Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Big data analytics in telecommunications: Governance, architecture and use cases

Mohamed Zouheir Kastouni, Ayoub Ait Lahcen *

Laboratoire Sciences de l'Ingénieur, Ecole Nationale des Sciences Appliquées (ENSA), Université Ibn Tofail, Kenitra, Morocco

ARTICLE INFO

Article history:
Received 21 July 2020
Revised 6 November 2020
Accepted 9 November 2020
Available online xxxx

Keywords:
Big data analytics
Big data project's governance methodology
Big data architecture
Data governance methodology
Big data project's team
Big data telecommunications use cases

ABSTRACT

With the upsurge of data traffic due to the change in customer behavior towards the use of telecommunications services, fostered by the current global health situation (mainly due to Covid-19), the telecommunications operators have a golden opportunity to create new sources of revenues using Big Data Analytics (BDA) solutions. Looking to setting up a BDA project, we faced several challenges, notably, in terms of choice of the technical solution from the plethora of the existing tools, and the choice of the governance methodologies for governing the project and the data. The majority of research documents related to the telecommunications industry have not addressed BDA project implementation from start to finish. The purpose of this study focuses on a BDA telecommunications project, namely, Project's Governance, Architecture, Data Governance and the BDA Project's Team. The last part of this study presents useful BDA use cases, in terms of applications enabling revenue creation and cost optimization. It appears that this work will facilitate the implementation of BDA projects, and enable telecommunications operators to have a better understanding about the fundamental aspects to be focused on. It is therefore, a study that will contribute positively toward such goal.

Contents

1. Introduction . . . 00
2. History of data analytics in the telecommunications industry . . . 00
3. Big data analytics challenges and benefits for telecommunication operators . . . 00
   3.1. BDA challenges . . . 00
      3.1.1. Technological challenges . . . 00
      3.1.2. Organisational challenges . . . 00
   3.2. BDA benefits . . . 00
4. Big data analytics project: The main pillars . . . 00
   4.1. Project governance methodology . . . 00
   4.2. Architecture design and infrastructure . . . 00
   4.3. Data governance . . . 00
   4.4. Data team . . . 00
5. Big data analytics applications in the telecommunications industry . . . 00
   5.1. Consumer churn prediction . . . 00
   5.2. Offer propensity . . . 00

* Corresponding author.
E-mail address: ayoub.aitlahcen@univ-ibntofail.ac.ma (A. Ait Lahcen).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2020.11.024
1319-1578/

Please cite this article as: Mohamed Zouheir Kastouni and A. Ait Lahcen, Big data analytics in telecommunications: Governance, architecture and use cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.11.024

5.3. Revenue leakage prevention . . . 00
5.4. Customer experience improvement . . . 00
5.5. Proactive care . . . 00
5.6. Mobile location data for COVID-19 prevention . . . 00
5.7. Social Network Analysis (SNA) and multi-SIM detection . . . 00
5.8. BDA for public security . . . 00
5.9. Energy saving . . . 00
5.10. Real time road traffic analysis . . . 00
6. Conclusion and future works . . . 00
Declaration of Competing Interest . . . 00
References . . . 00

1. Introduction

The covid-19 pandemic, the quarantines and stay at home orders broke data consumption records. In fact, the usage of telecom customers has evolved drastically. Certain practices which were considered exceptional, have become common, namely, telework, e-learning, online gaming, IpTv, and streaming. The bandwidths are more and more saturated, and the data generated is growing exponentially every minute. Telecom operators became among the richest companies in terms of data volume, however, they still don't know how to capitalize on efficiently. The question is to understand how they can leverage this data to lower operating costs, deliver a personalized customer experience, reduce the churn rate, and develop new sources of revenue. The answer to this question turns out not to be so obvious.

During the last decade, several Telecom operators have initiated BDA projects but have not been able to achieve the expected results. Indeed, McKinsey has conducted a survey on 80 telecom operators that have invested on BDA platforms, less than 8% of them have realized profit exceeding 10% (Bughin, 2016b), and almost the third has realized around zero percent profit. In 2015, Gartner predicted that 60% of BDA project will fail (Gartner, 2015), mainly due to the lack of proper management, combined with lack of clear vision and skills shortage. Jacques Bughin has found that the return on investment of BDA projects can be highly and positively influenced by Bughin (2016a)): First, the choice of the architecture as it will impact the performance and the scalability of the solution. Second, the projects ownership as it must come from the highest level of the organization. Lastly, the governance model, which must cover all aspects related to project and data governance. Other researches (Otto, 2011; Zahid et al., 2019), have stated that there is, still, a lack of guidelines in terms of governing BDA projects and data, as well as the absence of reference architecture for telecom projects.

The objective of this study is to give the telecom players a framework based on the best practices, enabling them to secure the most critical aspects for the success of their BDA projects implementation, which are, project and data governance, solution architecture and the project's required competencies (see Fig. 1). To realize this, we conducted a literature review related to BDA implementation within the telecom sector. This review focuses precisely on, BDA's project and data governance methodologies, BDA's architectures, and BDA's skills requirements. Next, we describe and analyse the most popular methodologies and architectures already implemented within several telecom operators, and depict the relevant competencies required for the success of this kind of initiatives. In the last section of this work, we present a list of BDA use cases that have been implemented successfully in Telecom sector.

2. History of data analytics in the telecommunications industry

The concept of data analytics and its applications may appear new to some, however, reviewing the literature, we found that data analytics may be summarized to the use of data to help decisions-making and business actions, which, suggests that the concept is not new (Davenport, 2005). Indeed, stories of Roman leader Caesar who receives analysts' prediction that March will be a "down month," but disregards the data (Inman, 2008), or Michelangelo using an advanced abacus to estimate the amount of paint needed to cover the Sistine Chapel (Inman, 2008), are two examples that shows us the roots of analytical thinking to predict outcomes and support business decisions.

In the telecommunications industry, during the period of the first generation mobile networks 1G, the data analytics focused, mainly, on business and operational efficiencies. The data generated was related to simple data transactions, such as texting and voice calls. Device penetration which will foster the creation of data platforms to use analytics was extremely limited. Consequently analytics software developments consisted of in-house and proprietary initiatives.

In the early 90s, the second-generation 2G networks, employed digital communication over TDMA and CDMA, and brought a set of new services, such as text messages, picture messages, MMS (multimedia messages), faxing, and voicemail. The second-generation devices were designed with limited storage and processing capabilities. Combining all this, the Telecommunications providers were able to perform some data-intensive operations, such as automating frequently generated reports and dashboards (e.g. sales, revenues ...etc.). The technical solutions were built upon the old conventional data bases and data warehouses, which used distinctive modes of information accumulation, extraction and analysis technologies. The analytical capabilities were based on statistical methods from the 1970s.

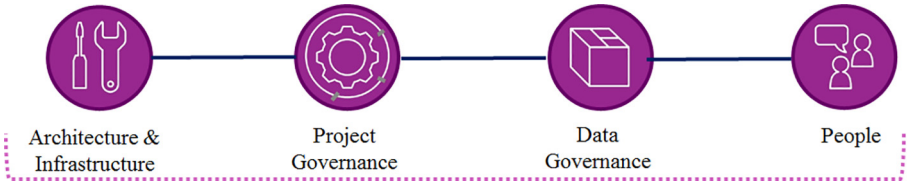

Fig. 1. Big Data Analytics Four Pillars.

Since 2000, 2G mobile devices have been gradually replaced by 3G products, the 3G network and handset were designed to have 2 Mbps speed to meet the demand of multimedia through the cellular system. During this phase, the Telecommunications operators switched from focused asset behaviour analytics to focused customer behaviour analytics. New data types have been made available, notably, graphics and videos, widening the spectrum of exploited data, and providing telecommunication operators with sophisticated analysis capabilities.

The first deployment of the 4G LTE (Long Term Evolution) network was in Stockholm, Sweden in 2009, allowing 100 Mbps download and 50 Mbps upload (Ericsson, 2009). This technology permitted 50 percent reduction in round-trip latency compared to its predecessor technology, making possible real-time applications such as high quality video calls, and online gaming (Ménard et al., 2012). The 4G network also facilitated the development of services via mobile applications, which led to a massive data generation.

The advent of the 4th generation mobile networks marked a turning point in analytics usage in the mobile ecosystem. The Telecommunications providers started producing new insights on network performance and users behaviours, which contributed to the creation of new revenues streams, improvement of customer experience, and boosting of customer retention (Ott, 2014).

Likewise, the introduction of the Hadoop ecosystem (Oussous et al., 2018) in the communication service providers' (CSPs) portfolio solutions, revolutionized the data analytics realm, as it provided new tools and technics to overcome the limitations of the old conventional data bases and data warehouses, in terms of performance, scalability and analytical tools.

