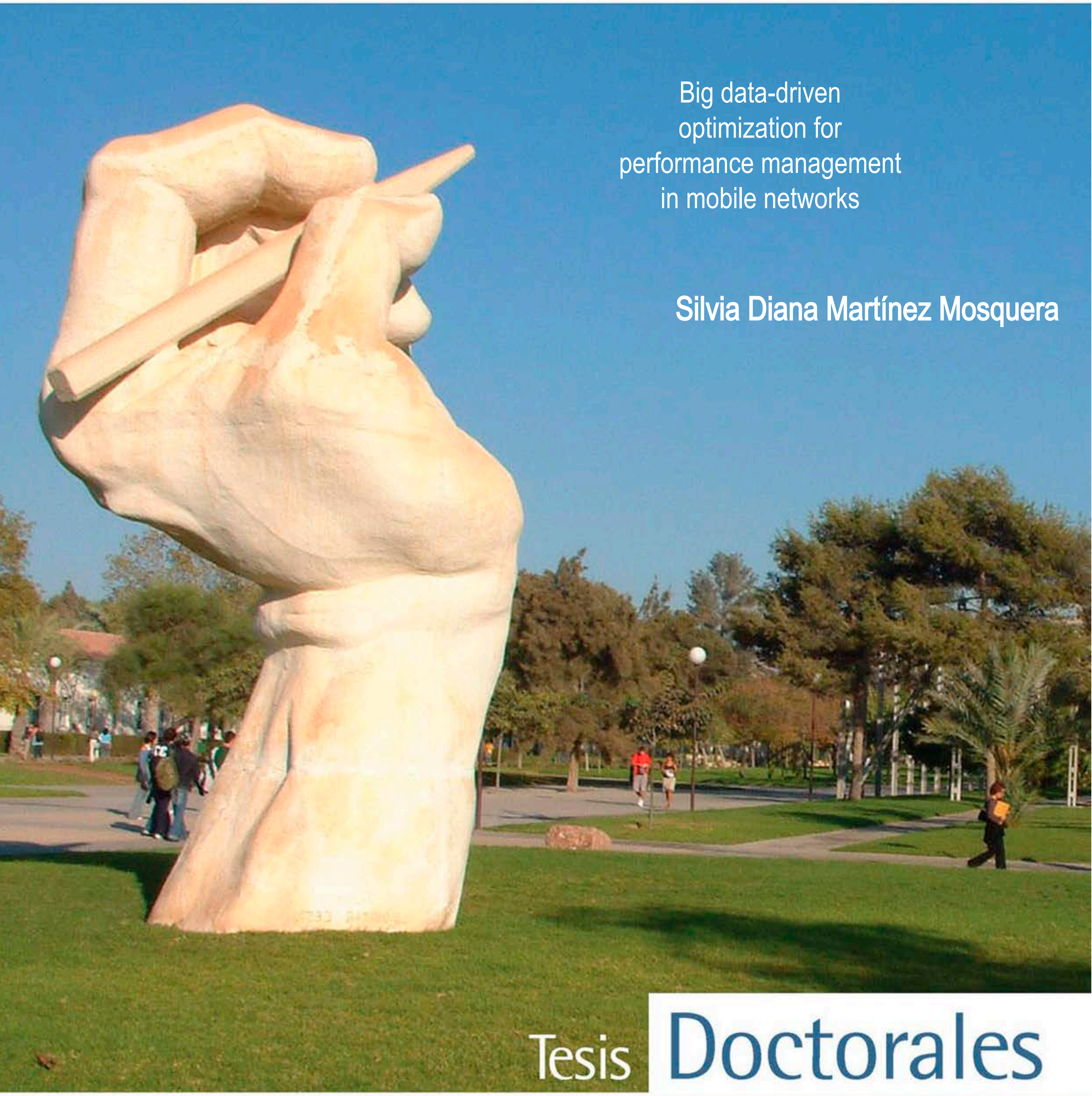


Big data-driven optimization for performance management in mobile networks

Silvia Diana Martínez Mosquera

Universitat d'Alacant
Universidad de Alicante

Instituto Universitario de Investigación Informática

Escuela Politécnica Superior

Big data-driven optimization for performance management in mobile networks

Silvia Diana Martínez Mosquera

Tesis presentada para aspirar al grado de

Doctora por la Universidad de Alicante

Doctorado en Informática

Dirigida por:

Dr. Sergio Luján Mora

Noviembre 2021

DOCTORAL THESIS IN THE FORM OF COMPENDIUM OF PUBLICATIONS

Big data-driven optimization for performance management in mobile networks

This document contains a summary of the work performed by Silvia Diana Martínez Mosquera, under the supervision of Dr. Sergio Luján Mora, to opt for the degree of Doctor in Informatics. It is presented at the University of Alicante and is structured according to the regulations established for the presentation of doctoral theses in the form of a compendium of publications. It contains the first part with the background and synthesis, the second part with the scientific publications, and the third part with conclusions and future work.

November 2021

TESIS DOCTORAL EN FORMA DE COMPENDIO DE PUBLICACIONES

Optimización basada en big data para la gestión del rendimiento en redes móviles

El presente documento contiene una síntesis del trabajo realizado por Silvia Diana Martínez Mosquera, bajo la dirección del Dr. Sergio Luján Mora, para optar por el grado de Doctora en Informática. Se presenta en la Universidad de Alicante y se estructura según la normativa establecida para la presentación de tesis doctorales en forma de compendio de publicaciones. Contiene una primera parte con los antecedentes y la síntesis, una segunda parte con las publicaciones científicas realizadas, y una tercera parte con las conclusiones y trabajo futuro.

Noviembre 2021

Dedication

This thesis is dedicated, with all my love, to my family who was always present offering me all their support and understanding from the beginning of this project. To my children Victoria and Nicolás, who understood, at a young age, that her mother had to occupy part of her family time to carry out her doctorate studies. To my husband Carlos Castillo, who always trusted me and with his trust and support is part of this dream come true. To my mother Rosa Mosquera, who allowed me to pave the way to reach this last step, and to all my family and friends who contributed in different ways to the realization of this project.

This work is also dedicated to my mentors Sergio Luján and Rosa Navarrete, who accompanied me unconditionally in this process and knew how to guide me with wisdom. And finally, this achievement is dedicated to everyone who inspired me to start and finish this journey, those people who despite their restrictions, both work, and family, reached the goal of obtaining their doctorate, making me understand that everything is possible with effort and dedication.

Esta tesis está dedicada, con todo mi cariño, a mi familia que estuvo siempre presente ofreciéndome todo su apoyo y comprensión desde el inicio de este proyecto. A mis hijos Victoria y Nicolás quienes comprendieron a su corta edad que su madre tenía que ocupar parte del tiempo familiar a realizar sus estudios de doctorado. A mi esposo Carlos Castillo, quien siempre confió en mi y con su confianza y apoyo es parte de este sueño hecho realidad. A mi madre Rosa Mosquera, quien me permitió labrar el camino para llegar a este último peldaño, y a todos mis familiares y amigos que aportaron de diferentes maneras a la realización de este proyecto.

Quiero también dedicar este trabajo a mis mentores Sergio Luján y Rosa Navarrete que me acompañaron incondicionalmente en este proceso y supieron guiarme con sabiduría. Y finalmente, dedico este logro a todos quienes me inspiraron para empezar y culminar este viaje, a aquellas personas que a pesar de sus restricciones, tanto laborales como familiares, llegaron a la meta de obtener su doctorado, haciéndome comprender que todo es posible con esfuerzo y dedicación.

Acknowledgments

"Yours they are, Lord,
greatness and power,
glory, victory and majesty.
Yours is all there is
in heaven and on earth."
Chronicles 29:11

Thank you, God, for giving me life, health, people, circumstances, skills, and everything that allowed me to fulfill this goal in my life.

I would like to thank the University of Alicante for accepting me as a student of the Doctoral School and allowing me to obtain an education of excellence. I owe my most sincere appreciation to my advisor Dr. Sergio Luján Mora for being my mentor inspired me, motivated me through all these years, and teaching me how to do scientific research and always seeking quality work. I also thank Dr. Jorge Azorin and Dr. Jose García for allowing me to carry out certain activities from my country and to all the academic and administrative personnel that make education for foreigners possible.

I also thank the National Polytechnic School, especially the Faculty of Systems Engineering and its head of Department Dr. Rosa Navarrete for co-authored the publications and for her invaluable intellectual contribution to the research. Likewise, thanks to the Advanced Data Analytics, Laboratory and its director Dr. Edison Loza for allowing me to carry out the tests at their facilities.

Acknowledgments

*"Tuyos son, Señor,
la grandeza y el poder,
la gloria, la victoria y la majestad.
Tuyo es todo cuanto hay
en el cielo y en la tierra."
Crónicas 29:11*

Agradezco a Dios por darme la vida, la salud, las personas, las circunstancias, las habilidades y todo cuánto me permitió cumplir esta meta en mi vida.

Agradezco a la Universidad de Alicante por aceptarme como estudiante de la Escuela de Doctorado y permitirme obtener una educación de excelencia. Debo mi más sincero agradecimiento a mi director Dr. Sergio Luján Mora por ser mi mentor, inspirarme y motivarme a lo largo de todos estos años, y por enseñarme cómo hacer investigación científica y buscar siempre un trabajo de calidad. También agradezco al Dr. Jorge Azorin y al Dr. José García y a todo el personal académico y administrativo que hace posible la educación a extranjeros.

Agradezco también a la Escuela Politécnica Nacional, en especial, a la Facultad de Ingeniería de Sistemas y su jefa de Departamento Dra. Rosa Navarrete por su co-autoría en las publicaciones y su invaluable aporte intelectual en la publicación de los mismos. De igual manera, gracias al Laboratorio de Analítica de Datos Avanzados y su director Dr. Edison Loza por permitirme llevar a cabo mi investigación en sus instalaciones.

Quito, 14 de octubre de 2021
Silvia Diana Martínez Mosquera

Abstract

Humanity, since its inception, has been interested in the materialization of knowledge. Various ancient cultures generated a lot of information through their writing systems. The beginning of the increase of information could date back to 1880 when a census performed in the United States of America took 8 years to be tabulated. In the 1930s the demographic growth exacerbated this increase of data. Already in 1940, libraries had collected a large amount of writing and it is in this decade when scientists begin to use the term "information explosion". The term first appears in the Lawton (Oklahoma) Constitution newspaper in 1941.

Currently, it can be said that we live in the age of big data. Exabytes of data are generated every day; therefore, the term big data has become one of the most important concepts in information systems. Big data refer to large amounts of data on a large scale that exceeds the capacity of conventional software to be captured, processed, and stored in a reasonable time. As a general criterion, most experts consider big data to be the largest volume of data, the variety of formats and sources from which it comes, the immense speed at which it is generated, the veracity of its content, and the value of the information extracted/processed.

Faced with this reality, several questions arise: How to manipulate this large amount of data? How to obtain important results to gain knowledge from this data? Therefore, the need to create a connecting bridge between big data and wisdom is evident. People, machines, applications, and other elements that make up a complex and constantly evolving ecosystem are involved in this process.

Each project presents different peculiarities in the development of an framework based on big data. This, in turn, makes the landscape more complex for the designer since multiple options can be selected for the same purpose. In this work, we focus on an framework for processing mobile network performance management data. In mobile networks, one of the fundamental areas is planning and optimization. This area analyzes the key performance indicators to evaluate the behavior of the network. These indicators are calculated from the raw data sent by the different network elements.

The network administration teams, which receive these raw data and process them, use systems that are no longer adequate enough due to the great growth of networks and the emergence of new technologies such as 5G and 6G that also include equipment from the Internet of things.

For the aforementioned reasons, we propose in this work a big data framework for processing mobile network performance management data. We have tested our proposal using performance files from real networks. All the processing carried out on the raw data with XML format is detailed and the solution is evaluated in the ingestion and reporting components.

This study can help telecommunications vendors to have a reference big data framework to face the current and future challenges in the performance management in mobile networks. For instance, to reduce the processing time data for decisions in many of the activities involved in the daily operation and future network planning.

Resumen

La humanidad, desde sus inicios, se ha interesado por la materialización del conocimiento. Varias culturas antiguas generaron mucha información a través de sus sistemas de escritura. Ya en la década de los 40's, las bibliotecas recopilaron una gran cantidad de escritura y el interés en la producción y el procesamiento de estos grandes volúmenes de datos comenzó a crecer.

Actualmente, se puede decir que vivimos en la era del big data. Todos los días se generan exabytes de datos; por lo tanto, el término big data se ha convertido en uno de los conceptos más importantes en los sistemas de información. Big data se refiere a grandes cantidades de datos a gran escala que exceden la capacidad del software convencional para ser capturados, procesados, y almacenados en un tiempo razonable. Como criterio general, la mayoría de los expertos consideran que el big data es el gran volumen de datos, la variedad de formatos y fuentes de donde proviene, la inmensa velocidad a la que se genera, la veracidad de su contenido y el valor de la información extraída/procesada.

Ante esta realidad, surgen varias preguntas: ¿Cómo manipular esta gran cantidad de datos? ¿Cómo obtener resultados importantes para ganar sabiduría a partir de estos datos? Por lo tanto, es evidente la necesidad de crear un puente de conexión entre big data y sabiduría. En este proceso intervienen personas, máquinas, aplicaciones y demás elementos que conforman un ecosistema complejo y en constante evolución.

Cada proyecto presenta diferentes peculiaridades en el desarrollo de un ecosistema basado en big data. Esto, a su vez, hace que el paisaje sea más complejo para el diseñador, ya que se pueden seleccionar múltiples opciones para el mismo propósito. En este trabajo, nos enfocamos en un ecosistema para procesar datos de gestión del desempeño de redes móviles.

En las redes móviles, una de las áreas fundamentales es la planificación y optimización que analiza los indicadores clave de rendimiento para evaluar el comportamiento de la red. Estos indicadores se calculan a partir de los datos brutos enviados por los diferentes elementos de la red.

Los equipos de administración de redes utilizan sistemas que ya no son lo suficientemente adecuados debido al gran crecimiento de las redes y la aparición de nuevas tecnologías como 5G y 6G que también incluyen equipos de Internet de las cosas.

Por las razones antes mencionadas, proponemos en este trabajo un ecosistema de big data para procesar datos de gestión del desempeño de redes móviles utilizando archivos de desempeño de redes reales. Se detalla todo el procesamiento realizado sobre los datos brutos con formato XML y se evalúa la solución en los componentes de ingesta y reportes de datos.

Este estudio puede ayudar a los proveedores de telecomunicaciones a tener un marco de referencia de big data para enfrentar los desafíos actuales y futuros en la gestión del desempeño en redes móviles. Por ejemplo, para reducir el tiempo de procesamiento de datos para decisiones en muchas de las actividades involucradas en la operación diaria y la planificación futura de la red.

Contents

List of Figures

List of Tables

Part I

BACKGROUND AND SYNTHESIS

1 Introduction

1.1 Motivation and Work Undertaken

The telecommunications industry, in recent decades, has become one of the most varied and is growing at a rapid pace in the world (Kovačević, Krajnović, & Čičin Šain, 2017). We have witnessed the emergence of several generations of mobile networks from 2G to 5G today and 6G soon and, according to the Global System for Mobile Communications, for the year 2025, there are estimated about 8.8 billion subscriptions (Global System for Mobile Communications, 2021).

With this demand, to provide an adequate service to users, mobile network operators must constantly monitor and measure the performance of the thousands of network elements (NEs) that have been deployed around the world. Therefore, files with the respective data on the status of each NE are sent to the network management systems, generally every 15 minutes. Therefore, a large amount of raw data are generated and must be processed and analyzed to monitor the behavior of the network service in the shortest possible time (3rd Generation Partnership Project, 2005).

On the other hand, research on the processing of large volumes of data has been of enormous interest for a long time. Already in 1944, the rapid growth of libraries began to be studied (Rider, 1944) and for 1997 the term “big data” was introduced for the first time in a scientific study (Cox & Ellsworth, 1997). Big data analysis has allowed many companies to improve their competitive advantages and mobile network operators (MNOs) have not been the exception.

As presented in chapter 3 Literature Review, since 2012 the results of the research on mobile networks, related to big data, have been presented. The studies have focused on the analysis of techniques, tools, frameworks, user and network data. Several mobile network vendors have designed systems that exploit the benefits of big data platforms to support the network planning and optimization area (Nokia, 2020; Skračić & Bodrušić, 2017). This area is one of the most important in an MNO whose main role is to detect, correct, and predict the state of the network through the measurements sent by the NEs.

In this work, we have concentrated on a big data framework for mobile network performance measurements (PM) data with the premise of optimizing the required computational resources and the processing time. Optimization in mobile network management is a very complex and important issue due to the large number of devices that must monitor to determine problems or deterioration in the level of quality of the service provided. The downtime of a NE on the mobile network has a direct impact on revenue and operating expenses (OPEX). Therefore, the longer it takes to analyze the mobile network PM, the higher the OPEX for the MNO.

Based on this information, our proposed framework has evaluated the ingestion and reporting times in a cluster implemented over the cloud and bare metal, with real data sets collected measurements of 3G, 4G, and 5G cells from four different MNOs.

1.2 Objectives

The main objective of this thesis is to design a big data framework that optimizes the management of PM for mobile networks in terms of computational resources at the ingestion and reporting components. The specific objectives (SO) are the following:

SO. 1. **Determine the trends and gaps** in the research of the modeling and management of big data in databases through a systematic literature review.

SO. 2. **Present the state-of-the-art** of the research of big data and mobile networks.

SO. 3. **Define a method** for the selection of the big data tools.

SO. 4. **Define a method** to design a big data framework for PM in mobile networks.

SO. 5. **Propose a method** to process complex eXtensible Markup Language (XML) files used in PM for Hive.

SO. 6. **Propose a method** to process complex XML files used in PM for Spark.

SO. 7. **Propose a big data framework** for PM in mobile networks and its implementation.

SO. 8. **Evaluate the big data framework** for ingestion and query execution times with the big data tools selected.

The objectives of this thesis have been fully achieved during the research carried out through the years of the doctoral program. The results obtained have been evidenced in the publications in journals with Impact Factor. In addition, previous works were performed that do not contribute directly to the thematic unit, but are related to big data; these were presented in different peer-reviewed conferences. Table 1.1 shows the contribution of each of the articles to the fulfillment of these objectives.

Objective	**Articles**
SO. 1	An approach to big data modeling for key-value NoSQL (Martinez-Mosquera, Luján-Mora, Navarrete, Mayorga, & Vivanco, 2019). Modeling and management big data in databases — A systematic literature review (Martinez-Mosquera, Navarrete, & Luján-Mora, 2020b).
SO. 2	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera, Navarrete, & Luján-Mora, 2020a).
SO. 3	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera et al., 2020a).
SO. 4	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera et al., 2020a).
SO. 5	Efficiently processing complex XSD using Hive and Spark (Martinez-Mosquera, Navarrete, & Luján-Mora, 2021).
SO. 6	Efficiently processing complex XSD using Hive and Spark (Martinez-Mosquera et al., 2021).
SO. 7	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera et al., 2020a). Framework for big data integration in e-government (Martinez-Mosquera & Luján-Mora, 2019).
SO. 8	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera et al., 2020a). Efficiently processing complex XSD using Hive and Spark (Martinez-Mosquera et al., 2021).

Table 1.1: Articles that contribute to the objectives of this thesis

1.3 Hypothesis

The following hypothesis has been raised in this thesis: It is feasible to design, implement, and optimize the performance management of a mobile network at ingestion and reporting components, with a big data framework.

1.4 Method

The methodology used in this thesis is the Design Science Research, which is widely used in Information Systems, and whose fundamental postulate is that knowledge, understanding of a problem, and its solution is acquired in the application and construction of an artifact. In this work, the artifact constitutes the big data framework for performance management in mobile networks. For the application of this methodology we considered these main seven guides:

1. **Relevance of the problem:** To know the relevance of the problem, several literature reviews were done. The first review was performed during the 2018-2019 period and allowed us to know the state of the art of modeling and management of big data in databases. The next reviews performed in the 2019-2020 and 2020-2021 periods allowed us to know the relevant works presented on big data and mobile networks, focusing on the management of PM.

2. **Rigor of research:** The research is based on the application of rigorous methods. For the literature review study, the guidelines proposed by Kitchenham were used, focusing on three main phases: planning, conducting, and reporting (Kitchenham, 2004). The collection of the studies was performed from the most important scientific digital libraries.

   Furthermore, the initial requirements of a big data framework for PM in mobile networks were collected from a telecommunications vendor with the help of the IEEE 29148 standard (Institute of Electrical and Electronics Engineers, 2018). The results are presented in Appendix A.

   For the tests, data from real 2G, 3G, 4G, and 5G networks were used and several tests were performed over bare metal equipment, a solution in the cloud, and tests over different version tools.

3. **Design as a search process:** Different iterative methods were used to test design alternatives against requirements or constraints.

4. **Design as an artifact:** Our research produced a viable artifact (in this case, a framework).

5. **Design evaluation:** The solution is presented as independent layers, and the ingestion and reporting layers were evaluated to present the usefulness, quality, and effectiveness of a design artifact.

6. **Research contributions:** This research provides clear and verifiable contributions in the areas of design artifact, design fundamentals, and/or design methodologies.

7. **Communication of the investigation:** The investigation results have been presented effectively in journals and conferences related to the research area.

1.5 Structure of the Thesis

This thesis is structured in four parts. Part I Background and Synthesis includes the Introduction, Theoretical Foundations, Literature Review, Publications and Visibility, and the Description of the Work with results obtained in the research that comply with the objectives of the thesis. Part II Compendium includes Compendium of Publications with the details of all the work published during the investigation period. Finally, Part III Conclusions and Outlook includes the discussion, contribution, and future work.

This thesis is composed by the following chapters:

Chapter 1 Introduction includes motivation and work undertaken, objectives, hypothesis, method, the structure of the thesis, and typographic conventions.

Chapter 2 Theoretical foundations includes descriptions of the main topics of this research.

Chapter 3 Literature review includes the state of the art of the study of big data and mobile networks, specifically about performance management.

Chapter 4 Publications and visibility includes the articles published in journals, other publications, and the academic profiles of the author.

Chapter 5 Description of the work contains the most important results of the research obtained during the doctoral studies period.

Chapter 6 Compendium of publications contains the published studies.

Chapter 7 Framework for Big Data Integration in E-government, article published in the DYNA journal, including reference, contribution, and full text.

Chapter 8 An Approach to Big Data Modeling for Key Value NoSQL, article published in Iberian Journal of Information Systems and Technologies, including reference, contribution, and full text.

Chapter 9 Modeling and Management Big Data in Databases — A Systematic Literature Review, article published in the Sustainability journal, including reference, contribution, and full text.

Chapter 10 Development and Evaluation of a Big Data Framework for Performance Management in Mobile Networks, article published in the IEEE Access Journal, including reference, contribution, and full text.

Chapter 11 Efficiently Processing Complex XSD using Hive and Spark, article published in the PeerJ Computer Science journal, including reference, contribution, and full text.

Chapter 12 Conclusions include the discussion, contribution of the work, and the planned future research.

1.6 Typographic Conventions

In this thesis, the terms and abbreviations of the 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Telecommunication management; Performance Management; Concepts and requirements have been used mostly. Figures and tables are elaborated by the author and some of them come from the publications that make up this compendium. Additionally, the point has been used as a separator between the integer and decimal part of a number according to the International Organization for Standardization 80000 and the International System of Units.

