con el fin de resolver un problema similar al del presente trabajo de titulación, buscando suplir algunas limitaciones que, según su autor, existen en la metodología CRISP-DM (la cual también se utiliza como referencia en el presente trabajo).

La Figura 2.11 representa las fases que conforman SMAM.

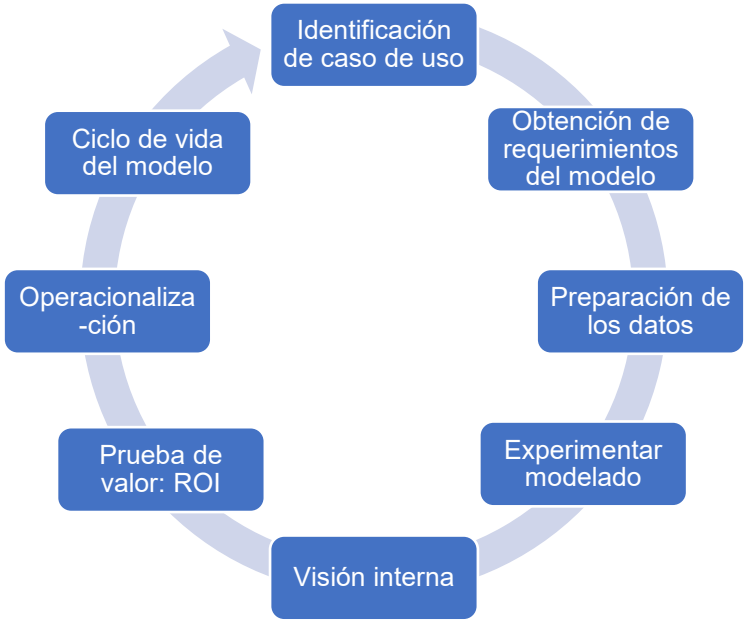

Figura 2.11. Descripción del proceso SMAM.

2.6.6. GB979

GB979 es el código que identifica al conjunto de documentos agrupados bajo la denominación "*Big Data Analytics Solution Suite*" de la empresa TM-Forum. La primera publicación de este conjunto de documentos se realizó en el año 2014 y la última actualización se realizó en junio y agosto de 2017 (Framework reléase 16.5 y 17, respectivamente). Los documentos que se agrupan bajo el código GB979 son: RN346 Big Data Analytics Release Notes, GB979 Big Data Analytics Guidebook, GB979A Big Data Analytics Use Cases, GB979B Big Data Analytics Building Blocks, GB979C Big Data Analytics Privacy Risk Score Details y GB979D Big Data Analytics Big Data Repository. Estos documentos no están disponibles libremente pues se requiere de una membresía en TM-Forum para poder obtenerlos [11].

El conjunto de documentos GB979 agrupa las mejores prácticas para Big Data de TM-Forum y las relaciona con su sistema Frameworx, que a su vez es el conjunto de mejores prácticas y estándares de TM-Forum para la operación efectiva y eficiente del negocio de un proveedor de servicios digitales y su respectivo ecosistema.

GB979 se desarrolla alrededor del documento denominado "Analytics Guidebook" donde consta el modelo de referencia de analítica Big Data, y se complementa con otros documentos que exponen, entre otras cosas, casos de uso del modelo, el valor que genera para el negocio y una lista de métricas que podrían utilizarse en el análisis de los datos.

TM-Forum ha publicado cada 6 meses, desde la primera divulgación de los documentos GB979, nuevas versiones de los mismos, por lo que son la referencia más actualizada que existe para guiar un proyecto de analítica.

La Figura 2.12 muestra el modelo de referencia en el que se basa la documentación GB979. Dicho modelo de referencia expone diferentes bloques con consideraciones que deben tomarse en cuenta al momento de desarrollar soluciones de analítica de datos en un proveedor de servicios digitales. Los bloques centrales son:

a) El Análisis de datos, que incluye temas como modelado de los datos, gestión de eventos complejos, alertas y reportes.
b) La Gestión de datos, que incluye temas como la transformación, la correlación y la manipulación de los datos.
c) La ingesta de datos, que hace referencia a la integración, importación y formateo de los datos.

Los bloques mencionados interactúan con los datos que se encuentran disponibles en un Repositorio de datos. Estos repositorios son alimentados con datos que provienen de varias fuentes, como, por ejemplo, de una red de telecomunicaciones, de sistemas de soporte de operación y negocio (conocido como OSS/BSS) o de redes sociales.

De forma general, los datos deben ser gobernados con medidas de privacidad, seguridad y en conformidad con las leyes y con las políticas internas de la empresa. Finalmente, el resultado del análisis de los datos permite generar distintos tipos de aplicaciones, como, por ejemplo, reducción de gastos de inversión y operación, gestión de experiencia de usuarios, generación de ingresos, entre otros.

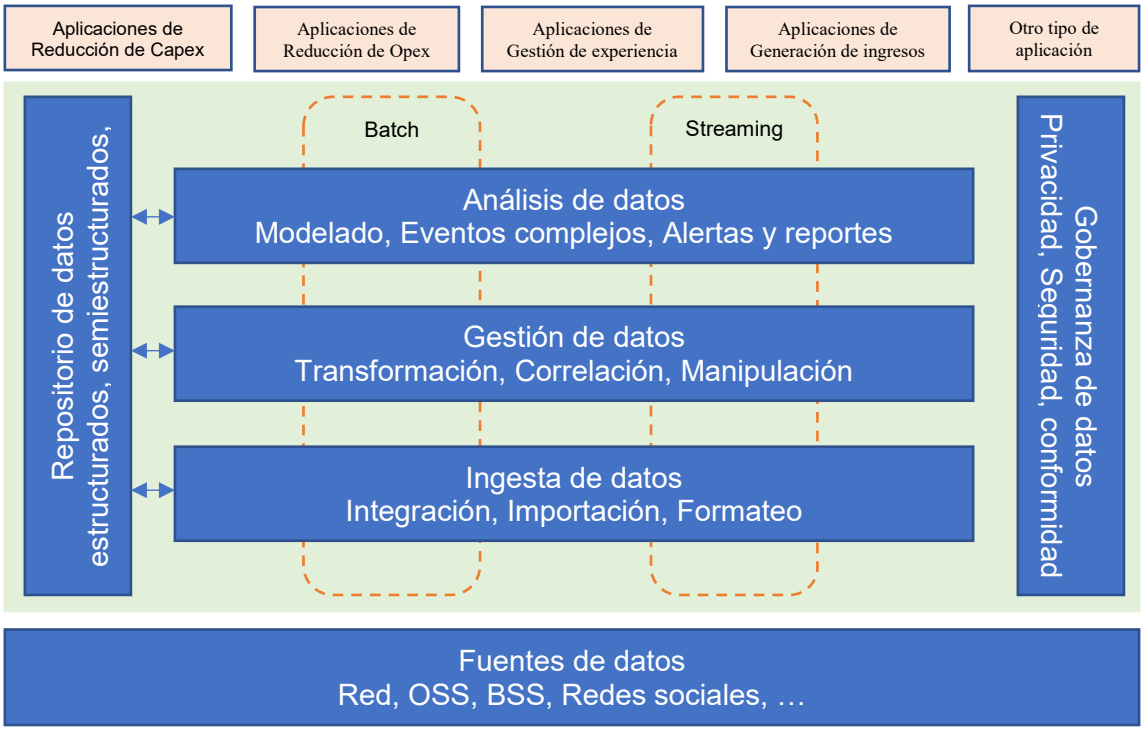

Figura 2.12. Modelo de Referencia de TM-Forum para Big Data Analytics

2.6.7. Características comunes

Si bien cada documento mencionado anteriormente aborda su tema con un enfoque particular y con distintos niveles de profundidad y detalle, todos coinciden en que la analítica de datos consiste en un proceso que es llevado a cabo mediante una serie de pasos, fases, actividades o tareas secuenciales iterativas. Por esta razón, se ha recopilado mediante una tabla comparativa los diferentes componentes principales de cada proceso, encontrando los puntos de coincidencia y las características únicas de cada uno de ellos. Esta tabla se ha registrado como Anexo 1 al presente trabajo de titulación y se ha utilizado como una de las bases del proceso de diseño de la propuesta metodológica, tal como se explicará en la sección 3.3.

2.6.8. Aplicabilidad para el estudio y análisis de tráfico de redes de telecomunicaciones.

De la revisión realizada a los diferentes documentos presentados se pudo obtener las siguientes conclusiones:

- Los documentos CRISP-DM y ASUM-DM constituyen metodologías de analítica de datos que abarcan de forma complementaria temas relacionados con “gestión de proyectos” y “gestión, operación y mantenimiento de infraestructura y soluciones tecnológicas”. Estos documentos exponen las actividades que conforman sus respectivas metodologías con suficiente detalle como para conformar también guías de usuario.

- Los documentos KDD process, SEMMA y SMAM no constituyen metodologías de analítica de datos, pero si describen en mayor o menor detalle procesos de analítica de datos. El nivel de detalle no permite que estos documentos conformen guías de usuario.

- Ninguno de los documentos CRISP-DM, ASUM-DM, KDD Process, SEMMA o SMAM fueron elaborados pensando en el sector de las telecomunicaciones o en el estudio y análisis de tráfico de redes de telecomunicaciones.

- El documento GB979 si fue elaborado pensando en el sector de las telecomunicaciones, pero no constituye una metodología de analítica de datos propiamente hablando. El documento GB979 es un conjunto de documentos que busca guiar la implementación de analítica de datos Big Data en proveedores de servicios de comunicación desde un enfoque de alto nivel, es decir, sin entrar en detalles de cómo realizar la implementación sino abordando de forma macro los componentes que se deben considerar.

- GB979 aporta de forma significativa con la elaboración de un documento que recopila los casos de uso que podrían emplear analítica de datos Big Data. Sin embargo, por ser casos de uso de alto nivel (por ejemplo, monitoreo del mercado, gestión de fraude, planificación de capacidad de red), no se menciona temas especializados como la analítica de los datos que cursan por la red de telecomunicaciones.

Considerando estos factores, la propuesta metodológica que se busca entregar mediante el presente trabajo de titulación será un aporte importante al sector de las telecomunicaciones.

3. METODOLOGÍA

El desarrollo del trabajo de titulación se realizó siguiendo la metodología "*Design Science Research in Information Systems*" (DSRIS).

La metodología DSRIS fue elaborada para ser aplicada en el campo tecnológico (sistemas de información) en la búsqueda de nuevo conocimiento mediante la creación de artefactos innovadores [63]. DSRIS considera que una metodología puede ser desarrollada con el fin de resolver un problema y su aplicación puede encontrarse en los campos de la ingeniería, de la computación [64] y de los negocios [22]. Puesto que el presente trabajo de titulación busca realizar una propuesta metodológica a ser aplicada en el sector tecnológico de las telecomunicaciones, que utilizará sistemas de información como soporte y que tendrá como fin solventar una actual carencia metodológica en dicho sector, la utilización de DSRIS resultó ser una guía válida de referencia.

La metodología DSRIS empleada define las siguientes 6 actividades [63]: 1.- Identificación del problema y motivación, 2.- Definición de objetivos para la solución del problema, 3.- Diseño y desarrollo, 4.- Demostración, 5.- Evaluación y 6.- Comunicación. En el caso del presente trabajo se decidió reemplazar la actividad "Comunicación" definida en [63] por la actividad denominada "Conclusión" en [64]. Esta adaptación de la metodología se realizó debido a que la actividad "Comunicación" definida en [63] consiste en difundir el trabajo realizado mediante diferentes vías paralelas, tanto durante su desarrollo como después de ser concretado, actividad que en el presente caso se realiza exclusivamente mediante este documento. La actividad "Conclusión" definida en [64] consiste, en cambio, en consolidar y resumir los resultados de un trabajo, tarea que es necesaria para la redacción de este documento y que, por lo tanto, resulta ser más adecuada para el presente caso.

Las actividades definidas en esta metodología se pueden agrupar en las siguientes cuatro fases: "Fase teórica", "Fase de diseño, análisis o implementación metodológica", "Fase de simulación y/o implementación" y "Fase de validación, análisis de resultados o pruebas de funcionamiento".

La fase teórica comprende a la actividad 1 de la metodología DSRIS. En esta actividad, denominada "Identificación del problema y motivación", se definirá de forma específica el problema tratado en el trabajo de titulación, se propondrá una solución y se justificará el valor de dicha solución.

La fase de diseño, análisis o implementación metodológica comprende a las actividades 2 y 3 de la metodología DSRIS. En la actividad 2, denominada “Definición de objetivos para la solución del problema”, se definirá qué características y objetivos se debe incluir en el diseño de la solución propuesta. En la actividad 3, denominada “Diseño y desarrollo”, se creará un artefacto que, en el caso de este trabajo de titulación, corresponde a una metodología con métodos, técnicas y herramientas.

La fase de simulación y/o implementación comprende a la actividad 4 de la metodología DSRIS denominada “Demostración”. En esta actividad ayudará a probar que el uso de la solución propuesta puede ayudar a resolver una o más instancias del problema de acuerdo a lo identificado en las actividades 1 y 2.

La fase de validación, análisis de resultados o pruebas de funcionamiento se ha recopilado en el capítulo final de este documento y comprende a las actividades 5 y 6 de la metodología DSRIS. En la actividad 5, denominada “Evaluación”, se observará y medirá el grado de efectividad que se tuvo al emplear la metodología propuesta al durante el proceso de evaluación realizado en la actividad 4. En la actividad 6, denominada “Conclusión”, se consolidará el resultado obtenido y las lecciones aprendidas en el proceso.

El proceso aquí descrito se ha resumido y representado en la Figura 3.1.

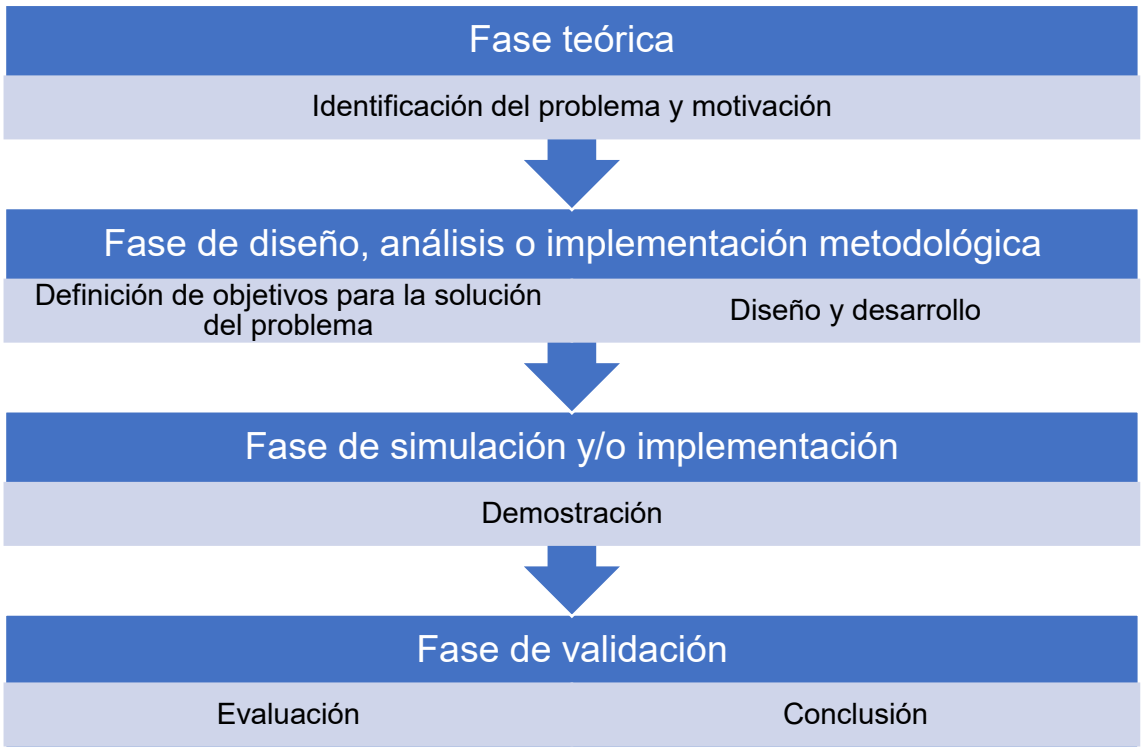

Figura 3.1: Proceso de desarrollo de la propuesta metodológica.

3.1. Identificación del problema y motivación.

En la introducción y el marco teórico de este trabajo se realizó una exposición de la situación actual que está atravesando tanto la sociedad como la industria de las telecomunicaciones. Esta situación puede resumirse de la siguiente forma: Estamos atravesando procesos de cambio que ponen en riesgo la posición de liderazgo que ostentan las empresas de telecomunicaciones en la economía.

Adicionalmente, se explicó que el interés que tienen las empresas de telecomunicaciones es el de mantener su actual posición de liderazgo en la economía, que su objetivo es incrementar el valor que ofrecen al mercado a través de sus servicios y que su estrategia consiste en adaptar sus modelos de negocio aprovechando la posición de ventaja que tienen en el mercado por disponer de una infraestructura que transporta una gran cantidad de datos con un valor potencial enorme. El principal reto que afrontan estas empresas para satisfacer su interés es encontrar formas eficientes de explotar sus datos, algo que implica superar los desafíos y servirse de las ventajas que se expusieron en la sección 2.4.

Una forma de explotar los datos de las redes de telecomunicaciones es hacer uso de las tecnologías Big Data y Analítica de Datos, expuestas en la sección 2.2. En lo que respecta a la analítica de datos, hemos visto en la sección 2.6 que existen varios documentos que buscan guiar el desarrollo de proyectos de este tipo, pero que ninguno de ellos constituye una metodología detallada de analítica de datos que haya sido elaborada considerando las características específicas del sector de las telecomunicaciones.

Por lo tanto, este trabajo pretende contribuir con el problema que genera la carencia de una metodología avanzada que permita realizar proyectos de analítica de datos aplicable al estudio y análisis de tráfico de redes en el sector de las telecomunicaciones y que incluya técnicas y herramientas aplicables a este tipo de proyectos.

Realizar una propuesta que supla esta carencia constituirá un aporte valioso para la industria y, dada la situación actual, resulta ser necesaria para afrontar los cambios que estamos viviendo como sociedad y como economía.

3.2. Definición de objetivos para la solución del problema.

Para solucionar el problema descrito, se ha decidido realizar una “propuesta metodológica para analítica de datos para estudio y análisis de tráfico en redes de telecomunicaciones”.

Siguiendo las definiciones recopiladas en la sección 2.1, se puede definir a este trabajo de titulación como esfuerzo por armonizar un conjunto de métodos y procedimientos que permitan realizar una investigación científica con el fin de estudiar y analizar el tráfico en redes de telecomunicaciones. Estos métodos se organizarán en forma de un proceso iterativo y sistemático que abarque varias disciplinas, modelos, técnicas, procedimientos, recursos, herramientas e instrumentos con el fin de permitir o facilitar el entendimiento, aprendizaje o desarrollo de habilidades relacionadas con el tráfico que cursa por dichas redes de telecomunicaciones.