Indeed, Hadoop runs on clusters, allowing the storage of a huge volume of data on thousands of servers. The resources scalability is enabled simply by increasing the number of the clusters nodes. Among the major components of the Hadoop ecosystem we quote Hadoop Distributed File System (HDFS), which is the main data storage system used by Hadoop applications. It stores large files by dividing them into blocks, and replicating them on several servers. In 2017, Hadoop 3.0 brought several HDFS enhancements enabling the support of additional NameNodes, and better data compression through the erasure coding function. The two other major components of Hadoop ecosystem are MapReduce (Mohammad et al., 2020), which is a programming model for large scale data processing, and YARN which is a framework for job scheduling and cluster resource management. With regards to the real time data collection, Apache Kafka enables the collection of data from heterogeneous sources and conveying the results to multiple consumers. These data are then processed by Apache Spark which is stream processing framework, suitable for high volume, high-reliability stream processing workloads. Finally, for faster query results, Hive or presto are appropriate engines to cater to the users' requirements. Unlike Hadoop, Relational Database Management System (RDBMS) are a structured databases in which data is stored in rows and columns and presented in tables. The operations on data are done through SQL language. This structured approach of RDB limits its capabilities in terms of:

- Data Volume: RDBMS works better when the volume of data is low to medium (Gigabytes to Terabytes). But when the data size is huge (Petabytes and more), RDBMS fails to give good results.
- Data Variety: Inability to process unstructured data.
- Throughput (the speed of data processing): The conventional databases fails to achieve a higher throughput as compared to the Apache Hadoop.
- Scalability: RDBMS enables the vertical scalability, which consist on adding resources (CPU, memory ...etc) to the same machine whereas Hadoop enables the horizontal scalability which consist on adding new machines to the cluster.
- Cost: Hadoop is a free and open source software framework. In contrary, RDBMS is a licensed software that needs to be bought to have the full use of functionalities.

Table 1 shows a capabilities comparison between the conventional solutions and Hadoop ecosystem.

3. Big data analytics challenges and benefits for telecommunication operators

Before starting any BDA initiative, it's very important to identify what would be the challenges that may hamper the project implementation, and also, the benefits that can be reaped once the solution delivered. In this section, we introduced the major challenges and benefits of big data implementation within telecom sector.

3.1. BDA challenges

Telecom operators are facing difficulties in dealing with the avalanche of data produced by connected devices, customer behaviours, social media networks, call data records, government portals, and billing information.

I. Malaka and I. Brown, based on their study related to the implementation of big data analytics in South Africa, classified the challenges into three sections (Malaka and Brown, 2015), technological, Organizational and Environmental. In this review, we will address only the first two sections which are, in my point of view, the more impactful on the BDA's implementation.

3.1.1. Technological challenges

a) **Absence of a reference architecture for telecom BDA implementation:** BDA architecture can be a brain teaser for the data architects, since it requires a multitude of integration of different data sources. In fact, the data integration represents a major challenge, because of the organizational silos' mode of operators, centred on products and services developments. Bringing this fragments of data into one single centralized platform, can be a challenging task (Malaka and Brown, 2015).

b) **Bad data quality:** According to McKinsey survey (Bughin, 2016b), conducted on 273 telecom players worldwide, the author announced, that the main reason of BDA projects failures is due to bad data quality. This can be explained by the multitude of systems and functions included within telecom operators solutions' portfolio, and by the volume of data managed.

Table 1
Hadoop Ecosystem vs Conventional RDBMS.

Capabilities	Conventional Solutions	Hadoop ecosystem
Ingestion speed	Low	High
Data Variety	Structured	Structured and Unstructured
Volume	Terabytes	Petabytes and more
Complex Queries response Time	Hours/Days	Minutes
Data Processing	Offline	Offline and Real-Time
Data Objects	Works on Relational Tables	Works on Key/Value Pair
Storage Cost	High	Low
Hardware Profile	High-End Servers	Commodity Hardware
Cost of maintenance	High	Low

c) **Performance and Storage:** The increasing demand of data traffic driven by social media, over-the-top (OTT) and mobile applications, is pushing the operators to find new ways to manage and leverage their data. Indeed, the traditional solutions based on the conventional data bases (RDBMS), has shown their limitations in terms of performance, storage and handling the different types of data, precisely, the unstructured one, which is outside the purview of RDBMS.

3.1.2. Organisational challenges

a) **Ownership and Control:** P. Russom, in his survey revealed that, the most common owner of BDA solutions is the Business Intelligence (BI) team (Russom, 2011). This is due to the fact that, the majority of organizations centralize as many BI and data warehouse (DWH) functions as possible through a single technical team, which is not the right configuration to have. In fact, according to T. Pearson and R. Wegener, BDA project is not seen as a technology initiative, they view it more as, a business program that requires technical savvy (Pearson and Wegener, 2013).
b) **Skills Shortage:** Indeed, the most challenging in BDA projects, is finding a qualified team. This can be explained by the fact that BDA is always considered as new technology compared to Business Intelligence, which most organisations have built over decades. Advanced analytics requires staff with deep knowledge in different domains, from data science to worldwide privacy laws, along with an understanding of the telecommunications business(Pearson and Wegener, 2013).

3.2. BDA benefits

In the telecom sector, BDA is a game-changer, as it gives operators the opportunity to exploit new data sets, and extract valuable information to better understand customer behaviour. Consequently, operators will offer more targeted offers, thereby improving revenues and reducing costs (Nwanga et al., 2015).

Undeniably, BDA solutions give telecom operators the means to process the different types of data, structured and unstructured, whatever the speed of generation. This data can be transformed into exploitable customer insights. Communication service providers (CSPs) can then develop, accurate customer profiles, better customer segmentation and proper customer indicators. In addition, BDA can help prevent certain cases of income leakage by enabling real-time fraud detection applications (Chen, 2016).

There are many others areas where BDA can have a great impact, such as:

- Service quality improvement (Jain et al., 2016). Operators are able to gain actionable insights into their networks in order to make them endurable, optimized, and scalable.
- Quality of experience (Rueda et al., 2018) can be improved at every touch point through high-performance services, fast feedback, and personalized offers.
- Real-Time call data records (CDRs) monitoring to detect abnormal behaviours.
- Network proactive care and anomaly detection (Parwez et al., 2017).
- Network traffic combined with real-time call drop rate analysis, to provide call routing optimization.
- Automated offers generation based on customers preferences. For example, Globe Telecom (telecommunications company in the Philippines) used big data analytics to improve effectiveness of promotions by 600% (Pearson, 2010). The typical required data for this use case were subscribers' location data, subscribers' demographic data, social media data and previous campaigns data.
- Leveraging social media and web data fused with the marketing initiatives to achieve better return on investment from marketing campaigns (Pearson, 2010).

4. Big data analytics project: The main pillars

The implementation of a BDA project requires special attention, whether on the technical side or in the governance side. Indeed, it is imperative for the success of the project to identify, on the one hand, the most suitable methodologies for the management of the project as well as the data, and on the other hand, the target technical and functional architectures to be implemented. It is also essential, to identify the skills required for each phase of the project, so that they are available when needed. Based on a literature review, and in order to cover the challenges mentioned above, we describe in this chapter, several examples of methodologies and architectures that are popular in the big data telecom field, in order to give the readers an overview of the existing solutions, and enable them to choose the ones that can best meet their needs.

4.1. Project governance methodology

As more and more communication service providers (CSPs) are developing BDA projects, project managers are facing new challenges. Big data analytics projects have much larger scope than standard software development project. There are both, data and analytics modelling technics to consider during the project implementation. The other challenging aspects of BDA projects governance stems from the enormous pressure of business decision makers to obtain quick results.

A. Tokuç et al. studied the possibility of implementing PMI methodology for the governance of BDA projects by bringing some adaptations to PMI processes (Tokuç, 2019). The study made detailed each knowledge area of the PMI methodology and provided tangibles positives results in managing BDA projects. Below, the adaptations of the PMI knowledge areas, proposed by the authors to fit the BDA projects specificities:

Project scope management: A. Tokuç et al. added four subphases: data generation, data acquisition, data storage and data analysis.
Project Schedule Management and Project Cost Management: Due to the lack of internal competencies, the authors advises to use iterative techniques consisting of giving only, an overview of the required cost and the project planning at

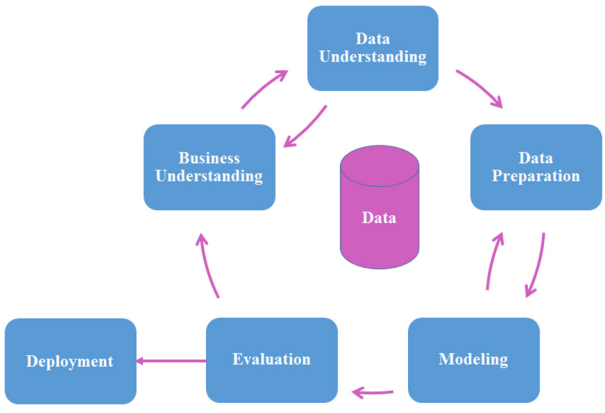

Fig. 2. CRISP-DM Methodology.

the beginning of the project. More details should be provided during the project progression.