2 Theoretical Foundations

This chapter includes the main basic concepts related to this research. First, the basis of big data are described to provide a general overview of the technology. Then, concepts related to performance management in mobile networks are presented and the process to collect the performance measurements and to create the Key Performance Indicators (KPIs).

2.1 Big Data

Since the beginning of humanity, the transfer of knowledge has been one of the greatest purposes of human beings. Many ancient cultures provided information to their descendants through their writing systems. Currently, it can be said that we live in the era of big data. The exabytes of data are generated every day; therefore, the term big data has become one of the most important concepts in information systems.

Big Data is the term that has been used in recent years to refer to a set of large and complex amounts of data that combine certain characteristics known as the Vs of big data (Ribeiro & da Silva, 2015). The processing and analysis of big data require non-traditional computer systems (Turck, 2020) developed to make the business takes timely and strategic decisions (Tekiner & Keane, 2013).

2.1.1 Big Data Characterization

As mentioned, to consider a large volume of data as big data, other characteristics such as velocity and variety must also comply. These three characteristics were well accepted for several years to define big data. However, by 2014, 10 characteristics had already been defined as shown in Table 2.1. For the year 2017, 42 V's were already discussed as shown in the Figure 2.1 and explained in Farooqi et al. (2019).

Characteristic	Brief Description
Volume	Large data sets
Variety	Different type of data formats
Velocity	High data generation rate
Value	Useful data to retrieve info
Veracity	Accuracy of data
Variability	Consistent data
Viscosity	Data velocity variations
Virality	Data transmission rate
Validity	Assessment of data
Visualization	Data symbolization

Table 2.1: The 10 Vs of big data (Martinez-Mosquera et al., 2020b)

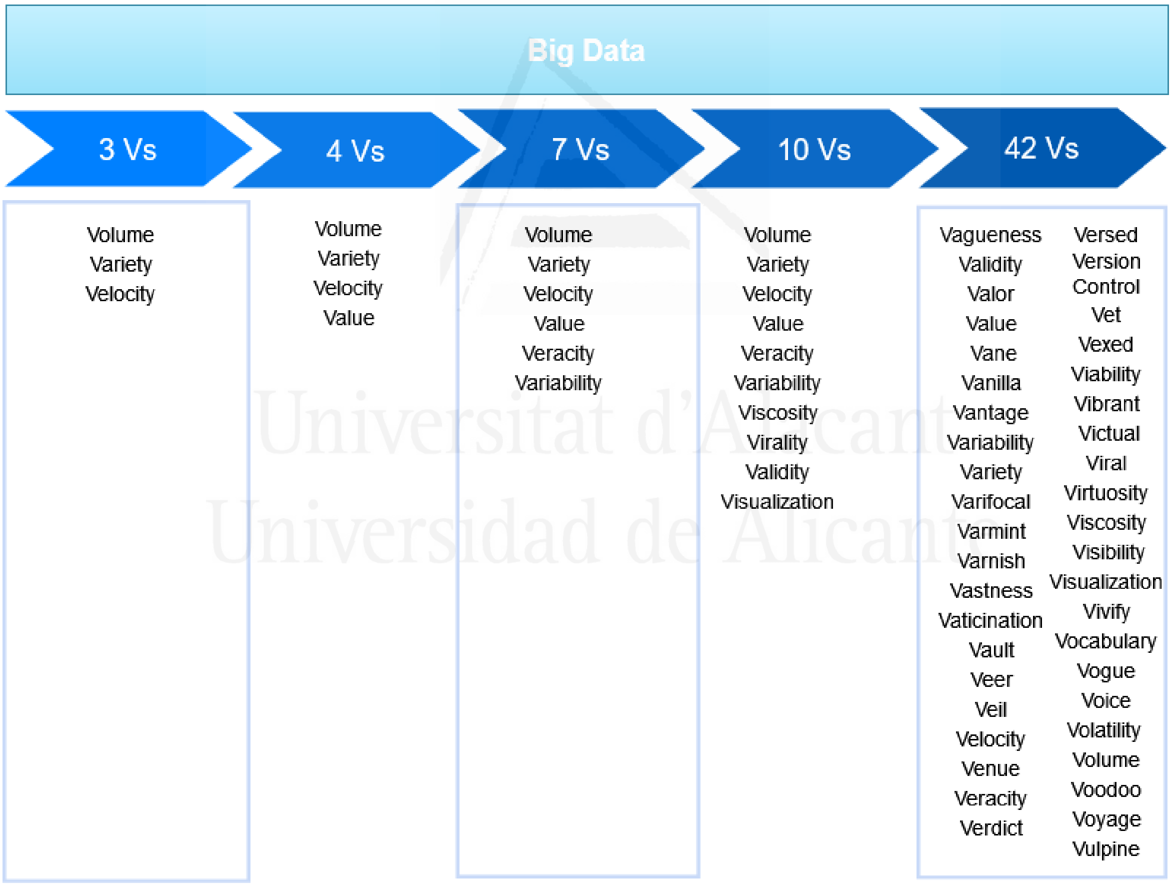


Figure 2.1: The evolution of the big data V's from three onwards

To deal with all these characteristics, the data have been classified into structured, semi-structured, and unstructured. According to this classification, there are processes, tools, and techniques that allow storage, manipulation, and management.

2.1.1.1 Structured Data

Structured data have been called those that are represented in tabular form, in spreadsheets or relational databases (Davoudian, Chen, & Liu, 2018). According to the report presented by the CISCO company (Barton & Henry, 2020), more than 90 % are unstructured data. Therefore, these type of data are explained in the next sections.

2.1.1.2 Semi-structured Data

Semi-structured data are considered to be data that do not have a formal structure, such as data within a relational database. Comma Separated Values (CSV) formats, Extensible Markup Language (XML), JavaScript Object Notation (JSON) and Binary JSON (BSON) formats are some of those considered semi-structured since they present a certain organization in its structure (Baek, Ahn, & Kim, 2016; P. Li, Gong, & Wang, 2020; Papadakis, 2018).

2.1.1.3 Unstructured Data

Those data that do not have a pre-defined or organized structure are known as unstructured. For example, text documents, emails, sensor data, audio files, image files, video files, website data, chats, electronic medical records, social network data, among others (Costa & Santos, 2017; O'Sullivan, Thompson, & Clifford, 2014).

Every data type can be processed in a big data architecture according to the business requirements.

2.1.2 Big Data Architecture

According to the IEEE 1471-2000 standard, architecture is the fundamental design of the organization of a computational system and its components (Institute of Electrical and Electronics Engineers, 2020). For big data, there are two available architectures: lambda and kappa (Ounacer, Talhaoui, Ardchir, Daif, & Azouazi, 2017).

Lambda architecture λ is composed of the results x from historical data from the batch layer (BL) and live streaming data from the serving layer (SRL) and the speed layer (SL). The operational sequencing is described in equation (2.1).

$$\lambda = \{x|x \in BL \vee x \in SRL\}. \tag{2.1}$$

Kappa architecture κ is composed of the results y from live streaming data from the SRL with the SL. The operational sequencing is described in equation (2.2).

$$\kappa = \{y|y \in SRL\}. \tag{2.2}$$

According to the architecture requirements presented in Appendix A, in this thesis, we follow a lambda architecture. The main components that any big data implementation must provide (Microsoft, 2020) are summarized in Table 2.2. The big data framework is composed of software architecture, ingestion, data lake, processing, reporting, and deployment.

Components	**Description**	**Options**
Software Architecture	"The fundamental organization of a system embodied in its components, their relationships to each other and to the environment, and the principles guiding its design and evolution" (Institute of Electrical and Electronics Engineers, 2020)	Lambda, Kappa
Ingestion	Mediator system for a massive collection of raw data from NEs to the data lake independently of format	Flume, Kafka, NiFi
Data Lake	Distributed file system that allows storing all kinds of data in an efficient way	Hadoop Distributed File System (HDFS)
Processing	Platform to provide a real-time processing of network data	Spark, Storm
Reporting	Framework to provide the use of SQL language to query data stored on the data lake	Hive, Impala, Pig, Spark SQL
Deployment	Platform to install and configure Hadoop cluster nodes and enabling agile application deployment	Cloudera, Hortonworks, IBM InfoSphere Big Insights, MapR, Pivotal HD

Table 2.2: Features and tools for a big data framework (Martinez-Mosquera et al., 2020a)

Since our study is focused on the development of a framework for performance management in mobile networks, the following section will explain the basic telecommunications concepts.

2.2 Mobile Network

Mobile networks are those composed of a set of NEs including base stations that offer wireless transmission and reception of digital information in one or several geographical areas called cells (Olsson, Sultana, & Mulligan, 2009). Figure 2.2 presents a common topology of a mobile network composed by the radio access and core networks. The wireless NEs to connect the end-users are included in the radio access and the NEs used for the management are located in the core.

Mobile network technologies have evolved rapidly in recent decades, from Global System for Mobile Communications (GSM) to General Packet Radio Service (GPRS), Enhanced Data Rates for GSM Evolution (EDGE), Universal Mobile Telecommunication System (UMTS), High-Speed Uplink Packet Access (HSUPA), High-Speed Downlink Packet Access (HSDPA), Long Term Evolution (LTE), and currently the Fifth-Generation (5G) (3rd Generation Partnership Project, 2021).

Any generation of a mobile network requires collecting performance data recorded by its NEs. This network procedure is called performance management. The purpose is to obtain information from the network to verify the physical and logical configurations and to locate network problems. The type of data that is collected is called PM.

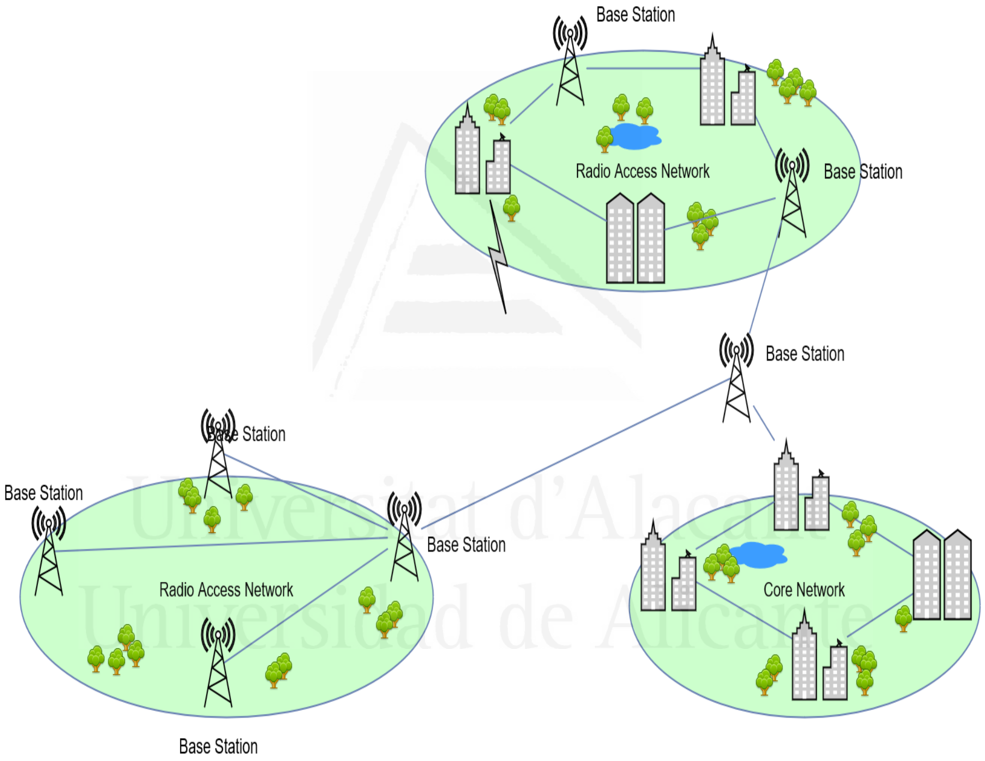


Figure 2.2: Example of a common mobile network topology

PM are data produced by the NEs that are collected and processed by a network Element Manager (EM) through open and well-standardized interfaces that support multi-vendor and multi-technology management (3rd Generation Partnership Project, 2005). PM files contain relevant data, for instance:

- User and signaling traffic that facilities the planning and operation of the network.
- Network settings to evaluate the effectiveness of the plan or planned changes.
- Measurements of access to resources for accurate evaluation.

- Service availability.
- Quality of service (QoS), for example: call establishment times, packet transfer times, among others. QoS allows determining the network performance experienced by the user.

PM files produced by the NEs are transferred to an EM and then to an Network Management System (NMS). The NMS is responsible for storage, post-processing, and presentation to determine the KPIs of the network for further evaluation. With the growth of the mobile networks, the number of NEs are also increasing and, therefore, the amount of these PM files.