La propuesta desarrollada en este trabajo se estructurará de forma que cumpla con los siguientes objetivos:

- Considerar los desafíos impuestos por el sector de las telecomunicaciones, descritos en la sección 2.4.1.
- Considerar las ventajas que ofrecen las empresas de telecomunicaciones, descritas en la sección 2.4.2.
- Incluir actividades que cubran todas las interacciones descritas en la sección 2.5.
- Usar como base la información provista por las metodologías, tanto maduras como recientes, presentadas en la sección 2.6 y que, a manera de recordatorio, se listan en la Tabla 3.1:

Tabla 3.1. Resumen de referencias expuestas en la sección 2.6.

Nombre	Autor	Año publicación	Sector de aplicación objetivo
KDD Process	AAI	1996	Uso general
SEMMA	SAS	1997	Uso general
CRISP-DM	CRISP-DM	1996	Uso general
ASUM-DM	IBM	2015	Uso general
SMAM	Olav Laudy	2015	Uso general
GB979	TM-Forum	2017	Proveedores de servicios digitales

3.3. Diseño y desarrollo.

3.3.1. Proceso utilizado.

El diseño y desarrollo se realizó mediante un proceso iterativo que incluyó cuatro fases, representadas en la Figura 3.2: selección de actividades, mapeo de características, estructuración de la metodología y desarrollo de la metodología.

Figura 3.2. Fases del proceso iterativo de diseño de la propuesta metodológica

A continuación, se explica cada paso del proceso y los resultados obtenidos.

3.3.2. Selección de actividades

La sección 2.6 expuso varios documentos de referencia que pueden guiar un proyecto o proceso de analítica de datos. Cada uno de estos documentos propone una lista de actividades a ejecutarse. Con el fin de estructurar la propuesta metodológica correspondiente a este trabajo de titulación, se realizó un análisis comparativo de estas actividades, lo que permitió determinar que existen actividades equivalentes o similares a lo largo de todos los documentos, así como actividades que solo son propuestas en algunos de ellos. El resultado de esta comparación se consolidó de forma tabular en el Anexo 1.

Una vez que se encontró las equivalencias y particularidades del conjunto de documentos analizados, se identificó que todas las actividades pueden agruparse dentro de cuatro dimensiones, obteniendo los datos recopilados en la Tabla 3.2.

Las dimensiones identificadas se han denominado “Analítica”, “Gestión”, “Infraestructura” y “Proyectos”.

- Analítica: Esta dimensión agrupa todas las actividades que están relacionadas con un proceso de analítica de datos.
- Gestión: Esta dimensión agrupa todas las actividades que están relacionadas con la gestión, operación y mantenimiento de un producto, servicio o solución tecnológica.
- Infraestructura: Esta dimensión agrupa todas las actividades que están relacionadas con el diseño, ingeniería, preparación, instalación y configuración

de la infraestructura que se usará para implementar un producto, servicio o solución tecnológica.

- Proyectos: Esta dimensión agrupa todas las actividades que están relacionadas con la formulación, planificación, ejecución, control y evaluación de proyectos.

Tabla 3.2. Cantidad de actividades por dimensión en cada documento referencial

Dimensión	Total	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Analítica	38	10	5	25	36	6	1
Gestión	49	1	0	3	46	2	3
Infraestructura	30	0	0	0	30	0	1
Proyectos	125	0	0	2	125	0	0
Total	**242**						

En total se detectó 242 actividades y tareas distintas, de las cuales 125 pertenecen a la dimensión Proyectos, 49 pertenecen a la dimensión Gestión, 30 pertenecen a la dimensión Infraestructura y 38 pertenecen a la dimensión Analítica.

La metodología ASUM-DM, si bien es la guía de referencia más extensa y detallada, es la única que abarca en gran detalle todas las dimensiones. Esto tiene sentido dada la naturaleza de esta metodología, la cual se explicó en la sección 2.6.4.

Las metodologías CRISP-DM y ASUM-DM, por su procedencia y nivel de detalle, abarcan el proceso de analítica de datos de forma bastante extensa y completa, cubriendo casi todas las actividades y tareas que son propuestas en los otros documentos en relación con la dimensión correspondiente. Los procesos KDD y SMAM aportan de forma adicional con actividades relacionadas con la detección y análisis de patrones de los datos obtenidos a partir del modelado de los mismos. El proceso KDD y el modelo GB979 aportaron de forma adicional con actividades relacionadas con la definición de alertas o eventos que requieran de acciones por parte de la compañía. El modelo GB979 y el proceso SMAM aportaron de forma adicional con el concepto de “casos de uso”.

Considerando que la propuesta metodológica que se realiza en el presente trabajo de titulación será aplicado al sector de las telecomunicaciones y que la gestión de proyectos, la gestión de infraestructura y la gestión, operación y mantenimiento de productos y servicios forman parte de las áreas de dominio de este tipo de empresas, se decidió que solamente se deben abordar en la propuesta metodológica actividades relacionadas con la analítica de datos, y que para el resto de dimensiones se deberá

referenciar a las mejores prácticas de la industria que sean pertinentes a cada dimensión y que sean utilizadas por cada empresa, de acuerdo a lo expuesto en la sección 2.4.2.

3.3.3. Mapeo de requerimientos

Una vez que se identificaron las actividades de los documentos referenciales presentados en la sección 2.6 a conservarse en la propuesta metodológica de este trabajo de titulación, el siguiente paso consistió en determinar si la lista de interacciones levantada en la sección 2.5 puede ser mapeada en su totalidad dentro de las actividades identificadas previamente o si es necesario agregar nuevas actividades a la lista. El resultado del mapeo se muestra en la Tabla 3.3, donde se especifica para cada una de las interacciones levantadas si se requirió o no definir nuevas actividades.

Tabla 3.3. Mapeo de interacciones

Interacción necesaria	Requerimiento adicional
Generar y/o capturar tráfico.	Nuevas actividades requeridas
Generar y recolectar datos estadísticos.	No
Identificar, clasificar y/o filtrar el tráfico.	Nuevas actividades requeridas
Anonimizar y/o censurar el contenido del tráfico.	Nuevas actividades requeridas
Analizar y visualizar el tráfico capturado o las estadísticas generadas.	Nuevas actividades requeridas
Modelar el tráfico capturado o las estadísticas generadas.	No
Transformar, reorganizar, estructurar o pre-procesar el contenido del tráfico.	No
Complementar o completar el tráfico o las estadísticas recolectadas.	No
Correlacionar varias fuentes de tráfico.	No

Posteriormente se analizó si las características del sector de las telecomunicaciones descritas en la sección 2.4 requieren de algún tipo de consideración especial con respecto a las actividades ya identificadas, obteniendo como resultado el mapeo que se indica en la Tabla 3.4, donde se especifica para cada una de las características identificadas si se requirió o no definir nuevas actividades.

Tabla 3.4. Mapeo de características del sector de telecomunicaciones

Características del sector de telecomunicaciones	Requerimiento adicional
Convergencia, neutralidad de la red y privacidad de los datos.	Nuevas actividades requeridas
Áreas de dominio de las empresas de telecomunicaciones.	Estructuración especializada de la metodología.

De esta forma, se identificó que existen cuatro tipos de interacciones que deben ser consideradas con especial atención dentro de procesos de analítica de datos en el sector de las telecomunicaciones. Estas interacciones tienen relación con las características del sector de telecomunicaciones que se expuso en la sección 2.4, y son 1) la generación y/o captura de tráfico, 2) la identificación, clasificación y/o filtrado de tráfico, 3) el proceso de anonimizar y/o censurar el contenido de tráfico, y 4) el proceso de analizar y visualizar el tráfico capturado.

Estas interacciones resultan especiales pues no son consideradas dentro de las guías estudiadas en la sección 2.6 y representan desafíos particulares que se deben tratar dentro de la metodología. De igual forma se identificó que las características propias del sector de telecomunicaciones hacen necesario agregar actividades relacionadas con la evaluación del entorno legal, regulatorio y político interno que afectarán la implementación de soluciones de analítica de datos sobre el tráfico de las redes de telecomunicaciones, y con el estudio previo de las características técnicas de dicho tráfico.

El resultado de las fases de selección y mapeo se muestra en la siguiente lista de actividades que integrarán la metodología propuesta. La estructura y la mayor parte de actividades provienen de la metodología ASUM-DM y están alineadas con la metodología CRISP-DM. En cursiva se ha destacado las actividades que se tomaron de otros documentos de referencia y en negrita-cursiva se ha destacado las actividades que se han agregado como parte del aporte propio del presente trabajo de titulación en base a lo expuesto en el Marco Teórico:

1. Comprender el negocio
   - 1.1. Determinar los objetivos del negocio
     - 1.1.1. Compilar el trasfondo del negocio
     - 1.1.2. Identificar los objetivos del negocio
     - 1.1.3. Determinar los criterios de éxito del negocio
   - ***1.2. Evaluar el entorno***
     - ***1.2.1. Evaluar factores legales y regulatorios***
     - ***1.2.2. Evaluar políticas y códigos internos***
   - 1.3. Evaluar la situación
     - 1.3.1. Levantar un inventario de recursos
     - 1.3.2. Determinar requerimientos, supuestos y restricciones
     - 1.3.3. Compilar un glosario de términos
   - 1.4. Determinar los objetivos del proceso de analítica de datos
     - *1.4.1. Revisar casos de uso previamente desarrollados.*
   - 1.5. Crear un reporte de entendimiento del negocio
2. Comprender los datos
   - ***2.1. Levantar información referencial***
     - ***2.1.1. Levantar topología de red***
     - ***2.1.2. Levantar y documentar pila de protocolos***
     - ***2.1.3. Levantar lógica de servicios***

2.2. Recolectar muestra de datos iniciales
2.3. Describir los datos
2.4. Explorar los datos
2.5. Verificar la calidad de los datos
2.6. Crear reporte de entendimiento de los datos

3. Preparar los datos
   3.1. Seleccionar los datos
   3.2. Capturar datos
   3.3. Clasificar y filtrar datos
   3.4. Proteger el contenido de los datos
   3.5. Limpiar los datos
   3.6. Construir los datos
   3.7. Integrar los datos
   3.8. Dar formato a los datos
   3.9. Crear reporte de preparación de los datos
4. Construir los modelos
   4.1. Determinar los requerimientos del modelo
   4.2. Seleccionar las técnicas de modelado
   4.3. Generar el diseño de pruebas
   4.4. Elaborar los modelos
   4.5. Experimentar con los modelos
   4.6. Valorar los modelos
   4.7. Crear el reporte de construcción de modelos
5. Evaluar los resultados
   5.1. Buscar patrones
   5.2. Interpretar patrones
   5.3. Definir eventos o alertas
   5.4. Evaluar los resultados
   5.5. Revisar el proceso
   5.6. Determinar los siguientes pasos
   5.7. Retroalimentar los casos de uso
   5.8. Crear el reporte de evaluación de los modelos

La descripción detallada de cada una de las tareas mencionadas se encuentra disponible en el Anexo 2, de acuerdo con lo explicado en las siguientes dos secciones.

3.3.4. Estructura del modelo de referencia de la metodología propuesta.

El siguiente paso de diseño que se realizó fue la estructuración de la metodología. Si bien la metodología abarcará solamente la descripción de actividades relacionadas con analítica de datos, no sería adecuado dejar por fuera de la estructuración metodológica las dimensiones de Gestión, Infraestructura y Proyectos. Estas dimensiones son necesarias debido a que el trabajo metodológico deberá organizarse en forma de un proyecto, y seguramente requerirá la puesta en producción de una solución tecnológica de analítica de datos, la cual requerirá de la gestión de su infraestructura subyacente y sufriría el ciclo de vida de un producto o servicio, por lo que requerirá ser gestionada, operada y mantenida. Es necesario por tanto definir cómo y cuándo interactúan estas dimensiones en relación con las actividades de analítica de datos.

Así, la propuesta metodológica contendrá las cuatro dimensiones mencionadas, y para cada una de ellas se referenciará a guías de implementación. Para las dimensiones Gestión, Infraestructura y Proyectos se mencionará a las mejores prácticas de la industria, y para la dimensión Analítica se realizará una propuesta de guía de usuario.

Con el fin de estructurar correctamente estas cuatro dimensiones, se decidió seguir el ejemplo de las metodologías CRISP-DM y ASUM-DM, definiendo un modelo de proceso jerárquico de tres niveles de abstracción. El primer nivel de abstracción contiene a cada una de las cuatro dimensiones identificadas previamente. El segundo nivel alberga las etapas y fases de desarrollo correspondientes a cada dimensión. El tercer nivel hace referencia a las guías que podrían usarse para organizar el trabajo respectivo.

El modelo jerárquico resultante se muestra en la Figura 3.3.

Dimensión Analítica					Dimensión Infraestructura					Dimensión Gestión				
Comprensión del Negocio	Comprensión de los Datos	Preparación de los datos	Modelado	Evaluación	Diseño de solución tecnológica	Instalación de infraestructura	Instalación de software	Prueba y validación de solución	Entregar solución	Lanzamiento	Desarrollo	Madurez	Declive	Retiro
Guía de usuario propuesta					ITIL, eTOM, ISO 20000, ISO 27001 e ISO 9001 Suplemento 40 de la serie Y de recomendaciones ITU									

Dimensión Proyecto														
Formulación					Implementación					Evaluación ex – post				
Identificación	Formulación	Evaluación ex – ante	Diseño e Ingeniería	Decisión de Inversión	Inicio	Planificación	Ejecución	Control	Cierre	Revisión	Análisis	Comparación	Lecciones aprendidas	Toma de decisiones
MoP, ISO 2505, Políticas internas, metodología CEPAL					Políticas internas, PMBOK, PRINCE2, ISO21500					MoP, ISO 2505, Políticas internas, metodología CEPAL				

Figura 3.3. Niveles de abstracción y fases del modelo de referencia propuesto.

En lo que respecta a las etapas y fases de cada dimensión, para aquella denominada Gestión se realizó una relación con el ciclo de vida de un producto o servicio, el cual

está constituido por las etapas de “Lanzamiento”, “Desarrollo”, “Madurez”, “Declive” y “Retiro” (adaptado de [65]). Para la dimensión Infraestructura se decidió establecer etapas de “Diseño de la solución tecnológica”, “Instalación de infraestructura”, “Instalación de software”, “Prueba y validación de solución” y “Entrega de solución”. La organización aquí detallada responde a una secuencia lógica requerida para desplegar una solución tecnológica, resume de forma simplificada las actividades relacionadas con Gestión e Infraestructura encontradas en las referencias de la Sección 2.6 y se acopla de forma simple a la dimensión Proyectos que se describe más adelante.

Estas dos dimensiones, sin embargo, comparten en gran medida actividades que están asociadas a las áreas de dominio de las empresas de telecomunicaciones y cuentan con las mejores prácticas de la industria recopiladas en documentos como ITIL, eTOM, ISO 20000, ISO 27001 e ISO 9001, entre otras. Adicionalmente, en lo que respecta a la infraestructura Big Data, tal como se expuso en la introducción a este trabajo de titulación, existen varios organismos internacionales de estandarización que están trabajando sobre este tema en la actualidad. Un documento clave de referencia es el suplemento 40 de la serie Y de recomendaciones de la ITU, ya que este documento corresponde a una hoja de ruta de los estándares relacionados con el ecosistema Big Data para el sector de las telecomunicaciones. Es por tanto altamente probable que cada empresa de telecomunicaciones decida gestionar estas dos dimensiones de acuerdo a las mejores prácticas de su elección y que, en la práctica, se gestionen de forma conjunta basándose, por ejemplo, en el modelo de Estrategia de Servicios de ITIL y en el Frameworx de TM-Forum, los cuales son altamente especializados, detallados y complejos, por lo que su incorporación en el modelo aquí desarrollado dificultaría el entendimiento de la propuesta metodológica planteada.

La dimensión Proyectos, mencionada previamente, se dividió en tres etapas: Formulación del proyecto, Implementación del proyecto y Evaluación Ex Post del proyecto [66]. Esta dimensión también corresponde a una de las áreas de dominio de las empresas de telecomunicaciones, las cuales se apoyan en las mejores prácticas de la industria recopiladas en documentos como PMBOK, PRINCE2 e ISO 21500.

- La etapa de Formulación del proyecto está constituida a su vez por las fases: “Identificación”, “Formulación”, “Evaluación ex ante”, “Diseño e ingeniería” y “Decisión de inversión”.
- La etapa de Implementación del proyecto está constituida por las fases “Inicio”, “Planificación”, “Ejecución”, “Control” y “Cierre”.

- La etapa de Evaluación Ex Post de proyecto está constituida por las fases “Revisión”, “Análisis”, “Comparación”, “Lecciones aprendidas” y “Decisión”.

La dimensión “Analítica de datos”, corresponde a las actividades que están directamente relacionadas con la materia de este trabajo, y está dividida en cinco fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado y Evaluación del modelo. Estas etapas fueron recogidas de la metodología CRISP-DM, y cuentan con una detallada descripción de cada tarea en la Guía de Usuario que se elaboró en el Anexo 2.

Es importante recalcar que la dimensión Proyecto soporta a las dimensiones Analítica, Infraestructura y Gestión, y que, de forma más específica, la etapa de Formulación de proyecto soporta a la dimensión Analítica, la etapa de Implementación de proyecto soporta a la dimensión Infraestructura y la etapa de evaluación ex–post se desarrolla en paralelo a la dimensión Gestión. Esto se debe a que la dimensión analítica se encargará de identificar los requerimientos de un proyecto de analítica de datos, de formular una propuesta de solución, de diseñar los modelos a implementar y de tomar la decisión de realizar o no la implementación. Los modelos generados se implementarán en la infraestructura que soporte la solución final de analítica de datos y dicha solución deberá ser evaluada una vez que entre en operación y se encuentre atravesando el ciclo de vida de productos o servicios.

3.3.5. Desarrollo de la propuesta

Una vez estructurado el modelo de referencia, el siguiente paso consistió en desarrollar a detalle cada una de las actividades que conforman la metodología. Debido a que dicho desarrollo consiste simplemente en describir el modelo de referencia expuesto en la Sección 3.3.4 y cada una de las tareas presentadas en la Sección 3.3.3, y considerando que dicha descripción da como resultado la totalidad de la Metodología elaborada en este trabajo, se decidió consolidar dicho resultado en el Anexo 2, estructurado a manera de un modelo de referencia y una guía de usuario, que servirá como documento referencial independiente facilitando así su aplicación práctica por cuanto sintetiza el resultado de este trabajo sin entrar en los detalles teóricos y fundamentos de referencia que son a su vez, expuestos a lo largo del documento principal. La guía de usuario contiene una descripción específica de cada una de las actividades y tareas que integran el modelo de referencia. Cada actividad y cada tarea, cuando aplica, incluye una lista de métodos, técnicas y herramientas que pueden usarse para el estudio y análisis de tráfico de redes de telecomunicaciones.

3.4. Demostración.

Una vez que la propuesta metodológica fue definida, estructurada y desarrollada, correspondió avanzar con la actividad de demostración, que no solo ayuda a probar que la propuesta es aplicable para el fin planteado, sino que también ayuda a entender de mejor manera cómo se debe aplicar la metodología elaborada.