Project Quality Management: The authors recommended having a quality expert in order to manage efficiently the data quality field.

Project Risk Management: The authors stressed the importance of mitigating the risks related to infrastructure, security and team know-how.

Another study has been led by French researchers B. Ponsard et al., who looked at the specific governance methodologies which are more adapted to Data Science projects (Ponsard et al., 2018). Based on their literature review, they identified a series of methodologies that can be applied to BDA projects. We quote, Knowledge Discovery in Data bases (Fayyad, 1996), Sample Explore Modify Model Assess (SEMMA), and Cross-Industry Standard Process for Data Mining (CRISP-DM) (Eckerson et al., 2000). Below, a brief summary of these three methodologies.

1. **Knowledge discovery in data bases (KDD)**
   KDD is defined as the nontrivial process of identifying valid, novel, potentially useful patterns within a dataset in order to make important decisions. The KDD methodology is considered as the seminal approach to data mining, that gave birth to the other ones (SEMMA, CRISP-DM). The knowledge discovery process is iterative and interactive, consisting of nine steps: *a) Developing an understanding of the application domain,* consists in understanding the goals of the end-user and the environment in which the knowledge discovery process will take place. *b) Selecting and creating a data set,* consists in identifying the data that will be used for the knowledge discovery. *c) Preprocessing and cleansing,* covers the enhancement of the data reliability by cleaning data. *d) Data transformation,* covers the preparation of better data for data mining. *e) Choosing the appropriate Data Mining task,* is related to the selection of the data mining type (classification, regression, or clustering). *f) Choosing the Data Mining algorithm,* consists in selecting the specific method to be used for searching patterns. *g) Tuning the model,* consists in tuning the model parameters until reaching the optimal results. *h) Evaluation,* consists in evaluating and interpreting the mined patterns with respect to the goals defined in the first step. *i) Consolidating discovered knowledge,* consists in incorporating the knowledge into another system for further action.
2. **SEMMA methodology**
   The third popular project governance methodology is SEMMA. It has been conceived by SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model and Assess. *a) Sample,* consists in analyzing a small portion of a large data set. *b) Explore,* consists in looking for patterns in the data with the purpose of gaining some information. *c) Modify,* consists in creating, modifying or eliminating the variables for the study. *d) Model,* consists in creating a valid model that best fits the project objectives. *e) Assess,* consists in evaluating the usefulness and reliability of the results.
3. **CRISP-DM methodology**Cross Industry Standard Process for Data Mining CRISP-DM (see Fig. 2), organizes the governance of data mining project process into six phases: Business understanding, data understanding, data preparation, modelling, evaluation, and deployment. These phases help organizations understand the project implementation process and provide a road map to follow while planning and carrying out a big data analytics project.

CRISP-DM, SEMMA and KDD share a lot of similarities. The Table 2 shows the analogy between the different phases of these three methodologies.

Table 2
Data Mining Methodologies Comparison.

KDD	SEMMA	CRISP-DM
Developing an understanding of the application domain	Not Covered	Business Understanding
Selecting and creating a data set on which discovery will be performed	Sample	Data Understanding
Pre-processing and cleansing	Explore	
Data transformation	Modify	Data Preparation
Choosing the appropriate Data Mining task	Model	Modeling
Choosing the Data Mining algorithm		
Tuning the model		
Evaluation	Assessment	Evaluation
Consolidating discovered knowledge	Not covered	Deployment

When analysing the described methodologies, the authors, B. Ponsard et al., have highlighted some limitations in terms of the absence of the coverage of the infrastructure and operations activities, and the weak communication related to the project progress. The third limitation is the lightness on tasks in the deployment phase, correlated with the absence of templates or guidelines.

With an aim of providing a more complete governance model, B. Ponsard et al. proposed a methodology that addresses the infrastructure aspects from design to installation, improve communication on the project's progress, and finally, incorporates all the agility rules, which will allow a better success rate of BDA projects. The methodology proposed is composed of three main phases: first *Context and awareness*, which is an introductory phase to the concepts of BDA and where the stakeholders' maturity level is gauged. Secondly, *Understanding the use cases* phase, which consist on identifying the requirements for which a BDA solution is envisaged. Thirdly, *Pilot implementation* phase, which encompasses the activities of data understanding, data modelling, model evaluation and solution deployment. Further to the analysis of the two previously described studies, it turns out that, both proposed methodologies by A. Tokuç et al. and B. Ponsard et al., could be improved in order to insure a more complete coverage of the different areas of BDA projects. The PMI adaptation proposal enables a quite large coverage of big data specificities, but it lacks addressing some important data mining aspects not present in standard software projects. We quote for instance, the data exploration phase which is not addressed on any knowledge area. This phase allows the project team to describe the data quality, the data completeness and the data value. Another example to mention, not properly addressed in the adapted PMI methodology, is the integration of the iterative aspect of the data mining evaluation process, which can create unpleasant surprises during the execution of the project. With regards to cost management and planning management, a big number of project managers will not do well with a rolling wave of progressive elaboration technics, as a workaround, due to the absence of expert opinion. It may results in budget overspending and non-adherence to schedule. Thus, starting the project after the whole needed experts are gathered would be a wise decision, otherwise, the risk of failure would be much significant. Concerning the Quality Management, the proposed methodology addresses only the processes enabling the high level project quality, but does not dive into the data quality management. Consequently, it will be of a great importance to set standardized framework for handling, solely, the issues related to data quality, while describing the technics and tools to provide high quality, relevant and valuable data. The Risks Management area is a critical field that should be followed very closely in order to avoid any disastrous consequences. We will cite, hereafter, two examples of risks, very frequent in the telecom sector, that need to be integrated and monitored by the

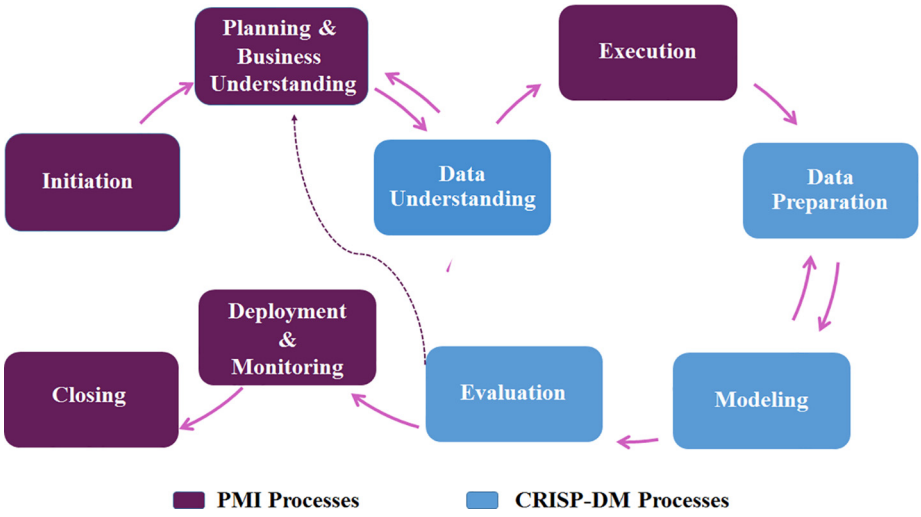

Fig. 3. Combination of PMI and CRISP-DM methodologies for BDA Projects' Governance.

project managers. The first one is the resource scarcity and volatility. The second one, is related to the rapid swelling of traffic due to special offers or changes in customer behaviors. The Procurement Management should also be tracked by the project manager, and a special attention has to be given to open source solutions, in order to lower the cost of ownership. The last remark is related to the lack of coverage of the agility aspect by the 'Adapted PMI methodology'. This needs to be addressed, to enable a more flexibility in terms of accepting business requirements changes during the project's execution.

The methodology proposed by B. Ponsard et al., was better suited to data science projects, but lacks details related to certain aspects that are greatly covered and depicted by the PMI methodology, notably, in terms of integration, planning and communication management. In addition, we didn't notice any mention related to resource management or team skills development. With regards to the infrastructure part, giving its complexity and the multitude of interactions with the other external solutions, its management processes should be further detailed.