PM files follow the XML format from the major part of vendors. Figure 2.3 presents the structure of the XML PM file according to (3rd Generation Partnership Project, 2005). Below, an example of a XML PM file is presented. Appendix B maps the tags presented in the Figure 2.3 and the XML file format definition.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="MeasDataCollection.xsl"?>
<measCollecFile
xmlns="http://www.3gpp.org/ftp/specs/latest/rel-5/32_series/32401-540
  #measCollec">
  <fileHeader
    fileFormatVersion="32.401 V5.0"
    vendorName="Company NN"
    dnPrefix="DC=a1.companyNN.com,SubNetwork=1,IRPAgent=1">
    <fileSender
        localDn="SubNetwork=CountryNN,
        MeContext=MEC-Gbg-1,
        ManagedElement=RNC-Gbg-1"
        elementType="RNC"/>
    <measCollec beginTime="2000-03-01T14:00:00+02:00"/>
  </fileHeader>
  <measData>
    <managedElement
                localDn="SubNetwork=CountryNN
                ,MeContext=MEC-Gbg-1,
                ManagedElement=RNC-Gbg-1"
                userLabel="RNC Telecomville"/>
    <measInfo>
      <granPeriod duration="PT900S"
      endTime="2000-03-01T14:14:30+02:00"/>
      <measTypes>attTCHSeizures succTCHSeizures
      attImmediateAssignProcs succImmediateAssignProcs
      </measTypes>
        <measValue measObjLdn="RncFunction=RF-1,
        UtranCell=Gbg-997">
            <measResults>234 345 567 789
            </measResults>
        </measValue>
        <measValue measObjLdn="RncFunction=RF-1,UtranCell=Gbg-999">
        <measResults>456 567 678 789
        </measResults>
        <suspect>true</suspect>
        </measValue>
    </measInfo>
  </measData>
  <fileFooter>
    <measCollec endTime="2000-03-01T14:15:00+02:00"/>
  </fileFooter>
</measCollecFile>
```

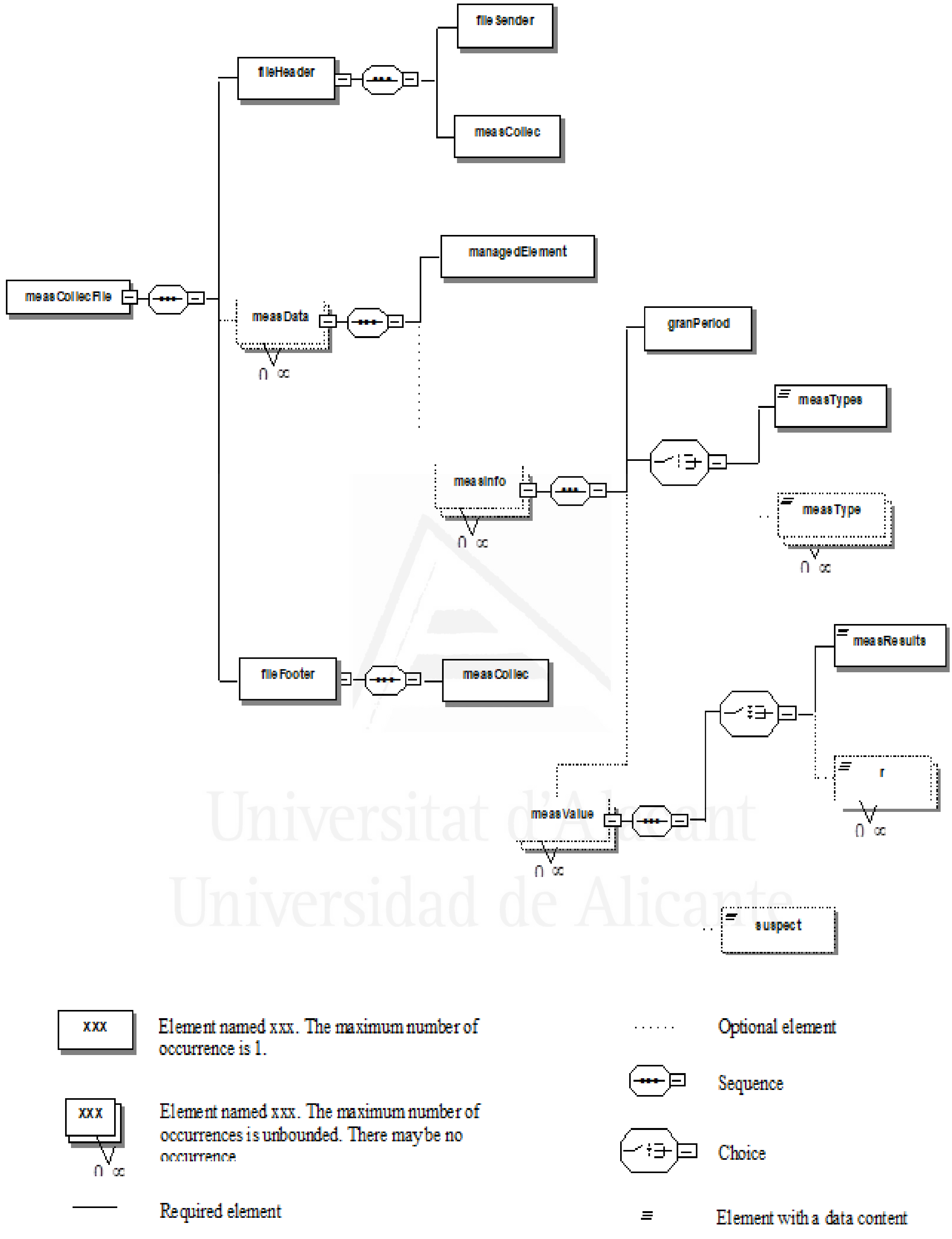


Figure 2.3: XML PM file format definition
(3rd Generation Partnership Project, 2005)

3 Literature Review

Based on the procedure proposed by Kitchenham (2004), the Figure 3.1 summarizes the main aspects considered in the performed systematic literature review (SLR).

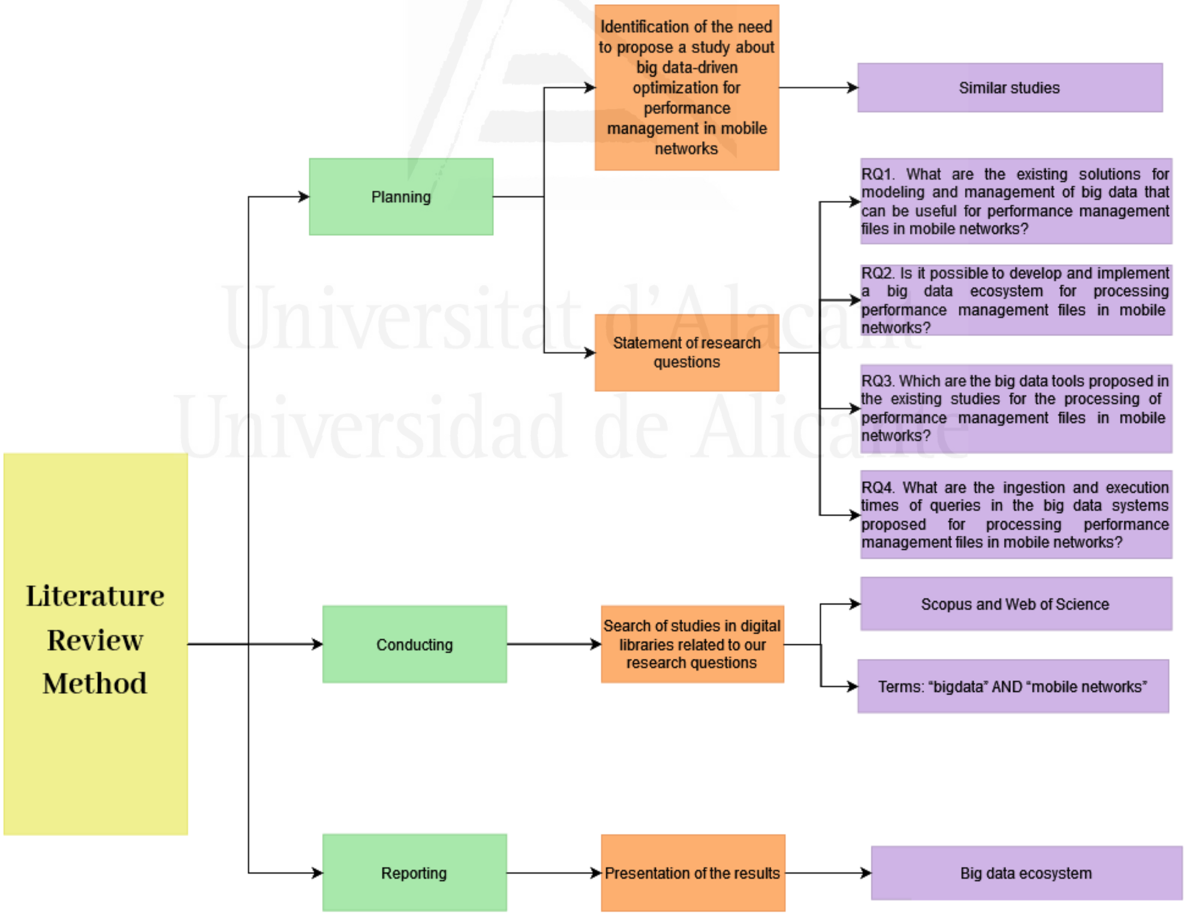

Literature Review Method
Planning
Conducting
Reporting
Identification of the need to propose a study about big data-driven optimization for performance management in mobile networks
Similar studies
Statement of research questions
RQ1. What are the existing solutions for modeling and management of big data that can be useful for performance management files in mobile networks?
RQ2. Is it possible to develop and implement a big data ecosystem for processing performance management files in mobile networks?
RQ3. Which are the big data tools proposed in the existing studies for the processing of performance management files in mobile networks?
RQ4. What are the ingestion and execution times of queries in the big data systems proposed for processing performance management files in mobile networks?
Search of studies in digital libraries related to our research questions
Scopus and Web of Science
Terms: "bigdata" AND "mobile networks"
Presentation of the results
Big data ecosystem

Figure 3.1: Procedure performed in the literature review

The main research questions posed were the following:

RQ1. What are the existing solutions for modeling and management of big data that can be useful for performance management files in mobile networks?

RQ2. Is it possible to develop and implement a big data architecture for processing performance management files in mobile networks?

RQ3. Which are the big data tools proposed in the existing studies for the processing of performance management files in mobile networks?

RQ4. What are the ingestion and execution times of queries in the big data systems proposed for processing performance management files in mobile networks?

The revision was performed over the main scientific libraries Scopus and Web of Science (WoS). As a first result, we found research related to mobile networks and big data is very extensive today. The terms “big data” AND “mobile networks” were searched in the scientific digital libraries mentioned before and studies were obtained since 2012 according to Scopus data and since 2014 in the WoS. Finally, 200 articles were found in WoS and 375 in Scopus.

In total, 575 articles were obtained; however, 159 studies were duplicated. Therefore, our corpus of primary studies consisted of 416 articles. After checked their titles, abstracts, and keywords, 28 studies were directly related to our research topic; therefore, they were considered as potentially relevant studies. After a complete review of each one, 7 studies were considered directly related to our research questions. The results are presented in Table 3.1.

For RQ1, the big data models used more frequently are entity-relationship in three studies, column-oriented in two studies, and key-value in two studies. For RQ2, in every case, it is possible to develop and implement a big data architecture for processing performance management files in mobile networks. For RQ3, Flume, Kafka, and NiFi are the tools most used in the ingestion layer; HDFS is the data lake used in almost all the cases; and Hive, PostgreSQL, HBase, Spark, Druid, and Flink are the databases tested. For RQ4, only two studies present the ingestion and reporting component times, and only one study presents the times for both components.

Furthermore, not all the studies present the used architecture in their proposals nor the number of processed records. We can highlight the use of high computational resources for RAM and CPUs.

ID	Title and Reference	RQ1	RQ2	RQ3	RQ4 Ingestion Time [seconds]	RQ4 Reporting Time [seconds]	Observation
A1	A Big Data Solution for Troubleshooting Mobile Network Performance Problems (Skračić & Bodrušić, 2017)	Entity Relationship	yes	Flume HDFS Hive Postgresql	NA	NA	2 namenodes (256 GB RAM) 8 datanodes (512 GB RAM)
A2	A quadri-dimensional approach for poor performance prioritization in mobile networks using Big Data (Mampaka & Sumbwanyambe, 2019)	Entity Relationship	yes	HDFS Impala	NA	2.3	11 millions raw records 4 VCPUs 16 GB RAM
A3	Applying Big Data, Machine Learning, and SDN/NFV to 5G Traffic Clustering, Forecasting, and Management Entity Relationship (Le, Sinh, Lin, & Tung, 2018)	Entity Relationship	yes	Kafka HDFS HBase Spark	NA	NA	
A4	Big Data Streaming Analytics for QoE Monitoring in Mobile Networks: A Practical Approach (Rueda, Vergara, & Reniz, 2018)	Column-oriented	yes	NiFi HDFS Druid	28.7	NA	20 billions raw records
A5	Characterizing Flow, Application, and User Behavior in Mobile Networks: A Framework for Mobile Big Data (Qiao, Xing, Fadlullah, Yang, & Kato, 2018)	key-value	yes	Flume Kafka HDFS Spark	NA	NA	1.792 TB RAM 224 VCPUs
A6	Distributed Big Data Driven Framework for Cellular Network Monitoring Data (Suleykin & Panfilov, 2019)	key-value	yes	Kafka Spark	0.014	0.464	61500 raw records 918 GB RAM 96 VCPUs
A7	Towards Adopting Big Data Technologies by Mobile Networks Operators: a Moroccan Case Study (Daki et al., 2016)	Column-oriented	yes	Kafka HDFS Flink	NA	NA	

Table 3.1: Results of the literature review

4 Publications and Visibility

4.1 Publications

This chapter presents the works that have been published during the development of this doctoral thesis. From nine articles published, five articles support the fulfillment of the objectives of this thesis, and four do not contribute directly to the subject of the thesis but address supplementary big data issues.

The results obtained during the development of this thesis have been published in five scientific journals. Two were published in journals with the Journal Citation Report (JCR) Q2 index and SJR Scimago Journal & Country Rank (SJR) Q1, one with the JCR Q3 index and SJR Q1, and two with SJR Q4. These articles depict the main content of this research. The details of the publications are presented in Table 4.1. The information in this table contains the identifier of the journal, the name of the journal and its ISSN, the Impact Factor (IF) of the JCR, the indexing and the IF of the SJR.

1. Martinez-Mosquera, D., & Luján-Mora, S. (2019). Framework for Big Data integration in e-government. In DYNA, vol.86(209), pp. 215–224. (Martinez-Mosquera & Luján-Mora, 2019). Chapter 7 presents the complete details of this article.

2. Martinez-Mosquera, D., Luján-Mora, S., Navarrete, R., Mayorga, T. C., & Vivanco, H. R. (2019). An approach to Big Data Modeling for Key-Value NoSQL Databases. In Iberian Journal of Information Systems and Technologies RISTI, vol. E19, pp. 519–530. (Martinez-Mosquera, Luján-Mora, Navarrete, et al., 2019). Chapter 8 presents the complete details of this article.

3. Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2020). Modeling and management big data in databases—A systematic literature review. In Sustainability, vol. 12, pp. 1–41. (Martinez-Mosquera et al., 2020b). Chapter 9 presents the complete details of this article.

4. Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2020). Development and evaluation of a big data framework for performance management in mobile networks. In IEEE Access, vol. 8, pp. 226380–226396. (Martinez-Mosquera et al., 2020a). Chapter 10 presents the complete details of this article.

5. Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2020). Efficiently Processing Complex XSD using Hive and Spark. In PeerJ Computer Science, vol. 8, pp. 1–33. (Martinez-Mosquera et al., 2021). Chapter 11 presents the complete details of this article.

ID	Journal	JCR IF	Indexing	SJR IF
J1	DYNA ISSN: 00127353	NA	COLCIENCIAS	Q4 0.16
J2	RISTI ISSN: 16469895	NA	SCOPUS, WOS, JCR, EBSCO, DOAJ	Q4 0.14
J3	Sustainability ISSN:20711050	Q2 3.251	SCOPUS, WOS, JCR, EBSCO, DOAJ	Q1 0.61
J4	IEEE Access ISSN: 21693536	Q2 3.367	SCOPUS, WOS, JCR, Google Scholar, DOAJ	Q1 0.59
J5	PeerJ Computer Science ISSN: 23765992	Q3 1.392	PMC, SCOPUS, WOS, JCR, Google Scholar, DOAJ, DBLP	Q1 0.81

Table 4.1: Journal articles included in the compendium of publications

4.2 Other Publications

This section presents other four articles published during the doctoral course that allowed the investigation of big data despite not being directly related to the specific objectives of this thesis. The following list presents the details of each article:

1. Martinez-Mosquera, D., Luján-Mora, S., & Recalde, H. (2017). Conceptual modeling of big data extract processes with UML. In International Conference on Information Systems and Computer Science (INCISCOS), pp. 207-211. (Martinez-Mosquera, Luján-Mora, & Recalde, 2017).

2. Martinez-Mosquera, D., & Luján-Mora, S.(2017). Data Cleaning Technique for Security Big Data Ecosystem. In International Conference on Internet of Things, Big Data and Security (IoTBDS), pp. 380-385. (Martinez-Mosquera & Luján-Mora, 2017).

3. Martinez-Mosquera, D., Luján-Mora, S., López, G., & Santos, L.(2017). Data cleaning technique for security logs based on Fellegi-Sunter Theory. EuroSymposium on Systems Analysis and Design, pp. 3-12. (Martinez-Mosquera, Luján-Mora, Lopez, & Santos, 2017).

4. Martinez-Mosquera, D., Luján-Mora, S., Reyes R., & Paredes M. (2019). Pillars for Big Data and Military Health Care: State of the Art. In International Conference on Advances in Emerging Trends and Technologies (ICAETT), pp. 125-135. (Martinez-Mosquera, Luján-Mora, Reyes, & Paredes, 2019)

4.3 Visibility

The main objective of a scientific publication is the generation of knowledge and its transfer to the world of research. In order to make the work visible and achieve a greater impact on our research, several academic profiles have been created that can be seen in Table 4.2. Additionally, four of the five articles are open access, which allows researchers to access the results in a less restrictive way.

ID	**Academic Profile**	**URL**
P1	ORCID	https://orcid.org/0000-0002-0573-8640
P2	Google Scholar	https://scholar.google.com/citations?user=KCpjL3oAAAAJ&hl
P3	Research Gate	https://www.researchgate.net/profile/Diana-Martinez-Mosquera
P4	Scopus	https://www.scopus.com/authid/detail.uri?authorId=57194945843
P5	SciProfiles	https://sciprofiles.com/profile/899748

Table 4.2: Academic profiles

5 Description of the Work

To fulfill the main objective of our doctoral study, the research was divided into two parts:

1. The first part concerns big data technology, which allows knowing about the available tools, applications, and modeling methods.

2. The second part focused on the study of the PM of mobile networks and the orchestration of big data services for the implementation of the architecture.

5.1 Big Data Modeling

From the beginning of our research, we found that big data analysis has been used in different areas such as healthcare, marketing, transportation, education, environment, among other applications. For example, the article Martinez-Mosquera and Luján-Mora (2019) presents a proposal of a big data framework for e-government. Big data can produce valuable information after the correct implementation of a framework that allows data to be loaded, cleaned, processed to finally generate storage that allows its mining, reporting, or manipulation.

One of the important aspects to manipulate big data is how modeling the data. Among the existing data modeling types of big data we find:

- Relational.

- Key value.

- Document oriented.

- Column oriented.

- Graphs.

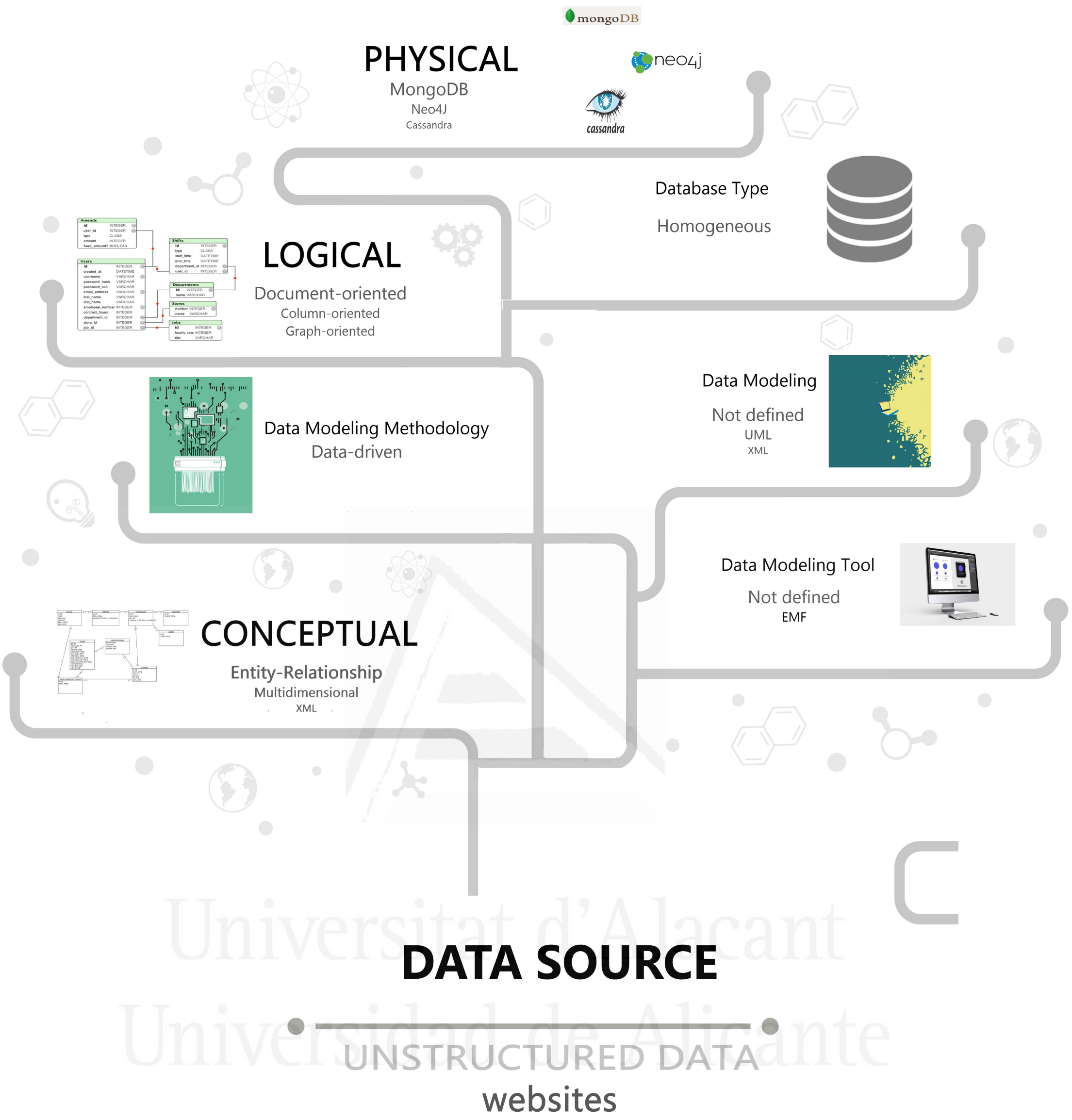


Figure 5.1: Trends in big data modeling according to the results of the SLR (Martinez-Mosquera et al., 2020b)

In the study Martinez-Mosquera, Luján-Mora, Navarrete, et al. (2019), we propose an approach for modeling big data in key-value databases. Furthermore, the results of the literature review in Martinez-Mosquera et al. (2020b) performed for big data modeling (2010-2019) allowed us to know the trends and gaps in the big data modeling topic. These results are summarized in Figure 5.1 and Figure 5.2.

The most important aspects presented in the figure Figure 5.1 are the following: the most researched data sources are unstructured data, the Entity-Relationship model is the most used in the approaches at the conceptual abstraction level, the most researched big data modeling is document-oriented, the implementations are focused in the MongoDB Data Base Management Systems (DBMSs), followed by Neo4j and Cassandra, the most proposed modeling methodology is data-driven. Furthermore, most scientific studies analyzed websites, sensors, and electronic documents. For NoSQL databases, the use of Polyglot Persistence System (PPS) is recommended; namely, the capability of querying data from different NoSQL databases (Karamjit & Rinkle, 2013).

Figure 5.2 presents the gaps where future studies can concentrate their efforts, between them: solutions for a variety of data, the use of case studies with real data sets, proposals for hybrid databases, query-driven data modeling, and standardization.

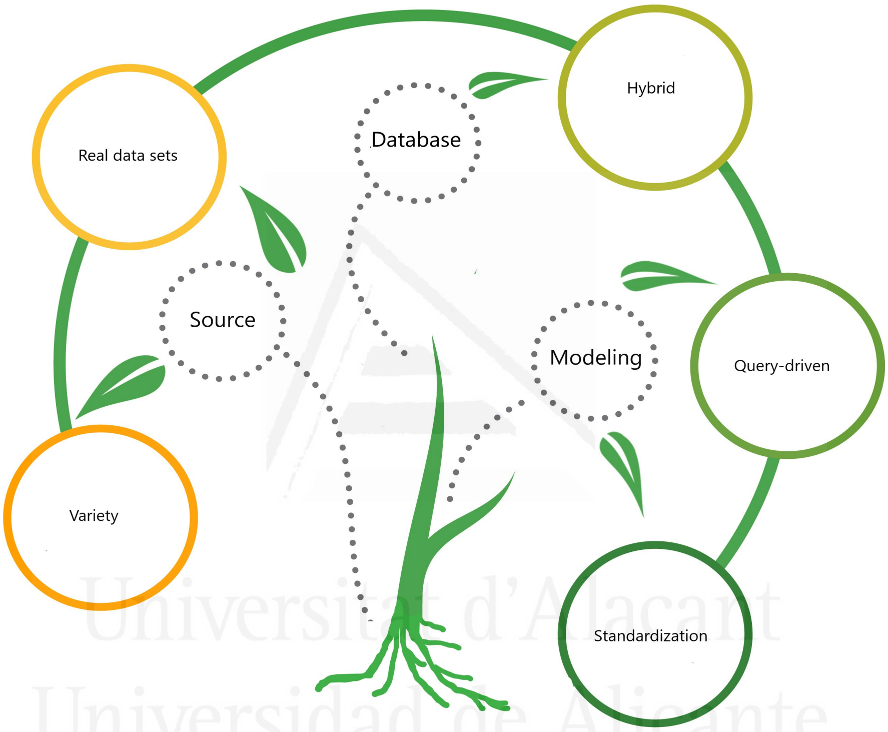


Figure 5.2: Gaps in big data modeling according to the results of the SLR (Martinez-Mosquera et al., 2020b)

5.2 Performance Measurements in Mobile Networks

PM transfer from NEs to EMS is usually performed through SFTP in pull mode; namely, the EMS connects to the NEs and collects new raw data. Then, the data are transferred to NMS. NMS will be responsible for storing, processing, reporting, and visualizing the data.

According to Khalifa et al. (2016), among the main pillars for building a big data framework are: deployment, storage or data lake, processing, and interface. In our research, we were able to recognize other important components such as software architecture, ingestion, reporting. Within each component, there are many tools that can be selected according to the requirements of the project. These tools can be found in the data landscape available from Turck (2020).

Among the big data tools analyzed in this study, we have the following:

1. **Software Architecture:** Lambda and Kappa.
2. **Ingestion:** Flume, Kafka, NiFi, Sqoop.
3. **Storage or Data Lake:** Azure, AWS, IBM Power Systems, HDFS.
4. **Reporting:** HBase, Hive, Impala, Pig.
5. **Processing:** Spark, Storm.
6. **Interface:** Power BI, Tableau.
7. **Deployment:** Cloudera Distribution Hadoop (CDH), Hortonworks, IBM InfoSphere Big Insights, MapR, Pivotal HD.

5.2.1 Software Architecture

The procedure to gather PM data from mobile networks and process them in big data architecture is presented in Figure 5.3. Lambda architecture is used and the BL, SRL, and SL layers are identified. Next, the tools used in each of the components of the architecture and the results obtained are presented.

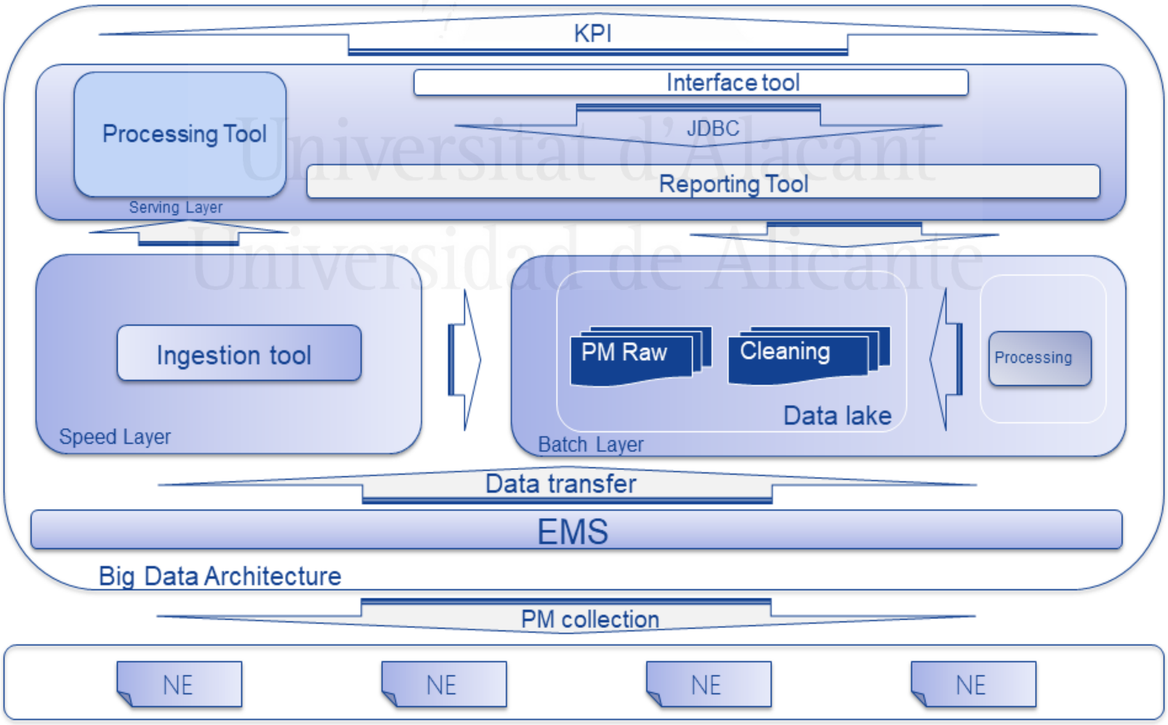


Figure 5.3: Collection and processing of PM data from NEs.

5.2.2 Ingestion

The ingestion component is responsible for collecting data from the source and transferring them to the target. In this case, the ingestion tool must collect PM raw data in XML format from EMS and transferring to a data lake.

Each PM file is depicted as a vector $P = \{C \cup M_i\}$; $i = 1, ..., N$, where:

C corresponds to the configuration data related to the measurement data; for instance, according to the example of (3rd Generation Partnership Project, 2005): fileFormatVersion, vendorName, dnPrefix, localDn, elementTYpe, among others.

M_i corresponds to the measurement data; for instance, according to the example of 3rd Generation Partnership Project (2005): attTCHSeizures, succTCHSeizures, attImmediateAssignProcs, and succImmediateAssignProcs metrics and their apparent numeric values.

N relates to the number objects of NEs; for instance, by the example of 3rd Generation Partnership Project (2005): N = 3 because of the objects Gbg-997, Gbg-998, and Gbg-999 from the NE RNC-Gbg-1.

In this thesis, we consider the following big data tools that support collecting and transferring XML files: Flume, Kafka, NiFi, and Sqoop. Kafka and Nifi need for an enterprise version of a deployment framework and other services installation. Sqoop works only for relational data bases transfer. Therefore, Flume was selected with the SFTP support. Flume works with three tiers: Tier1 sends data to Tier2, Tier2 moves the data to the storage, and Tier3 is the final storage (Apache, 2020a).

After implementing a test environment where the NE was simulated by a virtual machine (VM) and Flume was installed in another VM with CDH, the results obtained were the following: the fitted regression line complies with the equation (5.1), where y_1 corresponds to the time of ingestion variable, and x_1 the number of measurement data.

$$y_1 = 60.38x_1 + 30.83 \tag{5.1}$$

The determination coefficient R^2 and the Pearson correlation coefficient R are equal to 0.9999. Therefore, the variables y_1 and x_1 are linearly dependent. As a result, it is possible to reach around one billion metrics M_i per hour with 512 GB RAM in this test environment(Martinez-Mosquera et al., 2020a).

5.2.3 Data Lake

This component must comply with linear scalability and performance, as architecture requirements mentioned in Appendix A. Scalability means, it must be possible to increase the storage capacity by adding new nodes to the cluster. For performance, the storing process must be fast as possible. These features are fulfilled by several solutions; however, HDFS is an open tool, which also supports batch processing and is supported by several deployment solutions and it is compatible with many big data tools (Borthakur, 2008).

5.2.4 Processing

Spark (Apache, 2020c) and Storm (Apache, 2020d) are some of the stream processing tools that are supported in the SL of big data architecture. Spark allows a faster data analysis of the XML data; therefore, it is selected in this component. As a result of this research, Apache Spark process around 3 million rows in 14.25 seconds (Martinez-Mosquera et al., 2021). The complete process for process the PM data in Apache Spark is summarized in Algorithm 1.

XML file denoted by A_i. XSD that belongs to A_i is named vector X_i. E is composed of the elements in the catalog proposed in Table 5.1. $CreateDataFrame$ is applied to vector X_i using the deserialization method. The output of $CreateDataFrame$ is the input of the $CreateColumnSeparatedDataFrame$. E@ is the number of elements. $E\#$ is the number of columns and elements in vectors $E[]$ and $E\{\}$ for the positional explode method.

Element Types	Notation
Root	<>
Array	[]
Structure	{}
Attribute	@
Value	#

Table 5.1: Catalog.

Algorithm 1: Processing XML format for PM with Apache Spark