Para esta actividad se utilizó el modelo referencial y la guía de usuario elaborados para realizar el estudio y análisis de tráfico extraído de una red de telecomunicaciones real.

El resultado metodológico se presenta a continuación a través de los reportes que se generan al finalizar cada fase de la metodología analítica propuesta. Se recomienda que, antes de proceder con la lectura de dichos reportes, se revise la metodología resultante que se desarrolló en este trabajo y que se encuentra disponible de forma íntegra en el Anexo 2 a este documento, con el fin de facilitar el entendimiento del contenido de los reportes.

3.4.1. Reporte de entendimiento del negocio

El siguiente reporte abarca el resultado de cumplimiento de las siguientes actividades de la propuesta metodológica:

1.1. Determinar los objetivos del negocio

1.2. Evaluar el entorno

1.3. Evaluar la situación

1.4. Determinar los objetivos del proceso de analítica de datos

1.5. Crear un reporte de entendimiento del negocio

Objetivos del negocio:

Para el presente caso se puede plantear que la elaboración de la propuesta metodológica del presente trabajo de titulación corresponde al “negocio” que debe ser considerado en la primera fase de la metodología analítica. El trasfondo corresponde a la expuesto en los capítulos de Introducción y Marco Teórico, así como a la sección 3.1 del presente documento. Los objetivos del negocio son los ya enunciados en la sección 3.2 y por tanto el criterio de éxito desde el punto de vista del “negocio” que se deberá manejar para evaluar el resultado de la demostración será: “que se haya demostrado que la utilización de la propuesta metodológica elaborada efectivamente permite realizar el estudio y análisis de tráfico de telecomunicaciones a través de la ejecución de los

pasos planteados y de los métodos, técnicas y herramientas incluidas en la Guía de Usuario del Anexo 2.

Situación actual:

En este caso, el "negocio" no cuenta con una red de telecomunicaciones propia, razón por la cual es necesario obtener una fuente de datos para realizar un proceso de analítica de datos en base a la propuesta metodológica planteada.

Esta situación permite delimitar el alcance a considerar durante el desarrollo de la actividad: el trabajo deberá avanzar solamente dentro de la dimensión analítica del modelo de referencia propuesto en el trabajo ya que no se realizará ningún tipo de implementación de solución tecnológica. No es necesario por tanto desarrollar actividades relacionadas con las dimensiones Infraestructura, Gestión o Proyectos.

Evaluación del entorno:

Para la selección de la fuente de datos, se consideró, como una opción, trabajar con alguna de las empresas de telecomunicaciones ecuatorianas de forma que se pueda tener acceso a sus datos e infraestructura. Considerando que las redes de telecomunicaciones transportan datos personales de sus usuarios, que la Ley Orgánica de Telecomunicaciones del Ecuador obliga a las empresas de telecomunicaciones a garantizar la privacidad e intimidad de sus usuarios y que el presente trabajo será de libre acceso al público, se determinó que emplear esta opción podría generar grandes demoras en la aprobación de las actividades necesarias para su realización. Por lo tanto, se decidió trabajar con tráfico que provenga de una red de telecomunicaciones real pero que pueda ser accedido y analizado de forma pública, de acuerdo con lo detallado en la siguiente sección del reporte.

Evaluación de la situación:

Dada la situación actual descrita y el entorno en el que se debe desarrollar esta actividad, se buscó fuentes de datos apropiadas con las que se pudiera trabajar.

La organización CAIDA, que fue mencionada en la sección 2.5, provee muestras reales anonimizadas de tráfico de telecomunicaciones para fines de investigación tanto de forma pública como restringida. Estos conjuntos de datos cumplen con los requisitos de accesibilidad, legalidad y origen necesarios para poder avanzar con la demostración de la propuesta metodológica planteada.

El conjunto de datos seleccionado para ser utilizado en este trabajo fue el denominado como "The CAIDA UCSD Anonymized OC48 Internet Traces 2002-2003" [67].

Este conjunto es de acceso público y contiene tráfico real, truncado a 48 bits (es decir sin *payload* y por tanto sin datos personales de los usuarios), anonimizado con Crypto-Pan y capturado en el año 2003 por la organización CAIDA. El tráfico provino de un enlace de *peering* bidireccional tipo OC48, de aproximadamente 2.5 Gbps de capacidad, perteneciente a un ISP grande de la costa oeste de Estados Unidos de América. Los datos pueden ser utilizados para fines de investigación, incluyendo el análisis de las características del tráfico, su distribución topológica y geográfica, así como el volumen y duración de sus flujos, de acuerdo al documento de acuerdo de uso denominado "*CAIDA Acceptable Use Agreement (AUA) for Publicy Accessible Datasets*", que se debe aceptar para acceder a los datos públicos provistos por CAIDA y disponible en [68]

Como políticas para el desarrollo del análisis de los datos se estableció:

- "no extraer o exponer en el trabajo datos o información personal que pudiera haberse filtrado en el archivo provisto por CAIDA".
- "mantenerse dentro de las situaciones permitidas descritas en el portal de CAIDA y que corresponden a la investigación de las características de tráfico en cuanto a su distribución topológica y geográfica, así como volumen y duración de flujos".

Los archivos de trazas descargados fueron los siguientes: oc48-mfn.dirA.20030424-070000.UTC.anon.pcap y oc48-mfn.dirB.20030424-070000.UTC.anon.pcap, ambos en formato PCAP (*Paquet CAPture*), con un tamaño total de 1.08 GBytes y con una duración de 5 minutos de tráfico.

El inventario de recursos a emplearse fue el siguiente:

- Computadora marca Dell, con sistema operativo Linux (distribución ArchLinux) de 64bits, 3GBytes de memoria RAM disponibles y procesador Intel Core i5 de cuatro núcleos de procesamiento.
- Herramientas de software: Wireshark, BRO, TCPDump, TCPFlow, TCPRewrite, GeoIPDatabase, PcapPlusPlus, R, Python.

Como requerimientos, y con el fin de no incurrir en costos no previstos para el desarrollo del trabajo, se definió que el proceso de analítica de datos debe realizarse utilizando exclusivamente los recursos inventariados y no puede afectar el desempeño del sistema operativo ni de su interfaz de usuario (entorno de escritorio) por cuanto la aplicación de la metodología propuesta no debería requerir el uso de grandes recursos, como por ejemplo aquellos provistos por el ecosistema Big Data, que no forma parte del alcance de este trabajo. Adicionalmente, el proceso de analítica de datos debería poder

implementarse de forma que se pueda automatizar la extracción de conocimiento de más muestras de datos sin mayor involucramiento humano.

Como supuestos, se consideró que las herramientas de software disponibles serían capaces de analizar el 100% de los paquetes almacenados en el archivo de captura con el que se trabajará.

Como restricciones, se estableció que no sería posible realizar mejoras en la computadora, ni expansiones en el código del software utilizado, ni hacer la adquisición de software adicional para completar el trabajo.

Glosario de términos:

A continuación, se presenta el glosario de términos importantes que surgieron a lo largo del trabajo de demostración:

- HDLC: High Speed Data Link Control. Protocolo de encapsulación de capa 2.
- Peering: enlace directo establecido entre dos empresas de telecomunicaciones donde ninguna de las dos empresas factura a su contraparte por el tráfico cursado hacia su red.
- Payload: Datos entregados por una capa superior para ser encapsulados y gestionados por la capa inferior que recibe los datos. El payload suele contener datos de usuario.

Objetivos del proceso de analítica de datos:

Alineado con lo definido en la evaluación del entorno, se estableció que el objetivo del proceso de analítica de datos será: “Obtener las características de tráfico en cuanto a su distribución topológica y geográfica, así como volumen y duración de flujos”.

Se considerará este proceso de analítica como el primer caso de uso del “negocio” por lo que no se dispone de referencias previas del “negocio” que analizar.

3.4.2. Reporte de entendimiento de los datos

El siguiente reporte abarca el resultado de cumplimiento de las siguientes actividades de la propuesta metodológica:

2.1. Levantar información referencial

2.2. Recolectar muestra de datos iniciales

2.3. Describir los datos

2.4. Explorar los datos

2.5. Verificar la calidad de los datos

2.6. Crear reporte de entendimiento de los datos

Levantamiento de información referencial:

Considerando la descripción de la fuente de datos realizada en el reporte de entendimiento del negocio, se obtuvo la topología representada en la Figura 3.4 y que muestra a dos ISPs interconectados por un enlace OC48 de aproximadamente 2.5 Gbps de capacidad.

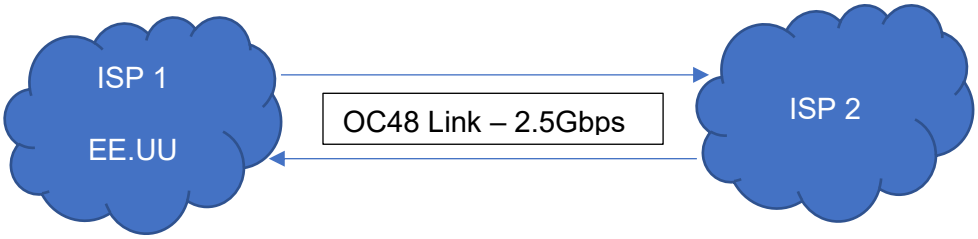

Figura 3.4. Topología de la red de telecomunicaciones a analizar.

Al tratarse de una conexión tipo *peering* entre ISPs, se espera que el tráfico que cursa por el enlace OC48 contenga la pila de protocolos TCP/IP. Dentro de este reporte no se realizará un levantamiento de cada uno de los protocolos ni de los servicios que se espera encontrar debido a la gran extensión de dicha información. Sin embargo, se espera encontrar de forma dominante tráfico que emplee el protocolo HTTP sobre el protocolo TCP, cuyas especificaciones se encuentran disponibles en los RFCs 2616 [69] y 793 [70] de la IETF (*The Internet Engineering Task Force*), respectivamente.

Muestras de datos iniciales:

Como ya se indicó, se obtuvieron a manera de muestra dos archivos en formato PCAP (*Paquet CAPture*), con un tamaño total de 1.08 GBytes y con una duración de 5 minutos de tráfico:

1. oc48-mfn.dirA.20030424-070000.UTC.anon.pcap; y
2. oc48-mfn.dirB.20030424-070000.UTC.anon.pcap.

Estos dos archivos contienen el tráfico de cada uno de los sentidos del enlace que se analizará. Con el fin de poder analizar correctamente el tráfico fue necesario realizar las siguientes transformaciones previas:

- Combinar los archivos de forma cronológica. Para esta transformación se utilizó la herramienta mergecap provista por el software Wireshark. A manera de

ejemplo, el comando siguiente permite combinar los archivos “origen1.cap” y “origen2.cap” en un nuevo archivo denominado “archivo_combinado.cap”: *mergecap -w archivo_combinado.cap origen1.cap origen2.cap*.

- Verificar que los paquetes combinados estén en orden cronológico. Para esta verificación se utilizó la herramienta reordercap provista por el software Wireshark. A manera de ejemplo, el comando siguiente permite ordenar cronológicamente los paquetes del archivo “archivo_combinado.cap” y almacenar el resultado en un nuevo archivo denominado “salida.cap”: *reordercap archivo_combinado.cap salida.cap*

Una vez combinados los flujos de paquetes, se probó si la herramienta Wireshark podía abrir y procesar el archivo combinado. El resultado fue que la memoria RAM de la computadora se saturó. Por esta razón, se procedió a almacenar los datos en varios archivos de menor tamaño:

- Dividir el flujo total en archivos de menor tamaño. Para esta transformación se utilizó la herramienta tcpdump. A manera de ejemplo, el comando siguiente permite dividir la información almacenada en el archivo “salida.cap” en archivos cuyo tamaño no excede 50Mbytes y cuyos nombres iniciarán con la palabra “bloque” y seguidos por un número que identifica la secuencia del archivo resultante: *tcpdump -r salida.cap -w bloque -C 50*

Descripción de los datos:

El proceso de división del flujo de datos en archivos de menor tamaño al original dio como resultado 24 archivos, 23 de 48,829 KBytes cada uno y 1 de 16,055 KBytes.

Una vez que se consiguieron muestras de tamaño manejable en la infraestructura disponible para realizar el trabajo, se pudo abrir uno de los archivos segmentados con la herramienta Wireshark.

Con el fin obtener una descripción general de los datos disponibles, se empleó la funcionalidad “*Capture File Properties”* de la herramienta Wireshark, la cual permite obtener información general resumida del archivo analizado como, por ejemplo, nombre del archivo, tamaño del archivo, encapsulación de los paquetes, fecha registrada del primer y último paquete almacenado en el archivo, y estadísticas simples del tráfico registrado.

A continuación, se muestra la información encontrada para el archivo décimo cuarto que se generó al momento de dividir el flujo total en archivos de menor tamaño:

File

Name:	oc48_part13.pcap
Length:	50 MB
Format:	Wireshark/tcpdump/... - pcap
Encapsulation:	Cisco HDLC
Snapshot length:	48

Time

First packet:	2003-04-24 02:02:51
Last packet:	2003-04-24 02:03:03
Elapsed:	00:00:12

Capture

Hardware:	Unknown
OS:	Unknown
Application:	Unknown

Interfaces

Interface	Dropped packets	Capture filter	Link type	Packet size limit
Unknown	Unknown	Unknown	Cisco HDLC	48 bytes

Statistics

Measurement	Captured	Displayed	Marked
Packets	797246	797246 (100.0%)	N/A
Time span, s	12.403	12.403	N/A
Average pps	64277.9	64277.9	N/A
Average packet size, B	439.5	439.5	N/A
Bytes	350711146	350711146 (100.0%)	0
Average bytes/s	28 M	28 M	N/A
Average bits/s	226 M	226 M	N/A

Figura 3.5. Descripción resumida de la muestra de datos analizada de forma manual

Exploración de los datos:

Una vez que se obtuvo información general descriptiva del archivo a analizar, se procedió a explorar sus datos utilizando funcionalidades más avanzadas de la herramienta Wireshark como, por ejemplo, "*Protocol Hierarchy*", "*Conversations*", "*Endpoints*", "*Expert Information*", entre otras. Al explorar la muestra descrita anteriormente, se pudo realizar las siguientes determinaciones:

- El método de encapsulación es HDLC para todos los paquetes.
- El 100% de los paquetes corresponden al protocolo de capa 3 IPv4. Más del 99% de los paquetes corresponden a los protocolos de capa 4 TCP y UDP.
- Se detectaron 55188 conversaciones IPv4, 44365 conversaciones TCP y 15837 conversaciones UDP. Por conversación se hace referencia a la totalidad de un

intercambio horizontal de información realizado entre dos protocolos correspondientes pertenecientes a entidades distintas. Una conversación IPv4 hace referencia al intercambio de datos entre dos direcciones IPv4 mientras que una conversación TCP o UDP hacer referencia al intercambio de datos entre dos sockets TCP o entre dos sockets UDP.

- La información personal de los usuarios ha sido removida pues cada paquete ha sido truncado en sus 48 Bytes iniciales retirando el *payload* de cualquier paquete que haya sido capturado. Las cabeceras de cada paquete han sido, sin embargo, conservadas en los 48 Bytes iniciales de los paquetes y permiten conocer el tamaño original de cada paquete.
- Para la Capa Aplicación del modelo OSI, se detectó una gran diversidad de protocolos para los cuales el analizador muestra constantemente errores en su estructura debido a que los paquetes fueron truncados al momento de capturarlos y la información necesaria para que el analizador pueda procesar correctamente su estructura no se encuentra disponible.

Verificación de calidad de los datos:

Todos los archivos de muestra generados pueden ser abiertos correctamente con Wireshark y analizados por lo que no se tuvo problemas de integridad al momento de descargar los archivos provistos por CAIDA.

El proceso de anonimización de los datos realizado por CAIDA previo a publicar los archivos analizados fue exitoso y no es posible recuperar a partir de los datos provistos información personal de ningún usuario. Dicho proceso también se encargó de modificar aleatoriamente, pero de forma congruente, las direcciones IPs de origen y de destino, por lo que resulta imposible rastrear a los usuarios que generaron el tráfico analizado. El proceso de anonimización eliminó por lo tanto datos importantes, lo cual genera errores en el software de análisis de protocolos y podría haber cambiado información necesaria para recuperar conocimiento verídico de la topología de la red y de los flujos de conversaciones.

3.4.3. Reporte de preparación de los datos

El siguiente reporte abarca el resultado de cumplimiento de las siguientes actividades de la propuesta metodológica:

3.1. Seleccionar los datos

3.2. Capturar datos

3.3. Clasificar y filtrar datos

3.4. Proteger el contenido de los datos

3.5. Limpiar los datos

3.6. Construir los datos

3.7. Integrar los datos

3.8. Dar formato a los datos

3.9. Crear reporte de preparación de los datos

Seleccionar los datos:

Una vez que se ha recolectado, descrito y explorado una muestra de datos inicial, corresponde seleccionar los datos definitivos con los que se realizará el proceso sistemático de analítica de datos. En el presente caso, los datos seleccionados corresponden a la totalidad de la muestra provista por CAIDA en los archivos descargados originalmente.

Considerando que la cantidad de información que se analizará es superior a lo que la herramienta utilizada anteriormente (Wireshark) puede gestionar con los recursos disponibles, y que el procesamiento de información se debe realizar de forma sistemática, se decidió emplear a partir de este momento la herramienta BRO [41], que es un sistema de análisis de tráfico de red potente, adaptable, eficiente, flexible, escalable y de libre acceso y utilización. BRO incorpora su propio lenguaje de programación, el cual permite desarrollar *scripts* con el fin de procesar grandes volúmenes de datos en línea, por lo que resulta adecuada para trabajar con los datos seleccionados.

Puesto que BRO solo soporta el procesamiento de protocolos TCP, UDP e ICMP sobre IP, el proceso de analítica se realizará solo para los paquetes que correspondan a estos protocolos. Si bien esto significa que no se podrá procesar el 100% de los paquetes, como se supuso originalmente, de acuerdo con lo observado en la sección 3.4.2 estos protocolos corresponden a más del 99% de los datos y por tanto permitirán extraer el conocimiento adecuado para estudiar de forma confiable el enlace OC48 y las redes interconectadas que están bajo análisis.

Capturar datos:

El proceso de captura fue realizado por CAIDA y los datos fueron descargados previamente de forma legal desde el portal web de dicha organización. Los archivos

donde se encuentra los datos capturados son los descritos anteriormente: oc48-mfn.dirA.20030424-070000.UTC.anon.pcap y oc48-mfn.dirB.20030424-070000.UTC.anon.pcap. Es importante recalcar que en la actividad de descripción y exploración de datos que se describe en el reporte previo se trabajó únicamente con una muestra de 50Mbytes de información extraída de los archivos citados, pero que en esta ocasión se trabajará con la información completa. El procesamiento sistemático de los datos se podría replicar a cualquier archivo similar provisto por CAIDA de forma que se extraiga información de forma continua a medida que se dispongan de nuevas capturas de la interfaz analizada.