BDA project implementation in telecommunications sector can be considered as an IT project, but it needs absolutely to consider the aspects related to data science projects execution. This is why, for managing BDA telecommunications project, we propose a fusion between the PMI methodology, which covers perfectly the different phases of software development project, and the Agile CRISP-DM that addresses quite well the data science projects' governance steps. The combination of these two methods has to be thought intelligently, in order to keep a certain flexibility during processes implementation, and avoid bringing new levels of complexity to projects (see Fig. 3). One option of the fusion of the two methodologies would consist on: first, the inclusion of the CRISP-DM's Business Understanding phase in the PMI's Planning phase, since the latter already covers all the elements related to customers needs' collection. Second, the positioning of the CRISP-DM's Data Understanding phase between the PMI's Planning and Execution phases. Indeed, before starting any realisation step, it's of major importance to explore the existing data in order to understand their content, to gauge their complexity, and to decide whether they are sufficient for reaching the project's objectives. Third, placing the remaining phases of CRISP-DM after the PMI Execution phase, with the possibility to loop to the planning phase in case of an underperforming model or an unsatisfied business requirements.Finally, the execution of the PMI Deployment and Closing phases immediately after the CRISP-DM Evaluation phase.

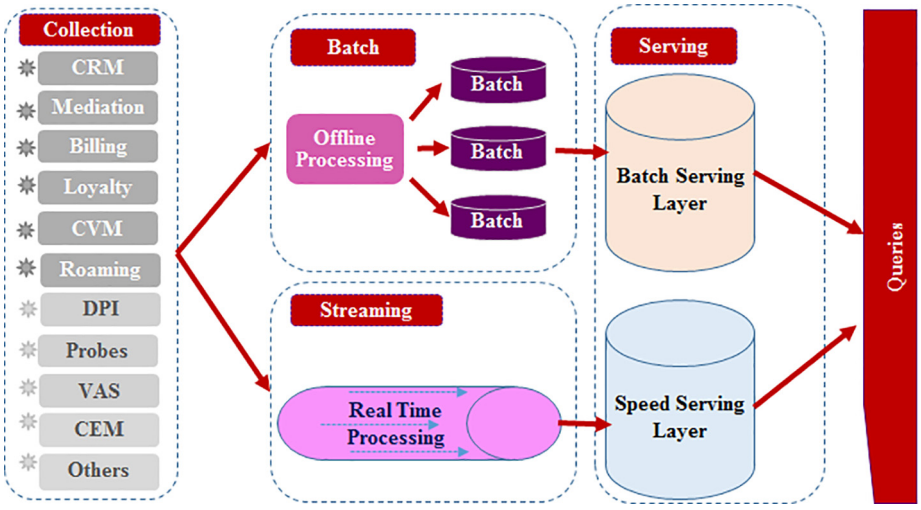

Fig. 4. Lambda Architecture.

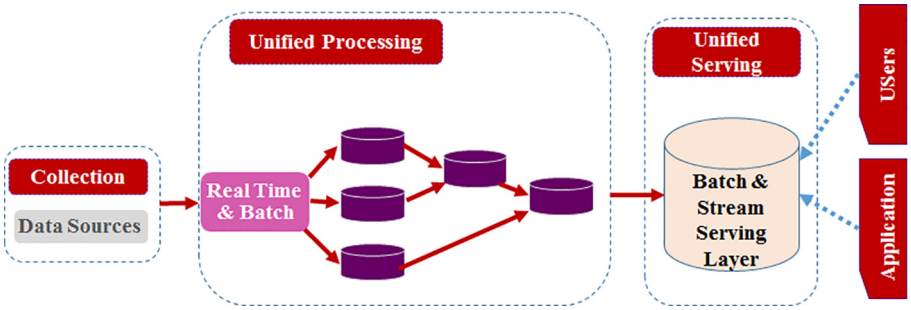

Fig. 5. Kappa Architecture.

Another option that is worth to be explored consist on building a 'Vertical' methodology encompassing the formalized projects' governance best practices in the telecommunications sector. It will certainly bring a valuable contribution to CSPs.

4.2. Architecture design and infrastructure

Telecommunications operators are sitting on a gold mine. They generate a huge volume of data, which can reach billions of CDRs/events per day. This data can be customer, usage or network data. Efficiently collecting, storing, processing, and analyzing this quantity of data can represent a real challenge for the telecommunications operators. The infrastructure needs to have high computational capabilities and storage space. It also needs flexibility to analyze different data formats. Hence, it is of great importance to design the most adequate architecture that will address these technical complexities and enable to cater to the business requirements. Two BDA architectures are widely implemented for these purposes, but in different forms: Lambda Architecture and Kappa Architecture. Lambda's first implementation was in 2011. It allowed the processing of streaming and batch data in parallel. This architecture is the most implemented within the telecommunications operators, as it enables them to fulfil the business requirements in terms of real time KPI, and also provides insights based on historical data. Lambda architecture (see Fig. 4) comprises three main layers: *a) The batch layer:* It focuses on fault tolerance. It treats data as immutable and compute views from the datasets. The duration of batch cycles depends of the volume of data processed. *b) The speed layer:* Its objective is to make up for the high latency of the batch layer, by providing an efficient way for querying most recent data. *c) The serving layer:* It merges the results from the two precedent layers and proceed to data indexation to provide low latency views and an easy users' access. Conversely, Kappa architecture, is not as popular as its predecessor in the telecom industry. It started in 2014, it was conceived in order to tackle the complexity of Lambda architecture by considering everything as streaming data (see Fig. 5). This architecture first store temporarily the data in a messaging system like Kafka, then, a stream processing engine read the data and transform it into an analyzable format; that will be stored into the analytic database.

To tackle the BDA solutions complexity, F. Su et al. proposed a big data architecture designed on five layers (Su et al., 2016): *a) Data Collection Layer:* Enabling the collection of the different types of data, from the different network elements. *b) Data Storage layer:* Based on distributed storage architecture in HDFS Cluster. *c) ETL Layer:* Enabling the extraction of the different types of data, and transforming it in a proper format before loading on the data base. *d) Data Analysis Layer:* Enable to discover knowledge and insights from data, by deploying specific models and algorithms based on the characteristics of the data. The analysis cluster adopted in the study was based on Hadoop/Spark and MPP data base. The idea is to enable batch processing through Spark and fast querying based on the data warehouse of MPP data base. *e) Data Visualization/application layer:* It enables the user to visualize the data that have been prepared by the four previous layers. The technologies applied were GIS and Echarts.

Other studies have been carried out in order to propose a reference architecture for BDA projects in telecommunications sector. A significant literature review done by H. Zahid et al. in which the authors have proposed an architecture called LambdaTel (Lambda refers to Lambda architecture (see Fig. 4), and TEL to Telecommunications), already implemented in a telecom solution called Darbi, enabling both, batch and streaming data processing to be executed simultaneously (Zahid et al., 2019). The architecture is composed of seven layers:

1. **Connection Layer:** This layer is responsible for the collection of the different data sources within the telecom operator. The data collection is done, either through the implementation of an application programming interface (API), or through connectors for data bases (SQL and NoSQL), IoT feeds, and other telecom data.
2. **Integration Layer:** This layer is responsible for integrating the data collected, and inserting the outcomes into a data lake. The logic proposed is to store each data source in its relevant data base (e.g. Social Network Data in Neo4J, CDRs in MangoDB) and then coordinating between the different stores using API Controller. The authors recommends using MongoDB, because the document data base is a powerful way to store and retrieve data and allows developers to move fast.
3. **Batch Layer:** This layer is responsible for batch processing of telecom data received from the master data base. The authors propose to differentiate between tasks based on the criteria of speed requirement. The ones that needs to run faster will be processed via Spark, and the others via Map Reduce.
4. **Streaming Layer:** This layer is responsible for processing real-time streams. The authors recommend using Apache Kafka for data's ingestion, and Apache Spark Streaming feature for data's processing.
5. **Serving Layer:** This layer prepares processed data (Batch and Streaming) for displaying the end-user dashboards.
6. **Interface Layer:** This layer combines all the precedent layers (back-end layers) with the Dashboard layer (front-end layer).
7. **Dashboard Layer:** This layer involves displaying dashboards to be seen by various telecom end-users. It comes on the top of the architecture, and it's connected to the serving Layer through standard connectors.