```
Input: XML documents $A_i$
Output: Apache Spark Data Frame
Create $X_i \leftarrow A_i$;
Create $E \leftarrow$ catalog;
while $X$ do
    CreateDataFrame;
    for $X_i$ do
        rowTag=<>" $\leftarrow E$ <>;
        location=/hdfs;
        Deserialization ;
    end
    return DataFrame;
    Create ColumnSeparatedDataFrame;
    for $DataFrame$ do
        Select Expression $\leftarrow E$@, $E\#$;
        Positional Explode $\leftarrow E[]$, $E\{\}$;
        return ColumnSeparatedDataFrame;
    end
end
```

Algorithm 2: Processing XML format for PM with Apache Hive

```
Input: XML documents $A_i$
Output: Apache Hive Table
Create $X_i \leftarrow A_i$;
Create $E \leftarrow$ catalog;
while $X$ do
    if internal table then
        Create RawTable;
        for $X_i$ do
            xmlinput.start=<>" ← $E$ <>;
            xmlinput.end=< / >" ← $E$ <> ;
            location=/hdfs;
            Deserialization ;
            load data into table;
        end
        return RawTable;
        Create ColumnSeparatedInternalTable;
        for $RawTable$ do
            XPATH strings ← $E$@, $E$#;
            Positional Explode ← $E$[], $E${};
            return ColumnSeparatedInternalTable;
        end
    else
        Create ExternalRawTable;
        for $X_i$ do
            xmlinput.start=<>" ← $E$ <>;
            xmlinput.end=< / >" ← $E$ <>;
            location=/hdfs;
            Deserialization ;
        end
        return ExternalRawTable;
        Create ColumnSeparatedExternalTable;
        for $ExternalRawTable$ do
            XPATH strings ← $E$@, $E$#;
            Positional Explode ← $E$[], $E${};
            return ColumnSeparatedExternalTable;
        end
    end
end
```

5.2.5 Reporting

For this component, several aspects must be considered. First, the type of data, since XML format is used, the reporting tool must process semi-structured data. The data model is presented in Figure 5.4.

Second, the data lake selected was HDFS; therefore, the reporting tool must be compatible with it. Third, as the initial requirements mention, the selected reporting tool must be compatible with SQL.

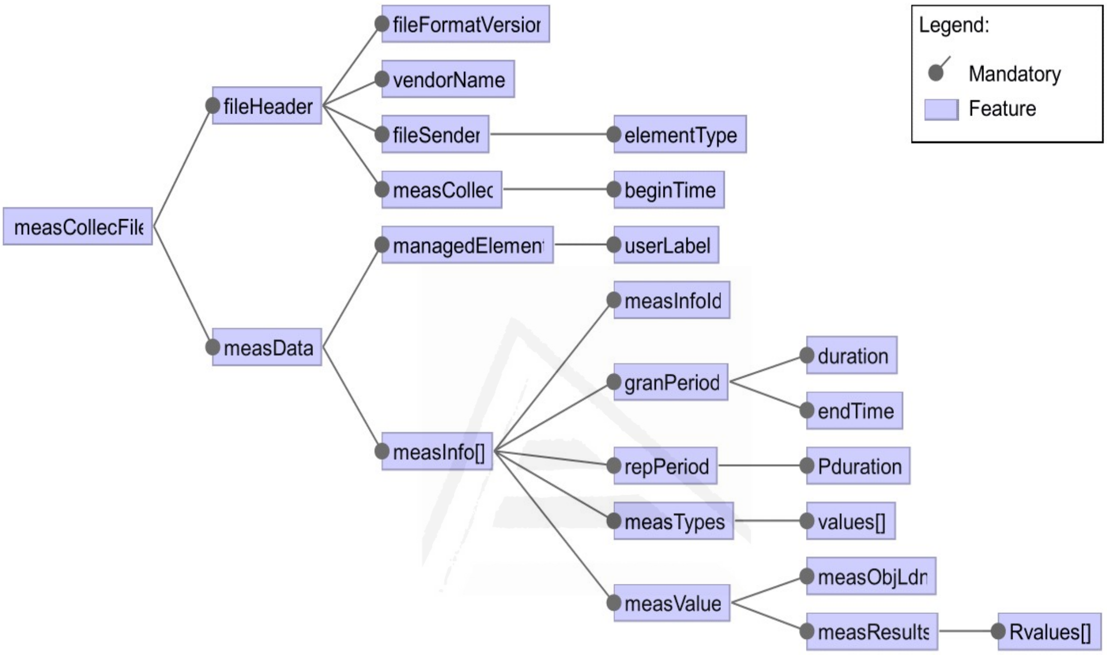


Figure 5.4: Data model for XML PM
(Martinez-Mosquera et al., 2020a)

As candidates, we have HBase, Hive, Impala, and Pig, that are compatible with HDFS. Hive and Pig perform batch processing. Hive (Apache, 2020b) offers a query language based on the SQL, called HiveQL. However, Hive provides native support to XML files and the benefits mentioned in a study (X. Li & Zhou, 2015) in comparison with other tools. Furthermore, Hive can use the XML parsing serializer deserializer of IBM to create the tables. The complete process for process the PM data in Apache Hive is summarized in Algorithm 2.

The results obtained in our test environment for this component was to reach around one billion metrics M_i per hour with 72 GB (Martinez-Mosquera et al., 2020a).

5.2.6 Interface

For the interface component, there are several tools for allowing the visualization of big data, among them, the most known: Tableau and Power BI (Carlisle, 2018). Since our goal is to offer a turnkey solution, Power BI (Microsoft, 2021) has been selected because it has a free version, while in Tableau (Tableau, 2021) the reports made in the free version are not private. Also, Power BI is for now the tool used compared to Tableau (Machiraju & Gaurav, 2018).

Power BI (Microsoft, 2021) is a software tool that allows the creation of customizable report dashboards, which allows you to take data from different sources such as text, csv files, local databases, as well as network databases, be they structured and documentary type. And that in turn allows easy manipulation of the data by using its included tool called Power Query.

Moreover, it has the main types of graphic objects that can be used in a general way within a board, which have a large number of customization options for colors, traffic lights, filters, etc .; and if you need other visual objects with additional properties, you can turn to the Power BI AppSource where you can find a large number of specialized visuals that can be accessed mostly for free.

Another advantage of Power Bi is that there is a Cloudera Open Database Connectivity (ODBC) driver that allows the connection between PowerBI and Hive in CDH. Once the connection has been established, the tables to be used for creating the dashboard are loaded. Power Bi facilitates to perform the aggregations to the data, through the DAX function, allowing the calculation of KPIs. Appendix C presents an example of a report from PM files obtained from PowerBI.

5.2.7 Deployment

In order to provide a turnkey solution, the CDH 5.16.0 machine was used since supports the selected tools in each component. CDH is an open-source environment with the support of Flume, HDFS, Hive, and Spark, among others. Later, by requiring the payment of a license for CDH versions 6 or higher, a solution based on Hadoop was installed with the updated versions of the tools.

5.3 Comparison with Other Solutions

As we can see in the literature review, we were able to find seven different studies about the implementation of a big data framework for PM in mobile networks. We compared the results of these studies with ours to determine the computational complexity which concerns the time and resources required to solve a given problem.

For performing the comparison possible, we have taken the data from section 5.2.1 Ingestion, where the determination coefficient R^2 and the Pearson correlation coefficient R are very close to the value 1; therefore, there is a strong linear dependence between the time and the number of record files.

In Table 5.2 the results of the comparison for the ingestion layer are presented. There, it has compared the results obtained in this thesis and study A5 (Qiao et al., 2018). Other studies could not be compared since no enough data about the resources and times are available. According to the results, the study A5 allows processing one billion counters in 228 seconds; however, the RAM capacity needed is oversize resulting in the need for many nodes from file ingestion.

ID	Big data tool	Time [seconds]	RAM [GB]	Number of nodes with 32 [GB]
Thesis	Flume	3600	512	10
A5	Kafka	228	14926830	466463

Table 5.2: Comparison of studies for processing one billion counters at ingestion layer

ID	Big data tool	Time [seconds]	RAM [GB]	Number of nodes with 32 [GB]
Thesis	Hive	3600	72	2
A2	Impala	209	1455	45
A5	Impala	7545	14926829	466463

Table 5.3: Comparison of studies for processing one billion counters at reporting layer

Table 5.3 presents the results of the comparison for reporting layer. There, it is compared the results obtained in this thesis, with the study A2 (Mampaka & Sumbwanyambe, 2019), and the study A5 (Qiao et al., 2018). According to the results, the study A2 (Mampaka & Sumbwanyambe, 2019) allows processing one billion counters in less time 209 seconds; however, it needs 15 times the RAM capacity as ours. Therefore, with the equal RAM capacity as study A2 (Mampaka & Sumbwanyambe, 2019), our proposal achieved better results.

Part II

COMPENDIUM

6 Compendium of Publications

The development of this thesis resulted in five articles that have been published in indexed journals with important contributions to the scientific community. The different publications made it possible to achieve the objectives proposed in the research. The five publications have an impact factor. Two of them were published in journals classified in the first quartile (Q1) of the SJR and the second quartile (Q2) of the JCR (Martinez-Mosquera et al., 2020a, 2020b). One article was published in a journal classified in the first quartile (Q1) of the SJR and the third quartile (Q3) of the JCR (Martinez-Mosquera et al., 2021). Finally, two articles were published in journals classified in the fourth quartile (Q4) of the SJR (Martinez-Mosquera & Luján-Mora, 2019; Martinez-Mosquera, Luján-Mora, Navarrete, et al., 2019). The publications are detailed in Figure 6.1.

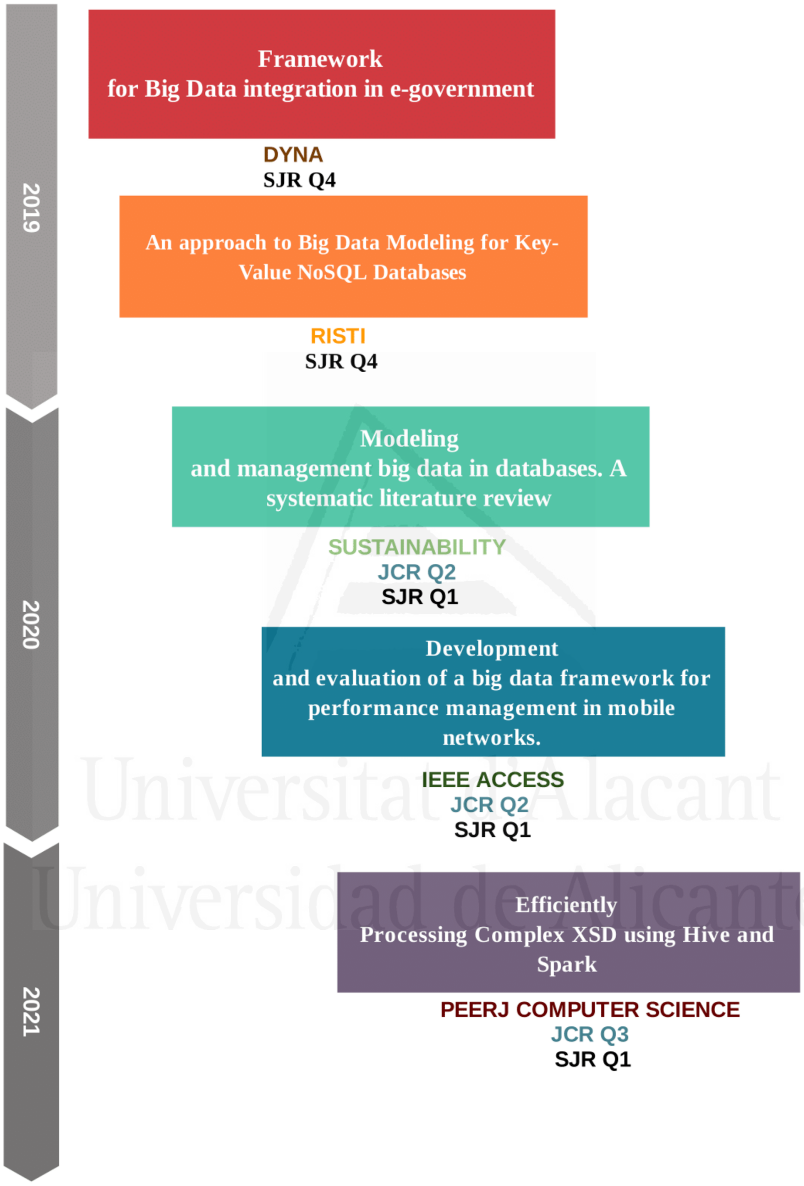

2019
Framework for Big Data integration in e-government
DYNA
SJR Q4
An approach to Big Data Modeling for Key-Value NoSQL Databases
RISTI
SJR Q4
2020
Modeling and management big data in databases. A systematic literature review
SUSTAINABILITY
JCR Q2
SJR Q1
Development and evaluation of a big data framework for performance management in mobile networks.
IEEE ACCESS
JCR Q2
SJR Q1
2021
Efficiently Processing Complex XSD using Hive and Spark
PEERJ COMPUTER SCIENCE
JCR Q3
SJR Q1

Figure 6.1: Timeline of compendium of publications

7 Framework for Big Data Integration in E-government

Reference:

Martinez-Mosquera, D., & Luján-Mora, S. (2019). Framework for Big Data integration in e-government. In DYNA, 86(209), pp. 215–224.

Available on:

- https://revistas.unal.edu.co/index.php/dyna/article/view/77902
- https://doi.org/10.15446/dyna.v86n209.77902

Contributes to the topic:

SO. 7. Propose a big data ecosystem for PM in mobile networks and its implementation.

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

Framework for Big Data integration in e-government

Diana Martinez-Mosquera & Sergio Luján-Mora

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, España, sdmm1@alu.ua.es, sergio.lujan@ua.es

Received: February 15th, 2019. Received in revised form: March 27th, 2019. Accepted: April 15th, 2019.

Abstract
This article describes research regarding Big Data integration in e-government decision-making, for instance, in areas such as solar energy provisioning, environmental protection, agricultural and natural resources exploitation, health and social care, education, housing and transportation management, among others. These studies refer to regions that have integrated Big Data in e-government, where South America is still in the early adoption stages. Hence, this study proposes three stepping-stones for Big Data integration in e-government decision-making: production, management and application. The proposed framework aims to be a reference in South America for Big Data adoption in e-government and, thus, help to mitigate the technology delay regarding other regions. Finally, the article presents a case study with open data obtained from the Instituto Nacional de Estadística y Censos of Ecuador (Ecuadorian Statistics and Census Agency).

Keywords: big data; e-government; integration; framework; reference.

Marco de referencia para la integración de Big Data en gobierno electrónico

Resumen
En el presente artículo se describen algunos estudios que incorporan Big Data en la toma de decisiones de gobierno electrónico como, por ejemplo, provisión de energía solar, protección del ambiente, producción agrícola, explotación de petróleo, gestión de salud, educación, vivienda y transporte, entre otros. Estos estudios corresponden a regiones que han integrado Big Data en gobierno electrónico. Sudamérica se encuentra aún en proceso de adopción. Por esta razón, el presente estudio propone un marco de integración de Big Data en la toma de decisiones de gobierno electrónico, que consta de tres etapas: la producción, la gestión y la aplicación de Big Data. El marco propuesto pretende servir de referencia y así ayudar a disminuir el retraso de Sudamérica con respecto a otras regiones. Finalmente, se presenta un caso de estudio en Ecuador, con datos abiertos del banco de datos del Instituto Nacional de Estadística y Censos.

Palabras clave: big data; e-gobierno; *e-government*; integración; marco; referencia.

1. Introduction

Currently, Big Data is one of the most popular terms in information systems and is used to describe large volumes of data that are exploited to gain competitive advantages in the case of private companies or achieve efficiencies in the case of public organizations. The main challenge of such organizations is the management of these vast data, their collection, processing and analysis. However, the information obtained from the Big Data analysis depicts that the function of insights is key to competitive advantage, allowing them to make timely decisions, optimize resources and even prevent disasters [1].

For a set of information to be considered as Big Data, it must comply with certain properties. Among others, these are the most frequently mentioned: volume, velocity, variety, veracity and value [2]. The main goal of Big Data is to obtain value from the data analyzed.

The continuous generation of data in smart devices, sensors, social networks, web browsing and the deployment of the Internet of Things (IoT), among other technologies, has prompted researchers' interest in the development of new solutions for the management of this massive quantity of data. Through these solutions, it has been possible to add value to the gathered data, and one of its applications has enabled social problems to be tackled, such as overcrowding,

How to cite: Martinez-Mosquera, D. and Luján-Mora, S., Framework for Big Data integration in e-government. DYNA, 86(209), pp. 215-224, April - June, 2019.

Revista DYNA, 86(209), pp. 215-224, April - June, 2019, ISSN 0012-7353
DOI: http://doi.org/10.15446/dyna.v86n209.77902

8 An Approach to Big Data Modeling for Key Value NoSQL

Reference:

Martinez-Mosquera, D., Luján-Mora, S., Navarrete, R., Mayorga, T. C., & Vivanco, H. R. (2019). An approach to Big Data Modeling for Key-Value NoSQL Databases. In Iberian Journal of Information Systems and Technologies RISTI, E19, pp. 519–530.

Available on:

- http://www.risti.xyz/issues/ristie19.pdf

Contributes to the topic:

SO. 1. Perform a systematic literature review of the modeling and management big data in databases..

Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 24/11/2018
Aceitação/Acceptance: 22/01/2019

An approach to Big Data Modeling for Key-Value NoSQL Databases

Diana Martinez-Mosquera[1], Sergio Lujan-Mora[2], Rosa Navarrete[3], Tannia Cecilia Mayorga[4], Henry Rodrigo Vivanco Herrera[5]

sdmm1@alu.ua.es, sergio.lujan@ua.es, rosa.navarrete@epn.edu.ec, tmayorga@uisrael.edu.ec, hvivanco@uisrael.edu.ec

[1,2] University of Alicante, 03690, San Vicente del Raspeig, Spain.

[3] Escuela Politécnica Nacional, 170525 , Quito, Ecuador.

[4,5] Universidad Tecnológica Israel, 170522, Quito, Ecuador.

Pages:519–530

Abstract: The scientific community has a special interest in providing solutions to deal with a huge amount of data generated by the Internet, mobile devices, and sensors, among others. One of the foremost research approaches has been in not only SQL (NoSQL) databases, mainly used to handle Big Data. A state of the art review presented in this article let us argue on the need to define modeling techniques to depict how data will be structured in NoSQL databases. The majority of studies have focused on structured data and column oriented databases; thus, we propose an approach for semi-structured data at conceptual and logical modeling levels, using the Unified Modeling Language and key-value databases. The logical model is attained from a class diagram, with the use of transformation rules based on some aspects of the Query View Transformation. Furthermore, our proposal presents a case study concerning data in security log files.

***Keywords*:** big data; key-value; modeling; UML; semi-structured.

1. Introduction

The permanent growing of data generated and collected by companies is in the range of terabytes of information; this high amount of data is known as Big Data (Martinez-Mosquera, Lujan-Mora & Parra, 2017). Big Data is a concept widely used to describe a big amount of data that comply with some specific features like volume, velocity, variety, variability, value, among others (Qaiyum, Aziz & Jaafar, 2016). Big Data researchers have commendably focused in different approaches, one of them about not only SQL (NoSQL) databases, since relational database paradigms are not adequate and NoSQL, by contrast, handles Big Data with high performance (Abdelhedi et al., 2016).

Although this type of database was designed to operate without schema, there is an underlying need for defining the data structure in the database (Abdelhedi et al., 2016; Santos & Costa, 2016). The reason is modeling helps to design the data processing flow (Da Silva et al., 2014). Many researches have proposed Big Data modeling, but the major part

9 Modeling and Management Big Data in Databases — A Systematic Literature Review

Reference:

Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2020). Modeling and management big data in databases - A systematic literature review. In Sustainability, Vol. 12 (2), pp. 1–41.

Available on:

- https://www.mdpi.com/2071-1050/12/2/634
- https://doi.org/10.3390/su12020634

Contributes to the topic:

SO. 1. Perform a systematic literature review of the modeling and management big data in databases..

sustainability

MDPI

Review

Modeling and Management Big Data in Databases—A Systematic Literature Review

Diana Martinez-Mosquera [1,*], Rosa Navarrete [1] and Sergio Lujan-Mora [2]

[1] Department of Informatics and Computer Science, Escuela Politécnica Nacional, 170525 Quito, Ecuador; rosa.navarrete@epn.edu.ec

[2] Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain; sergio.lujan@ua.es

* Correspondence: diana.martinez@epn.edu.ec; Tel.: +593-02-2976300

Received: 25 November 2019; Accepted: 9 January 2020; Published: 15 January 2020

check for updates

Abstract: The work presented in this paper is motivated by the acknowledgement that a complete and updated systematic literature review (SLR) that consolidates all the research efforts for Big Data modeling and management is missing. This study answers three research questions. The first question is how the number of published papers about Big Data modeling and management has evolved over time. The second question is whether the research is focused on semi-structured and/or unstructured data and what techniques are applied. Finally, the third question determines what trends and gaps exist according to three key concepts: the data source, the modeling and the database. As result, 36 studies, collected from the most important scientific digital libraries and covering the period between 2010 and 2019, were deemed relevant. Moreover, we present a complete bibliometric analysis in order to provide detailed information about the authors and the publication data in a single document. This SLR reveal very interesting facts. For instance, Entity Relationship and document-oriented are the most researched models at the conceptual and logical abstraction level respectively and MongoDB is the most frequent implementation at the physical. Furthermore, 2.78% studies have proposed approaches oriented to hybrid databases with a real case for structured, semi-structured and unstructured data.

Keywords: big data; management; modeling; literature review

1. Introduction

The Big Data modeling term became widespread in 2011, as is visible in Figure 1. This figure shows searches in Google Trends related to Big Data modeling, which are intensified from 2011 onwards. Searches before 2004 are not presented, since Google Trends does not store earlier data. In recent years, researchers have consolidated their efforts to study new paradigms to deal with Big Data. Thus, novel Big Data modeling and management in databases approaches have emerged, in line with the new requirements. In consequence, new techniques in the database context have evolved towards Not Only SQL (NoSQL).

The work presented in this paper is motivated by the acknowledgement that a complete systematic literature review (SLR) that consolidates all the research efforts for Big Data modeling and management in databases is missing. An SLR is the best way to collect, summarize and evaluate all scientific evidence about a topic [1]. It allows for the description of research areas shown the greatest and least interest by researchers. Considering the exposed issues, the SLR conducted in this work can contribute to solving this lack by collecting and analyzing details about the research published from 2010 to 2019. As a basis for our SLR, we adhered to the guidelines proposed by Kitchenham [1]. Moreover, this paper presents a complete bibliometric analysis and summarizes existing evidence about research

10 Development and Evaluation of a Big Data Framework for Performance Management in Mobile Networks