Para casos reales de analítica de datos de tráfico de telecomunicaciones, la actividad de captura de datos correspondería a la extracción de datos en tiempo real de la red analizada, lo cual se debería realizar en coordinación con las áreas técnicas pertinentes, siguiendo todos los procedimientos y procesos adecuados de la empresa dueña de la infraestructura de telecomunicaciones y respetando el entorno legal y regulatorio que corresponda.

Transformar los datos:

Los pasos de construcción de datos fueron realizados previamente de forma parcial al momento de preparar las muestras de datos iniciales. Sin embargo, al intentar procesar los archivos con BRO se obtuvo un error indicando que no se soporta el protocolo de encapsulación de capa 2 HDLC. Por lo tanto, fue necesario hacer una transformación de encapsulamiento.

Se recopila por tanto a continuación todos los pasos que fueron necesarios seguir para construir los datos que posteriormente serían procesados. Considerando que la metodología propuesta es un proceso iterativo y sistemático, y que los siguientes pasos deberán aplicarse a cada archivo que se quiera procesar, se incluyen aquellos pasos realizados en fases previas y que, a final de cuentas, corresponden a esta actividad dentro de la metodología.

- Combinar los archivos de forma cronológica. Para esta transformación se utilizó la herramienta mergecap provista por el software Wireshark.
- Verificar que los paquetes combinados estén en orden cronológico. Para esta verificación se utilizó la herramienta reordercap provista por el software Wireshark.
- Reemplazar el protocolo de encapsulamiento HDLC por el protocolo IEEE 802.3 (utilizado en redes Ethernet). Para esta transformación se utilizó la herramienta

tcprewrite provista por el software tcpdump. Los protocolos HDLC e IEEE 802.3 son protocolos de capa 2 y, al momento de reemplazar un protocolo por otro, se pierde información correspondiente a su capa. Considerando, sin embargo, que el enlace analizado es un enlace punto a punto tipo *Peering* y que el tráfico ha sido anonimizado, el proceso de reemplazo a nivel de capa 2 de un protocolo por otro implica pérdida de información, pero no pérdida de conocimiento importante para el ejercicio en curso. A manera de ejemplo, el siguiente comando permite reemplazar el tipo de encapsulamiento de capa 2 del archivo "entrada.cap" para que corresponda al protocolo IEEE 802.3, y almacenar el resultado en el archivo "salida.cap": *tcprewrite –dlt=enet --enet-dmac= 00:11:12:13:14:15 --enet-smac=00:21:22:23:24:25 --infile=entrada.cap --outfile=salida.cap*

Estos pasos abarcan las actividades de Integración, Construcción y Formateo de datos.

Adicionalmente, como parte de la actividad de Integración de los datos, se obtuvo la última versión de la base de datos GeoIP [71] de la empresa MaxMind y se la incorporó al paquete de software BRO, con el fin de permitir determinar la información geo-referenciada de ubicación de las direcciones IP observadas durante el análisis de los datos.

Proteger el contenido de los datos:

Puesto que los datos que fueron capturados ya fueron anonimizados por CAIDA y no se conservó información de usuarios finales al eliminar los *payloads* de los paquetes y modificar aleatoriamente sus direcciones de origen y destino, no se requiere realizar un proceso de protección de contenido de los datos.

Sin embargo, si se estuviera trabajando con datos extraídos directamente de una red de telecomunicaciones sería necesario efectuar las acciones realizadas ya por CAIDA, modificando de forma aleatoria las direcciones IPs de los paquetes, y suprimiendo o alterando el *payload* de cada paquete antes de proceder con el análisis de los datos. Para realizar este trabajo, se puede recurrir a las herramientas descritas en la Sección 2.5.3. Resulta importante aclarar que, en un caso real, sería altamente probable que no se pueda anonimizar completamente el tráfico, puesto que, por ejemplo, casi siempre se conocerá cuál es la red que proveyó los datos, lo cual revela en cierta medida información sobre los usuarios finales.

Limpiar los datos:

Una vez construidos y transformados los datos, se procedió a procesarlos con BRO, con el comando *bro -r archivo.pcap*. No se obtuvo ningún error que indicara que hubiera datos que tuvieran que ser retirados del archivo procesado, lo cual podría ocurrir, por ejemplo, por incompatibilidad del software con el tipo de información procesada o porque la información procesada haya sido almacenada en un formato inadecuado para la herramienta utilizada.

Clasificar y filtrar datos:

Los protocolos distintos a UDP, TCP e ICMP fueron ignorados por el paquete de software y por tanto tampoco se tuvo que realizar un proceso previo de clasificación y filtrado de datos.

3.4.4. Construcción de modelos

El siguiente reporte abarca el resultado de cumplimiento de las siguientes actividades de la propuesta metodológica:

4.1. Determinar los requerimientos del modelo

4.2. Seleccionar las técnicas de modelado

4.3. Generar el diseño de pruebas

4.4. Elaborar los modelos

4.5. Experimentar con los modelos

4.6. Valorar los modelos

4.7. Crear el reporte de construcción de modelos

Requerimientos del modelo

En base a los objetivos y criterios establecidos en el reporte de entendimiento del negocio, tanto por parte del negocio como por parte del proceso de analítica de datos, se establecieron los siguientes requerimientos para el modelo:

- El modelo debe ser capaz de extraer la información necesaria para determinar la estructura topología y geográfica de la red que genere el tráfico analizado.
- El modelo debe ser capaz de extraer la información necesaria para determinar el volumen y duración de los flujos que componen el tráfico analizado.

- El modelo debe permitir el procesamiento sistemático de un gran volumen de datos sin que sea necesario realizar procesamientos parciales.
- La información que se requiere extraer del tráfico analizado es: las direcciones IP de origen y destino de cada paquete, su ubicación geográfica, su distancia relativa respecto al punto donde se capturaron los paquetes, las conversaciones existentes con sus respectivos volúmenes de paquetes y duraciones.

Técnicas de modelado

Para poder modelar el tráfico analizado de acuerdo con los requerimientos establecidos, se decidió emplear modelos simbólicos de tipo gráficos y descriptivos.

A nivel gráfico, se buscará representar el volumen de los flujos detectados en el tráfico mediante un diagrama Sankey [72] y la estructura geográfica mediante un mapamundi coloreado. En cada caso se realizará un conteo de la totalidad de los paquetes que conforman los flujos detectados y los países de origen y destino de dichos flujos, lo cual permitirá determinar el ancho de cada línea del diagrama Sankey y el color de cada país representado en el mapamundi.

Cada esquema y diagrama será acompañado de una descripción de los hallazgos considerados más relevantes. Estas descripciones se han registrado en el reporte de evaluación de los modelos.

Diseño de pruebas

Con el fin de probar que el modelo generado sea correcto se realizarán las siguientes verificaciones:

- Aplicar el modelo a una muestra pequeña de tráfico y comparar el resultado con el que se obtenga de realizar el análisis manual de dicho tráfico. El tamaño y duración de la muestra deberá ser lo suficientemente pequeña como para ser analizada con herramientas similares a Wireshark [55] pero suficientemente grande como para ser representativa. El tamaño y duración dependen de los recursos disponibles, y para el presente caso, se estableció en 50 MBytes.
- Verificar que la sumatoria de los paquetes detectados por cada modelo corresponda a la totalidad de paquetes analizados en el tráfico de red.

Modelos generados

La generación de los modelos tuvo que realizarse en forma iterativa, es decir, experimentando con diferentes opciones mediante ejercicios de ensayo y error. Parte

del proceso aquí descrito consistió en generar *scripts* para Bro que extraigan información que pueda servir para obtener conocimiento a partir de los datos procesados. Con la información extraída, se procedió a experimentar con diferentes métodos de visualización de datos (gráficos, tablas, esquemas) con el fin de evidenciar de forma intuitiva el conocimiento extraído. El proceso descrito se repitió por varias ocasiones hasta obtener los modelos finales que se presentan a continuación:

a) Tabla de concentración de comunicaciones realizadas entre tipos de redes de origen y destino.

A continuación, en la Tabla 3.5, se muestra la distribución porcentual de intercambio de paquetes por tipos de redes de origen y destino, donde las columnas y las filas referencian diferentes rangos de IP (IPs públicas, IPs privadas, IPs loopback, etc) y los datos mostrados indican, en relación con el total de paquetes detectados, la cantidad porcentual de paquetes cursados entre los rangos de IP determinados por la fila (IP origen) y la columna (destino) a la que pertenece cada dato. Queda en evidencia, por ejemplo, que el 36.63% de las comunicaciones se realizan entre direcciones públicas, que 0% de las comunicaciones involucran direcciones privadas, y que el resto de comunicaciones involucran, sea en origen o en destino, direcciones que no son ni públicas ni privadas.

Tabla 3.5. Distribución porcentual de intercambio de paquetes por tipos de redes

Origen/Destino	FutureUse.	IP Multicast.	IP Privada.	IP Pública.	This_Net.	Total Result
FutureUse	1.01 %	2.37 %	0.00 %	12.38 %	2.57 %	**18.33 %**
IP Loopback				0.00 %		**0.00 %**
IP Multicast	1.17 %	0.87 %	0.00 %	8.30 %	2.08 %	**12.42 %**
IP Privada				0.00 %		**0.00 %**
IP Pública	6.82 %	6.22 %		36.63 %	7.32 %	**56.99 %**
This_Net	2.10 %	1.96 %	0.00 %	7.49 %	0.72 %	**12.27 %**
Total Result	**11.10 %**	**11.42 %**	**0.00 %**	**64.80 %**	**12.69 %**	**100.00 %**

b) Diagrama Sankey que representa gráficamente las interacciones existentes entre distintos tipos de direcciones IP.

La Figura 3.6 muestra de forma gráfica lo descrito en la Tabla 3.5. Este tipo de diagramas permiten visualizar cómo diferentes elementos (en este caso, tipos de rangos de IPs) interactúan entre sí (en este caso, mediante el intercambio de paquetes). En dicha figura, se tiene a la izquierda diferentes tipos de rangos de IPs correspondientes a las direcciones de origen de los paquetes, y a la derecha diferentes tipos de rangos de IPs correspondientes a las direcciones de destino de los paquetes. Cada tipo de rango de IPs de la izquierda se conecta por líneas a los rangos de IPs de la derecha,

reflejando la existencia de intercambio de paquetes y evidenciando mediante el grosor de cada línea la cantidad relativa de paquetes intercambiados.

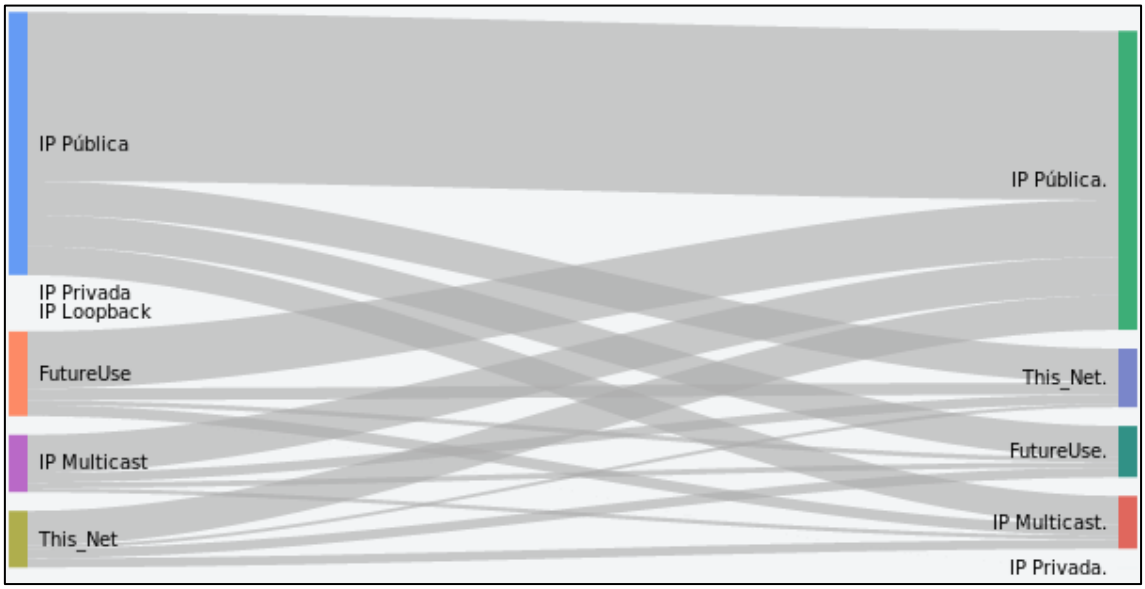

Figura 3.6. Diagrama Sankey de interacciones entre tipos de direcciones IP.

c) Mapamundis coloreados que representan la cantidad relativa de paquetes originados o recibidos, por país.

La Figura 3.7 muestra un Mapamundi coloreado, donde cada país fue coloreado de acuerdo a la cantidad de paquetes que haya originado para el conjunto de datos analizado. Los países en blanco no originaron paquetes, y los países en rojo originaron la mayor parte de paquetes. Las tonalidades varían de blanco, a amarillo, a naranja y a rojo, de acuerdo al aporte relativo de paquetes asociados.

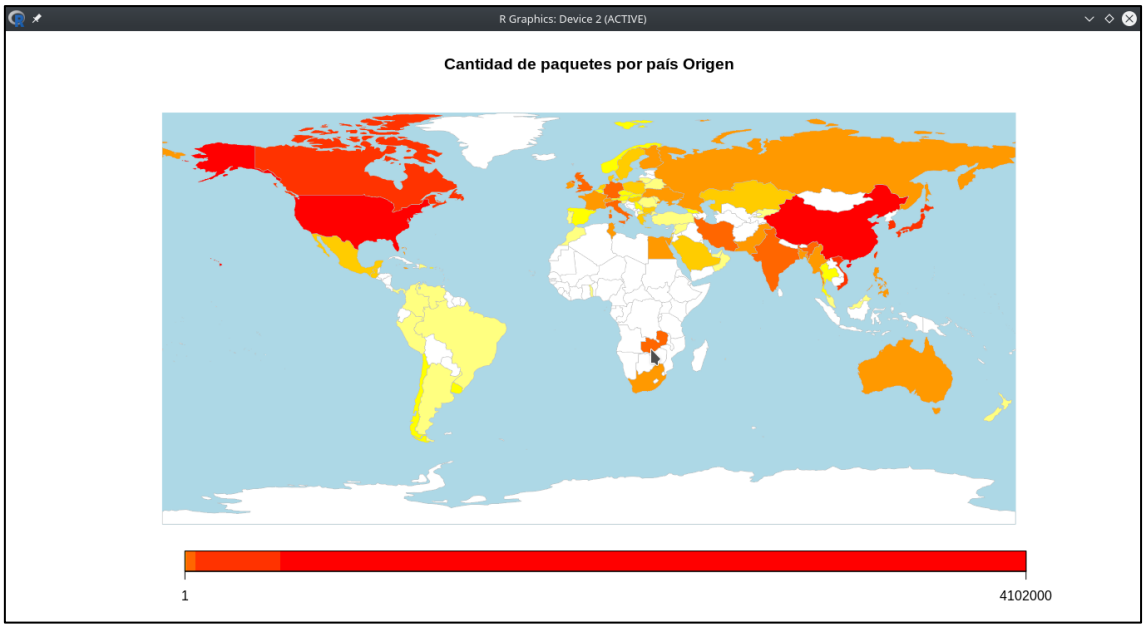

Figura 3.7. Mapamundi coloreado con cantidad de paquetes de origen por país.

La Figura 3.8 muestra un Mapamundi coloreado, similar al anterior, pero donde los colores permiten visualizar el aporte relativo de paquetes recibidos por país.

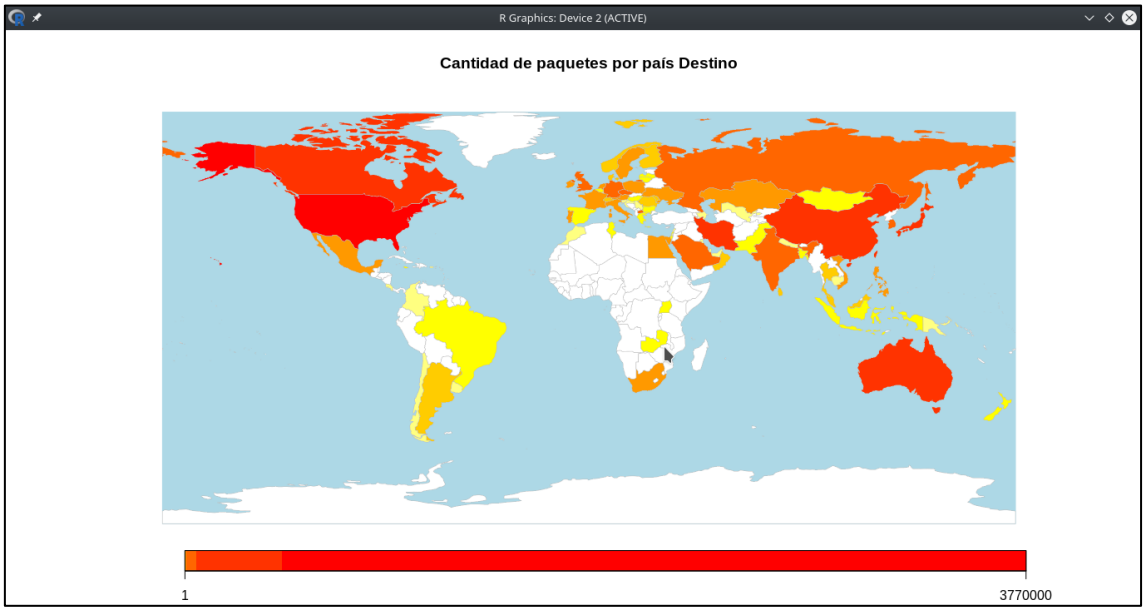

Figura 3.8. Mapamundi coloreado con cantidad de paquetes de destino por país.

Las escalas de colores o leyenda de los Mapamundis, si bien utilizan una escala logarítmica, no permiten apreciar la existencia de las tonalidades amarillas. Esto se debe a que la tonalidad roja y naranja abarca casi en su totalidad el tráfico analizado, mientras que la tonalidad amarilla contribuye evidencia contribuciones poco significativas.

La siguiente sección evaluará los resultados obtenidos a partir de los modelos indicados.

3.4.5. Evaluación de resultados

El siguiente reporte abarca el resultado de cumplimiento de las siguientes actividades de la propuesta metodológica:

5.1. Buscar patrones

5.2. Interpretar patrones

5.3. Definir eventos o alertas

5.4. Evaluar los resultados

5.5. Revisar el proceso

5.6. Determinar los siguientes pasos

5.7. Retroalimentar los casos de uso

5.8. Crear el reporte de evaluación de los modelos

Búsqueda e interpretación de patrones

Los modelos generados permiten observar que los flujos más voluminosos corresponden a comunicaciones realizadas entre IPs públicas, conformando así el 36.63% del tráfico representado en la Figura 3.6 y mostrado en la Tabla 3.5. Si se considera flujos de comunicaciones donde el origen o el destino corresponde a una IP pública, se observa que aproximadamente el 57% de las comunicaciones tienen como origen una IP pública y el 64.80% de las comunicaciones tienen como destino una IP pública. En conjunto, aproximadamente el 85% de las comunicaciones involucran direcciones públicas. Esta información es coherente con el tipo de tráfico de telecomunicaciones analizado, pues fue extraído de un enlace tipo *peering* perteneciente a un ISP.