Reviewing the two described architectures, we noticed that they all share a core part of collection, integration and processing

layers. However, the F. Su et al. proposed a convergent architecture including a data warehouse (DWH) module, which would be appealing to the Telecom operators. This will have several positive financial and technical impacts. On one hand, there will be no additional needs in terms of investing on the acquisition of new DWH Platform. On the other hand, the technical performances would be enhanced, due to the elimination of the network latencies that can take place in case the DWH and Data Lake were located in two separated machines. Nevertheless, the architecture lacks a real time stream process, which will require adding of new components in the collection and processing modules, and also, the boosting of the hardware resources.

The second architecture proposed by H. Zahid et al. (LambdaTEL) is using the Lambda architecture model in order to insure streaming and batch processing modes. Certainly, this type of architecture can cover most of the telecommunication operators' requirements, but still present areas that can be improved in order to optimize the operators' data, without going through enormous investments or complicated conceptions. The idea is to design an architecture that will combine the best from Lambda and Kappa architectures (Feick et al., 2018), while being efficient and cost effective.

The inconvenient of Lambda architecture is its high cost, due to the maintenance of batch and streaming APIs. The question that may cross the mind is, would it be possible to use only one processing mode (Streaming or Batch) instead of two in order to process all data types (real-time and historical), and guarantee the same or better results? One option could be to remove the real time processing from Lambda and to keep the batch mode only to be used for streaming too. The second option could be to use Kappa capacities for both streaming and batches. In this paper we will focus only on the second option.

Indeed, the use of the streaming mode for the processing of historical data will not be so obvious, for several reasons. On one hand, there will be a need to extend Kafka storage capacity, which can be very expensive compared to Hdfs storage. On the other hand, beyond a certain size of storage scaling, it will become impossible to add more capacity, due to the bijection between partition and broker (single partition for single broker). A workaround would be to expand the storage on tierce platform (e.g. Hdfs or Hive), and decompose the data set into many small pieces, before injection into Kafka. This scenario will require an important amount of effort and more resources in order to recreate the original order of data (during the retrieval from Hdfs/Hive to Kafka injection), which can be, also, very costly.

In 2019, Uber came up with a proposition to tackle the above depicted limitations, and proposed a new 'boosted' version of Kappa architecture (see Fig. 5), named Kappa+ (Naik, 2019), enabling the processing of offline data, directly from the data warehouse, while using the streaming API only. This was built upon the following assumptions:

- Data in warehouse has to be partitioned by time.
- Job classification has to be done (e.g. Stateless, Aggregation . . .etc.)
- Differentiation between processing models: 'Partially ordered reads' (Only one partition will be processed at the time), 'Emit watermark at end of partition', and 'Lockstep progression' in case of multi-sources (The processing of all the sources have to be finished in one partition, in order to move to the following one e.g. 'joins queries').

The benefits of the Kappa+ architecture are:

- Cost effective solution. Only one API to maintain.
- Optimal usage of resources. Only one reserved partition to process the historical data.
- One single Job can process all the partitions, unlike batch mode where we have to do the split into a several smaller jobs and coordinate them.
- Results starts showing right after the end of the partition's processing. We don't need to wait until the end of processing of all the data.
- Fast recovering when there is failure during a partition processing. Only the partition having the problem would be reprocessed and not all the batch.

Therefore, due to the high similarities between Uber data and Telecom operators' data, it would be appropriate to implement this new architecture within Telecom operators, and evaluate the outcomes in terms of processing performance and cost effectiveness.

4.3. Data governance

In the literature, data governance refers to the policies and procedures adopted in order to define how data assets could be accessed and manipulated and by whom in the organization. A Data Governance Framework (DGF) is defined as 'a set of processes that ensures that important data assets are formally managed at the enterprise Level' (Isson and Harriott, 2012). However, data quality is typically determined by the capability of data to meet business requirements (Olson, 2003). A common mistake that telecom operators makes is that they confuse Data Quality management, which comprises activities for the improvement of data quality, and Data Governance. This can be explained by the close connection between Data Governance and Data Quality management results from a data perspective as a company asset.

Data governance is very essential to protect the data and assets of telecom operators. Knowing how quickly the production and use of data is developing in the telecommunications sector, a special consideration needs to be allocated to data governance.

Mark Newman (Newman, 2019) has conducted a survey of people working in data analytics roles within CSPs about the challenges encountered during governing and leveraging data. The main reasons identified by the participants were:

- **Lack of a consistent data model:** Half of the respondents to the study have raised the inconsistency of data models as a main impediment for leveraging data. This is explained by the fact that each CSPs' owned solutions use its own specific data model. This configuration requires a lot of efforts to make the data reusable in other solutions.
- **Weak progress on leveraging data:** This point doesn't mean that the managers are not aware of the benefits that can be achieved through harnessing the data, but rather, they are unaware of the small progress realized on this field. Hence the importance of measuring the gain obtained when undertaking actions to leverage data.
- **Talents scarcity:** Lack of talented resources is another key challenge. The CSPs are aware of the importance of having internal expertise, but finding them remain a difficult task.
- **Bad data quality:** Having a clean, correct and complete data will have a great effect on the AI and ML outcomes. Boasman-Patel TM Forum's VP of AI and Customer experience said that during his experience with CSPs, the data quality is ranked the number one challenge for the operators.
- **Data governance:** There are two main aspects to shed the lights on: first, the compliance with the regulations, and second, the rules defined to ensure the correct approach to collect, process and distribute data.

The strategy proposed by Mark Newman to tackle these challenges is based on seven axis:

- **Create a vision:** Having a vision is the basis for a successful data governance program. The top management has to be fully involved in the vision definition and has to take the ownership.
- **Set a governance::** The author proposes to set a framework for data governance that will contribute to the improvement of customer centricity and operational efficiency. It can also play the role of an innovation catalyst.
- **Ask for others' contribution:** Learning from the experiences of partners and suppliers can be very helpful to define the right data model that can be applied for the best use of telecom operators' data.
- **Avoid silo organization:** Concentrating all the analytics expertise on a single team will surely slow the deployment of analytics initiatives according to the author. The analytics skills should be distributed across the organization, and the central team should be very small and responsible for the architecture and the data governance.
- **Track the journey:** For a successful deployment of the vision, it is required to set milestones in order to measure the progress realized. Failures and difficulties should be studied in order to be avoided in the future.
- **Consider the edge:** CSPs has to consider the usage of edge computing as a major axis on the data analytics strategy and not a secondary one. Hiring the required skills for it has to be also a priority.
- **Move to the cloud:** The telecom operators should think seriously about the cloud option as it can have several positives impacts in terms of flexibility, scalability and efficiency. In July 2019, AT&T has invested $2 billion in order to migrate non-network applications to the Azure cloud.

In the Telecommunications industry, working with mastered and high-quality data is widely seen as a competitive advantage. Boris Otto has decomposed the Data Governance within telecommunications sector into three subsets (Otto, 2011): *a) Organizational Goals:* Includes formal and functional goals. *b) Organizational Form:* Includes all the necessary activities to attain the organization's business objectives. *c) Organizational Transformation:* Includes transformation processes and organizational change.

Boris Otto has led a comparative study (Otto, 2011) about the implementation of the above-mentioned data governance framework within two big players in the Telecommunications industry, namely British Telecommunications (BT Group) and Deutsche Telekom. The author emphasized on the particularity of data governance in the Telecommunications industry in comparison with other industries, which is characterized by the volume of data generated due to the intensive consumer interactions, as well as the regulatory requirements in terms of data security and privacy. British Telecom started its data governance initiative in 1997, where they initiated an Information Management (IM) program sponsored by CIO group who was a business function at that time. The aim of this program was to identify opportunities to better leverage investments in information systems in BT Wholesales. It emerged that data quality was the priority number one to tackle. In 1998, a one-year data quality software license has been purchased. The licence fees have been recovered within the first three months post going live. In 1999, British Telecom gives rise to the IM Forum, which covers the management of data quality projects: Identifying opportunities for data quality projects, planning and budgeting data quality activities/processes, and insuring the alignment with BT's overall business goals. In 2000, the project team developed a Data Quality Methodology, composed of five phases:

1. **Problem and Opportunity Identification Phase:** It aims at identifying data quality problems hindering the business objectives.
2. **Diagnosis Phase:** It aims at assessing the data quality level through data discovery and data profiling.
3. **Proposals Phase:** It aims at delivering a commercial proposition, and ensuring the ownership of the project by the business teams.
4. **Reengineering or realization phase:** It covers the design and the implementation of the solution.
5. **Consolidation Phase:** It insured a durable solution.