Reference:

Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2020). Development and evaluation of a big data framework for performance management in mobile networks. In IEEE Access, Vol. 8, pp. 226380–226396.

Available on:

- https://ieeexplore.ieee.org/abstract/document/9296310
- https://doi.org/10.1109/ACCESS.2020.3045175

Contributes to the topic:

SO. 2. Perform a study of the state of the art of big data and mobile networks.

SO. 3. Define a method to design a method for the selection of the big data tools.

SO. 4. Define a method to design a big data ecosystem for PM in mobile networks.

SO. 7. Propose a big data ecosystem for PM in mobile networks and its implementation.

SO. 8. Evaluate the big data ecosystem for ingestion and query execution times.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 25, 2020, accepted December 11, 2020, date of publication December 16, 2020,
date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3045175

Development and Evaluation of a Big Data Framework for Performance Management in Mobile Networks

DIANA MARTINEZ-MOSQUERA[1], ROSA NAVARRETE[1], AND SERGIO LUJÁN-MORA[2]
[1]Department of Informatics and Computer Science, Escuela Politécnica Nacional, Quito 170525, Ecuador
[2]Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain

Corresponding author: Diana Martinez-Mosquera (diana.martinez@epn.edu.ec)

This work was supported by the Unidad de Gestión de Investigación y Proyección Social from the Escuela Politécnica Nacional.

ABSTRACT In telecommunications, Performance Management (PM) data are collected from network elements to a centralized system, the Network Management System (NMS), which acts as a business intelligence tool specialized in monitoring and reporting network performance. Performance Management files contain the metrics and named counters used to quantify the performance of the network. Current NMS implementations have limitations in scalability and support for volume, variety, and velocity of the collected PM data, especially for 5G and 6G mobile network technologies. To overcome these limitations, we proposed a Big Data framework based on an analysis of the following components: software architecture, ingestion, data lake, processing, reporting, and deployment. Our work analyzed the PM files' format on a real data set from four different vendors and 2G, 3G, 4G, and 5G technologies. Then, we experimentally assessed our proposed framework's feasibility through a case study involving 5G PM files. Test results of the ingestion and reporting components are presented, identifying the hardware and software required to support up to one billion counters per hour. This proposal can help telecommunications operators to have a reference Big Data framework to face the current and future challenges in the NMS, for instance, the support of data analytics in addition to the well-known services.

INDEX TERMS Big data, framework, mobile networks, network management system, performance management.

I. INTRODUCTION

The mobile network industry is one of the most extensive and heterogeneous worldwide [1]. In the last two decades, it has evolved from 2G, 3G, and 4G technologies to the rollout of 5G and 6G technologies in the near future [2]. With the advent of 5G and 6G, mobile networks will be challenged to respond to the dramatic rise of connected devices.

Since the first generations were developed, mobile networks count using a centralized system that acts as a business intelligence tool specialized in monitoring and reporting the network's performance. This system is known as the Network Management System (NMS). Essentially, an NMS allows operators to manage the network and visualize its behavior by collecting Performance Management (PM) data from Element Management Systems (EMS) [3]. The purpose of PM is to continuously monitor and measure the performance of various subsystems [4].

As shown in [5], during 2020, there were around eight billion mobile connections in the world, and for 2025 it is estimated that there will be 8.8 billion mobile connections. The IDC Corporation forecasts that the global data will grow to 175 zettabytes (ZB) by 2025 [6] (1 ZB is equivalent to a trillion gigabytes). Therefore, due to the projected increase in the number of users, more extensive networks will need to analyze larger amounts of data in shorter periods of time in order to prevent network outages as soon as possible. With PM data, it is possible to plan and optimize the network to avoid outages and provide a better user experience. With this premise, we have focused our research on PM data.

The procedure to gather these PM data is shown in Fig. 1, where NMS collects PM data from each network element (NE). Mediation components are responsible for data

The associate editor coordinating the review of this manuscript and approving it for publication was Dominik Strzalka.

11 Efficiently Processing Complex XSD using Hive and Spark

Reference:

Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2020). Efficiently Processing Complex XSD using Hive and Spark. In PeerJ Computer Science, Vol. 8, pp. 1–33.

Available on:

- https://peerj.com/articles/cs-652/
- https://doi.org/10.7717/peerj-cs.652

Contributes to the topic:

SO. 5. Propose a method to process complex eXtensible Markup Language (XML) files used in PM for Hive.

SO. 6. Propose a method to process complex XML files used in PM for Spark.

SO. 8. Evaluate the big data ecosystem for ingestion and query execution times.

PeerJ Computer Science

Efficient processing of complex XSD using Hive and Spark

Diana Martinez-Mosquera[1], Rosa Navarrete[1] and Sergio Luján-Mora[2]

[1] Department of Informatics and Computer Science, Escuela Politecnica Nacional, Quito, Ecuador
[2] Department of Software and Computing Systems, University of Alicante, Alicante, Spain

ABSTRACT

The eXtensible Markup Language (XML) files are widely used by the industry due to their flexibility in representing numerous kinds of data. Multiple applications such as financial records, social networks, and mobile networks use complex XML schemas with nested types, contents, and/or extension bases on existing complex elements or large real-world files. A great number of these files are generated each day and this has influenced the development of Big Data tools for their parsing and reporting, such as Apache Hive and Apache Spark. For these reasons, multiple studies have proposed new techniques and evaluated the processing of XML files with Big Data systems. However, a more usual approach in such works involves the simplest XML schemas, even though, real data sets are composed of complex schemas. Therefore, to shed light on complex XML schema processing for real-life applications with Big Data tools, we present an approach that combines three techniques. This comprises three main methods for parsing XML files: cataloging, deserialization, and positional explode. For cataloging, the elements of the XML schema are mapped into root, arrays, structures, values, and attributes. Based on these elements, the deserialization and positional explode are straightforwardly implemented. To demonstrate the validity of our proposal, we develop a case study by implementing a test environment to illustrate the methods using real data sets provided from performance management of two mobile network vendors. Our main results state the validity of the proposed method for different versions of Apache Hive and Apache Spark, obtain the query execution times for Apache Hive internal and external tables and Apache Spark data frames, and compare the query performance in Apache Hive with that of Apache Spark. Another contribution made is a case study in which a novel solution is proposed for data analysis in the performance management systems of mobile networks.

Submitted 23 Aperil 2021
Accepted 6 July 2021
Published 17 August 2021

Corresponding author
Diana Martinez-Mosquera,
diana.martinez@epn.edu.ec

Academic editor
Yilun Shang

Additional Information and Declarations can be found on page 30

DOI 10.7717/peerj-cs.652

Subjects Algorithms and Analysis of Algorithms, Computer Architecture, Computer Networks and Communications, Programming Languages
Keywords Hive, Spark, Performance management, Mobile network, Complex XSD, XML

INTRODUCTION

The eXtensible Markup Language (XML) is now widely used on the Internet for different purposes. There are numerous XML-based applications that utilize tag-based and nested data structures (*Chituc, 2017*; *Debreceny & Gray, 2001*; *Hong & Song, 2007*) due to greater flexibility in the representation of different types of data: these can be customized by the user. However, the main constraint is that XML representation is inefficient in terms of processing and with respect to query times; for this reason, agile and intelligent search and

How to cite this article Martinez-Mosquera D, Navarrete R, Luján-Mora S. 2021. Efficient processing of complex XSD using Hive and Spark. *PeerJ Comput. Sci.* 7:e652 http://doi.org/10.7717/peerj-cs.652

Part III

CONCLUSIONS

12 Conclusions and Outlook

12.1 Discussion

In this thesis, several important aspects were discovered. In the systematic literature review of the modeling and management of big data in databases, some main trends were identified: the Entity-Relationship model is the most used, followed by the multidimensional model, and XML at the conceptual level. The most researched model is the document-oriented at the level of logical abstraction, followed by the column-oriented model and the graphical model in third place. MongoDB is the most researched DBMS, followed by Neo4j and Cassandra. MongoDB can be thought of as the most used document-oriented data model, Cassandra for the column-oriented data model, and Neo4j for the graphics data model. Another finding is that the most proposed modeling methodology is based on data and 55.55 % of the studies focus on homogeneous database types.

Among the gaps identified in the systematic literature review, we can highlight, neither the NoSQL system nor language modeling has emerged as standards. In addition, a few studies present a solution for hybrid databases, that is, for structured and unstructured data.

Regarding the literary review of articles that study PM processing in mobile networks with big data tools, only seven related articles were found. However, in order to make a comparison of the results obtained, only two articles presented complete data: A2 (Mampaka & Sumbwanyambe, 2019), and A5 (Qiao et al., 2018).

For this study, PM files from four different providers, and four telecommunications technologies 2G, 3G, 4G, and 5G were used. We were able to determine that from 3G the main providers use the XML format defined by 3GPP TS 32.401. In addition, the systems for the data ecosystem were selected based on the support of this format.

Additionally, we were able to verify that PM files use a complex XML format. Therefore, a processing method based on cataloging, deserialization and the positional explosion was also proposed. For cataloging, the elements of the XML schema were mapped to root, arrays, structures, values, and attributes. Then deserialization and positional blasting were implemented. The proposed method was implemented and evaluated for Hive and Spark. Another important point found was the lower query execution time that was obtained in the scenario where Hive was used with internal tables, and recent versions of the tools reduced execution times.

Once the format XML to be processed was analyzed, the research determined the appropriate tools for implementing the framework. The selected components of the Big Data framework were implemented in bare metal and in a cloud solution. We evaluated the resources required to process until one billion records per hour for ingestion and reporting components.

12.2 Contribution

Based on the results of this research, the hypothesis has been verified both in the literary review and in the carried out implementation. Therefore, it is feasible to optimize the performance management of a mobile network through a big data framework.

Furthermore, each requirement for stakeholder and architecture identified in Appendix A has been complied:

- It was implemented a solution based on big data components.
- It was possible to reduce the times of ingestion and execution of queries with a turnkey solution and fewer computational resources, especially for RAM requirements. This feature of the proposal will allow reducing the times of the root cause analysis of network problems.
- It was carried out tests with PM files of a 5G network; thus, the proposal ensures the support for this new technology.
- It was performed the implementation of the solution in a bare-metal solution and in the cloud, obtaining similar results. Additionally, using Hive ensures that operators can continue to use the well-known SQL language for queries.
- It was proved that the use of HDFS allows offering a linear scalable solution.
- It was verified that Hive allows batch processing while Spark allows near-real-time.
- It was ensured that the solution is aligned with current trends in academic research and industry as it was evidenced in the literature review.

Regarding the specific objectives, a systematic literature review of the modeling and management of big data in databases was presented, which allowed determining trends and gaps in the subject. The study of the state of the art of big data and mobile networks was also presented, where the studies that present ecosystems to analyze PM data were identified and their results obtained in our topic of interest ingestion and reporting. The components of a big data architecture that should be considered during deployment were identified as software architecture, ingestion, data lake, processing, reporting, interface, and deployment. Each component was analyzed and, according to the type of data to be processed, the tools of the big data ecosystem were selected. Methods for processing complex XML in Hive and Spark-based on cataloging, deserialization, and positional exploding were proposed, since PM data from the mobile networks has this format. Finally, a big data framework was proposed for PM in mobile networks, and ingestion and query execution times were evaluated until to reach 1 billion records and compared with the studies A2 (Mampaka & Sumbwanyambe, 2019) and A5 (Qiao et al., 2018) identified in the SLR.

12.3 Future Work

As future work, we plan to evaluate data aggregations for KPI calculations from the records. Additionally, it is planned to evaluate a kappa architecture that allows reports to be presented in near-real-time. In this thesis, it has been presented how to query the PM files through Spark for streaming reports. Concerning this, we pursuit to test that through a query model the query execution times can improve remarkably.

The interface component is also considered as work to be done in the future as it allows the reports to be visualized in a friendly form. It has been considered to analyze which are the most used visualization tools, the most useful and their respective advantages and disadvantages.

Moreover, we consider this work can be used as a basis for several studies; for instance, the proposed methods could be extended to other areas not only PM in mobile networks, and the gaps found in the SLR can be addressed in future research.

An important finding was the continuous improvements in the new versions of the big data tools; thus, to keep up to date the methods for these new versions would be an excellent contribution in future research.

APPENDIX

A Initial Requirements

IEEE 29148

APPENDIX A

INITIAL REQUIREMENTS

Stakeholders
1. Network Operation Center staff
2. Network and Planning Optimization staff
3. Research and Development staff
4. Operation and Maintenance staff

Parameters

Stakeholder requirements specification (StRS)					
BUSINESS REQUIREMENTS					
#	**REQUIREMENT**	**PRIORITY**			**VERIFICATION**
		High	**Medium**	**Low**	
StRS 1	*Replacing it's current centric architecture with a more scalable architecture based on Big Data components.*	☒	□	□	
OPERATIONAL REQUIREMENTS					
StRS 2	*Lower data latencies at ingestion and reporting layers*	□	☒	□	
StRS 3	*Support to 5G and Internet of Thing technologies that are expected to increase the amount and variety of network data that must be collected and analyzed*	□	☒	□	
StRS 4	*Where possible reuse existing infrastructure and offer more freedom on data warehouse supplier choice*	☒	□	□	
USER REQUIREMENTS					
StRS 5	*Improving end use experience*	□	☒	□	
StRS 6	*Shorter times in the root cause analysis of the network problems*	☒	□	□	

ARCHITECTURE REQUIREMENTS					
#	**REQUIREMENT**	**PRIORITY**			**VERIFICATION**
		High	**Medium**	**Low**	
ARS 1	*Linear scalability and performance*	☒	□	□	*Scale on all this layers by simple physical/virtual node/resource addition.*
LOGICAL					
ARS 2	*Batch and near-real time supporting*	□	☒	□	*Provide high frequency data ingestion and transformation, efficient and simple storage, batch and near real time analytics and reporting*
DESIGN					
ARS 3	*Low latency reporting*	☒	□	□	*Fast reporting over raw and historical data and several running reports*
SOFTWARE					
ARS 4	*Platform independency*	□	☒	□	*Alignment with industry trends and state of the art large scale analytics processing strategies*
HARDWARE					
ARS 5	*Flexible*	☒	□	□	*Bare metal, virtual or containerized hardware*

B Measurement Report File Format XML Tags

A.1 Parameter description and mapping table

Table A.1 Mapping of ASN.1 Measurement Report File Format tags to XML tags

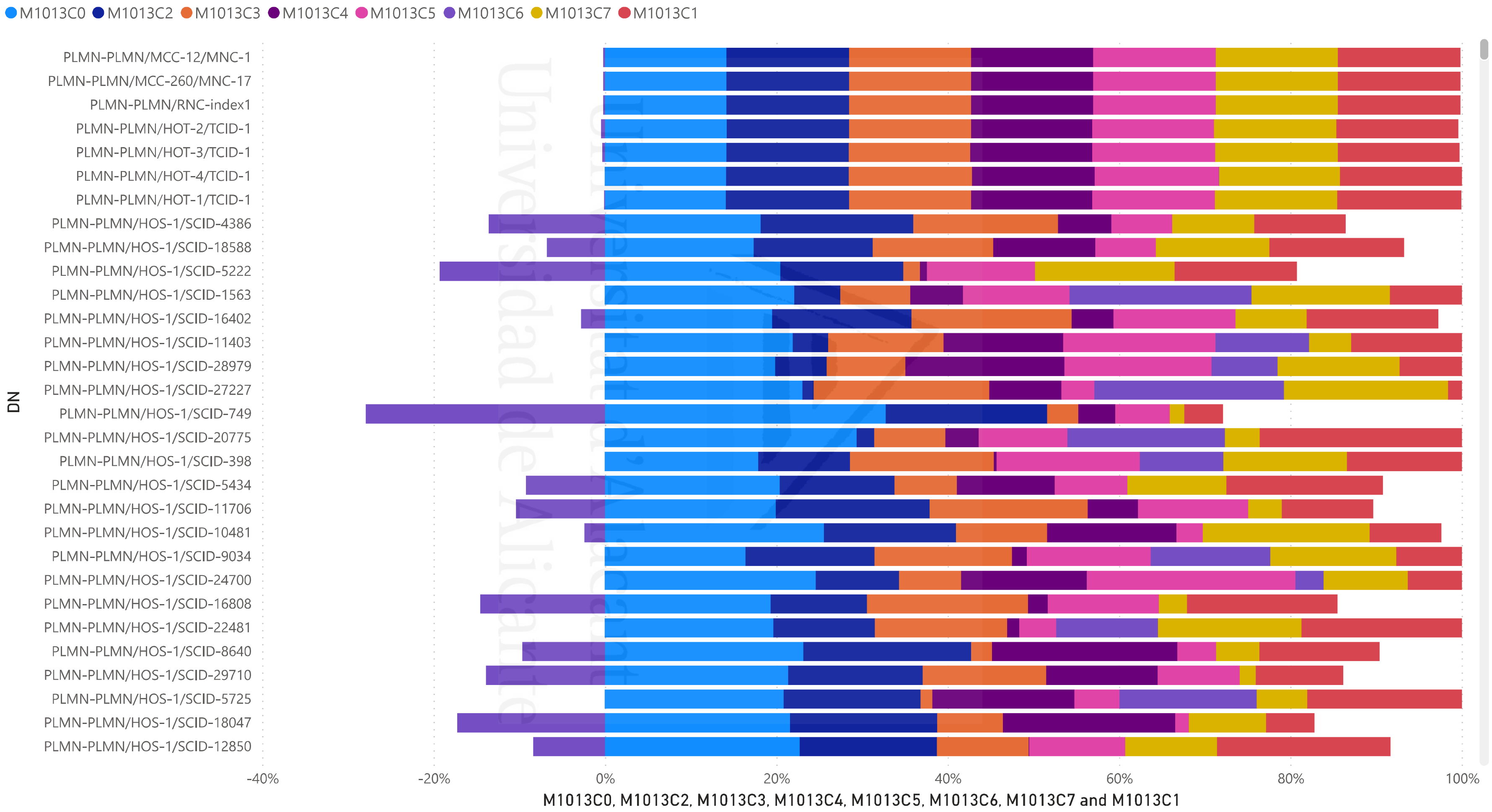
ASN.1 Tag	DTD based XML tag	XML schema based XML tag	Description
MeasDataCollection	mdc	measCollecFile	This is the top-level tag, which identifies the file as a collection of measurement data. The file content is made up of a header ("measFileHeader"), the collection of measurement result items ("measData"), and a measurement file footer ("measFileFooter").
measFileHeader	mfh	fileHeader	This is the measurement result file header to be inserted in each file. It includes a version indicator, the name, type and vendor name of the sending network node, and a time stamp ("collectionBeginTime").
measData	md	measData	The "measData" construct represents the sequence of zero or more measurement result items contained in the file. It can be empty in case no measurement data can be provided. The individual "measData" elements can appear in any order. Each "measData" element contains the name of the NE ("nEId") and the list of measurement results pertaining to that NE ("measInfo").
measFileFooter	mff	fileFooter	The measurement result file footer to be inserted in each file. It includes a time stamp, which refers to the end of the overall measurement collection interval that is covered by the collected measurement results being stored in this file.
fileFormatVersion	ffv	fileHeader fileFormatVersion	This parameter identifies the file format version applied by the sender. The format version defined in the present document shall be the abridged number and version of this 3GPP document (see below) for XML formats and the ASN.1 format alike. The abridged number and version of a 3GPP document is constructed from its version specific full reference "3GPP […] (yyyy-mm)" by: - removing the leading "3GPP TS" - removing everything including and after the version third digit, representing editorial only changes, together with its preceding dot character - from the resulting string, removing leading and trailing white space, replacing every multi character white space by a single space character and changing the case of all characters to uppercase.
senderName	sn	fileHeader dnPrefix and fileSender localDn	The senderName uniquely identifies the NE or EM that assembled this measurement file by its Distinguished Name (DN), according to the definitions in 3GPP TS 32.300 [10]. In the case of the NE-based approach, it is identical to the sender's "nEDistinguishedName". For ASN.1 and DTD based XML format, the string may be empty (i.e. string size =0) in case the DN is not configured in the sender. For the XML schema based XML format, the DN is split into the DN prefix and the Local DN (LDN) (see 3GPP TS 32.300 [10]). XML attribute specification "dnPrefix" may be absent in case the DN prefix is not configured in the sender. XML attribute specification "localDn" may be absent in case the LDN is not configured in the sender.