El análisis permite detectar que existe presencia de paquetes cuyas IP de origen son direcciones privadas o direcciones de loopback, así como paquetes cuyas IP de destino son IP privadas. La cantidad de paquetes con estas características son insignificantes en comparación con el volumen total de datos analizados, pues representan aproximadamente 0% del tráfico. Esta información también es coherente con el tipo de tráfico de telecomunicaciones capturado.

Por otro lado, existen varios flujos de comunicación que llaman la atención por utilizar direcciones correspondientes a rangos de IP reservados para futuro uso (*FutureUse*) o para acciones de uso local (*This_Net*). Este comportamiento es significativo y representa aproximadamente en conjunto un 21% de las comunicaciones.

Finalmente se observa tipos de comunicación que involucran direcciones multicast y que en conjunto representan aproximadamente un 23% del tráfico.

El tráfico que involucra direcciones reservadas o de multicast podría requerir un mayor análisis debido a que representa un volumen significativo de las comunicaciones existentes en el enlace *peering*. Este análisis más profundo podría ayudar a determinar si dicho tráfico corresponde a un ataque realizado en la red o si el *peering* transporta tráfico que no solo corresponde a datos de Internet.

En cuanto a la estructura geográfica del tráfico, los mapamundis graficados en la Figura 3.7 y en la Figura 3.8 permiten observar que la mayor parte del tráfico se concentra en los Estados Unidos de América, información coherente con el tipo de tráfico de telecomunicaciones analizado, pues fue extraído de un enlace tipo *peering* perteneciente a un ISP de dicho país.

Países donde el idioma oficial es el inglés, así como países cuyas economías son fuertes tienen una presencia importante en el tráfico analizado. Sudamérica y algunos países de Europa y Asia tienen una presencia baja en el tráfico y África está mayoritariamente ausente. Este tipo de información permite observar de qué forma la penetración de Internet se pudo haber desarrollado en el año en el que se capturó el tráfico analizado.

Definición de eventos y alertas

Si bien no se implementará una solución tecnológica que permita realizar un análisis constante del tráfico del enlace OC48 que fue estudiado como demostración de la metodología propuesta, en caso de que se realizara dicha implementación se podría definir qué eventos deberían generar alertas. Estos eventos, en un caso real, se definirían en base a las mejores prácticas de la industria, a requerimientos regulatorios o compromisos con los clientes, y podrían ser, a manera de ejemplo:

- Presencia de tráfico relacionado con IPs privadas superior al 1%.
- Presencia de tráfico relacionado con IPs reservadas superior al 20%.

Evaluación de resultados

Considerando que:

- el objetivo propuesto para el proceso de analítica de datos fue: “Obtener las características de tráfico en cuanto a su distribución topológica y geográfica, así como volumen y duración de flujos”.
- el proceso de analítica que se realizó al tráfico provisto por CAIDA permitió elaborar modelos esquemáticos y gráficos claros que muestran las características deseadas.
- el proceso de analítica se realizó usando exclusivamente los recursos inventariados sin experimentar disminución en el desempeño de la computadora disponible.
- los requerimientos establecidos para los modelos, así como las políticas definidas para el proceso de analítica, fueron cumplidos a cabalidad.
- la utilización de las actividades que forman parte de la propuesta metodológica permitió desarrollar un proceso de analítica de datos aplicado a tráfico real de telecomunicaciones alcanzando los objetivos establecidos.

Se puede concluir que los resultados son favorables, que los objetivos fueron alcanzados en su totalidad y que el proceso de analítica de datos realizado a manera de demostración fue exitoso.

Revisión del proceso

La experimentación con los modelos y el diseño de pruebas permitieron ir detectando fallas en la lógica empleada para la extracción de información hasta llegar a los modelos finales mostrados. Algunos de los problemas que se presentaron fueron:

- Las herramientas disponibles para realizar un procesamiento sistemático de tráfico de telecomunicaciones tienen limitaciones en cuanto a soporte de protocolos distintos a IP, ICMP, UDP y TCP. Para efectos de la presente demostración, esto no representó un inconveniente práctico, pero en caso de querer analizar datos provenientes de otro tipo de redes, como por ejemplo redes celulares, sería necesario conseguir herramientas adecuadas.

- Se intentó graficar flujos de comunicación usando IPs, rangos de IPs y direcciones de redes como origen y destino, lo cual debido al gran volumen de datos existente hacía ininteligible el resultado. Un desafío importante que se debe superar al realizar analítica de tráfico de redes de telecomunicaciones es encontrar la forma de modelar de forma adecuada la complejidad de dicho tráfico.

- El script generado originalmente en Bro para la extracción sistemática de datos no observó la buena práctica de programación que consiste en no reutilizar variables a lo largo del código, lo cual provocó que los resultados generen flujos inexistentes. El diseño de pruebas permitió detectar el problema y corregir el código del script. Se evidenció que un proceso de analítica se puede desarrollar exitosamente solamente si se dispone del conocimiento correcto para afrontar los distintos retos que imponen las disciplinas involucradas en dicho proceso. Resulta por tanto importante contar con un equipo profesional preparado.

- Los flujos inexistentes observados provocaron originalmente que se muestren paquetes cuyas direcciones IP de origen y destino eran las mismas. Esto derivó en que se cuestione la integridad de la información disponible en las capturas de tráfico provistas por CAIDA, suponiendo que la condición observada se debía a que el proceso de anonimización de la información había alterado las direcciones IP de origen y destino de forma inadecuada. Una vez corregido el código del script Bro se pudo verificar que el problema desaparecía y que la información

provista no mostraba incoherencias en su estructura. Este problema permitió entender la importancia de registrar y tener en cuenta los supuestos que se van generando a lo largo del proceso de analítica para evitar encasillar el análisis de forma equivocada. Por ejemplo, el problema en el código del script Bro pudo ser detectado solamente tras cuestionar el supuesto descrito previamente, permitiendo así analizar otras opciones y encontrando correctamente la fuente del problema. Mientras dicho supuesto no fue cuestionado, la búsqueda de una solución se realizaba asumiendo que era necesario encontrar una forma de superar una alteración de información inadecuada de las direcciones IP que en realidad no existía.

- Con el fin de entender el tráfico analizado se buscó discriminar el origen y destino de cada paquete en base a los rangos de IPs definidos en el RFC 6890 de la IETF (*Internet Engineering Task Force*). Esta clasificación permitió elaborar un diagrama Sankey con información entendible y que muestra los flujos de comunicación que se generan entre clases de redes. Este tipo de información se recopila como parte de la actividad 2.1 de la propuesta metodológica “Levantar información referencial”, y el trabajo demostrativo permitió dejar en evidencia la importancia de realizar esta actividad.

Retroalimentación de casos de uso

La documentación generada tras el proceso de analítica de datos debería servir para alimentar una base de conocimiento sobre los casos de uso o de aplicación de la metodología en procesos de analítica de datos aplicados a redes de telecomunicaciones. El “negocio” podría a futuro servirse de dicha documentación para facilitar y acelerar nuevos requerimientos de analítica.

Para el presente caso, por ejemplo, se podría crear un caso de uso denominado “Distribución geográfica de tráfico” donde se describa el proceso utilizado para generar los Mapamundis de la Figura 3.7 y de la Figura 3.8, los scripts generados, los problemas encontrados, y las lecciones aprendidas. De esta forma, si posteriormente se quisiera generar un proceso de analítica similar en otro enlace de la red de telecomunicaciones, la información provista en el caso de uso “Distribución geográfica de tráfico” permitirá al nuevo equipo encargado del proceso avanzar de forma más rápida y eficiente.

4. CONCLUSIONES

El presente trabajo de titulación dio como resultado una propuesta metodológica de analítica de datos para el estudio y análisis de tráfico de redes de telecomunicaciones. Dicha propuesta se apoyó en diferentes documentos referenciales, como por ejemplo CRISP-DM (*Cross Industry Standard Process for Data Mining*) [13] y ASUM-DM (*Analytics Solutions Unified Method for Data Mining*) [14]), los cuales fueron creados para guiar proyectos de analítica de datos con un ámbito de aplicación general, pero realizando las adaptaciones necesarias para que se adapte a las características propias del sector de las telecomunicaciones, considerando así los desafíos, ventajas e interacciones descritas en las secciones 2.4.1, 2.4.2. y 2.5, respectivamente.

La propuesta metodológica resultante de este trabajo fue recopilada en el Anexo 2 y se conforma por un modelo referencial y una guía de usuario. El Anexo 2 permite tener acceso directo a la metodología elaborada facilitando así su aplicación práctica.

Las adaptaciones que fueron necesarias realizar a las metodologías CRISP-DM y ASUM-DM para elaborar la propuesta metodológica de este proyecto estuvieron relacionadas con la estructuración y reorganización, tanto del modelo referencial de la metodología ASUM-DM como del conjunto de las actividades provistas por los documentos referenciales analizados, conformando así la dimensión Analítica del modelo referencial propuesto.

El modelo referencial considera cuatro dimensiones que pueden o no ser requeridas en un proceso de analítica de datos. Estas dimensiones son Analítica, Proyectos, Infraestructura y Gestión. De esta forma, el modelo referencial se adapta a casos de estudio y análisis puntuales de analítica de datos donde solo la dimensión Analítica es empleada, así como a casos donde los modelos generados por el proceso de analítica se deben implementar en forma de una solución tecnológica, la cual debe ser organizada en forma de un proyecto, implementada en infraestructura apropiada y gestionada, operada y mantenida a lo largo de su vida. El modelo referencial sugiere el uso de las mejores prácticas y estándares de la industria, de acuerdo con cómo opere la empresa que se encuentre haciendo uso de dicho modelo.

Este trabajo también mencionó diferentes métodos, técnicas y herramientas que pueden servir al momento de analizar el tráfico proveniente de una red de telecomunicaciones. El ejercicio demostrativo realizado en la sección 3.4 hace uso de algunas de estas herramientas (wireshark, tcpdump, Bro, diagramas sankey, mapamundis, etc), y permitió evidenciar que su selección dependerá del tipo de tráfico que se requiera procesar. Si bien existen herramientas disponibles de forma libre en Internet, es posible

que las empresas de telecomunicaciones requieran adquirir o desarrollar tecnología especializada que les permita procesar ciertos protocolos especiales.

El desarrollo del ejercicio demostrativo se realizó con facilidad, sin sobretiempos atribuibles a la metodología propuesta y alcanzando exitosamente los objetivos planteados para el ejercicio.

Se puede por lo tanto concluir que la propuesta metodológica planteada en este trabajo de titulación podrá ser utilizado de forma exitosa en el sector de las telecomunicaciones para procesos de analítica de tráfico de redes de telecomunicaciones de cualquier tamaño o tecnología. Este tipo de proyectos podrá servir, por ejemplo, en el campo de la ingeniería para diseñar, adecuar u optimizar redes de telecomunicaciones, en el campo técnico para encontrar y resolver problemas existentes en redes de telecomunicaciones, en el campo académico para estudiar y profundizar el conocimiento sobre el funcionamiento de redes de telecomunicaciones, en el campo regulatorio para analizar el nivel de cumplimiento de una red de telecomunicación con los parámetros esperados, o en el campo empresarial para ofrecer servicios o soluciones a operadores de telecomunicaciones.

Una aplicación futura que se desprende de este trabajo de titulación es la utilización de la metodología desarrollada en el estudio y análisis del tráfico internacional de una operadora de telecomunicaciones con el fin de facilitar, optimizar y guiar el establecimiento de acuerdos comerciales estratégicos con proveedores de contenido y aplicaciones internacionales, y de apalancar la estrategia de desarrollo de los recursos de conectividad internacional de la empresa.

REFERENCIAS BIBLIOGRÁFICAS

[1] S. Crawford, «The origin and development of a concept: the information society,» *Bulletin of the Medical Library Association 71 (4),* pp. 380-385, octubre 1983.

[2] K. Schwab, The Fourth Industrial Revolution, World Economic Forum, 2016.

[3] UNESCO, Towards Knowledge Societies, UNESCO, 2005.

[4] R. Davies, «Industry 4.0, Digitalisation for productivity and growth,» European Union, septiembre 2015. [En línea]. Available: http://www.europarl.europa.eu/RegData/etudes/BRIE/2015/568337/EPRS_BRI(2015)568337_EN.pdf.

[5] Naciones Unidas - CEPAL, «La nueva revolución digital: de la Internet del consumo a la Internet de la producción,» CEPAL, agosto 2016. [En línea]. Available: http://www.cepal.org/es/publicaciones/38604-la-nueva-revolucion-digital-la-internet-consumo-la-internet-la-produccion.

[6] F. Groene, A. Navalekar y M. Kramer Coakley , «An industry at risk: Commoditization in the wireless telecom industry,» 22 febrero 2017. [En línea]. Available: http://www.strategyand.pwc.com/reports/industry-at-risk.

[7] World Economic Forum, «Digital Transformation Initiative - Telecommunications Industry,» enero 2017. [En línea]. Available: http://reports.weforum.org/digital-transformation.

[8] GSMA, «Industry 4.0 - MWC17 Big Themes: The Fourth Industrial Revolution,» 2016. [En línea]. Available: http://cc.gsma.com/fourth-industrial-revolution/.

[9] K. Butner, «Operating in the fourth industrial revolution,» 11 diciembre 2015. [En línea]. Available: http://www.ibmbigdatahub.com/blog/operating-fourth-industrial-revolution.

[10] ITU, «Recomendación Y.Sup40,» ITU, 28 septiembre 2016. [En línea]. Available: https://www.itu.int/rec/T-REC-Y.Sup40-201607-I/es.

[11] TM-Forum, «GB979 Big Data Analytics Solution Suite R17.0.0,» 2017. [En línea]. Available: https://www.tmforum.org/resources/best-practice/gb979-big-data-analytics-solution-suite-r17-0-0/.

[12] ITU, «Y.3600 : Big data - Requisitos y capacidades basados en la computación en la nube,» ITU, 23 diciembre 2015. [En línea]. Available: https://www.itu.int/rec/T-REC-Y.3600-201511-I/es.

[13] IBM, «CRISP-DM,» 2015. [En línea]. Available: https://www-01.ibm.com/marketing/iwm/iwm/web/pick.do?source=swerpba-basimext&lang=en_US.

[14] J. Haffar, «Have you seen ASUM-DM?,» IBM, 16 octubre 2015. [En línea]. Available: https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/.

[15] R. T. de Sousa Júnior y B. C. da Rocha, «IDENTIFYING BANK FRAUDS USING CRISP-DM AND DECISION TREES,» *International journal of computer science & information Technology ,* vol. 2, nº 5, pp. 162-169, 2010.

[16] K. A.-A. JR, «PREDICTING CUSTOMER CHURN IN THE MOBILE TELECOMMUNICATION INDUSTRY, A CASE STUDY OF MTN GHANA, KUMASI,» 2011.

[17] European Comission DG Connect, «A European strategy on the data value chain,» 07 11 2013.

[18] J. Liebowitz, Business Analytics: An Introduction, 2014.

[19] T. Ravindra Babu, M. Narasimha Murty y S.V. Subrahmanya, Compression Schemes for Mining Large Datasets: A Machine Learning Perspective, 2013.

[20] Asamblea Nacional del Ecuador, «Ley Organica de Telecomunicaciones,» 2015.

[21] A. R. Figueiras Vidal, Una panorámica de las telecomunicaciones, 2002.

[22] M. Dara, «VALUE NETWORKS AND BUSINESS MODELS - FORMULATING AND DEMONSTRATING A METHODOLOGY FOR THE DEVELOPMENT OF VALUE NETWORKS AND ALIGNMENT OF BUSINESS MODELS BASED ON DESIGN SCIENCE RESEARCH METHODOLOGY,» University of Twente, School of Management and Governance , 2013.

[23] J. M. Cavanillas, E. Curry y W. Wahlster, New Horizons for a Data-Driven Economy, SpringerOpen, 2016.

[24] CISCO, «The Zettabyte Era: Trends and Analysis,» 07 junio 2017. [En línea]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html.

[25] CISCO, «Cisco Visual Networking Index: Forecast and Methodology, 2016–2021,» 07 junio 2017. [En línea]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html.

[26] prnoticias, «Tendencias Tech 2016: IBM trabaja en la Computación Cognitiva,» 20 enero 2016. [En línea]. Available: http://prnoticias.com/tecnologia/tendencias-de-tecnologia/20148492-tendencias-tecnologicas-2016-ibm-computacion-cognitiva.

[27] European Comission DG Connect, «Communication on data-driven economy,» 02 julio 2014. [En línea]. Available: https://ec.europa.eu/digital-single-market/en/news/communication-data-driven-economy.

[28] Investopedia, The Industry Handbook, Investopedia, 2010, pp. 51-55.

[29] M. Wohlers, Convergencia tecnológica y agenda regulatoria de las telecomunicaciones en América Latina, CEPAL, 2008.

[30] C. Cullell March, El principio de neutralidad tecnológica y de servicios en la UE: la liberalización del espectro radioeléctrico, Revista de los Estudios de Derecho y Ciencia Política de la UOC, 2010.

[31] J. García Falconí, «DERECHO A LA INTIMIDAD PERSONAL Y FAMILIAR,» 02 febrero 2011.

[32] ASIET, Manual regulatorio, marco de actuación sobre los temas de agenda de políticas públicas de la Asociación Interamericana de Empresas de Telecomunicaciones (ASIET), Montevideo: ASIET, 2015.

[33] S. Reid y S. Kamen, «End-user experience traffic data capture and segmentation,» 10 julio 2016. [En línea]. Available: https://docs.bmc.com/docs/display/tsavm105/End-user+experience+traffic+data+capture+and+segmentation.

[34] P. Offord, «Packet Capture Techniques,» 2015. [En línea]. Available: https://sharkfest.wireshark.org/assets/presentations15/15.pdf.

[35] Microsoft, «Virtual Network, Your private network in the cloud,» [En línea]. Available: https://azure.microsoft.com/en-us/services/virtual-network/.

[36] A. Cecil, «A Summary of Network Traffic Monitoring and Analysis Techniques,» [En línea]. Available: http://www.cse.wustl.edu/~jain/cse567-06/ftp/net_monitoring/index.html.

[37] Wireshark, «Ethernet capture setup,» 12 abril 2017. [En línea]. Available: https://wiki.wireshark.org/CaptureSetup/Ethernet.

[38] iQsim, «iQsim R400 Series Mobile Robot,» [En línea]. Available: http://www.iqsim.com/v2/mobile_robot.fr.htm.

[39] Sandvie, «Identifying and measuring internet traffic: techniques and considerations,» 2015. [En línea]. Available: https://www.sandvine.com/downloads/general/whitepapers/identifying-and-measuring-internet-traffic.pdf.