Regarding Deutsche Telekom, they decided to set up organizational units in 2006, with the sole responsibility of addressing the issue of data quality management. In April 2007, Deutsche Telecom established two departments for data quality management. The first department was within the business function and was in charge of consolidating the business requirements related to data. The second department was within the central IT department and was in charge of putting forward concepts for data quality management, such as defining standards for Data Governance, developing guidelines and rules to ensure high data quality. At the same period, a sustainable Data Quality project has been initiated to set a framework defining the responsibilities related to data activities. The conclusion of Boris Otto analysis was that, the two studied telecom operators were able to estimate the business gain resulting from the implementation of data governance initiatives. The author added that organizational design of data governance at British Telecom seems to be more favorable to generate profits for the company.

In 2016, Tele Management Forum (TM Forum), which is the organization that works for setting the standards for the Telecommunications industry, has proposed a framework for implementing data governance (Wray, 2016), based on six steps:

1. **Define and align:** This first step consist on defining a data governance strategy that will perfectly address the business problems through a well-defined objectives and approaches.
2. **Roles and responsibilities:** This second step consist on, first, identifying the different data types and assigning them to their owners. Second, defining the role and responsibilities of each owner in terms of monitoring and controlling the data.
3. **Policy and process:** First, this step covers the development of the various processes required for data governance. Secondly, the assignment of these processes to their owners.
4. **Measure and monitor:** This step consist on defining the measurement methods and metrics.
5. **Select technology:** This step consist on identifying the tools enabling the application of the data governance framework.
6. **Close the loop:** This last step consist on checking the contribution of the application of data governance framework to the business objectives.

During our literature review, we noticed that the subject of data governance has not been sufficiently explored by the researchers' community. Boris Otto's research has shown that each operator has its own understanding of Data Governance, with a major confusion between Data Governance and Data Quality. The latter is the most popular within telecom operators, given its direct impact on the analytical projects. TM Forum framework came to fill this vacuum, by converging best practices of data management methods and processes within the telecommunication sector, contributing thus, to business objectives' achievement.

4.4. Data team

People are the most important asset of any big data analytics project, and the quality of the team members is crucial for the project success.

With talent scarcity related to big data analytics profiles, companies are struggling to hire the right employees with the right skills. Some companies turned to internal recruitment to form big data analytics team. However in the majority of cases, the existing employees lacked of the necessary skills (Brown et al., 2017). Therefore, an initial analysis of the business requirements have to take place in order to identify the key skills required for the realization. Indeed, each analytical project, may require specific set of skills, that may not be significant to other ones (Hammerström, 2018). Once the competencies identified, the recruitment (internally or externally) can be then initiated. The BDA project should start, only if the entire team is formed.

The big data organization part has barely been addressed in the literature. Also, there was no agreement with respect to a standard team members profiles required for the implementation of big data Telco projects. However Lennart Hammerström proposed a list of competencies that, according to him, needs to be present on each Big Data projects (Hammerström, 2018): *Data analytics expert* is responsible for applying statistical models in order to detect patterns in the data. *Data scientist* is responsible for retrieving data from data sources and customizing it to meet the data analytics expert requirements. *Systems architect* is responsible for designing the optimal solution that will respond to business requirements. *Data software expert* is responsible for creating and establishing the computer system to carry out the Big Data operations. *Operations expert* is responsible for the management of workflow throughout all the involved departments. *Data project manager* is a key role for the success of the project. He has to be specialized in big data, must have deep knowledge in cognitive and behavioural sciences, and have the ability to manage highly educated people. The proposal of Lennart Hammerström can be supplemented by the presence of additional competencies which could highly improve the success rate of BDA projects. Those players are the Data Governance Manager, the Change Manager, and also the Project Sponsor who plays a pivotal role in this kind of project. The recent researches and publications related to BDA projects' organizations, proposes data teams that mostly exhibits the following competencies:

- Wide and deep knowhow on all technical and functional aspects related to big data analytics projects.
- Strong experience in BDA implementation.
- Flexible team with high capability to address fluctuations in demand due to the agile mode of management.
- Capability to address unforeseen results due to the uncertain aspects of the project.

The key roles, commonly shared by these different publications, and which should be part of the project team are: *a) Data engineer,* his role is to develop algorithms to help make raw data more useful to the enterprise. *b) Data scientist,* is responsible for organizing and analyzing the data in order to provide stakeholders with findings to make informed business decisions. *c) System engineer,* is responsible for the installation, management and monitoring of the systems and infrastructure. *d) Data and system architect,* his role is to design the optimal solution that will respond to business requirements. *e) Project manager,* is responsible for ensuring the completion of the project on time and according to budget. *f) Project sponsor,* is the one who promotes the project, and hold the overall responsibility for the project's success. *g) Data governance Manager,* he is in charge of the implementation of the company's data management goals, standards and processes. Knowing the complexity of the business in the telecommunication sector, it's mandatory to have the right competent people who master the telecommunication business in order to ensure the long term success of the project. The identified profiles are:

Business experts: They are the ones who know the company's products and services, and also, master the different business processes.
Data owners' representatives: They are members of different business departments who are accountable for the different types of data generated by the telecom operator.

Now that we have identified the profiles for the realization of BDA projects, we must remember, that they will not be necessarily required for any kind of data analytics projects. A workshop must be held during the project initiation phase in order to identify the required profiles. Their intervention will be determined by project manager and will be timely bound.

5. Big data analytics applications in the telecommunications industry

After having listed, to the best of our knowledge, all the fundamentals elements enabling the implementation of BDA project, from the start to the deployment, we continue in this article, presenting few examples of BDA implementation projects that brought financial and operational gain to their telecom operators.

5.1. Consumer churn prediction

The churn prediction is among the most popular BDA use cases developed in the Telecommunications field. This is due to the high cost of acquisition of new customer in comparison to the cost of retaining an existing one. Several research papers have treated the subject under different aspects.

The first use case we quote (Xia et al., 2018) was built using the Hidden Markov Model (HMM). The training set was composed of five months of historical data that included customer data, billing data and network data. The application was built on Hadoop Platform of a telecom operator. Hive was used for operations on statistical data. For the sake of comparison, the project team has developed the churn application using two other popular classifiers: Random Forest and LIBLINEAR. The results have shown a shy difference in favour of the Hidden Markov Model with an Area under the Curve (AUC) equal to 0.855.

The second example of Churn prevention is little more exotic, as it uses modelling technics from the health sector which have been applied to the Telecommunications sector (Kurt et al., 2019). The aim of the authors was to emphasize on the importance of having multidisciplinary skills in Data Science. The authors found similarities between churn and survival. Thus the survival vector machine algorithm was selected for predicting telecom churners. The resulting application performed quite well with an AUC value equal to 0.82.

Another interesting churn case to quote, is the case of SyriaTel (Ahmad et al., 2019), where the authors used Extreme Gradient Boosting 'XGBOOST' model to build the churn prevention application. In addition to the classical data usually used for churn prediction (CRM, Billing and Network KPI), the team used the customer social network data to get a better performance. The authors experimented three others algorithms: Decision Tree, Random Forest and Gradient Boosted Machine Tree 'GBM'. However, the best results were obtained by the application of XGBOOST algorithm, which attained an AUC of 0.933.

A last example to quote has been developed in Pakistan (Khan et al., 2019) using Deep Learning technics, specially Artificial Neural Network. The model used multiple attributes like demographic data from CRM, billing information and usage patterns from Telecom Company. An AUC of about 0.79 has been obtained.

5.2. Offer propensity

The use of usage traffic, loyalty points, event-based promotion, and demographics data, combined with analytics technics makes it possible to design targeted offers or services to fit customers' requirements. A telecom operator in the Asia-Pacific region (Fox, 2015) needed to monitor its marketing performance, so he rolled out a predictive analytics solution that develops propensity models to track customer preferences and identify business opportunities. The results were excellent, as the operator realized 10 percent higher net revenues by improving productivity and competitiveness, and also, increased the speed of its ad hoc reporting by 190x.

5.3. Revenue leakage prevention

During the past few years, telecom operators have been suffering from different types of revenues leakage. SIM box fraud was among the most popular ones. The losses for telecommunications and government was huge. It was estimated to be averaging 150 million US dollars every year in Africa. SIM boxing is a practice in Telecommunications where some people set-up an equipment that can handle several SIM cards and use it to terminate international calls that have been received through voice over IP. Chung-Min Chen evoked two approaches for SIM Box detection: Proactive test calls and passive CDR analysis (Chen, 2016). The first method requires the operator to make calls from foreign countries to its own country, and check their termination type (local on international). This method is too heavy to implement, because, it requires the operator to ensure full coverage in order to detect all possible fraudsters. On the other hand, traffic analysis will cost much lesser and insure more satisfying results. For that reasons, some rules have been identified in order to detect suspicious SIM cards based on the number of calls generated and the area of calls' generation. Kahsu Hagos has developed a solution based on CDRs analysis using three classifiers: Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM) for the sake of comparison (Hagos and University, 2018). The first model was showing better performance compared to the two others, with an accuracy of 95.98%. Another similar study (Elmi et al., 2013) based on Artificial Neural Network (ANN) model allowed a better accuracy of 98.71%. The 9 features selected were subscriber identity, total calls, total number called, total minutes, total night calls, total numbers called at night, total minutes at night, number of calls per day, called numbers divided by total calls ratio and average minutes.