senderType	st	fileSender elementType	This is a user configurable identifier of the type of network node that generated the file, e.g. NodeB, EM, SGSN. The string may be empty (i.e. string size =0) in case the "senderType" is not configured in the sender. For the XML schema based XML format, XML attribute specification "elementType" may be absent in case the "senderType" is not configured in the sender.

ASN.1 Tag	DTD based XML tag	XML schema based XML tag	Description
vendorName	vn	fileHeader vendorName	The "vendorName" identifies the vendor of the equipment that provided the measurement file. The string may be empty (i.e. string size =0) if the "vendorName" is not configured in the sender. For the XML schema based XML format, XML attribute specification "vendorName" may be absent in case the "vendorName" is not configured in the sender.
collectionBeginTime	cbt	measCollec beginTime	The "collectionBeginTime" is a time stamp that refers to the start of the first measurement collection interval (granularity period) that is covered by the collected measurement results that are stored in this file.
nEId	neid	managedElement	The unique identification of the NE in the system. It includes the user name ("nEUserName"), the distinguished name ("nEDistinguishedName") and the software version ("nESoftwareVersion") of the NE.
nEUserName	neun	managedElement userLabel	This is the user definable name ("userLabel") defined for the NE in 3GPP TS 32.622 [24]. The string may be empty (i.e. string size =0) if the "nEUserName" is not configured in the CM applications. For the XML schema based XML format, XML attribute specification "userLabel" may be absent in case the "nEUserName" is not configured in the CM applications.
nEDistinguishedName	nedn	fileHeader dnPrefix and managedElement localDn	This is the Distinguished Name (DN) defined for the NE in 3GPP TS 32.300 [10]. It is unique across an operator's 3G network. The string may be empty (i.e. string size =0) if the "nEDistinguishedName" is not configured in the CM applications. For the XML schema based XML format, the DN is split into the DN prefix and the Local DN (LDN) (see 3GPP TS 32.300 [10]). XML attribute specification "localDn" may be absent in case the LDN is not configured in the CM applications.
nESoftwareVersion	nesw	managedElement swVersion	This is the software version ("`swVersion`") defined for the NE in 3GPP TS 32.622 [24]. This is an optional parameter which allows post-processing systems to take care of vendor specific measurements modified between software versions. For the XML schema based XML format, XML attribute specification "swVersion" may be absent in case the "nESoftwareVersion" is not configured in the CM applications.
measInfo	mi	measInfo	The sequence of measurements, values and related information. It includes a list of measurement types ("measTypes") and the corresponding results ("measValues"), together with the time stamp ("measTimeStamp") and granularity period ("granularityPeriod") pertaining to these measurements.
measTimeStamp	mts	granPeriod endTime	Time stamp referring to the end of the granularity period.
granularityPeriod	gp	granPeriod duration	Granularity period of the measurement(s) in seconds. For the XML schema based XML format, the value of XML attribute specification "duration" shall use the truncated representation "PT*n*S" (see [28]).
measTypes	mt	measTypes or measType	This is the list of measurement types for which the following, analogous list of measurement values ("measValues") pertains. The GSM only measurement types are defined in TS 52.402 [22]. The measurement types for UMTS and combined UMTS/GSM implementations are specified in TS 32.403 [23]. For the XML schema based XML format, depending on sender's choice for optional positioning presence, either XML element "measTypes" or XML elements "measType" will be used.
measValues	mv	measValue	This parameter contains the list of measurement results for the resource being measured, e.g. trunk, cell. It includes an identifier of the resource ("measObjInstId"), the list of measurement result values ("measResults") and a flag that indicates whether the data is reliable ("suspectFlag").

ASN.1 Tag	DTD based XML tag	XML schema based XML tag	Description
measObjInstId	moid	measValue measObjLdn	The "measObjInstId" field contains the local distinguished name (LDN) of the measured object within the scope defined by the "nEDistinguishedName" (see 3GPP TS 32.300 [10]). The concatenation of the "nEDistinguishedName" and the "measObjInstId" yields the DN of the measured object. The "measObjInstId" is therefore empty if the "nEDistinguishedName" already specifies completely the DN of the measured object, which is the case for all measurements specified on NE level. For example, if the measured object is a "ManagedElement" representing RNC "RNC-Gbg-1", then the "nEDistinguishedName" will be for instance "DC=a1.companyNN.com,SubNetwork=1,IRPAgent=1,SubNetwork=CountryNN,MeContext=MEC-Gbg-1,ManagedElement=RNC-Gbg-1", and the "measObjInstId" will be empty. On the other hand, if the measured object is a "UtranCell" representing cell "Gbg-997" managed by that RNC, then the "nEDistinguishedName" will be for instance the same as above, i.e. "DC=a1.companyNN.com,SubNetwork=1,IRPAgent=1,SubNetwork=CountryNN,MeContext=MEC-Gbg-1,ManagedElement=RNC-Gbg-1", and the "measObjInstId" will be for instance "RncFunction=RF-1,UtranCell=Gbg-997". The class of the "measObjInstId" is defined in item F of each measurement definition template.
measResults	r	measResults or r	This parameter contains the sequence of result values for the observed measurement types. The "measResults" sequence shall have the same number of elements, which follow the same order as the measTypes sequence. Normal values are INTEGERs and REALs. The NULL value is reserved to indicate that the measurement item is not applicable or could not be retrieved for the object instance. For the XML schema based XML format, depending on sender's choice for optional positioning presence, either XML element "measResults" or XML elements "r" will be used.
suspectFlag	sf	suspect	Used as an indication of quality of the scanned data. FALSE in the case of reliable data, TRUE if not reliable. The default value is "FALSE", in case the suspect flag has its default value it may be omitted.
timeStamp	ts	measCollec endTime	This tag carries the time stamp that refers to the end of the measurement collection interval (granularity period) that is covered by the collected measurement results that are stored in this file. The minimum required information within timestamp is year, month, day, hour, minute, and second.
Not Required	mt p	measType p	An optional positioning XML attribute specification of XML elements "mt" (DTD based) and "measType" (XML schema based), used to identify a measurement type for the purpose of correlation to a result. The value of this XML attribute specification is expected to be a non-zero, non-negative integer value that is unique for each instance of XML element "mt" or "measType" that is contained within the measurement data collection file.
Not Required	r p	r p	An optional positioning XML attribute specification of XML element "r", used to correlate a result to a measurement type. The value of this XML attribute specification should match the value of XML attribute specification "p" of corresponding XML element "mt" " (DTD based) or "measType" (XML schema based).

C Example of a Report of PM files with PowerBI

M1013C0, M1013C2, M1013C3, M1013C4, M1013C5, M1013C6, M1013C7 and M1013C1 by DN



D Resumen

En cumplimiento de la normativa de la Universidad de Alicante sobre “TESIS EN LENGUAS DISTINTAS A LAS OFICIALES EN LA COMUNIDAD AUTÓNOMA VALENCIANA” que indica:

En cualquier caso, la tesis deberá contener un resumen en una de las dos lenguas oficiales de esta Comunidad Autónoma. Este resumen deberá contener una introducción general, un resumen global de los resultados obtenidos, de la discusión de estos resultados y de las conclusiones finales. Este resumen deberá dar una idea bastante precisa del contenido de la Tesis. La extensión de la parte escrita en uno de los idiomas oficiales de esta Comunidad Autónoma no será inferior a 5.000 palabras.

Se incluye, a continuación, este apéndice que es un resumen del trabajo de investigación realizado.

D.1 Introducción

La industria de las telecomunicaciones, en las últimas décadas, se ha convertido en una de las que más evolucionan y crecen vertiginosamente en el mundo (Kovačević et al., 2017). Se ha sido testigo del surgimiento de varias generaciones de redes móviles desde 2G hasta 5G en la actualidad, y 6G en un futuro cercano. Según el Sistema Global de Comunicaciones Móviles, para el año 2025, se estima que existirán alrededor de 8,8 mil millones de suscripciones en el mundo (Global System for Mobile Communications, 2021).

Con esta demanda, para brindar un servicio adecuado a los usuarios, los operadores de redes móviles deben monitorear y medir constantemente el desempeño de los miles de elementos de red que se han desplegado en todo el mundo. Para este fin, se envían ficheros con los respectivos datos sobre el estado de cada elemento de red (*network element*, NE) hacia los sistemas de gestión de la red , generalmente cada 15 minutos. Esto genera una gran cantidad de datos crudos que deben procesarse y analizarse para monitorear el comportamiento del servicio de red en el menor tiempo posible (3rd Generation Partnership Project, 2005).

Por otro lado, la investigación sobre el procesamiento de grandes volúmenes de datos ha sido de gran interés durante mucho tiempo. Ya en 1944 se empezó a estudiar el rápido crecimiento de las bibliotecas (Rider, 1944) y para 1997 se introdujo por primera vez el término *"big data"* en un estudio científico (Cox & Ellsworth, 1997). El análisis de *big data* ha permitido a muchas empresas mejorar sus ventajas competitivas y los operadores de redes móviles (*mobile network operator*, MNO) no han sido la excepción.

Como se presenta en el capítulo 3 Revisión de la Literatura, desde el año 2012 se presentan resultados de investigación sobre redes móviles relacionadas con *big data*. Los estudios se han centrado en el análisis de técnicas, herramientas, marcos de trabajo (*framework*, datos de usuarios y redes. Varios proveedores de redes móviles (Nokia, 2020; Skračić & Bodrušić, 2017) han diseñado sistemas que aprovechan los beneficios de las plataformas de *big data* para respaldar el área de planificación y optimización de la red, una de las más importantes en un MNO cuya función principal es detectar, corregir y predecir el estado de la red a través de las mediciones enviadas por los NE.

En este trabajo, se propone un *framework* que utiliza herramientas de *big data* para mediciones de desempeño (*performance measurement*, PM) de redes móviles, después de una investigación exhaustiva de la arquitectura más adecuada, con la premisa de optimizar los recursos computacionales requeridos y el tiempo de procesamiento. La optimización en la gestión de redes móviles es un tema muy complejo e importante debido a la gran cantidad de dispositivos que se deben monitorear para determinar problemas o deterioro en el nivel de calidad del servicio brindado. El tiempo de inactividad de un NE en la red móvil tiene un impacto directo en los ingresos y gastos operativos (*operational expenditures*, OPEX). Por lo tanto, cuanto más se tarde en analizar el PM de la red móvil, mayor será el OPEX para el MNO.

Con base en esta información, el *framework* propuesto ha sido evaluado, a nivel de tiempos de ejecución requeridos, en los componentes de ingesta y de reportería sobre un clúster implementado en la nube y en un equipo físico. Las pruebas de evaluación fueron realizadas mediante conjuntos de datos reales recopilados desde celdas 3G, 4G y 5G de cuatro MNO diferentes.

El objetivo principal de esta tesis es diseñar un *framework* de *big data* que optimice la gestión de PM para redes móviles en términos de recursos computacionales requeridos en los componentes de ingesta y reportería. Los objetivos específicos (OE) son los siguientes:

OE. 1. **Determinar las tendencias y las brechas** en la investigación del modelado y gestión de *big data* en bases de datos, mediante una revisión de literatura sistemática.

OE. 2. **Conocer el estado de la cuestión** de la investigación de *big data* y redes móviles.

OE. 3. **Definir un método** para la selección de herramientas de *big data*.

OE. 4. **Definir un método** para diseñar un *framework* de *big data* para PM en redes móviles.

OE. 5. **Proponer un método** para procesar archivos XML de estructura compleja utilizados en PM para Hive.

OE. 6. Proponer un método para procesar archivos XML complejos utilizados en PM para Spark.

OE. 7. Proponer un *framework* de *big data* para PM en redes móviles y su implementación.

OE. 8. Evaluar el *framework* de *big data* para conocer los tiempos de ingesta y ejecución de consultas con las herramientas seleccionadas.

Los objetivos de esta tesis se han cumplido plenamente durante los años del programa de doctorado. Los resultados obtenidos se han evidenciado en las publicaciones en revistas indexadas. Además, se realizaron trabajos previos que no aportan a la unidad temática, sino únicamente a *big data* y fueron presentados en diferentes conferencias revisadas por pares. La tabla D.1 muestra la contribución de cada uno de los artículos al cumplimiento de estos objetivos.

En esta tesis se ha planteado la siguiente hipótesis: Es factible diseñar, implementar y optimizar la gestión del desempeño de una red móvil a nivel de componentes de ingesta y reportería, mediante un *framework* de *big data*.

D.2 Método

La metodología utilizada en esta tesis es la investigación en ciencias del diseño, la cual es ampliamente utilizada en sistemas de información, y cuyo postulado fundamental es que el conocimiento, comprensión de un problema, y su solución se adquiere en la aplicación y construcción de un artefacto. En este trabajo, el artefacto constituye el *framework* de *big data* para la gestión del rendimiento en redes móviles. Para la aplicación de esta metodología se han considerado estas siete guías principales:

1. **Relevancia del problema:** Para conocer la relevancia del problema, se realizaron varias revisiones de la literatura existente. La primera revisión se realizó durante el periodo 2018-2019 y permitió conocer el estado de la cuestión del modelado y gestión de *big data* en bases de datos. Las próximas revisiones realizadas en los periodos 2019-2020 y 2020-2021 permitieron conocer los trabajos relevantes presentados sobre *big data* y redes móviles, con énfasis en la gestión de PM.

2. **Rigor de la investigación:** La investigación se basa en la aplicación de métodos rigurosos. Para la revisión de la literatura se utilizaron las pautas propuestas por Kitchenham, centrándose en tres fases principales: planificación, realización, y presentación de informes (Kitchenham, 2004). La recopilación de los estudios se realizó a partir de las bibliotecas digitales científicas más importantes.

   Además, los requisitos iniciales de un *framework* de *big data* para PM en redes móviles se recopilaron de un proveedor de telecomunicaciones con la ayuda del estándar IEEE 29148 (Institute of Electrical and Electronics Engineers, 2018). Los resultados se presentan en el Apéndice A.

   Para las pruebas, se utilizaron datos de redes reales 2G, 3G, 4G y 5G de cuatro diferentes MNO y se realizaron varias pruebas sobre equipos físicos, una solución en la nube y sobre diferentes versiones de herramientas de *big data*.

Objetivo	**Artículos**
OE. 1	An approach to big data modeling for key-value NoSQL (Martinez-Mosquera, Luján-Mora, Navarrete, et al., 2019). Modeling and management big data in databases — A systematic literature review (Martinez-Mosquera et al., 2020b).
OE. 2	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera et al., 2020a).
OE. 3	Development and evaluation of a big data framework for performance management in mobile networks.
OE. 4	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera et al., 2020a).
OE. 5	Efficiently processing complex XSD using Hive and Spark.
OE. 6	Efficiently processing complex XSD using Hive and Spark (Martinez-Mosquera et al., 2021).
OE. 7	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera et al., 2020a). Framework for big data integration in e-government (Martinez-Mosquera & Luján-Mora, 2019).
OE. 8	Development and evaluation of a big data framework for performance management in mobile networks (Martinez-Mosquera et al., 2020a). Efficiently processing complex XSD using Hive and Spark (Martinez-Mosquera et al., 2021).

Table D.1: Artículos que contribuyen a los objetivos de esta tesis

3. **Diseño como proceso de búsqueda:** Se utilizaron diferentes métodos iterativos para probar las alternativas de diseño frente a requisitos o restricciones.

4. **Diseño como artefacto:** Nuestra investigación produjo un artefacto viable (en este caso, un *framework*).

5. **Evaluación del diseño:** La solución se presenta como capas independientes, y las capas de ingestión e informes se evaluaron para presentar la utilidad, calidad y efectividad de un artefacto de diseño.

6. **Contribuciones de investigación:** Esta investigación proporciona contribuciones claras y verificables en las áreas de artefactos de diseño, fundamentos de diseño y / o metodologías de diseño.

7. **Comunicación de la investigación:** Los resultados de la investigación se han presentado eficazmente en revistas y congresos relacionados con el área de investigación.

Característica	**Breve descripción**
Volumen	Grandes conjuntos de datos
Variedad	Diferentes tipos de formatos de datos
Velocidad	Alta tasa de generación de datos
Valor	Datos útiles para recuperar información
Veracidad	Precisión de los datos
Variabilidad	Datos consistentes
Viscosidad	Variaciones en la velocidad de datos
Viralidad	Velocidad de transmisión de datos
Validez	Evaluación de datos
Visualización	Simbolización de datos

Table D.2: Las 10 V de big data
(Martinez-Mosquera et al., 2020b)

D.3 Fundamentos Teóricos

D.3.1 Big Data

Desde los inicios de la humanidad, la transferencia del conocimiento ha sido uno de los principales propósitos del ser humano. Muchas culturas antiguas proporcionaron información a sus descendientes a través de sus sistemas de escritura. Actualmente, se puede decir que vivimos en la era de las grandes cantidades de datos. Exabytes de datos se generan todos los días; por lo tanto, el término *big data* se ha convertido en uno de los conceptos más importantes en los sistemas de información.

Big data es el término que se ha utilizado en los últimos años para referirse a grandes y complejos volúmenes de datos que combinan características conocidas como las V de *big data* (Ribeiro & da Silva, 2015). El procesamiento y análisis de *big data* requieren sistemas informáticos no tradicionales (Turck, 2020) desarrollados para que la organización tome decisiones estratégicas y oportunas (Tekiner & Keane, 2013).

D.3.2 Caracterización de Big Data

Como se mencionó, para considerar un gran volumen de datos como *big data*, también se deben cumplir otras características como la velocidad y la variedad. Estas tres características fueron bien aceptadas durante varios años para definir *big data*. Sin embargo, para 2014, ya se habían definido 10 características como se muestra en la Tabla D.2. Para el año 2017, ya se discutieron 42 V como se muestra en la Figure D.1 y se explica en Farooqi et al. (2019).

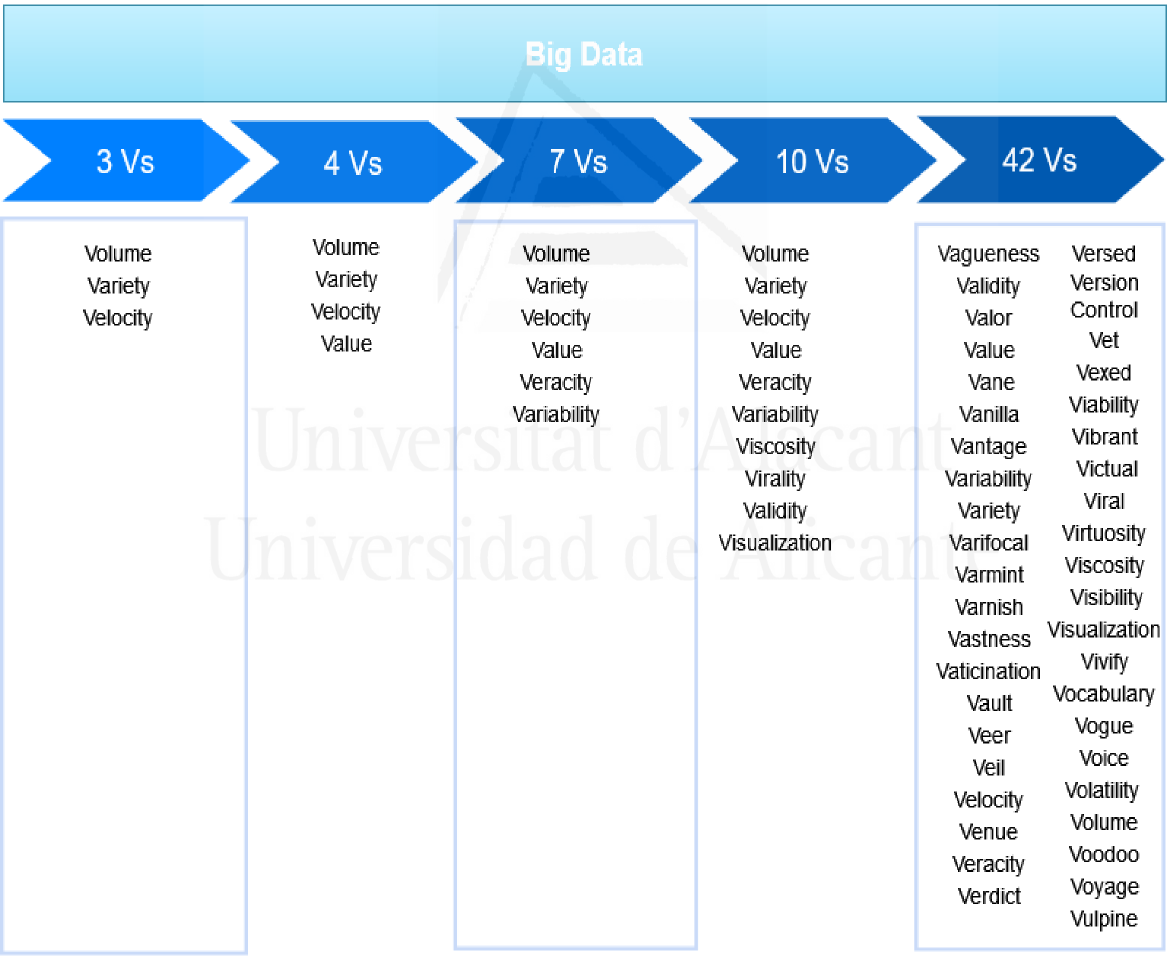


Figure D.1: La evolución de las V de *big data*

En cuanto a la característica de Variedad, los datos se han clasificado en estructurados, semiestructurados y no estructurados. Según esta clasificación, existen procedimientos, procesos, herramientas y técnicas que permiten el almacenamiento, la manipulación y la gestión.

D.3.2.1 Datos estructurados

Se han denominado datos estructurados a aquellos que se representan en forma tabular, en hojas de cálculo o bases de datos relacionales (Davoudian et al., 2018). Según el informe presentado por la empresa CISCO (Barton & Henry, 2020), más del 90 % son datos no estructurados. Este tipo de datos se explican en las siguientes secciones.