[40] Sandvine, «Internet Traffic Classification,» 2015. [En línea]. Available: https://www.sandvine.com/downloads/general/sandvine-technology-showcases/traffic-classification-identifying-and-measuring-internet-traffic.pdf.

[41] Bro, «Documentation and Training,» Bro, [En línea]. Available: https://www.bro.org/documentation/index.html.

[42] SNORT, «Snort,» [En línea]. Available: https://snort.org/.

[43] D. Koukis, S. Antonatos, D. Antoniades, E. Markatos y P. Trimintzios, A Generic Anonymization Framework for Network Traffic, Forth Institute of Computer Science, 2006.

[44] T. Farah, ALGORITHMS AND TOOLS FOR ANONYMIZATION OF THE INTERNET TRAFFIC, BRAC University, 2013.

[45] J. Fan, J. Xu, M. Ammar y S. Moon, «Cryptography-based Prefix-preserving Anonymization,» Georgia Tech, [En línea]. Available: https://www.cc.gatech.edu/computing/Telecomm/projects/cryptopan/.

[46] A. Slagell, K. Lakkaraju y K. Luo, FLAIM: A Multi-level Anonymization Framework for Computer and Network Logs, National Center for Supercomputing Applications, 2006.

[47] G. Minshall, «TCPDPRIV,» Ipsilon Networks Inc, [En línea]. Available: http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html.

[48] J. Han, M. Kamber y J. Pei, Data Mining, Concepts and Techniques, Morgan Kaufmann, 2012.

[49] F. Gorunescu, Data Mining, Concepts, Models and Techniques, Springer, 2011.

[50] C. C. Aggarwal, Data Mining, The Textbook, Springer, 2015.

[51] M. Brown y J. Brocklebank, «Data Mining,» 1997. [En línea]. Available: http://www2.sas.com/proceedings/sugi22/DATAWARE/PAPER128.PDF.

[52] O. Maimon y L. Rokach, Data Mining and Knowledge Discovery Handbook, Springer, 2010.

[53] T. Nguyen y G. Armitage, A Survey of Techniques for Internet Traffic Classification using Machine Learning, Swinburne University of Technology, 2008.

[54] A. Moore y D. Zuev, Internet Traffic Classification Using Bayesian Analysis Techniques, University of Cambridge, 2005.

[55] Wireshark, «Wireshark,» [En línea]. Available: https://www.wireshark.org/.

[56] QlikTech, «Business Intelligence | Data Visualization tools,» QlikTech, [En línea]. Available: http://www.qlik.com.

[57] Google, «Chart Gallery,» Google, [En línea]. Available: https://developers.google.com/chart/interactive/docs/gallery.

[58] A. Azevedo y M. F. Santos, «KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW,» *IADIS European Conference on Data Mining 2008,* 2008.

[59] U. Shafique y H. Qaiser, «A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA),» 2014.

[60] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «From Data Mining to Knowledge Discovery in Databases,» 1996. [En línea]. Available: https://aaai.org/ojs/index.php/aimagazine/article/view/1230.

[61] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, «CRISP-DM 1.0,» 2000. [En línea]. Available: https://www.the-modeling-agency.com/crisp-dm.pdf.

[62] O. Laudy, «Standard_methodology_for_analytical_models,» 2015. [En línea]. Available: http://olavlaudy.com/MediaWiki/index.php?title=Standard_methodology_for_analytical_models.

[63] K. Peffers, T. Tuunanen, M. Rothenberger and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems,* pp. 45-78, agosto 2007.

[64] V. Vaishnavi y B. Kuechler, «Design Science Research in Information Systems (IS),» DESRIST, 20 enero 2004. [En línea]. Available: http://www.desrist.org/design-research-in-information-systems/.

[65] T. Levitt, «Exploit the Product Life Cycle,» Harvard Business Review, 1965.

[66] ITU, «Project management guidelines,» noviembre 2013. [En línea]. Available: http://www.itu.int/en/ITU-D/Projects/Documents/ProjectManagementGuidelinesandTemplates.pdf.

[67] CAIDA, «The CAIDA UCSD Anonymized OC48 Internet Traces 2002-2003 - [20030424-070000], https://data.caida.org/datasets/oc48/oc48-original/».

[68] CAIDA, «CAIDA Acceptable Use Agreement (AUA) for Publicy Accessible Datasets,» 2015. [En línea]. Available: https://www.caida.org/home/legal/aua/public_aua.xml.

[69] IETF, Hypertext Transfer Protocol -- HTTP/1.1, 1999.

[70] IETF, TRANSMISSION CONTROL PROTOCOL, 1981.

[71] MaxMind, «Bases de datos GeoIP2,» [En línea]. Available: https://www.maxmind.com/es/geoip2-databases.

[72] Google, «Sankey Diagram,» Google, [En línea]. Available: https://developers.google.com/chart/interactive/docs/gallery/sankey.

ANEXO I. TABLA COMPARATIVA DE ACTIVIDADES PROPUESTAS POR DOCUMENTOS REFERENCIALES

El presente Anexo presenta una tabla comparativa de las actividades pertenecientes a seis documentos referenciales analizados a lo largo del presente trabajo de titulación y que buscan guiar proyectos o procesos de analítica de datos. Estos proyectos son los detallados en la Tabla I.1.

Tabla I.1. Documentos referenciales comparados en este Anexo.

Nombre	Autor	Año publicación	Sector de aplicación objetivo
KDD Process	AAI	1996	Uso general
SEMMA	SAS	1997	Uso general
CRISP-DM	CRISP-DM	1996	Uso general
ASUM-DM	IBM	2015	Uso general
SMAM	Olav Laudy	2015	Uso general
GB979	TM-Forum	2017	Proveedores de servicios digitales

La tabla comparativa muestra en la primera columna el nombre de lo que se ha denominado Dimensión y que se refiere a una descripción del tipo de actividades que son mencionadas por cada una de los documentos referenciales. El resto de columnas muestran las actividades propuestas por los documentos referenciales analizados, coincidiendo en cada fila aquellas que son similares entre sí.

En total se detectaron cuatro dimensiones:

- Analítica: la cual abarca actividades relacionadas con analítica de datos.
- Proyectos: la cual abarca actividades relacionadas con la Gestión de Proyectos.
- Infraestructura: la cual abarca actividades relacionadas con la Gestión de Infraestructura.
- Gestión: la cual abarca actividades relacionadas con la Gestión, Operación y Mantenimiento de soluciones tecnológicas.

El resultado se muestra a partir de la siguiente página, en la Tabla I.2.

Tabla I.2: Análisis comparativo de actividades presentes en documentos referenciales.

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Proyectos				Analyze-Design-Configure&Build		
Proyectos				Prepare for Implementation		
Proyectos				Liaise with Sales and Review Project Details		
Proyectos				Identify Resources		
Proyectos				Conduct Readiness Assessment		
Proyectos				Assess Customer Readiness for Implementation		
Proyectos				Conduct Project Kick-off		
Proyectos				Prepare Project Kick-off Deck		
Proyectos				Orient Project Resources		
Proyectos				Execute Project Kick-off		
Analítica	Comprensión del objetivo		Comprensión del negocio	Understand Business	Identificación de caso de uso	Casos de uso
Analítica			Determinar objetivos del negocio	Determine Business Objectives	Obtención de requerimientos del modelo	
Analítica				Compile the Business Background		
Analítica				Determine Business Success Criteria		
Analítica			Evaluar estado	Assess Situation		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Analítica				Conduct an Inventory of Resources		
Analítica				Determine Requirements, Assumptions, and Constraints		
Proyectos				Consider Risks and Contingencies		
Analítica				Compile a Glossary of Terminology		
Proyectos				Conduct a Cost Benefit Analysis		
Analítica			Determinar objetivos de la analítica	Determine Data Mining Goals		
Proyectos			Producir el plan de proyecto	Create Project Plan		
Analítica				Create Business Understanding Report		
Analítica			Comprensión de los datos	Understand Data		
Analítica		Muestreo	Recolectar datos iniciales	Collect Initial Data		
Analítica			Describir datos	Describe Data		
Analítica		Exploración	Explorar datos	Explore Data		
Analítica			Verificar calidad de los datos	Verify Data Quality		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Analítica				Create Data Understanding Report		
Infraestructura				Design and Validate Infrastructure		Repositorios de datos
Infraestructura				Design Technical Infrastructure		
Infraestructura				Design Analytical, QA, and Production Environments Architecture		
Infraestructura				Validate Technical Infrastructure		
Infraestructura				Design Security Infrastructure		
Infraestructura				Design Authentication and Authorization Strategies		
Infraestructura				Document the Security Model and Validate with Stakeholders		
Infraestructura				Set up Environments		
Infraestructura				Set Up Environments On Site		
Infraestructura				Set Up Analytical Environment		
Infraestructura				Install and Set up Server		
Infraestructura				Install IBM Software		
Infraestructura				Set up Access and Provide to IBM Team		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Infraestructura				Set Up QA Environment		
Infraestructura				Install and Set up Server		
Infraestructura				Install IBM Software		
Infraestructura				Set up Access and Provide to IBM Team		
Infraestructura				Set Up Production Environment		
Infraestructura				Install and Set up Server		
Infraestructura				Install IBM Software		
Infraestructura				Set Up Environments On Cloud		
Infraestructura				Set Up Analytical Environment		
Infraestructura				Set Up Instance		
Infraestructura				Set up Access and Provide to IBM Team		
Infraestructura				Set Up QA Environment		
Infraestructura				Set Up Instance		
Infraestructura				Set up Access and Provide to IBM Team		
Infraestructura				Set Up Production Environment		
Infraestructura				Set Up Instance		
Infraestructura				Set Up Access		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Analítica		Modificación	Preparación de los datos	Prepare Data	Preparación de los datos	
Analítica	Selección de datos		Seleccionar datos	Select Data		
Analítica	Pre-procesamiento		Limpieza de datos	Clean Data		
Analítica	Transformación		Construcción de datos	Construct Data		
Analítica			Integrar datos	Integrate Data		
Analítica			Formatear datos	Format Data		
Analítica				Create Data Preparation Report		
Analítica	Minería de datos	Modelado	Modelado	Build Model		
Analítica	Seleccionar Métodod		Seleccionar técnicas de modelado	Select Modeling Techniques		
Analítica			Generar diseño de pruebas	Generate Test Design		
Analítica	Explorar y Modelar		Construir modelo	Build Model	Experimentar modelado	
Analítica			Evaluar modelo	Assess Model		
Analítica	Buscar patrones				Visión interna	
Analítica	Interpretar patrones					
Analítica	Evaluación	Evaluación	Evaluación	Evaluate Model	Prueba de valor: ROI	
Analítica			Evaluar resultados	Evaluate Results		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Analítica			Revisar proceso	Review Process		
Analítica			Determinar siguientes pasos	Determine Next Steps		
Proyectos				Conduct Analytical Knowledge Transfer		
Proyectos				Orient and Transfer Knowledge to Client Project Analytical Team		
Proyectos				Define Deployment Approach		
Proyectos				Create Roll-out Plan		
Proyectos				Design Operational Testing Strategy		
Proyectos				Identify and Agree Testing Plans		
Proyectos				Create Test Plans		
Proyectos				Validate and Test in QA Environment		
Proyectos				Create QA Data Files		
Proyectos				Ensure QA Environment is Ready		
Proyectos				Migrate/Restore Analytical Model in QA		
Proyectos				Conduct System and Performance Tests		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Proyectos				Review and Refine System and Performance Tests Plans		
Proyectos				Schedule Tests		
Proyectos				Execute Tests		
Proyectos				Perform Fixes		
Proyectos				Conduct User Acceptance Test		
Proyectos				Review and Refine User Acceptance Test Plan		
Proyectos				Execute User Acceptance Test		
Proyectos				Gather User Feedback		
Proyectos				Perform Fixes		
Proyectos				Communicate Results		
Proyectos				Make Prodcution Deployment Decision		
Gestión			Despliegue	Deploy		
Gestión			Plan de despliegue			
Gestión				Conduct Operational Knowledge Transfer	Operacionaliza-ción	
Gestión				Orient and Transfer Knowledge to Client Project Operational Team		
Gestión			Plan de monitoreo y mantenimiento	Prepare for Ongoing Maintenance		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Gestión				Establish a Schedule for On-call Support		
Gestión				Schedule Maintenance Activities for the Production Environment		
Gestión				Schedule Monitoring Activities for the Production Environment		
Gestión				Deploy Solution		
Gestión				Create Production Data Files		
Gestión				Create and perform Operational Readiness Testing		
Gestión				Migrate/Restore QA Model Into Production		
Gestión				Transit to IBM Support		
Gestión				Orient Administrators to IBM Customer Support Resources		
Gestión				Hand Over to Support		
Gestión				Launch		
Gestión				Go Live		
Gestión				Implement Launch Communication Plan		
Gestión				Review Launch		
Analítica	Actuación					Procesamiento de eventos complejos

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Analítica						Alertas, Disparadores, KPIs y Reportes
Proyectos				Prepare for Project Closure		
Proyectos			Producir reporte final	Create Final Report		
Proyectos				Create Management Summary		
Proyectos			Revisar proyecto	Create Project Review		
Proyectos				Create Forward Assessment		
Proyectos				Conduct Review Meeting		
Proyectos				Follow up on the Review Meeting Actions		
Proyectos				Plan and Execute Immediate Remedial Actions		
Proyectos				Update Final Report With Additional Recommendations		
Proyectos				Communicate Follow-Up Actions To Client		
Proyectos				Obtain Consent to Close Project		
Proyectos				Obtain Consent for Reference Initiation		
Proyectos				Conduct Project Closure Team Event		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Proyectos				Prepare For Close Out Meeting		
Proyectos				Conduct Close Out Team Event		
Proyectos				Conduct Internal Project Review and Lessons Learned		
Proyectos				Prepare For Internal Team Meeting		
Proyectos				Conduct Internal Close Out Team Meeting		
Proyectos				Document Findings		
Proyectos				Complete Project Management Experience Form		
Gestión				Operate & Optimize		
Gestión				Monitor Model		
Gestión				Monitor Continuously Model Accuracy and Model Refresh		
Gestión				Operate, Optimize and Imrove System		
Gestión				Schedule and Operate System		
Gestión				Monitor Production Processes		
Gestión				Manage Maintenance Processes		
Gestión				Manage Defects		
Gestión				Manage Enhancements		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Gestión				Optimize System		
Gestión				Maintain Application Documentation		
Gestión				Manage Competency		
Gestión				Support User Community		
Gestión				Manage Help Desk Operations		
Gestión				Manage Issues		
Gestión				Manage Help Desk Resources		
Gestión				Manage Communication		
Gestión				Manage Infrastructure		
Gestión				Manage System Configuration		
Gestión				Manage Capacity		
Gestión				Manage System Availability		
Gestión				Manage Releases		
Gestión				Manage Security		
Gestión				Govern System Lifecycle Program	Ciclo de vida del modelo	Gobernar los datos
Gestión				Verify Benefits Realization		
Gestión				Review System Performance		
Gestión				Review System Life Cycle Plan		
Gestión				Manage Quality		
Proyectos				Initiate		
Proyectos				Conduct Handover Meeting		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Proyectos				Complete Project Initiation Meeting		
Proyectos				Review With Customer Scope and Expectations		
Proyectos				Conduct Project Readiness Assessment		
Proyectos				Review Business Case and Project Charter With Customer		
Proyectos				Review Methodology With Customer		
Proyectos				Complete Project Initiation Check List		
Proyectos				Create Project Success Plan Section 1		
Proyectos				Create Project Overview		
Proyectos				Create Project Strategy		
Proyectos				Create Project Management Plan		
Proyectos				Create Risk Assessment		
Proyectos				Document Pre-implementations Requirements		
Proyectos				Document Project Assumptions		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Proyectos				Create High-Level Resource Schedule		
Proyectos				Create Milestone Plan and Detailed Plan		
Proyectos				Create Appendices Section 1		
Proyectos				Create Project Financial Plan		
Proyectos				Identify Project Costs		
Proyectos				Identify Cost Schedule		
Proyectos				Baseline Budget		
Proyectos				Document Assumptions		
Proyectos				Obtain Consent for Project Success Plan Section 1		
Proyectos				Approve Transition At Initiate		
Proyectos				Plan		
Proyectos				Initiate Team Building		
Proyectos				Conduct Kickoff Meeting		
Proyectos				Train Project Team on Methodology		
Proyectos				Create Project Success Plan section 2		
Proyectos				Create Exception Statement		
Proyectos				Create Pre-Configure and Deployment Requirements		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Proyectos				Create Configure and Deployment Assumptions		
Proyectos				Create Detailed Configure and Deployment Resource Plan		
Proyectos				Create Detailed Configure and Deployment Schedule		
Proyectos				Create appendices Section 2		
Proyectos				Update Project Financial Plan		
Proyectos				Obtain Consent for Project Success Plan section 2		
Proyectos				Approve Transition At Plan		
Proyectos				Execute		
Proyectos				Manage Project Schedule		
Proyectos				Manage Financial Plan		
Proyectos				Manage risks		
Proyectos				Manage Communication		
Proyectos				Manage Acceptance		
Proyectos				Manage Change		
Proyectos				Manage Issues		
Proyectos				Manage Quality		
Proyectos				Manage Documentation and Configuration		
Proyectos				Manage Resources		
Proyectos				Obtain Acceptance Approval		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Proyectos				Approve Transition At Execute		
Proyectos				Close		
Proyectos				Create Go Forward Plan		
Proyectos				Create Management Summary		
Proyectos				Create Project Review		
Proyectos				Create Forward Assessment		
Proyectos				Conduct Review Meeting		
Proyectos				Follow up on the Review Meeting Actions		
Proyectos				Plan and Execute Immediate Remedial Actions		
Proyectos				Update Go Forward Plan With Additional Recommendations		
Proyectos				Communicate Follow-Up Actions To Client		
Proyectos				Obtain Consent to Close Project		
Proyectos				Obtain Consent for Reference Initiation		
Proyectos				Conduct Project Closure Team Event		
Proyectos				Prepare For Close Out Meeting		
Proyectos				Conduct Close Out Team Event		

Dimensión	KDD	SEMMA	CRISP-DM	ASUM-DM	SMAM	GB979
Proyectos				Conduct Internal Project Review and Lessons Learned		
Proyectos				Prepare For Internal Team Meeting		
Proyectos				Conduct Internal Close Out Team Meeting		
Proyectos				Document Findings		
Proyectos				Complete Project Management Experience Form		

ANEXO II. PROPUESTA METODOLÓGICA.

INTRODUCCIÓN:

La presente metodología de analítica de datos ha sido formulada para ser aplicada en el sector de las telecomunicaciones para el estudio y análisis de tráfico de redes de telecomunicaciones. Esta metodología utiliza como base referencial metodologías y modelos como CRISP-DM, ASUM-DM, KDD Process, SEMMA, SMAM y GB979.

Su estructura se divide en dos partes: un modelo referencial y una guía de usuario.

El modelo referencial consiste en un esquema que permite visualizar de forma macro cómo se organiza un proyecto de analítica de datos. La guía de usuario describe cada una de las tareas y actividades que corresponden a un proceso de analítica de datos.