5.4. Customer experience improvement

Big Data capabilities and analytical tools has revolutionized the way the telecom operators manages the relation with their customers. As an example, some operators developed services to alert their customers whenever they experienced network problem. This resulted in decreasing calls received by the call centres.

Other initiatives aiming at improving customer's experience were built using chatbots. The results were seen on different levels, notably, Opex optimization, Net Promoter Score (NPS) improvement, and better employees' satisfaction, since they have been reassigned to more motivating activities.

E. Diaz-Avile et al. have led a research about the implementation of real-time customer experience prediction for major telecommunication operator in Africa (Diaz-Aviles et al., 2015). The authors affirmed that determining the type of user experience (good or bad) at every moment, without interacting directly with the final user, is not an obvious task. Consequently, the principle on which the authors have modelled the solution was that, any poor customer experience results in calls to the telco's care center.

E. Diaz-Avile et al. worked on predicting the customers' calls toward the call centers. The results of the latter were combined with the other data feeds and used as inputs variables to the final customer experience model prediction. For the purpose of facilitating the application results' interpretation, the authors opted for an ensemble of decision trees that they named Restricted Random Forest (RRF). The model took the result of each decision tree and combined it with the results of the other ones. The final prediction represented simply, the most chosen class (call or not call the call center) from the different trees. The solution has proven to be successful by predicting accurately problems occurring on some services provided by the telecom operator. Consequently, the latter was able to anticipate calls toward call centers, by communicating proactively to customers, the current problems that may affect their service usage.

5.5. Proactive care

Several Telecommunications providers have developed big data solutions to detect network problems proactively, even before impacting the end customers. These solutions enable, inter alia, to provide the Telecom operator with the appropriate recommendations, allowing them to intervene in advance, and to avoid possible impacts on revenues and customer experience.

5.6. Mobile location data for COVID-19 prevention

During the COVID-19 pandemic period, governments of several countries (Doffman, 2020) in Asia, Europe and the United States of America have developed solutions based on mobile users anonymized location data, to track their movements, in order to control and limit the spread of the Corona virus. These solutions enabled them, on one hand, to track people infected by the Covid-19 by retracing their travel routes, places visited, as well as people met, and on the other hand, highlighted the areas where people didn't respect public health safety measures.

5.7. Social Network Analysis (SNA) and multi-SIM detection

With a concern to continuous revenues' improvements, some telecom operators have built analytical solutions, based on the Social Network principle, allowing them to map the relationships between the different users of the network. This helped them, to identify the most influential customers, and to target them with tailor-made offers. For the case developed by N. R. Al-Molhem et al., the implementation of SNA to target influencers has resulted in 30% increase of mobile traffic compared to traditional ways (Al-Molhem et al., 2019). On the same study, the authors have developed an application aiming to detect customers switching between multiple SIM Cards. The model was built based on the principal of 'mutual friends'. Two scores have been calculated to identify the pairs of SIMs that may represent the same user. The first one was similarity score based on the Jaccard and Cosine measures. The second one was the behavioural score based on CDRs analysis. The application achieved a 92% of accuracy for customers belonging to the same operator.

5.8. BDA for public security

Some governments uses BDA combined with cloud services and IoT technologies, in order to improve their citizens' security. This is done by predicting the areas with the high probability of crimes during each period of time. This have been made possible by leveraging telecommunications data combined with public security data.

The 'intelligent police' stared receiving insights enabling clue extraction, case analysis, crime types, crime localisation, crime early detection, suspects tracking, etc. Regarding the tools used, initially W. LIANG et al. have modelled the solution using Xgboost algorithm (Liang et al., 2018). The result was not satisfying as the accuracy obtained was about 48%. Then the authors decided to explore other angles of realization and tried the Self Excitation Point Process Model (SEPP) provided by PredPol. The accuracy has significantly improved reaching then 92%. The solution has been deployed and used in Public Security Bureau of Beijing in the Republic of China.

5.9. Energy saving

The Data centers' (DC) energy saving is an important field addressed by data scientist these last years. The best example is the case of J. Gao who implemented AI technics for the efficient use of energy in Google Data Centers (GAO, 2014). In china, W. LIANG et al. tackled the data centers power-saving solutions by the usage of deep learning (DL) technics, which has proven to be effective (Liang et al., 2018). A five layer DL model was implemented for predicting energy consumptions trend. The model variables were about 19 features, including total server IT load, resource usage, environmental index of DC (temperature and humidity)...etc. The logic behind the design of the energy saving application is based on the following steps:

1. Monitoring the load of each virtual machine (VM) running in every physical device.
2. Identifying the VMs running a fewer jobs and with low load.
3. Moving the VMs with fewer jobs from their physical device to another one that has redundant resources, in order to ensure service continuity.
4. Turning into idle the former physical device, allowing hence, the reduction of the energy consumption.

5.10. Real time road traffic analysis

The impact of telecommunications' data has not remained confined inside the telecommunication industry. Another use case leveraging telecom BDA has been developed by C. Costa, D. Zeinalipour-Yazti, G. Chatzimilioudis and M.F. Mokbel (Costa et al., 2017), aiming at the improvement of the travellers' quality of life through a real-time road traffic solution. Indeed, in order to deliver a flawless service quality, telecom operators must insure a wide radio coverage of all the areas of presence of their customers and prospects. The network's equipment used for this purpose combined with the IP backbone infrastructure, are generating an impressive volume of data, which contribute to the efficient real-time road traffic management. Unlike the crowdsourcing approach, the authors proposed an innovative approach based on micro-level traffic modelling, integrating also, the temporal dimension of the analysed traffic. This allowed traffic monitoring during different time slots. Customers' data privacy and solution cost efficiency were among the basic principles for the solution design. The Traffic Telco Big Data (Traffic-TBD) solution was built on a three layers architecture: data layer, processing layer and application layer. The role of the first layer consist on aggregating the data collected from the different sources, namely, telecom data, geospatial data and social data. The technology used for this purpose is HDFS combined with Apache Hive. Regarding the catalogue management, the authors opted for the use of an RDBMS. The second layer, is leveraging Apache Spark for online data processing. Its role is to map the incoming traffic with the road network, while insuring the data privacy. The last layer provides several functions enabling the users to have insights on: congested areas, optimized routes, travels durations and travellers' speed.

6. Conclusion and future works

BDA technologies have revolutionized the Telecommunications era. Hadoop ecosystem, streaming analytics tools, and machine learning tools, opened up new opportunities for telecom operators to gain insights from data sets that were not harnessed before. However, to reap the benefit of BDA solutions, it is mandatory to define, first, the most adapted methodologies for governing the project and the data. Second, select the architecture that will address all the specificities and requirements of telecom sector. This includes, processing of both batch and streaming data, and the ability to provide real time insights. At last, guarantee the data security and privacy.

In this document, and through a literature review, we highlighted, in one hand, the major challenges encountered during the implementation of BDA projects, as well as the scarcity of documentation related to BDA governance methodologies. On the other hand, we stressed the absence of standards in terms of architecture design as well as project organization.

To solve this problem, we have broken down the governance methodologies described in the literature, in order to keep only, those that are most popular and have proven their effectiveness, to describe and analyse them in this paper. Similarly for the architecture design, we described different types of the most implemented BDA architectures, which are mostly around lambda architecture, and we proposed implementing the new Kappa+ architecture used by Uber in order to benefit from its numerous advantages, notably in terms of, resource effectiveness, cost effectiveness, and the ease of maintenance. In the last section of this paper, we presented several use cases, to help telecom players built applications enabling business processes' improvement, effective network management, and revenue growth. Nonetheless, several aspects that can have a significant impact on the implementation of BDA projects are not yet, for the best of my knowledge, addressed by Telecom associations and the main players in the industry. We cite as an example the setting up of vertical methodologies for project and data governance, or even, the design of a BDA reference architecture, specific to the needs of the telecom sector.

As a future work, we intend to implement two proof of concepts based respectively on Lambda architecture and Kappa+ architecture, and develop the same use cases on both configurations in order to evaluate and compare the results obtained.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmad, A.K., Jafar, A., Aljoumaa, K., 2019. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data 6, 28. https://doi.org/10.1186/s40537-019-0191-6.