D.3.2.2 Datos semiestructurados

Los datos semiestructurados se consideran datos que no tienen una estructura formal, como los datos dentro de una base de datos relacional. Los formatos de valores separados por comas (*comma-separated values*, CSV), XML, notación de objetos JavaScript (*JavaScript object notation*, JSON) y JSON binario (*binary JSON*, BSON) son algunos de los que se consideran semiestructurados ya que presentan una determinada organización en su estructura (Baek et al., 2016; P. Li et al., 2020; Papadakis, 2018).

D.3.2.3 Datos no estructurados

Aquellos datos que no tienen una estructura predefinida u organizada se conocen como no estructurados. Por ejemplo, documentos de texto, correos electrónicos, datos de sensores, archivos de audio, archivos de imagen, archivos de video, datos de sitios web, chats, registros médicos electrónicos, datos de redes sociales, entre otros (Costa & Santos, 2017; O'Sullivan et al., 2014).

Cada tipo de datos se puede procesar en una arquitectura de *big data* de acuerdo con sus requisitos comerciales.

D.3.3 Arquitectura de Big Data

Según el estándar IEEE 1471-2000, la arquitectura es el diseño fundamental de la organización de un sistema computacional y sus componentes (Institute of Electrical and Electronics Engineers, 2020). Para *big data*, hay dos arquitecturas disponibles: lambda y kappa (Ounacer et al., 2017).

De acuerdo con los requisitos de arquitectura presentados en el Apéndice A, en esta tesis se ha utilizado una arquitectura lambda. Los componentes principales que debe proporcionar cualquier implementación de *big data* (Microsoft, 2020) se resumen en la Tabla D.3. El *framework* de *big data* se compone de arquitectura de software, ingesta, *data lake*, procesamiento, reportes e implementación.

Dado que este estudio se centra en el desarrollo de un *framework* para la gestión del rendimiento en redes móviles, la siguiente sección explicará los conceptos básicos de telecomunicaciones.

Componente	**Descripción**	**Opciones**
Arquitectura de Software	"La organización fundamental de un sistema implementado en sus componentes, sus relaciones entre sí y con el medio ambiente, y los principios que guían su diseño y evolución. " (Institute of Electrical and Electronics Engineers, 2020)	Lambda, Kappa
Ingesta	Sistema mediador para la recopilación masiva de datos sin procesar de los NE al *data lake* independientemente del formato	Flume, Kafka, NiFi
Data Lake	Sistema de archivos distribuido que permite almacenar todo tipo de datos de forma eficiente	Hadoop Distributed File System (HDFS)
Procesamiento	Plataforma para proporcionar un procesamiento en tiempo real de los datos de la red	Spark, Storm
Reportes	Herramienta para proporcionar el uso del lenguaje SQL para consultar datos almacenados en el *data lake*	Hive, Impala, Pig, Spark SQL
Implementación	Plataforma para instalar y configurar nodos de clúster de Hadoop y permitir una implementación ágil de aplicaciones	Cloudera, Hortonworks, IBM InfoSphere Big Insights, MapR, Pivotal HD

Table D.3: Características y herramientas para un *framework* de *big data* (Martinez-Mosquera et al., 2020a)

D.4 Red móvil

Las redes móviles son aquellas compuestas por un conjunto de NE que incluyen estaciones base que ofrecen transmisión y recepción inalámbrica de información digital en una o varias áreas geográficas llamadas celdas (Olsson et al., 2009).

Las tecnologías de redes móviles han evolucionado rápidamente en las últimas décadas, desde el sistema global para comunicaciones móviles (*global system for mobile communications*, GSM), el servicio general de papuetes de radio (*general packet radio service* , GPRS), las tasas de datos mejoradas para la evolución GSM (*enhanced data rates for GSM evolution*, EDGE), el sistema universal de telecomunicaciones móviles (*universal mobile telecommunications system*, UMTS), el acceso de paquetes de enlace ascendente de alta velocidad (*high-speed uplink packet access*, HSUPA), el acceso de paquetes de enlace descendente de alta velocidad (*high-speed downlink packet access*, HSDPA), la evolución a largo plazo (*long term evolution*, LTE) y, actualmente, la quinta generación (5G) (3rd Generation Partnership Project, 2021).

Cualquier generación de una red móvil requiere la recopilación de datos de rendimiento registrados por sus NE. Este procedimiento de red se denomina gestión del rendimiento. El propósito es obtener información de la red para verificar las configuraciones físicas y lógicas y localizar problemas de red. El tipo de datos que se recopilan se llaman PM.

Los PM son datos producidos por los NE que son recopilados y procesados por un administrador de elementos de red (*element manager*, EM) a través de interfaces abiertas y bien estandarizadas que admiten la gestión de múltiples proveedores y tecnologías (3rd Generation Partnership Project, 2005). Los archivos PM contienen datos relevantes, por ejemplo:

- Tráfico de usuarios y señalización que facilita la planificación y operación de la red.

- Configuración de red para evaluar la eficacia del plan o los cambios planificados.

- Medidas de acceso a recursos para una evaluación precisa.

- Disponibilidad del servicio.

- Calidad de servicio (*quality of service*, QoS), por ejemplo: tiempos de establecimiento de llamadas, tiempos de transferencia de paquetes, entre otros. QoS permite determinar el rendimiento de la red experimentado por el usuario.

Los archivos PM producidos por los NE se transfieren a un EM y luego a un Sistema de gestión de red (*network management system*, NMS). El NMS es responsable del almacenamiento, el posprocesamiento y la presentación para determinar los indicadores de rendimiento clave (*key performance indicator*, KPI) de la red para una evaluación adicional. Con el crecimiento de las redes móviles, también está aumentando el número de NE y, por tanto, la cantidad de estos archivos PM. Los archivos PM siguen el formato XML en la mayor parte de los MNO.

D.5 Revisión Literaria

Basados en el procedimiento propuesto por Kitchenham (2004), la Figure D.2 resume los principales aspectos considerados en la revisión literaria sistemática (*systematic literature review*, SLR) realizada.

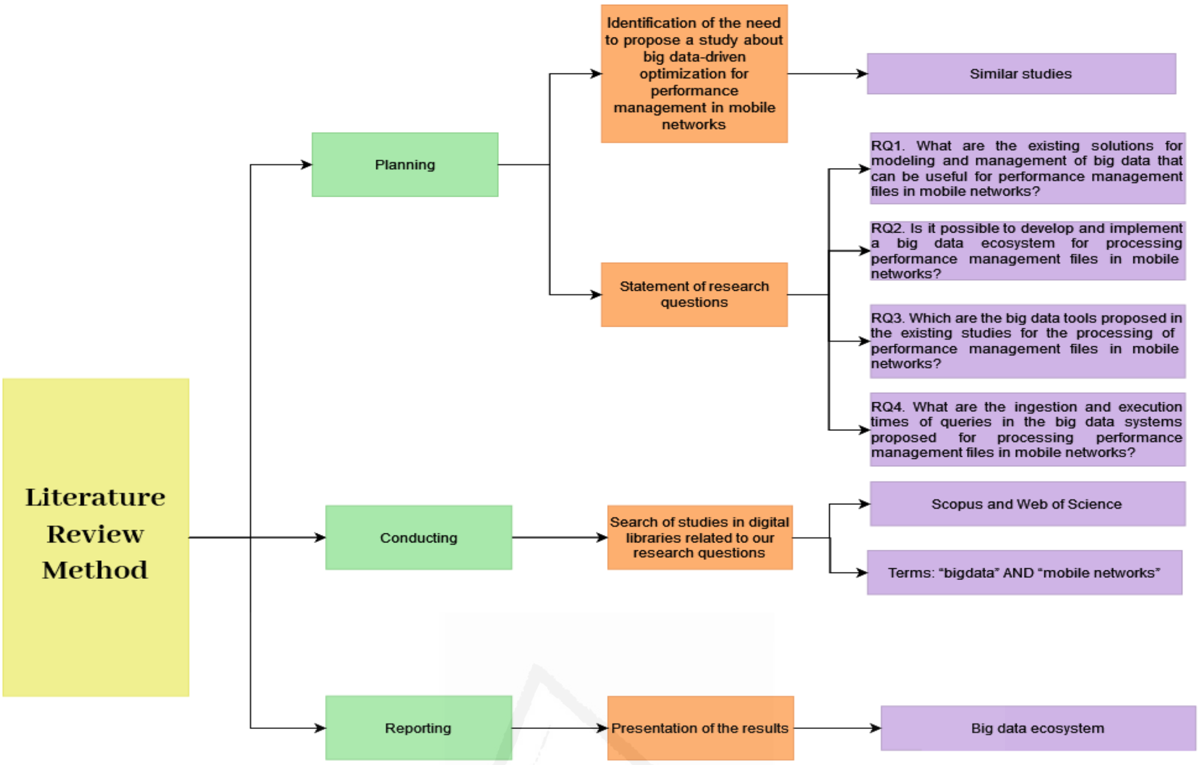

Literature Review Method
Planning
Conducting
Reporting
Identification of the need to propose a study about big data-driven optimization for performance management in mobile networks
Statement of research questions
Search of studies in digital libraries related to our research questions
Presentation of the results
Similar studies
RQ1. What are the existing solutions for modeling and management of big data that can be useful for performance management files in mobile networks?
RQ2. Is it possible to develop and implement a big data ecosystem for processing performance management files in mobile networks?
RQ3. Which are the big data tools proposed in the existing studies for the processing of performance management files in mobile networks?
RQ4. What are the ingestion and execution times of queries in the big data systems proposed for processing performance management files in mobile networks?
Scopus and Web of Science
Terms: "bigdata" AND "mobile networks"
Big data ecosystem

Figure D.2: Procedimiento realizado en la revisión literaria

En total, se obtuvieron 575 artículos de las bibliotecas científicas Scopus y Web of Science; sin embargo, se duplicaron 159 estudios. Por tanto, nuestro corpus de estudios primarios constaba de 416 artículos. Después de verificar sus títulos, resúmenes y palabras clave, 28 estudios se relacionaron directamente con nuestro tema de investigación; por tanto, se consideraron estudios potencialmente relevantes. Luego de una revisión completa de cada uno, se consideraron 7 estudios directamente relacionados con nuestras preguntas de investigación (*research question*, RQ). Los resultados se presentan en la Tabla D.4.

Para RQ1, los modelos de *big data* que se utilizan con más frecuencia son entidad-relación en tres estudios, orientados a columnas en dos estudios y clave-valor en dos estudios. Para RQ2, en todos los casos, es posible desarrollar e implementar una arquitectura de *big data* para procesar archivos de gestión del rendimiento en redes móviles. Para RQ3, Flume, Kafka y NiFi son las herramientas más utilizadas en el componente de ingesta; HDFS es el *data lake* utilizado en casi todos los casos; y Hive, PostgreSQL, HBase, Spark, Druid y Flink son las bases de datos probadas. Para RQ4, solo dos estudios presentan los tiempos de los componentes de ingesta y reportes, y solo un estudio presenta los tiempos para ambas capas.

Además, no todos los estudios presentan la arquitectura utilizada en sus propuestas ni el número de registros procesados. Podemos destacar de estos estudios el uso de altos recursos computacionales para RAM y CPU.

ID	Title and Reference	RQ1	RQ2	RQ3	RQ4 Ingestion Time [seconds]	RQ4 Reporting Time [seconds]	Observation
A1	A Big Data Solution for Troubleshooting Mobile Network Performance Problems (Skračić & Bodrušić, 2017)	Entity Relationship	yes	Flume HDFS Hive Postgresql	NA	NA	2 namenodes (256 GB RAM) 8 datanodes (512 GB RAM)
A2	A quadri-dimensional approach for poor performance prioritization in mobile networks using Big Data (Mampaka & Sumbwanyambe, 2019)	Entity Relationship	yes	HDFS Impala	NA	2.3	11 millions raw records 4 VCPUs 16 GB RAM
A3	Applying Big Data, Machine Learning, and SDN/NFV to 5G Traffic Clustering, Forecasting, and Management Entity Relationship (Le et al., 2018)	Entity Relationship	yes	Kafka HDFS HBase Spark	NA	NA	
A4	Big Data Streaming Analytics for QoE Monitoring in Mobile Networks: A Practical Approach (Rueda et al., 2018)	Column-oriented	yes	NiFi HDFS Druid	28.7	NA	20 billions raw records
A5	Characterizing Flow, Application, and User Behavior in Mobile Networks: A Framework for Mobile Big Data (Qiao et al., 2018)	key-value	yes	Flume Kafka HDFS Spark	NA	NA	1.792 TB RAM 224 VCPUs
A6	Distributed Big Data Driven Framework for Cellular Network Monitoring Data (Suleykin & Panfilov, 2019)	key-value	yes	Kafka Spark	0.014	0.464	61500 raw records 918 GB RAM 96 VCPUs
A7	Towards Adopting Big Data Technologies by Mobile Networks Operators: a Moroccan Case Study (Daki et al., 2016)	Column-oriented	yes	Kafka HDFS Flink	NA	NA	

Table D.4: Resultados de la revisión literaria

D.6 Descripción del Trabajo

Para cumplir con el objetivo principal de este estudio de doctorado, la investigación se dividió en dos partes:

1. La primera parte se refiere a la tecnología de *big data*, que permite conocer las herramientas, aplicaciones y métodos de modelado disponibles.

2. La segunda parte se centró en el estudio de las PM de redes móviles y la orquestación de servicios de *big data* para la implementación de la arquitectura.

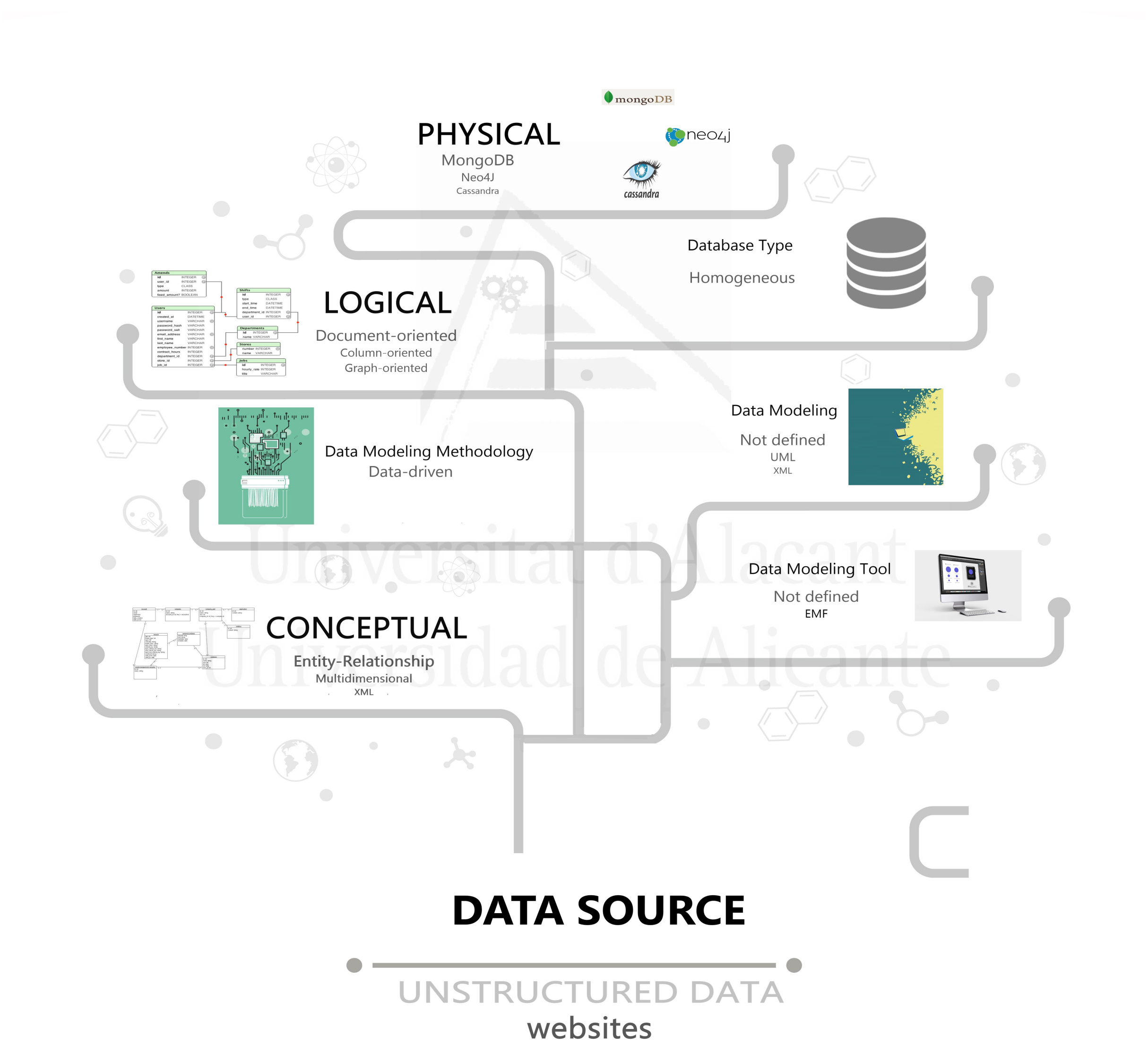


Figure D.3: Tendencias en el modelado de big data según los resultados de la SLR (Martinez-Mosquera et al., 2020b)

D.7 Modelamiento de Big Data

Desde el inicio de esta investigación, se encontró que el análisis de *big data* se ha utilizado en diferentes áreas como salud, marketing, transporte, educación, medio ambiente, entre otras aplicaciones. Por ejemplo, el artículo Martinez-Mosquera and Luján-Mora (2019) presenta una propuesta de un *framework* de *big data* para el gobierno electrónico. *Big data* puede producir información valiosa luego de la correcta implementación de un *framework* que permite que los datos sean procesados para después de su almacenamiento permitir su minería, reporte o manipulación.

En el estudio Martinez-Mosquera, Luján-Mora, Navarrete, et al. (2019), se propone un enfoque para modelar *big data* en bases de datos de clave-valor. Además, los resultados de la revisión de la literatura en Martinez-Mosquera et al. (2020b), realizada para el modelado de big data (2010-2019), permitió conocer las tendencias y brechas en el tema. Estos resultados se resumen en las Figure D.3 y Figure D.4.

Los aspectos más importantes presentados en la Figure D.3 son los siguientes: las fuentes de datos más investigadas son datos no estructurados, el modelo entidad-relación es el más utilizado a nivel de abstracción conceptual, el más investigado a nivel lógico es el modelado de datos orientado a documentos, las implementaciones se centran en los sistemas de gestión de bases de datos MongoDB, seguido de Neo4j y Cassandra, la metodología de modelado más propuesta está basada en datos. Además, la mayoría de los estudios científicos analizaron sitios web, sensores y documentos electrónicos. Para las bases de datos NoSQL, se recomienda el uso del sistema de persistencia políglota (*polyglot persistence system*, PPS).

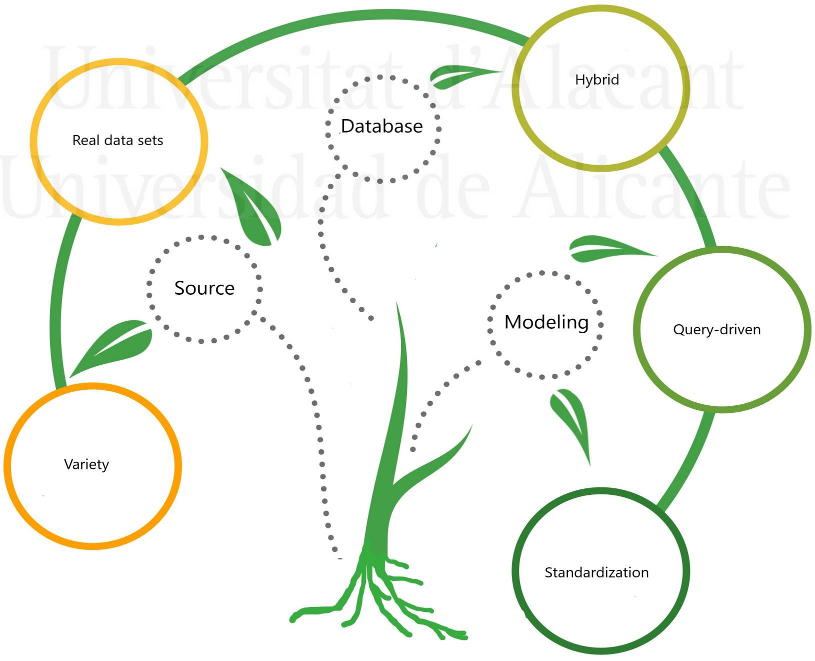


Figure D.4: Brechas en el modelado de big data según los resultados de la SLR (Martinez-Mosquera et al., 2020b)

Figure D.4 presenta las brechas donde los estudios futuros pueden concentrar sus esfuerzos, entre ellos: soluciones para variedad de datos, el uso de casos de estudios con conjuntos de datos reales, propuestas para bases de datos híbridas, modelado de datos basado en consultas y estandarización.

D.8 PM en Redes Móviles

La transferencia de PM desde NE al EM se realiza normalmente a través de SFTP en modo pull; es decir, el EM se conecta a los NE y recopila nuevos datos sin procesar. Luego, los datos se transfieren al NMS. El NMS será responsable de almacenar, procesar, informar y visualizar los datos.

Según Khalifa et al. (2016), entre los pilares principales para construir un ecosistema de big data se encuentran: implementación, almacenamiento o *data lake*, procesamiento e interfaz. En esta investigación, se pudo reconocer otros componentes importantes como la arquitectura del software, la ingesta y los reportes. Dentro de cada componente, hay muchas herramientas que se pueden seleccionar de acuerdo con los requisitos del proyecto. Estas herramientas se pueden encontrar en el panorama de datos disponible en Turck (2020). Entre las herramientas de *big data* analizadas en este estudio, se tienen las siguientes:

1. **Arquitectura de software:** Lambda and Kappa.
2. **Ingesta:** Flume, Kafka, NiFi, Sqoop.
3. **Almacenamiento o Data Lake:** Azure, AWS, IBM Power Systems, HDFS.
4. **Reportes:** HBase, Hive, Impala, Pig.
5. **Procesamiento:** Spark, Storm.
6. **Interfaz:** Power BI, Tableau.
7. **Implementación:** Cloudera Distribution Hadoop (CDH), Hortonworks, IBM InfoSphere Big Insights, MapR, Pivotal HD.

El procedimiento para recopilar datos de PM de redes móviles y procesarlos en arquitectura de *big data* se presenta en Figure D.5. Se utiliza la arquitectura Lambda y se identifican las capas *batch layer* (BL), *serving layer* (SRL) y *streaming layer* (SL). A continuación, se presentan las herramientas utilizadas en cada uno de los componentes de la arquitectura y los resultados obtenidos .

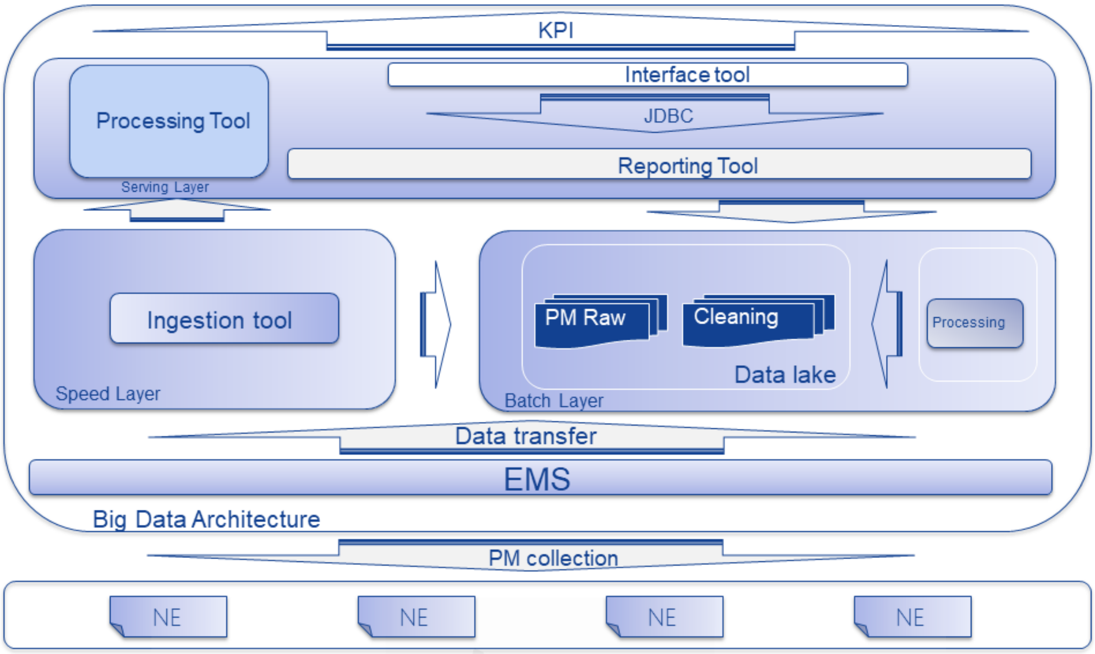


Figure D.5: Recolección y procesamiento de datos PM desde NE.

D.8.1 Ingesta

El componente de ingesta es responsable de recopilar datos de la fuente y transferirlos al destino. En este caso, la herramienta de ingesta debe recopilar datos sin procesar de PM en formato XML del EM y transferirlos a un *data lake*.

En esta tesis, se consideraron las siguientes herramientas de *big data* que admiten la recopilación y transferencia de archivos XML: Flume, Kafka y NiFi. Kafka y Nifi necesitan una versión empresarial de un marco de implementación e instalación de otros servicios. Por lo tanto, Flume fue seleccionado con el soporte SFTP. Flume funciona con tres niveles: Tier1 envía datos a Tier2, Tier2 mueve los datos al almacenamiento y Tier3 es el almacenamiento final (Apache, 2020a).

Luego de implementar un entorno de prueba donde se simuló el NE mediante una máquina virtual (*virtual machine*, VM) y se instaló Flume en otra VM con CDH, los resultados obtenidos fueron los siguientes: La línea de regresión ajustada cumple con la ecuación (D.1), donde y_1 corresponde al tiempo de la variable de ingesta y x_1 el número de datos de medición.

$$y_1 = 60.38x_1 + 30.83 \tag{D.1}$$

El coeficiente de determinación R^2 y el coeficiente de correlación de Pearson R son iguales a 0,9999. Por lo tanto, las variables y_1 y x_1 son linealmente dependientes. Como resultado, es posible alcanzar alrededor de mil millones de métricas por hora con 512 GB de RAM.

D.8.2 Data Lake

Este componente debe cumplir con la escalabilidad lineal y el rendimiento, como los requisitos de arquitectura mencionados en el Apéndice A. Escalabilidad significa que debe ser posible aumentar la capacidad de almacenamiento agregando nuevos nodos al clúster. Para el rendimiento, el proceso de almacenamiento debe ser lo más rápido posible. Varias soluciones cumplen estas características; sin embargo, HDFS es una herramienta abierta, que también admite el procesamiento por lotes y es compatible con varias soluciones de implementación y es compatible con muchas herramientas de *big data* (Borthakur, 2008).

D.8.3 Procesamiento

Spark (Apache, 2020c) y Storm (Apache, 2020d) son algunas de las herramientas de procesamiento de flujo que son compatibles con el SL de la arquitectura de big data. Spark permite un análisis de datos más rápido de los datos XML; por lo tanto, se selecciona en este componente. Como resultado de esta investigación, Apache Spark procesa alrededor de 3 millones de líneas XML en 14,25 segundos.

D.8.4 Reportes

Para este componente, se deben considerar varios aspectos, primero, el tipo de dato. Dado que se usa formato XML, se deben procesar datos semiestructurados.

En segundo lugar, el *data lake* seleccionado fue HDFS; por lo tanto, la herramienta de reportes debe ser compatible con este. En tercer lugar, como mencionan los requisitos iniciales, la herramienta de reportes seleccionada debe ser compatible con SQL.

Como candidatos, se tienen HBase, Hive, Impala y Pig, que son compatibles con HDFS. Hive y Pig realizan el procesamiento por lotes. Hive (Apache, 2020b) ofrece un lenguaje de consulta basado en SQL, llamado HiveQL. Sin embargo, Hive proporciona soporte nativo para archivos XML y los beneficios mencionados en un estudio (X. Li & Zhou, 2015) en comparación con otras herramientas. Además, Hive puede utilizar el deserializador del serializador de análisis XML de IBM para crear las tablas.

El resultado obtenido en el entorno de prueba para este componente fue alcanzar alrededor de mil millones de métricas por hora con 72 GB de RAM.

D.8.5 Interfaz

Para el componente de interfaz, existen varias herramientas que permiten la visualización de *big data*, entre ellas, las más conocidas: Tableau y Power BI (Carlisle, 2018). Dado que el ambiente de pruebas utiliza aplicaciones libres, se seleccionó la versión gratuita de Power BI (Microsoft, 2021), debido a que en Tableau (Tableau, 2021) los informes realizados en la versión gratuita no son privados. Adicionalmente, Power BI es por ahora la herramienta más utilizada en comparación con Tableau (Machiraju & Gaurav, 2018). Power BI (Microsoft, 2021) es una herramienta de software que permite la creación de paneles de informes personalizables, lo que le permite tomar datos de diferentes fuentes como texto, archivos csv, bases de datos locales y de red, ya sean estructuradas y documentales. Y eso, a su vez, permite una fácil manipulación de los datos mediante el uso de su herramienta incluida llamada Power Query.

Además, cuenta con los principales tipos de objetos gráficos que se pueden utilizar de forma general dentro de un tablero, los cuales cuentan con una gran cantidad de opciones de personalización de colores, semáforos, filtros, etc.; y si necesita otros objetos visuales con propiedades adicionales, puede recurrir a Power BI AppSource, donde puede encontrar una gran cantidad de elementos visuales especializados a los que se puede acceder principalmente de forma gratuita.

Otra ventaja de Power BI es que hay un controlador Cloudera Open Database Connectivity (ODBC) que permite la conexión entre PowerBI y Hive en CDH. Una vez establecida la conexión, se cargan las tablas que se utilizarán para crear el tablero. Power BI facilita realizar las agregaciones a los datos, a través de la función DAX, permitiendo el cálculo de KPI. El Apéndice C presenta un ejemplo de un informe de archivos PM obtenidos de PowerBI.