MODELO REFERENCIAL:

A continuación, se presenta un modelo referencial jerárquico estructurado en tres niveles de abstracción.

Un proceso de analítica de datos, dependiendo de su complejidad y alcance, puede abarcar tareas y actividades relacionadas con:

- Analítica de datos (Analítica).
- Gestión de proyectos (Proyectos).
- Gestión de Infraestructura (Infraestructura).
- Gestión, operación y mantenimiento de soluciones tecnológicas (Gestión).

Estos cuatro componentes se consideran dentro del modelo referencial como dimensiones en las que se desarrollará el trabajo relacionado con un proyecto de analítica de datos. Cada una de las dimensiones listadas se ubica en el primer nivel de abstracción del modelo referencial y está compuesta a su vez por diferentes etapas y fases. Cada etapa o fase conforma el segundo nivel de abstracción del modelo referencial y contiene una serie de actividades que deben desarrollarse. Las actividades correspondientes a cada dimensión se ubican en el tercer nivel de abstracción del modelo referencial y son determinadas por los documentos que contengan las mejores prácticas de la industria, estándares o guías de usuario que la empresa de telecomunicaciones donde se ejecute el proyecto de analítica de datos haya seleccionado para ser utilizados en sus actividades normales.

Para la dimensión “Analítica de datos”, la presente metodología propone cinco fases y un conjunto de actividades que han sido detalladas en lo que se ha denominado como guía de usuario.

El modelo jerárquico descrito previamente se muestra en la Figura II.1.

Dimensión Analítica					Dimensión Infraestructura					Dimensión Gestión				
Comprensión del Negocio	Comprensión de los Datos	Preparación de los datos	Modelado	Evaluación	Diseño de solución tecnológica	Instalación de infraestructura	Instalación de software	Prueba y validación de solución	Entregar solución	Lanzamiento	Desarrollo	Madurez	Declive	Retiro
Guía de usuario propuesta					ITIL, eTOM, ISO 20000, ISO 27001 e ISO 9001 Suplemento 40 de la serie Y de recomendaciones ITU									

Dimensión Proyecto														
Formulación					Implementación					Evaluación ex – post				
Identificación	Formulación	Evaluación ex – ante	Diseño e Ingeniería	Decisión de Inversión	Inicio	Planificación	Ejecución	Control	Cierre	Revisión	Análisis	Comparación	Lecciones aprendidas	Toma de decisiones
MoP, ISO 2505, Políticas internas, metodología CEPAL					Políticas internas, PMBOK, PRINCE2, ISO21500					MoP, ISO 2505, Políticas internas, metodología CEPAL				

Figura II.1. Modelo de referencia

Se puede observar que la dimensión Proyecto sostiene el desarrollo de las otras tres dimensiones. La dimensión Analítica se desarrolla a la par de la etapa de formulación de un proyecto. La dimensión Infraestructura se desarrolla a la par de la etapa de implementación de un proyecto. La etapa de Evaluación Ex-Post de un proyecto se desarrolla una vez que la dimensión Gestión se encuentra en ejecución.

La organización de las actividades correspondientes a cada una de las dimensiones identificadas deberá realizarse de acuerdo a las mejores prácticas, estándares, guías de usuario, políticas y reglamentos internos que estén aprobados por la empresa de telecomunicaciones que desarrolle un proceso de analítica de datos. El modelo de referencia aquí propuesto expone a manera de ejemplo algunos de los documentos más

empleados actualmente para desarrollar cada una de las dimensiones identificadas como Infraestructura, Gestión y Proyectos.

GUÍA DE USUARIO:

Para la dimensión de Analítica de datos se han definido 5 fases para la ejecución de una serie de actividades y tareas que se detallan a continuación, y que se organizan de acuerdo a lo descrito en la Figura II.2:

Dimensión Analítica				
Fase: Comprensión del Negocio	**Fase: Comprensión de los Datos**	**Fase: Preparación de los datos**	**Fase: Modelado**	**Fase: Evaluación**
Determinar los objetivos del negocio	Levantar información referencial	Seleccionar los datos	Determinar requerimientos del modelo	Buscar patrones
Evaluar el entorno	Recolectar muestra de datos iniciales	Capturar datos	Seleccionar técnicas de modelado	Interpretar patrones
Evaluar la situación	Describir los datos	Clasificar y filtrar datos	Generar diseño de pruebas	Definir eventos o alertas
Determinar los objetivos del proceso de analítica	Explorar los datos	Proteger el contenido de los datos	Elaborar los modelos	Evaluar los resultados
Crear un reporte de entendimiento del negocio	Verificar la calidad de los datos	Limpiar los datos	Experimentar con los modelos	Revisar el proceso
	Crear reporte de entendimiento de los datos	Construir los datos	Valorar los modelos	Determine siguientes pasos
		Integrar los datos		Retroalimentar los casos de uso
		Dar formato a los datos		
		Crear reporte de preparación de los datos		

Figura II.2. Estructura de fases y actividades correspondientes a la Dimensión Analítica

1. Comprender el negocio

La primera fase de la metodología servirá para conocer el punto de partida y el destino para el cual se deberá enfocar los esfuerzos de todas las actividades desarrolladas.

1.1. Determinar los objetivos del negocio

La primera actividad de la metodología consiste en conocer y comprender lo que el negocio desea conseguir y cómo un proceso de analítica de datos podría ayudar en ese sentido.

1.1.1. Compilar el trasfondo del negocio

Antes de poder determinar los objetivos del negocio es necesario comprender al negocio en sí. En esta actividad se debe estudiar al negocio para saber a qué se dedica, cuál es su forma de operar, qué desafíos está atravesando y en qué mercado opera.

1.1.2. Identificar los objetivos del negocio

El siguiente paso consiste en averiguar qué está intentando conseguir el negocio, sea de forma general (misión y visión, por ejemplo), sea de forma temporal (metas y objetivos empresariales anuales, por ejemplo), sea de forma puntual (metas y objetivos establecidos para el proyecto de analítica de datos, por ejemplo).

1.1.3. Determinar los criterios de éxito del negocio

Finalmente, se debe determinar en base a qué criterios el negocio evaluará los resultados obtenidos en el proceso de analítica de datos, qué expectativas existen al respecto y cómo se decidirá si estos resultados superan, cumplen o no satisfacen lo esperado.

1.2. Evaluar el entorno

Adicionalmente a conocer y comprender el negocio, se debe conocer y comprender el entorno en el que se desarrolla. Parte del entorno está conformado por el mercado, por la economía, por el marco legal y regulatorio, por las políticas nacionales y empresariales, etc.

1.2.1. Evaluar factores legales y regulatorios

En esta actividad se debe levantar y evaluar las consideraciones legales y regulatorias que pueden afectar, condicionar o influir en el proceso de analítica de datos. Debido a que el sector de telecomunicaciones está fuertemente regulado, esta actividad deberá realizarse de con mucha atención.

1.2.2. Evaluar políticas y reglamentos internos

En esta actividad se debe levantar y evaluar las políticas y los reglamentos propios de la empresa que pueden afectar, condicionar o influir en el proceso de analítica de datos.

1.3. Evaluar la situación

Previo a iniciar un proceso o proyecto de analítica de datos, es necesario conocer la situación de la empresa en cuanto a su capacidad para ejecutar tal proceso o proyecto.

1.3.1. Levantar un inventario de recursos

Es necesario conocer con qué recursos cuenta la empresa, tanto humanos, como financieros, tecnológicos, logísticos y su disponibilidad temporal.

1.3.2. Determinar requerimientos, supuestos y restricciones

Es importante registrar a lo largo del proceso analítico todo requerimiento, supuesto y restricción que se vaya estableciendo desde el inicio hasta el final del trabajo. Conocer a todo momento los requerimientos y restricciones ayudará a evitar concluir con resultados que no cumplen con lo esperado. Conocer a todo momento los supuestos permitirá detectar con agilidad aquellos problemas cuyo origen sea un supuesto que no corresponde a la realidad, evitando la pérdida de tiempo valioso.

En este punto es importante entender que algunos supuestos se harán de forma consciente y otros supuestos se harán de forma inconsciente. Adicionalmente, puede ser que algunos supuestos se realicen considerándolos como hechos fácticos incuestionables. Es importante por lo tanto contar con un equipo que sea capaz de identificar y reconocer las situaciones en las que el trabajo ha tomado rumbos equivocados que deben ser corregidos mediante cambios de paradigmas.

1.3.3. Compilar un glosario de términos

De igual forma es importante registrar a lo largo del proceso analítico aquellos términos que se relacionen con el proceso de analítica. Este proceso será ejecutado seguramente por un equipo multidisciplinario e involucrará una amplia gama de herramientas, técnicas, tecnologías y conocimientos. Contar con un glosario facilitará el trabajo en equipo y la exposición del avance del trabajo, así como la exposición de los resultados.

1.4. Determinar los objetivos del proceso de analítica de datos.

Alineados con los objetivos del negocio, los objetivos de la analítica de datos deberán establecerse con el fin de que se conozca claramente qué se debe hacer y qué se debe conseguir a nivel técnico, a lo largo del proceso. Se podría decir que los objetivos del

negocio definirán el “Qué”, mientras que los objetivos del proceso de analítica de datos definirán el “Cómo”.

1.4.1 Revisar casos de uso previamente desarrollados.

Si la empresa realiza procesos de analítica de datos con frecuencia, entonces debe contar con un registro de “casos de uso”. Este registro servirá como una guía histórica y base de conocimiento de todos los procesos de analítica que se hayan desarrollado y podrá ser consultada para averiguar de qué forma se habrían desarrollado anteriores procesos, cuales han sido las lecciones aprendidas en el pasado y si existe algún trabajo previo que permita conseguir los objetivos planteados. Al final de cada proceso de analítica de datos se deberá alimentar esta referencia.

1.5. Crear un reporte de entendimiento del negocio.

La información que se recabe en la primera fase de la metodología deberá plasmarse en un documento que servirá de reporte del trabajo realizado y de referencia a lo largo del resto de fases.

2. Comprender los datos

La segunda fase de la metodología consiste en realizar un primer acercamiento a los datos correspondientes al tráfico que se busca analizar. El objetivo de esta fase es conocer y comprender la estructura, conformación, características y componentes del tráfico con el que se va a trabajar.

2.1. Levantar información referencial

Dada la complejidad inherente a las redes modernas de telecomunicaciones y de su tráfico, así como el nivel de especialización que se podría requerir para analizar dicho tráfico, es importante levantar de forma previa información que facilite el trabajo analítico.

2.1.1. Levantar topología de red

Una referencia importante será la topología lógica y física, tanto de la red de telecomunicaciones, como de los servicios que serán analizados.

Existen herramientas que facilitan llevar a cabo esta actividad. Una de las más empleadas es Microsoft Visio, que permite realizar diagramas de red de forma manual. También existen herramientas de mapeo de red que pueden generar topologías de forma automática, como PRTG de la empresa Paessler o Network Topology Mapper de la empresa Solarwinds. Para el caso de empresas de telecomunicaciones también se

puede recurrir a las herramientas OSS/BSS (*Operational Support Systems / Business Support Systems*) que estén disponibles y que generalmente permiten visualizar los diagramas de las redes gestionadas por dicho software.

2.1.2. Levantar y documentar pilas de protocolos

Otra referencia importante será la información relacionada con los protocolos que sean utilizados por los servicios que hace uso de la red de telecomunicaciones analizada.

Dado que las tecnologías de telecomunicaciones están mayoritariamente estandarizadas, gran parte de los protocolos que se pudieran encontrar en una red de telecomunicaciones estarán disponibles en Internet. Sin embargo, es posible que se encuentren protocolos propietarios cuya descripción sea guardada con recelo por la empresa que la desarrolló. El análisis de los factores legales y regulatorios podría habilitar o no al equipo para tomar acciones que permitan descubrir cómo funcionan estos protocolos, mediante, por ejemplo, ingeniería inversa.

2.1.3. Levantar lógica de servicios

Complementariamente a la información de las pilas de protocolos, será necesario conocer la lógica que usan los servicios que serán analizados para comunicarse a través de la red de telecomunicaciones. Un ejemplo de esta lógica puede ser el proceso de establecimiento de conexiones TCP conocido como "*three way handshake*".

Para servicios estandarizados, la lógica de los servicios podría ser recuperada con facilidad. Sin embargo, en la actualidad existe un auge de nuevas aplicaciones y servicios que se sirven de topologías complejas de funcionamiento y cuyo comportamiento es conocido únicamente por los desarrolladores de dichas aplicaciones y servicios. Adicionalmente, existe una tendencia a encriptar el tráfico generado por las aplicaciones y por tanto conocer la lógica de los servicios subyacentes puede requerir trabajos de ingeniería inversa y la utilización de herramientas como Bro [https://www.bro.org/] o Snort [https://snort.org/], que permiten hacer una inspección profunda de los paquetes y analizar el comportamiento de los flujos de datos. Nuevamente, el análisis de los factores legales y regulatorios podrán habilitar o no al equipo a realizar este tipo de actividades.

2.2. Recolectar muestra de datos iniciales

El siguiente paso que se requerirá tomar para llevar a cabo un proceso de analítica de datos es recolectar muestras de datos iniciales. En redes de telecomunicaciones se suelen denominar a estas muestras "trazas" o "capturas" de tráfico. La cantidad de datos

que se debe capturar debe ser lo suficientemente pequeña como para que pueda ser analizada de forma manual sin utilizar infraestructura avanzada. Sin embargo, la muestra debe ser al mismo tiempo lo suficientemente grande como para que contenga información significativa y representativa de forma que permita obtener un buen acercamiento a lo que se obtendrá al analizar de forma sistemática la red.

La toma de muestras de una red de telecomunicaciones requiere realizarse en coordinación con las áreas de planificación, monitoreo, operación y mantenimiento de la empresa dueña de la red y siguiendo los procedimientos que determine dicha empresa. Hay que tener especial cuidado de no afectar el servicio al momento de tomar las muestras.

El proceso de recolección de información de una red de telecomunicaciones para llevar a cabo procesos de analítica de datos se puede efectuar, en primer lugar, en distintos puntos físicos de la red.

- El proceso puede realizarse en los puntos originales y finales de la red, es decir en los dispositivos de los usuarios, en los servidores o en los equipos de telecomunicaciones a los que se conectan los usuarios para acceder a ciertos servicios. En este caso se estaría haciendo un monitoreo en el extremo o borde de la red.
- El proceso también podría realizarse en puntos intermedios o internos de la red, como por ejemplo en medios de transmisión o en equipos de telecomunicaciones como ruteadores o switchs.
- Adicionalmente, están empezando a emerger soluciones de redes virtualizadas por lo que, en estos casos, el proceso mencionado podría ejecutarse en un entorno virtualizado.

Los procesos de recolección de información, independientemente de que se realicen en el borde o en el interior de una red, pueden basarse en contadores o en tráfico. En el primer caso se estaría generando y recolectando datos estadísticos mientras que en el segundo caso se estaría generando y/o capturando tráfico real proveniente de la red.

- Para la generación y recolección de datos estadísticos, existen diferentes herramientas que pueden emplearse, dependiendo de su disponibilidad. Estas herramientas corresponden a aquellos protocolos y paquetes de software que permiten generar los datos estadísticos deseados. Ejemplos de protocolos comunes para estas herramientas son SNMP (*Simple Network Management Protocol*), RMON (*Remote Monitoring*) o Netflow. Estas herramientas generan contadores en los

dispositivos, servidores o equipos de telecomunicaciones monitoreados y los actualizan en base a mediciones que luego son recuperadas para análisis. Estos contadores pueden ser, por ejemplo, cantidad de paquetes recibidos, cantidad de paquetes enviados, cantidad de errores detectados, etc.

- Para la generación y/o recolección de tráfico, se puede utilizar métodos activos, pasivos o una combinación de ambos.
    - Los métodos activos consisten en generar tráfico en la red de telecomunicaciones. La herramienta más conocida, básica y utilizada para este propósito es el protocolo ICMP junto el programa PING, disponibles en la mayoría de sistemas operativos y equipos de telecomunicaciones. Sin embargo, existen herramientas más especializadas como, por ejemplo, sondas activas capaces de generar y simular llamadas telefónicas celulares. El dispositivo Mobile Robot R400 de la empresa iQsim es una opción disponible en el mercado que puede emplearse para pruebas de calidad de servicio, de roaming y de aseguramiento de ingresos, entre otras cosas. Estas herramientas pueden generar y recolectar datos estadísticos y el tráfico en sí.
    - Los métodos pasivos consisten en capturar tráfico de la red de telecomunicaciones, sin que haya sido generado como parte de dichos métodos. El proceso de captura podría realizarse gracias a que existen medios de transmisión compartidos, como, por ejemplo, medios inalámbricos o redes basadas en hubs. En estos casos, la técnica consiste en ubicar herramientas apropiadas (por ejemplo, computadoras con software de captura como Wireshark) en lugares adecuados, de forma que reciban las señales de telecomunicaciones y las capturen. Sin embargo, las redes de telecomunicaciones más comunes son aquellas donde se utilizan medios de transmisión guiados no compartidos y en las cuales se deben seguir los métodos y usar las herramientas que se detallan más adelante.
    - Los métodos combinados usan de forma conjunta métodos activos y métodos pasivos. Un ejemplo puede ser utilizar métodos pasivos para capturar tráfico en la red proveniente de usuarios reales, y en caso de que no existan usuarios reales generando tráfico, activar métodos activos que generen el tráfico necesario para que sea capturado con los métodos pasivos en uso.

Los métodos de captura de tráfico en medios guiados no compartidos pueden realizarse en los equipos que están involucrados en las comunicaciones, como computadoras,

servidores o equipos de telecomunicaciones, o pueden realizarse interceptando el tráfico en los medios y equipos de transmisión/red.

- Un método que permite capturar tráfico en equipos es la utilización de interfaces internas virtuales provistas por los sistemas operativos de dichos equipos y el uso de software capaz de capturar tráfico a través de dichas interfaces. Se puede configurar el software que se desea monitorear y/o el sistema operativo para que curse tráfico por dichas interfaces virtuales. Posteriormente, en sistemas tipo Unix se puede emplear el software denominado dumpcap para capturar el tráfico deseado, mientras que en sistemas Windows se puede emplear el software rawcap.
- Otro método que permite capturar tráfico en equipos es la utilización del modo promiscuo en las tarjetas de red junto con software de captura como Wireshark o tcpdump.
- Por otro lado, para capturar tráfico en medios de transmisión guiados no compartidos existen los siguientes métodos:
    - Se puede configurar opciones de duplicación de tráfico en equipos de red como switchs y ruteadores. Estas opciones instruyen a los equipos de red a que generen copias del tráfico deseado y a que lo envíen a través de una interfaz específica, donde se puede conectar un equipo que se encargará de capturar el tráfico duplicado, generalmente usando interfaces de red en modo promiscuo. Estos métodos suelen conocerse bajo la denominación SPAN (Switched Port Analyzer), Port Mirroring o Port Monitoring.
    - También se puede insertar en el medio de transmisión dispositivos que permitan capturar el tráfico sin interferir en las comunicaciones, mediante técnicas conocidas como Machine in the Middle. Estos dispositivos pueden ser Hubs (que se encargarán de enviar el tráfico por todas las interfaces existentes), Network Taps (que permitirán “leer” el tráfico que cursa por el medio de transmisión interceptado), o equipos de captura especializados.