Al-Molhem, N.R., Rahal, Y., Dakkak, M., 2019. Social network analysis in Telecom data. Journal of Big Data 6, 99. https://doi.org/10.1186/s40537-019-0264-6.
Brown, T.J., Suter, T.A., Churchill, G.A., 2017. Basic Marketing Research. Cengage Learning.
Bughin, J., 2016a. Reaping the benefits of big data in telecom. Journal of Big Data 3, 14. https://doi.org/10.1186/s40537-016-0048-1.
Bughin, J., 2016b. Telcos: The untape promise of big data, p. 2.
Chen, C.M., 2016. Use cases and challenges in telecom big data analytics. APSIPA Transactions on Signal and Information Processing 5,. https://doi.org/10.1017/ATSIP.2016.20 e19.
Costa, C., Chatzimilioudis, G., Zeinalipour-Yazti, D., Mokbel, M.F., 2017. Towards real-time road traffic analytics using telco big data, in: Proceedings of the International Workshop on Real-Time Business Intelligence and Analytics, ACM, Munich Germany. pp. 1–5. DOI: 10.1145/3129292.3129296.
Davenport, 2005. Competing on Analytics.
Diaz-Aviles, E., Pinelli, F., Lynch, K., Nabi, Z., Gkoufas, Y., Bouillet, E., Calabrese, F., Coughlan, E., Holland, P., Salzwedel, J., 2015. Towards Real-time Customer Experience Prediction for Telecommunication Operators. ArXiv: 1508.02884 version: 2.
Doffman, Z., 2020. COVID-19 Phone Location Tracking: Yes, It's Happening Now-Here's What You Should Know. Library Catalog: www.forbes.com Section: Innovation.
Eckerson, W.W., Hanlon, N., Barquin, R., 2000. The crisp-dm model:the new blueprint for data mining. Journal of Data Warehousing 5, 15.
Elmi, A.H., Ibrahim, S., Sallehuddin, R., 2013. Detecting SIM box fraud using neural network, in: Kim, K.J., Chung, K.Y. (Eds.), IT Convergence and Security 2012. Springer, Netherlands, Dordrecht. vol. 215, pp. 575–582. DOI: 10.1007/978-94-007-5860-5_69. series Title: Lecture Notes in Electrical Engineering.
Ericsson, 2009. World's first 4G/LTE network goes live today in Stockholm. Last Modified: 2017-06-22T13:59:20+00:00 Library Catalog: www.ericsson.com.
Fayyad, U., 1996. From data mining to knowledge discovery in databases. AI Magazine 17 (3), 18.
Feick, M., Kleer, N., Kohn, M., 2018. Fundamentals of Real-Time Data Processing Architectures Lambda and Kappa. Lecture Notes in Informatics (LNI). Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, p. 12.
Fox, B., 2015. Keeping telecom on target how csps tap the transformative power of data and analytics. IBM Institute for Business Value 24.
GAO, 2014. Machine learning applications for data center optimization, p. 13.
Gartner, 2015. Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics. Library Catalog: www.gartner.com.
Hagos, K., University, A.A., 2018. SIM-Box Fraud Detection Using Data Mining Techniques: The Case of ethio telecom. Master's thesis. In: School of Electrical and Computer Engineering. Addis Ababa Institute of Technology.
Hammerström, L., 2018. Organizational design of big data and analytics teams. European Journal of Social Science Education and Research 5, 132–149. https://doi.org/10.2478/ejser-2018-0065.
Inman, 2008. Top 10 moments in the history of business analytics. Library Catalog: blogs.sas.com.
Isson, J.P., Harriott, J., 2012. Win with Advanced Business Analytics: Creating Business Value from Your Data. John Wiley & Sons. Google-Books-ID: nlrQbSiIcssC.
Jain, S., Khandelwal, M., Katkar, A., Nygate, J., 2016. Applying big data technologies to manage QoS in an SDN, in: 2016 12th International Conference on Network and Service Management (CNSM), IEEE, Montreal, QC, Canada. pp. 302–306. DOI: 10.1109/CNSM.2016.7818437.
Khan, Y., Shafiq, S., Naeem, A., Ahmed, S., Safwan, N., Hussain, S., 2019. Customers churn prediction using artificial neural networks (ANN) in telecom industry. International Journal of Advanced Computer Science and Applications 10. 10.14569/IJACSA.2019.0100918.
Kurt, K.K., Atay, H.T., Cicek, M.A., Karaca, S.B., Turkeli, S., 2019. The importance of multidisciplinary in data science: application of the method in health sector to telecommunication sector. In: 2019 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE, Malatya, Turkey, pp. 1–4. https://doi.org/10.1109/IDAP.2019.8875917.
Liang, W., Sun, M., He, B., Yang, M., Liu, X., Zhang, B., Wang, Y., 2018. New technology brings new opportunity for telecommunication carriers. International Telecommunication Union: ICT Discoveries, p. 7.
Malaka, I., Brown, I., 2015. Challenges to the organisational adoption of big data analytics: a case study in the south african telecommunications industry. In: Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists. ACM Press, Stellenbosch, South Africa, pp. 1–9. https://doi.org/10.1145/2815782.2815793.
Ménard, A., Travasoni, A., Begonha, D.B., Gropp, M., 2012. Seizing the 4G opportunity, 10.
Mohammad et al., Joshua Zhexue Huang, S.S.T.Z.E.K.S., 2020. A survey of data partitioning and sampling methods to support big data analysis. IEEE/ Big Data Mining and Analytics (Volume: 3, Issue: 2, June 2020), 85-10110.26599/BDMA.2019.9020015.
Naik, R., 2019. [Uber Seattle] Introduction to Kappa+ Architecture using Apache Flink.
Newman, M., 2019. How to Leverage Data Analytics. TM Forum.
Nwanga, M.E., Onwuka, E.N., Aibinu, A.M., Ubadike, O.C., 2015. Impact of big data analytics to nigerian mobile phone industry. In: 2015 International Conference on Industrial Engineering and Operations Management (IEOM). IEEE, Dubai, pp. 1–6. https://doi.org/10.1109/IEOM.2015.7093810.
Olson, J.E., 2003. Data Quality: The Accuracy Dimension. Elsevier. Google-Books-ID: x8ahL57VOtcC.
Ott, I., 2014. Monetising customer insights.
Otto, B., 2011. Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers. Communications of the Association for Information Systems 29. 10.17705/1CAIS.02903.
Ahmed Oussous, Fatima-Zahra Benjelloun, A.A.L.S.B., 2018. Big data technologies: A survey. Journal of King Saud University Computer and Information Sciences, 1810.1016/j.jksuci.2017.06.001.
Parwez, M.S., Rawat, D.B., Garuba, M., 2017. Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network. IEEE Transactions on Industrial Informatics 13, 2058–2065. https://doi.org/10.1109/TII.2017.2650206.
Pearson, Lesser, 2010. A new way of working, insights from global leaders. IBM Institute for Business Value.
Pearson, T., Wegener, R., 2013. Big Data: The organizational challenge. Big Data 8.
Ponsard, C., Touzani, M., Majchrowski, A., 2018. Synthèse des méthodes de conduite de projets Big Data et des retours collectés lors de pilotes industriels. Ingénierie des systèmes d'information 23, 9–33. https://doi.org/10.3166/isi.23.1.9-33.
Rueda, D.F., Vergara, D., Reniz, D., 2018. Big Data Streaming Analytics for QoE Monitoring in Mobile Networks: A Practical Approach, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, Seattle, WA, USA. pp. 1992–1997. 10.1109/BigData.2018.8622590.
Russom, P., 2011. Big Data Analytics. Big Data Analytics 38.
Su, F., Peng, Y., Mao, X., Cheng, X., Chen, W., 2016. The research of big data architecture on telecom industry. In: 2016 16th International Symposium on Communications and Information Technologies (ISCIT). IEEE, Qingdao, China, pp. 280–284. https://doi.org/10.1109/ISCIT.2016.7751636.
Tokuç, Uran, T., 2019. Management of Big Data Projects:PMI Approach for Success. ResearchGate. 10.4018/978-1-5225-7865-9.
Wray, S., 2016. The 6-step roadmap for good data governance. Library Catalog: inform.tmforum.org Section: AI & Data Analytics.
Xia, X., Zeng, L., Yu, R., 2018. HMM of telecommunication big data for consumer churn prediction. In: 2018 IEEE SmartWorld. IEEE, Guangzhou, China, pp. 1903–1910. https://doi.org/10.1109/SmartWorld.2018.00319.
Zahid, H., Mahmood, T., Morshed, A., Sellis, T., 2019. Big data analytics in telecommunications: literature review and architecture recommendations. IEEE/CAA Journal of Automatica Sinica, 1–22. https://doi.org/10.1109/JAS.2019.1911795.