D.8.6 Implementación

Para brindar una solución llave en mano, se utilizó la máquina CDH 5.16.0 que soporta las herramientas seleccionadas en cada componente. CDH es un entorno de código abierto con el apoyo de Flume, HDFS, Hive y Spark, entre otros. Para el uso de versiones más recientes de las herramientas, se instaló una solución basada en Hadoop, ya que las versiones 6 o superiores de CDH son pagadas.

D.9 Comparación con Otras Soluciones

Como se puede ver en la revisión de la literatura, se encontraron siete estudios diferentes sobre la implementación de un *framework* de *big data* para PM en redes móviles. Comparando los resultados de estos estudios con los obtenidos en esta tesis para determinar la complejidad computacional que se refiere al tiempo y los recursos necesarios para resolver un problema determinado, se han tomado los datos de la sección D.8.1 Ingesta, donde el coeficiente de determinación R^2 y el coeficiente de correlación de Pearson R son muy cercanos al valor 1; por lo tanto, existe una fuerte dependencia lineal entre el tiempo y el número de archivos de registro.

En la Tabla D.5 se presentan los resultados de la comparación para la capa de ingesta. Se han comparado los resultados obtenidos en esta tesis y el estudio A5 (Qiao et al., 2018). No se pudieron comparar otros estudios porque no se dispone de datos suficientes sobre los recursos y los tiempos. Según los resultados, el estudio A5 permite procesar mil millones de contadores en 228 segundos; sin embargo, la capacidad de RAM necesaria es demasiado grande, lo que resulta en la necesidad de muchos nodos para la ingesta de archivos.

ID	**Herramienta big data**	**Tiempo [segundos]**	**RAM [GB]**	**Número de nodos con 32 [GB]**
Tesis	Flume	3600	512	10
A5	Kafka	228	14926830	466463

Table D.5: Comparación de estudios para procesar mil millones de contadores en el componente de ingesta

La tabla D.6 presenta los resultados de la comparación para la capa de reportes. Se comparan los resultados obtenidos en esta tesis, con el estudio A2 (Mampaka & Sumbwanyambe, 2019), y el estudio A5 (Qiao et al., 2018). Según los resultados, el estudio A2 (Mampaka & Sumbwanyambe, 2019) permite procesar mil millones de contadores en menos de 209 segundos; sin embargo, necesita 15 veces más capacidad RAM que la nuestra. Por lo tanto, con la misma capacidad de RAM que el estudio A2 (Mampaka & Sumbwanyambe, 2019), los resultados de esta tesis son mejores.

ID	**Herramienta big data**	**Tiempo [segundos]**	**RAM [GB]**	**Número de nodos con 32 [GB]**
Tesis	Hive	3600	72	2
A2	Impala	209	1455	45
A5	Impala	7545	14926829	466463

Table D.6: Comparación de estudios para procesar mil millones de contadores en informes tardíos

D.10 Discusión

En esta tesis se descubrieron varios aspectos importantes. En primer lugar, en la revisión sistemática de la literatura sobre el modelado y manejo de *big data* en bases de datos, se identificaron algunas tendencias principales, por ejemplo: el modelo entidad-relación es el más utilizado, seguido del modelo multidimensional, y XML a nivel conceptual. El modelo más investigado es el orientado a documentos a nivel de abstracción lógica, seguido del modelo orientado a columnas y el modelo gráfico en tercer lugar. MongoDB es el sistema de administración de base de datos (*data base management systems, DBMS*) más investigado, seguido de Neo4j y Cassandra. MongoDB puede considerarse como el modelo de datos orientado a documentos más utilizado, Cassandra para el modelo de datos orientado a columnas y Neo4j para el modelo de datos gráficos. Otro hallazgo es que la metodología de modelado más propuesta se basa en datos y el 55,55 % de los estudios se centra en tipos de bases de datos homogéneas.

Entre las brechas, se puede destacar que ni el sistema NoSQL ni el modelado de lenguajes han emergido como estándares. Además, pocos estudios presentan soluciones para bases de datos híbridas, es decir, para datos estructurados y no estructurados.

En cuanto a la revisión literaria de artículos que estudian el procesamiento de PM en redes móviles con herramientas de *big data*, solo se encontraron siete artículos relacionados. Sin embargo, para hacer una comparación de los resultados obtenidos, solo dos artículos presentaron datos completos A2 (Mampaka & Sumbwanyambe, 2019) y A5 (Qiao et al., 2018).

Para este estudio, se utilizaron archivos PM de cuatro proveedores diferentes y cuatro tecnologías de telecomunicaciones 2G, 3G, 4G y 5G. Se pudo determinar que desde 3G los proveedores utilizan el formato XML definido por 3GPP TS 32.401. Por lo tanto, los sistemas se seleccionaron en base al soporte de este formato.

Además, se pudo verificar que los archivos PM utilizan un formato XML complejo; por tanto, también se propuso un método de procesamiento basado en la catalogación, deserialización y explosión posicional. Para la catalogación, los elementos del esquema XML se asignaron a raíz, matrices, estructuras, valores y atributos. Luego se implementaron la deserialización y la voladura posicional. El método propuesto se implementó y evaluó para Hive y Spark. Otro punto importante encontrado fue el menor tiempo de ejecución de consultas que se obtuvo en el escenario donde se usó Hive con tablas internas, y las versiones recientes de las herramientas reducen los tiempos de ejecución.

Una vez analizado el formato XML a procesar, la investigación determinó las herramientas adecuadas para implementar el *framework*. Los componentes seleccionados del *framework* de *big data* se implementaron en físico y en una solución en la nube. Se evaluaron los recursos necesarios para procesar hasta mil millones de registros por hora para componentes de ingesta y reportes.

D.11 Contribución

Con base en los resultados de esta investigación, la hipótesis ha sido verificada tanto en la revisión literaria como en la implementación realizada. Es decir, es factible optimizar la gestión del rendimiento de una red móvil a través de un *framework* de *big data*.

Además, se ha cumplido con todos los requisitos para las partes interesadas y la arquitectura identificados en el Apéndice A:

- Se implementó una solución basada en componentes de *big data*.
- Se logró reducir los tiempos de ingesta y ejecución de consultas con una solución llave en mano y menos recursos computacionales, especialmente para requerimientos de RAM. Esta característica de la propuesta permitirá reducir los tiempos de análisis de la causa raíz de los problemas de red.
- Las pruebas se realizaron con archivos PM de una red 5G; así, la propuesta asegura el soporte de esta nueva tecnología.
- La implementación de la solución se realizó en una solución física y en la nube, obteniendo resultados similares. Además, el uso de Hive garantiza que los operadores puedan seguir utilizando el conocido lenguaje SQL para las consultas.
- El uso de HDFS permite ofrecer una solución escalable lineal.

- Hive permite el procesamiento por lotes, mientras que Spark lo permite casi en tiempo real.

- La solución está alineada con las tendencias actuales en la investigación académica y la industria, como se evidencia en la revisión de la literatura.

En cuanto a los objetivos específicos, se presentó una revisión bibliográfica sistemática del modelado y manejo de big data en bases de datos, que permite determinar tendencias y brechas en el tema. También se presentó el estudio del estado del arte de big data y redes móviles, donde se identificaron los estudios que presentan *frameworks* para analizar datos de PM y se obtuvieron sus resultados en nuestro tema de interés ingesta y reportes.

Los componentes de una arquitectura de big data que deben tenerse en cuenta durante la implementación se identificaron como arquitectura de software, ingesta, *data lake*, procesamiento, reportes, interfaz e implementación. Se analizó cada componente y, según el tipo de datos a procesar, se seleccionaron las herramientas del *framework* de *big data.* Se propusieron métodos para procesar XML complejo en Hive y Spark basados en catalogación, deserialización y explosión posicional, ya que los datos de PM de las redes móviles tienen este formato. Finalmente, se propuso un ecosistema de big data para PM en redes móviles, y se evaluaron tiempos de ingesta y ejecución de consultas hasta llegar a mil millones de registros y se compararon con los estudios A2 (Mampaka & Sumbwanyambe, 2019) y A5 (Qiao et al., 2018) identificados en la SLR .

D.12 Trabajo Futuro

Como trabajo futuro, se planea evaluar agregaciones de datos para cálculos de KPI de los registros. Además, está planificado evaluar una arquitectura kappa que permita que los informes se presenten en tiempo cercano al real. En esta tesis, se ha presentado cómo consultar los archivos PM a través de Spark para informes de transmisión; sin embargo, pensamos que a través de un modelo basado en consultas, los tiempos de ejecución de reportes pueden mejorar notablemente.

El componente de interfaz también se considera un trabajo a realizar en el futuro ya que permite que los informes se puedan visualizar de forma amigable. Se ha considerado analizar cuáles son las herramientas de visualización más utilizadas y sus respectivas ventajas y desventajas.

Además, consideramos que este trabajo puede servir de base para varios estudios; por ejemplo, los métodos propuestos podrían extenderse a otras áreas, no solo PM en redes móviles, y las tendencias y brechas encontradas en el SLR pueden abordarse en investigaciones futuras.

Un hallazgo importante fueron las continuas mejoras en las nuevas versiones de las herramientas de *big data*; por lo tanto, mantener actualizados los métodos para estas nuevas versiones sería una excelente contribución en futuras investigaciones.

References

3rd Generation Partnership Project. (2005). *Technical Specification Group Services and System Aspects; Telecommunication Management; Performance Management (PM); Concept and Requirements 32.401 V5.5.0 .* (Available on https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1991) (cited on 3, 13, 14, 16, 29, 77, 85)

3rd Generation Partnership Project. (2021). *Generations of Mobile Systems.* (Available on https://www.3gpp.org/about-3gpp) (cited on 13, 85)

Apache. (2020a). *Apache Flume.* (Available on https://flume.apache.org/) (cited on 29, 91)

Apache. (2020b). *Apache Hive.* (Available on https://hive.apache.org/) (cited on 32, 92)

Apache. (2020c). *Apache Spark.* (Available on https://spark.apache.org/) (cited on 30, 92)

Apache. (2020d). *Apache Storm.* (Available on https://storm.apache.org/) (cited on 30, 92)

Baek, G., Ahn, K., & Kim, S. (2016). Dynamic Transform Method for Ontology DB from Semi-structured Datasets. *International Journal of Intelligent Computing Research*, *7*(3), 741-747. (cited on 11, 83)

Barton, R., & Henry, J. (2020). *Unlocking the Mystery of Machine Learning and Big Data Analytics.* (Available on https://www.ciscolive.com/global/on-demand-library.html?search=unlocking&search.event=ciscoliveus2020#/session/1573153548448001JuUB) (cited on 11, 83)

Borthakur, D. (2008). *HDFS Architecture Guide* (Vol. 53) (No. 2). (Available on https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html) (cited on 29, 92)

Carlisle, S. (2018). Software: Tableau and Microsoft Power BI. *Technology|Architecture + Design*, *2*(2), 256-259. doi: 10.1080/24751448.2018.1497381 (cited on 32, 92)

Costa, C., & Santos, Y. (2017). Big Data: State-of-the-art concepts, techniques, technologies, modeling approaches and research challenges. *International Journal* *of Computer Science*, *44*(3), 285-301. (cited on 11, 83)

Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. In *8th IEEE Conference on Visualization* (p. 235-244). doi: 10.1109/VISUAL.1997.663888 (cited on 3, 78)

Daki, H., Hannani, A. E., Aqqal, A., Haidine, A., Dahbi, A., & Ouahmane, H. (2016). Towards adopting Big Data technologies by mobile networks operators: A Moroccan case study. In *2nd International Conference on Cloud Computing Technologies and Applications* (p. 154-161). doi: 10.1109/CloudTech.2016.7847693 (cited on 19, 87)

Davoudian, A., Chen, L., & Liu, M. (2018). A Survey on NoSQL Stores. *ACM Computing Surveys*, *51*, 1-43. doi: 10.1007/978-3-319-96139-2_14 (cited on 11, 83)

Farooqi, M., Shah, M., Wahid, A., Akhunzada, A., Khan, F., Amin, N., & Ali, I. (2019). Big Data in Healthcare: A Survey. *Applications of Intelligent Technologies in Healthcare*, 143-152. doi: 10.1007/978-3-319-96139-2_14 (cited on 9, 82)

Global System for Mobile Communication. (2021). *The Mobile Economy 2020.* (Available on `https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_Global.pdf`) (cited on 3, 77)

Institute of Electrical and Electronics Engineers. (2018). *29148 ISO/IEC/IEEE Systems and software engineering. Life cycle processes. Requirements engineering.* (Available on `https://ieeexplore.ieee.org/document/8267470`) (cited on 6, 79)

Institute of Electrical and Electronics Engineers. (2020). *IEEE 1471-2000 Standard - IEEE Recommended Practice for Architectural Description for Software-Intensive Systems.* (Available on `https://standards.ieee.org/standard/1471-2000.html`) (cited on 11, 12, 83, 84)

Karamjit, K., & Rinkle, R. (2013). Modeling and querying data in NoSQL databases. In *1st IEEE International Conference on Big Data* (p. 1-7). doi: 10.1109/BigData.2013.6691765 (cited on 27)

Khalifa, S., Elshater, Y., Sundaravarathan, K., Bhat, A., Martin, P., Imam, F., & Statchuk, C. (2016). The six pillars for building big data analytics ecosystems. *ACM Computing Surveys*, *49*(2), 33:1-33:36. doi: 10.1145/2963143 (cited on 27, 90)

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. *Keele University: Keele*, *33*(2004), 1–26. (cited on 6, 17, 79, 85)

Kovačević, D., Krajnović, A., & Čičin Šain, D. (2017). Market Analysis of the Telecommunications Market – The Case of Croatia. In *Dubrovnik International Economic Meeting* (Vol. 3, p. 161-175). (cited on 3, 77)

Le, L., Sinh, D., Lin, P., & Tung, P. (2018). Applying big data, machine learning, and SDN/NFV to 5G traffic clustering, forecasting, and management. In *4th IEEE Conference on Network Softwarization and Workshops* (p. 168-176). doi: 10.1109/NETSOFT.2018.8460129 (cited on 19, 87)

Li, P., Gong, Y., & Wang, C. (2020). *Schema Extraction on Semi-structured Data.* (Available on `https://arxiv.org/abs/2012.08105`) (cited on 11, 83)

Li, X., & Zhou, W. (2015). Performance Comparison of Hive, Impala and Spark

SQL. In *7th International Conference on Intelligent Human-Machine Systems and Cybernetics* (p. 418-423). doi: 10.1109/IHMSC.2015.95 (cited on 32, 92)

Machiraju, S., & Gaurav, S. (2018). *Power BI Data Analysis and Visualization.* (De|G Press) doi: 10.1515/9781547400720 (cited on 33, 92)

Mampaka, M., & Sumbwanyambe, M. (2019). A quadri-dimensional approach for poor performance prioritization in mobile networks using Big Data. *Journal of Big Data*, *6*(1), 1-15. doi: 10.1186/s40537-019-0173-8 (cited on 19, 34, 61, 63, 87, 94, 95, 96)

Martinez-Mosquera, D., & Luján-Mora, S. (2017). Data Cleaning Technique for Security Big Data Ecosystem. In *1st International Conference on Internet of Things, Big Data and Security* (p. 380-385). doi: 10.5220/0006360603800385 (cited on 23)

Martinez-Mosquera, D., & Luján-Mora, S. (2019). Framework for Big Data integration in e-government. *DYNA*, *86*(209), 215–224. doi: 10.15446/dyna.v86n209.77902 (cited on 5, 21, 25, 37, 80, 89)

Martinez-Mosquera, D., Luján-Mora, S., Lopez, G., & Santos, L. (2017). Data cleaning technique for security logs based on Fellegi-Sunter Theory. In *6th EuroSymposium on Systems Analysis and Design* (p. 3-12). doi: 10.1007/978-3-319-66996-0_1 (cited on 23)

Martinez-Mosquera, D., Luján-Mora, S., Navarrete, R., Mayorga, T., & Vivanco, H. (2019). An approach to Big Data Modeling for Key-Value NoSQL Databases. *Iberian Journal of Information Systems and Technologies RISTI*, *E19*, 519–530. (cited on 5, 21, 26, 37, 80, 89)

Martinez-Mosquera, D., Luján-Mora, S., & Recalde, H. (2017). Conceptual modeling of big data extract processes with UML. In *1st International Conference on Information Systems and Computer Science* (p. 207-211). doi: 10.1109/INCISCOS.2017.18 (cited on 22)

Martinez-Mosquera, D., Luján-Mora, S., Reyes, R., & Paredes, M. (2019). Pillars for Big Data and Military Health Care: State of the Art. In *1st International Conference on Advances in Emerging Trends and Technologies* (p. 125-135). doi: 10.1007/978-3-030-32022-5_12 (cited on 23)

Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2020a). Development and Evaluation of a BigData Framework for Performance Management in Mobile Networks. *IEEE Access*, *8*, 226380–226396. doi: 10.1109/ACCESS.2020.3045175 (cited on 5, 12, 22, 29, 32, 37, 80, 84)

Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2020b). Modeling and Management Big Data in Databases—A Systematic Literature Review. *Sustainability*, *12*(634), 1–41. doi: 10.3390/su12020634 (cited on 5, 10, 21, 26, 27, 37, 80, 81, 88, 89)

Martinez-Mosquera, D., Navarrete, R., & Luján-Mora, S. (2021). Efficiently Processing Complex XSD using Hive and Spark. *PeerJ Computer Science*, *8*, 1-33. doi: 10.7717/peerj-cs.652 (cited on 5, 22, 30, 37, 80)

Microsoft. (2020). *Big data architecture style.* (Available on `https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data`) (cited on 12, 83)

Microsoft. (2021). *Power BI.* (Available on `https://powerbi.microsoft.com/es-es/`) (cited on 32, 33, 92)

Nokia. (2020). *Shaping the future of telecommunication.* (Available on `https://nokiawroclaw.pl/wp-content/uploads/2019/03/NOKIA_Book_2nd.pdf`) (cited on 3, 78)

Olsson, M., Sultana, S., & Mulligan, C. (2009). *SAE and the Evolved Packet Core.* (Academic Press) (cited on 12, 84)

Ounacer, S., Talhaoui, M., Ardchir, S., Daif, A., & Azouazi, M. (2017). A New Architecture for Real Time Data Stream Processing. *Journal of Advanced Computer Science and Applications*, *8*(11), 44-51. doi: 10.14569/IJACSA.2017.081106 (cited on 11, 83)

O'Sullivan, P., Thompson, G., & Clifford, A. (2014). Applying data models to big data architectures. *IBM Research and Development*, *58*(5/6), 18:1–18:11. doi: 10.1147/JRD.2014.2352474 (cited on 11, 83)

Papadakis, G. (2018). The return of JedAI: End-to-End Entity Resolution for Structured and Semi-Structured Data. In *5th VLDB Endowment* (Vol. 11, p. 1950-1953). doi: 10.14778/3229863.3236232 (cited on 11, 83)

Qiao, Y., Xing, Z., Fadlullah, Z., Yang, J., & Kato, N. (2018). Characterizing flow, application, and user behavior in mobile networks: A framework for mobile big data. *IEEE Wireless Communications*, *25*(1), 40-49. doi: 10.1109/MWC.2018.1700186 (cited on 19, 33, 34, 61, 63, 87, 93, 94, 95, 96)

Ribeiro, A., & da Silva, A. R. (2015). Data Modeling and Data Analytics: A Survey from a Big Data Perspective. *Journal of Software Engineering Applying*, *8*(12), 617-634. doi: 10.4236/jsea.2015.812058 (cited on 9, 82)

Rider, F. (1944). *The Scholar and the Future of the Research Library: A Problem and Its Solution.* (Hadham Press: New York, NY, USA) (cited on 3, 78)

Rueda, F., Vergara, D., & Reniz, D. (2018). Big data streaming analytics for QoE monitoring in mobile networks: A practical approach. In *5th IEEE International Conference on Big Data* (p. 1992-1997). doi: 10.1109/BigData.2018.8622590 (cited on 19, 87)

Skračić, K., & Bodrušić, I. (2017). A Big Data Solution for Troubleshooting Mobile Network Performance Problems. In *40th International Convention on Information and Communication Technology, Electronics and Microelectronics* (p. 472-477). doi: 10.23919/MIPRO.2017.7973471 (cited on 3, 19, 78, 87)

Suleykin, A., & Panfilov, P. (2019). Distributed big data driven framework for cellular network monitoring data. In *24th Conference of Open Innovations Association* (p. 430-436). doi: 10.23919/FRUCT.2019.8711912 (cited on 19, 87)

Tableau. (2021). *Tableau.* (Available on `https://www.tableau.com/trial/tableau-online`) (cited on 33, 92)

Tekiner, F., & Keane, J. A. (2013). Big Data Framework. In *25th IEEE International Conference on Systems, Man, and Cybernetics* (p. 1494-1499). doi: 10.1109/SMC.2013.258 (cited on 9, 82)

Turck, M. (2020). *Resilience and Vibrancy: The 2020 Data & AI Landscape.* (Available on `https://mattturck.com/data2020/`) (cited on 9, 28, 82, 90)