Otra opción más avanzada y riesgosa es la de utilizar técnicas lógicas de intercepción. Estas técnicas son, por ejemplo, ataques denominados *Man in the Middle* o *MAC Flooding*, las cuales son artilugios que consiguen que equipos de red como switchs, ruteadores, servidores o computadores tengan un comportamiento distinto al normal provocando que envíen su tráfico a un destino definido por el interceptor, donde se captura la información.

2.3. Describir los datos

Una vez que se hayan recopilado los datos iniciales, se deberá realizar un primer acercamiento mediante una descripción general de los mismos incluyendo información sobre el volumen, duración, origen y características del conjunto de datos.

2.4. Explorar los datos

La exploración del conjunto de datos recopilados consiste en un análisis más profundo de sus características. La exploración debe plantear las primeras hipótesis respecto a la información que puede contener los datos analizados, guiar el resto del proceso de analítica y suele hacer uso de herramientas de visualización y manipulación de los datos.

2.5. Verificar la calidad de los datos

El proceso de exploración de los datos debe permitir la verificación de la calidad de los datos analizados. La calidad dependerá de que se encuentre la totalidad de los datos necesarios para extraer el conocimiento deseado y de la presencia o ausencia de errores o valores faltantes en el conjunto de datos.

2.6. Crear reporte de entendimiento de los datos

La información que se recabe en la segunda fase de la metodología deberá plasmarse en un documento que servirá de reporte del trabajo realizado y de referencia a lo largo del resto de fases.

3. Preparar los datos

La tercera fase de la metodología consiste en realizar las transformaciones necesarias en los datos con los que se trabajará con el fin de que puedan ser procesados de forma sistemática extrayendo el conocimiento deseado.

3.1. Seleccionar los datos

Tras haber realizado las actividades necesarias para comprender los datos que serán analizados, es posible seleccionar de forma definitiva y precisa las fuentes y las características de los datos que se utilizarán a partir de esta fase, y que deberían ser aquellos a utilizar en caso de que se implemente una solución tecnológica de analítica de datos en la empresa de telecomunicaciones.

3.2. Capturar datos

Esta actividad es equivalente a aquella denominada “2.2. Recolectar muestra de datos iniciales”. La diferencia está en que esta vez se realizará una captura masiva y sistemática de los datos a analizar y es posible que la captura de datos se deba implementar de forma automatizada y permanente en la red de telecomunicaciones. Los métodos, técnicas y herramientas descritas en la sección 2.2 del presente Anexo tienen igual validez en esta actividad.

3.3. Clasificar y filtrar datos

Puesto que la captura de datos de una red de telecomunicaciones, por defecto, entrega todos los datos que cursan por los enlaces, interfaces o equipos donde se realiza la captura, es necesario clasificar y filtrar los datos que sirvan para el análisis que se requiera realizar.

Tanto la clasificación como el filtrado tienen por objeto: limitar la cantidad de datos que deberá ser procesada, retirar los datos que no sean útiles para el análisis, separar los flujos de datos de forma que puedan ser analizados de forma independiente y aislada, facilitar y agilizar el procesamiento de información, y permitir realizar una analítica de datos distribuida.

Una vez que se ha obtenido el tráfico, las mediciones o estadísticas deseadas de la red de telecomunicaciones analizada, puede ser necesario identificar, clasificar y/o filtrar los datos obtenidos. De acuerdo a la empresa Sandvine, [39] [40], la clasificación del tráfico de una red de telecomunicaciones abarca y requiere tanto su identificación, como su categorización, su medición y la extracción de su información.

La identificación del tráfico consiste en determinar el tipo de tráfico observado. De acuerdo a Sandvine, se puede utilizar tres técnicas para realizar esta identificación:

- Firmas (*signatures*): consiste en detectar patrones de comportamiento de un flujo de datos y compararlo con firmas de comportamiento conocidas. Si el comportamiento detectado corresponde a una firma conocida, entonces se determina que el tráfico correspondiente al flujo observado proviene del servicio conocido que produjo la firma equivalente. Una firma de tráfico correspondiente a una llamada de voz puede ser un flujo constante de tráfico de baja velocidad, mientras que una firma de tráfico correspondiente a una transmisión de video puede ser un flujo constante de tráfico de alta velocidad y una firma de tráfico correspondiente a navegación de páginas de internet puede ser la detección de ráfagas de tráfico espaciadas en el tiempo.

- Rastreadores (*trackers*): consiste en hacer un seguimiento del tráfico de control que gestiona un flujo de datos. Los rastreadores deben conocer el comportamiento básico de control de los servicios que rastrea, y cuando observa que el tráfico de control corresponde al de un servicio conocido, entonces se determina que el tráfico gestionado corresponde a dicho servicio. Un rastreador podría detectar una llamada celular al observar una secuencia específica de intercambio de mensajes entre un controlador de radio base (RNC), las bases de datos de visitantes y de base (VLR y HLR).
- Analizadores (*analyzers*): Los analizadores son similares a los rastreadores, pero tienen un nivel de conocimiento profundo de los protocolos y flujos que corresponden a distintos servicios. Los analizadores son capaces, no solamente de detectar patrones de comportamiento en los flujos de datos y en el tráfico de control correspondiente, sino que también son capaces de extraer información e incluso interactuar o interferir con los servicios.

El resultado de la identificación de tráfico debe revelar información importante sobre los siguientes aspectos que son relativos al tráfico observado. Ejemplos de esta información son Protocolo, Aplicación, Servicio, Proveedor y Red de origen/destino.

Una vez identificado el tráfico, se puede proceder a categorizarlos. Ejemplos de categorías son: Servicio de almacenamiento, Juegos en línea, Servicios de comunicación y Tráfico de navegación.

El proceso de identificación y clasificación puede aprovecharse para realizar tomar mediciones y métricas del tráfico observado. Se puede así conocer el volumen de tráfico, la velocidad de los flujos, la calidad del servicio experimentado, la cantidad de usuarios que usan un mismo servicio, entre otras cosas.

Finalmente, se puede aprovechar el proceso para extraer información como atributos y características propias del servicio detectado y tomar ciertas acciones. Se puede, por ejemplo, detectar los códecs y la resolución empleada para un servicio de streaming de video, definir eventos que disparen alertas en caso de que se detecte problemas de calidad en el servicio, asociar el flujo con un usuario específico de la red y alimentar una base de datos que sea utilizada por las áreas de inteligencia del negocio para contactarse con el usuario e investigar más afondo los problemas de servicio que experimentó.

El proceso de clasificación de tráfico, de acuerdo a Sandvine, debe superar una serie de retos que se listan a continuación:

- Los servicios pueden intentar ocultar su información mediante técnicas de encapsulamiento, cifrado, ofuscación, empleo de *proxies*, compresión y empleo de túneles de comunicación.
- El proceso puede generar falsos positivos y falsos negativos en la identificación y clasificación del tráfico.
- El tráfico puede cursarse a través de múltiples flujos y sesiones, por lo que el proceso debe conocer a fondo el diseño y funcionamiento de los protocolos y aplicaciones involucrados y puede requerir recursos de procesamiento importantes.
- El tráfico puede cursarse por distintos recursos físicos de red, por lo que la extracción de datos de la red puede ser tan compleja como la misma red.

Finalmente, en cuanto a herramientas se refiere, la tecnología más apropiada para llevar a cabo procesos de identificación y clasificación de tráfico es la denominada *Deep Packet Inspection* (DPI) o Inspección Profunda de Paquetes, y sus variaciones *Network Intrusion Detection* y *Network Intrusion Prevention*. Estos sistemas están diseñados para procesar en gran detalle el tráfico proveniente de una red y pueden emplear las técnicas aquí descritas para así determinar la naturaleza de los flujos de datos analizados. Ejemplos de paquetes de software DPI que pueden emplearse de forma libre son Bro [41] y Snort [42].

3.4. Proteger el contenido de los datos

Considerando que el sector de telecomunicaciones está fuertemente regulado y que las empresas de telecomunicaciones están obligadas a proteger la privacidad, intimidad y secreto de las comunicaciones de sus usuarios, una actividad crítica que se deberá realizar en un proceso de analítica de datos será la de Proteger el contenido de los datos.

La protección del contenido puede realizarse mediante la eliminación de información personal de los usuarios, mediante el reemplazo de la información personal de los usuarios por información de relleno, o mediante la modificación de datos relacionada con los usuarios por datos aleatorios equivalentes.

La necesidad de anonimizar y/o censurar el contenido del tráfico se explicó en la Sección 2.4.1 y consiste en evitar que información privada e íntima de los usuarios sea expuesta a terceros, así como prevenir que información que revele la infraestructura de una red sea revelada [43]. Esta información no solo se encuentra en el *payload* de cada paquete con datos generados por los usuarios, sino también en las cabeceras de cada paquete,

aunque estas no hayan sido generadas directamente por los usuarios. Las direcciones IP, por ejemplo, que se encuentran en las cabeceras, también son información sensible pues permiten rastrear a los usuarios y por tanto violar su derecho a la privacidad e intimidad.

El proceso de anonimización y/o censura de tráfico puede realizarse mediante la eliminación o transformación de la información que se requiere proteger. La eliminación de información es un método sencillo, pero puede remover datos indispensables para realizar un análisis acertado del tráfico. La transformación de información, en cambio, es un método complejo pues la modificación de datos debería realizarse de una forma congruente en el tiempo y con respecto a la naturaleza del tráfico procesado, con el fin de no alterar el resultado del análisis de dicho tráfico.

Para el caso de información no generada por los usuarios y que generalmente se encuentra en las cabeceras de los paquetes procesados, el método más utilizado es la transformación. Se puede, por ejemplo, modificar las direcciones MAC, las direcciones IP, los *time-stamps* de los paquetes, los contadores y los números de protocolos y puertos [44]. Para el caso de información generada por los usuarios y que generalmente se encuentra en los *payloads* de los paquetes procesados, el método más utilizado es la eliminación.

Herramientas que pueden servir para anonimizar tráfico son: Crypto-Pan [45], Anontool [43], FLAIM [46] y tcpdpriv [47].

3.5. Limpiar los datos

La actividad de limpieza de datos tiene por objetivo subsanar en la medida de lo posible los problemas detectados en la actividad “2.5. Verificar la calidad de los datos”. Esta actividad puede llevarse a cabo mejorando las fuentes de datos utilizadas para la actividad “3.2. Capturar datos”, o empleando métodos y técnicas de completado, estimación o interpolación de datos.

3.6. Construir los datos

La actividad de construcción de datos permite agregar información al conjunto de datos analizado. Esta construcción puede realizarse mediante fuentes de información externas a la red de telecomunicaciones, como por ejemplo bases de datos georreferenciadas o bases de datos comerciales.

3.7. Integrar los datos

Las redes de telecomunicaciones modernas pueden transportar datos correspondientes a un mismo flujo de comunicación a través de diferentes caminos o enlaces. El proceso de captura de datos de cada enlace de la red de telecomunicaciones debería pasar por un proceso de integración de datos, donde se correlacionen los paquetes de forma que se pueda tener acceso a la totalidad de datos correspondientes a cada flujo. Si bien esta actividad puede llegar a ser muy desafiante por cuestiones de diversidad de rutas geográficas o incluso por cuestiones de movilidad de los usuarios, es absolutamente necesaria para permitir obtener resultados correctos de los análisis realizados al tráfico de telecomunicaciones.

3.8. Dar formato a los datos

El proceso de formato de datos consiste en unificar y adecuar el formato de los datos analizados para que puedan ser procesados correctamente por las herramientas de análisis, visualización y computación que estén disponibles. Esta actividad puede ayudar a facilitar y agilizar el procesamiento de información.

3.9. Crear reporte de preparación de los datos

La información que se recabe en la tercera fase de la metodología deberá plasmarse en un documento que servirá de reporte del trabajo realizado y de referencia a lo largo del resto de fases.

4. Construir los modelos

La cuarta fase del proceso de analítica tiene por objetivo construir los modelos que servirán para extraer conocimiento de los datos que sean analizados. Estos modelos tendrían que implementarse en las soluciones tecnológicas que diseñe la empresa para sistematizar y automatizar el análisis del tráfico de la red, en caso de que así se decida.

4.1. Determinar requerimientos del modelo

El primer paso para construir los modelos necesarios será determinar qué requerimientos, características y criterios deberán aplicarse a los modelos que se construirán.

4.2. Seleccionar técnicas de modelado

El análisis, visualización y modelado de información proveniente de una red de telecomunicaciones puede realizarse mediante los métodos, técnicas y herramientas tradicionales descritas en libros de minería de datos [48] [49] [50] [51] [52], los cuales

se han aplicado exitosamente al análisis de tráfico de redes de telecomunicaciones [53] [54]. Sin embargo, herramientas de análisis como Wireshark [55] y de visualización como QlikView [56] o Google Charts [57], pueden resultar de gran utilidad para analizar, visualizar y modelar información proveniente de redes de telecomunicaciones, haciéndolas más fáciles e intuitivas de entender. La sección 3.4.4 muestra cómo el uso de estas herramientas visuales puede ayudar a visualizar y modelar información proveniente de una red de telecomunicaciones.

4.3. Generar diseño de pruebas

El tercer paso para construir los modelos necesarios será definir y generar métodos o procedimientos que permitan poner a prueba la calidad y efectividad de los modelos implementados. Los diseños de prueba permitirán validar que los modelos construidos realmente permitan alcanzar los objetivos definidos en la actividad “1.4. Determinar los objetivos del proceso de analítica de datos”. Para poner a prueba los modelos, es recomendable disponer de conjuntos de datos de entrenamiento y de prueba, que permitan validar tanto los modelos generados como los procedimientos de verificación diseñados.

4.4. Elaborar los modelos

El cuarto paso consiste en estructurar, definir y elaborar los modelos necesarios. Es aconsejable elaborar varios modelos.

4.5. Experimentar con los modelos

El quinto paso consiste en aplicar los modelos elaborados en el análisis de conjuntos de datos, poner a prueba los resultados mediante los métodos o procedimientos de pruebas establecidos y evaluar los resultados considerando lo que se haya determinado en las fases 1 y 2 de la metodología, relacionadas con el entendimiento del negocio y de los datos, así como la definición de los respectivos objetivos.

4.6. Valorar los modelos

Finalmente, los resultados obtenidos mediante los distintos modelos elaborados deberán ser comparados y valorados. Esta valorización dependerá del método de modelado utilizado y permitirá detectar aquellos modelos que no cumplen con las expectativas del negocio y del proceso de analítica, y aquellos que si cumplen. Aquellos que cumplan con los mejores puntajes podrán ser empleados e implementados en eventuales soluciones tecnológicas empresariales. La valoración de los modelos puede realizarse, por ejemplo, en base a niveles de precisión en el caso de modelos

matemáticos, facilidad de visualización en el caso de modelos esquemáticos o gráficos, y facilidad de comprensión en el caso de modelos descriptivos.

4.7. Crear el reporte de construcción de modelos

La información que se recabe en la cuarta fase de la metodología deberá plasmarse en un documento que servirá de reporte del trabajo realizado y de referencia a lo largo de la última fase.

5. Evaluar los resultados

La última fase del proceso de analítica de datos consiste en evaluar los resultados.

5.1. Buscar patrones

Puesto que se han construido y seleccionado modelos que cumplen con los objetivos del negocio y del proceso de analítica de datos, es posible ahora aplicar dichos modelos en el análisis de los datos determinados en la fase 3 de la metodología. Los modelos generarán resultados que deberán ser estudiados.

Una actividad importante correspondiente a este estudio es determinar patrones en los datos. Estos patrones permitirán obtener información extraída de los datos tras ser analizados con los modelos construidos previamente. A manera de ejemplo, se podría detectar que los picos de tráfico de una red de telecomunicaciones están directamente relacionados con la cantidad de usuarios conectados en la red.

5.2. Interpretar patrones

Una vez detectados los patrones, será necesario dar una interpretación a los mismos. Esta interpretación implica un proceso de razonamiento y entendimiento de la información extraída, y por tanto implica la generación de conocimiento. A manera de ejemplo, se podría interpretar el nivel de intensidad usado para colorear regiones de un mapa como zonas que concentran información de interés para los usuarios de una red o como zonas donde existe mayor penetración de usuarios de internet.

5.3. Definir eventos o alertas

A partir del conocimiento obtenido, será de interés de la empresa de telecomunicaciones definir eventos o alertas que permitan que las soluciones tecnológicas eventualmente implementadas permitan detectar instancias en las que se requiera tomar acciones. A manera de ejemplo, y siguiendo el caso descrito previamente, una empresa de telecomunicaciones podría querer mejorar su infraestructura en las zonas donde se

detecta mayor concentración de tráfico con el fin de brindar mejores servicios, mejorar sus ingresos o disminuir sus costos.

5.4. Evaluar los resultados

En la fase final de la metodología corresponde evaluar los resultados, para lo cual se debe contar con la participación de representantes del negocio, quienes ayudarán a validar si los objetivos planteados al inicio se han alcanzado y si los resultados están de acuerdo con los criterios de éxito definidos originalmente.

5.5. Revisar el proceso

Cualquiera que fuere el resultado obtenido, es necesario que se haga una revisión del proceso que se siguió durante la ejecución de la metodología analítica y se genere una retroalimentación que permita mejorar dicho proceso para futuras aplicaciones. A diferencia de la actividad previa, el presente trabajo consiste en determinar si la forma en que se lleva a cabo un proceso de analítica tiene oportunidades de mejora que puedan aplicarse a futuros procesos, obteniendo posteriormente un mejor desempeño del grupo de trabajo.

5.6. Determine siguientes pasos

Considerando la evaluación de resultados y la revisión del proceso, se deberá determinar los siguientes pasos a seguir. Un siguiente paso podría ser, por ejemplo, la implementación de una solución tecnológica que emplee los modelos construidos durante la ejecución de la metodología analítica para que extraigan conocimiento a partir del tráfico de la red de telecomunicaciones de forma sistemática y automática. Otro ejemplo podría ser actualizar los procedimientos y procesos que se siguen a lo largo de la metodología analítica para futuras ocasiones.

5.7. Retroalimentar los casos de uso

Finalmente, el resultado del proceso de analítica de datos deberá ser documentada y alimentar los casos de uso de la empresa de telecomunicaciones. Esta actividad consiste en crear y desarrollar una base de conocimiento que documenta todos los trabajos de analítica en los que ha participado el “negocio” y que puede utilizarse como referencia cada vez que se inicie un nuevo proceso de analítica. Las empresas grandes podrían iniciar en un momento dado un esfuerzo que ya fue realizado anteriormente sin saberlo. Los casos de uso permitirán llevar una memoria empresarial de todos los esfuerzos realizados y así evitar destinar recursos innecesarios repitiendo trabajos ya realizados.

ORDEN DE EMPASTADO