Big Data On Telco

Drilona Fatusha
University of Tirana
drilona.fatusha@gmail.com

Prof.Asoc.Dr. Ana Ktona
University of Tirana
ana.ktona@fshn.edu.al

Abstract

As businesses in every area face intense competition, advanced analytics will help improving their profitability and gain a competitive advantage by enhancing customer experience. Traditionally the analysis process has been done in off-line mode, by using Data Warehouse Technologies combined with BI tools. That is not enough anymore. Today, big data is becoming a business imperative. The benefits of big data have been deeply analyzed in many articles and reports during the past years. And it is evident that such increased value is something tangible in every area. Just to mention a few, this includes the energy sector, the financial services, the telecommunication, the transport, healthcare and education, etc. In this article, we intend to understand better what Big Data Technology is all about, the benefits that they bring to the society, focusing in particular to the telecom industry. The experimental environment is set up. Installation and configuration of the platform for data management and analytics will allow us to access all the sample data possessed by a telecom operator in a single platform. Big Data technology will be used to a telecom operator to find out pattern and reasons of call drop in real time and send customers' apology text message and also refund money for dropped calls resulting in improved customer satisfaction and brand value.

1. Introduction

The benefits of big data have been deeply analyzed in many articles and reports during the past years. And it is evident that such increased value is something tangible in every area. Just to mention a few, this includes the energy sector, the financial services, the telecommunication, the transport, healthcare and education, etc. The emerging challenge for an organization is to derive meaningful insights from available data and re-apply it intelligently. Knowledge management plays a crucial role in efficiently managing this data from all sources and synthesizing it along with relevant enterprise data, to derive meaningful information and intelligence, converting it into useful knowledge base, storing it and delivering it to the end users. Every minute 300 million emails are sent, 270K photos are uploaded, and 1.8 million likes are generated on Facebook. Based on the surveys performed among telecom companies, utilizing big data brings most value to operators in the areas of customer retention and segmentation and network optimization. The telecom companies need to have a full understanding of the customer's attitudes and behaviors. This would improve the customer retention and acquisition, would increase the revenue and would reduce costs.

As telecom operators[1] face high network and spectrum costs and intense competition, advanced analytics will help improving their profitability and gain a competitive advantage by enhancing customer experience and optimizing network usage. New sources of data that manufacturers are starting to mine vary from customer's feedback on social networks to sensors data that record the actual product usage. In order to produce better products, obtain comprehensive data about customers and gain holistic insights, manufacturers, distributors, retailers and other service providers must combine synergies to integrate their respective data sets to each other. Figure 1 explains the mains procedure of Big Data Process and Data Mining:

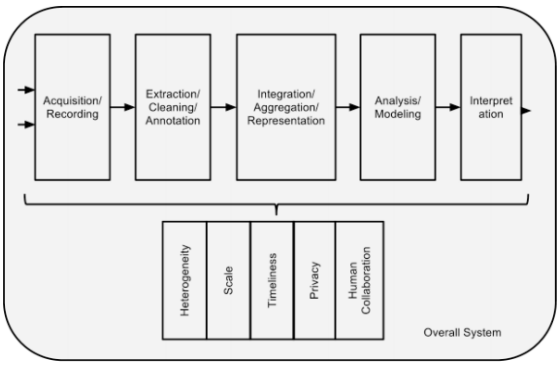

Figure 1: Big Data Process Flow

[1] Opportunities in Telecom Sector: Arising from Big Data - School of Business School of Data Science, School of Communication, Deloitte, November 2015.

2. What is Big Data and what are its features?

"Big Data" is a term which refers to the use of techniques to capture, process, analyze and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies" [CuSoDa15].

Big data tools and technologies offer opportunities and challenges in being able to analyze data efficiently to better understand customer preferences, gain a competitive advantage in the marketplace, and grow the value of the business. Data management architectures have evolved from the traditional data warehousing model to more complex architectures that address more requirements, such as real-time and batch processing; structured and unstructured data; high-velocity transactions; and so on.
Big Data technology has is based on the below main aspects:

- Infrastructure
- Data Storage
- Data Analytics Processing

2.1.1 Infrastructure

The key to big data infrastructure is scalability and flexibility to handle petabytes of data, so the cloud becomes a natural choice. Key public cloud providers include: Amazon Web Services, Gogrid and Rockspace.

2.1.2 Data Storage

Traditional, legacy systems and methods of storage are suboptimal due to price and scalability restrictions. They include relational databases, data marts, and data warehouse. The loading and storing of data include of ETL (Extract, Transform and Load). Thus data is cleansed and organized before loading them. While with Big Data, a Magnetic, Agile and Deep (MAD) approach is being used. This means that all data, no matter how clean and organized they are need to be captured and stored. This is called the Magnetic feature of the Big Data Storage process. Furthermore, given the growing numbers of data sources big data storage should allow analysts to easily produce and adapt data rapidly. This requires an agile database, whose logical and physical contents can adapt in sync with rapid data evolution. Finally, since current data analyses use complex statistical methods, a big data repository also needs to be deep, and serve as a sophisticated algorithmic runtime engine. New methods of storage, particularly NoSQL and DFS (Distributed File System) represent the paradigm shift in the storage arena. Amongst these, Hadoop is the most commonly used storage for Big Data. Key Big Data storage players include: mongoDB, Hadoop, Clustrix and Netezza.

2.1.3 Data Analytics Processing

This is the area which provides visualization and predictive analytics. There are four critical requirements for big data processing. The first requirement is fast data loading. Considering the high volume of data, and the network traffic that may interfere with the query executions during data loading, it is necessary to reduce the data loading time. The second requirement is fast query processing. Given the heavy workloads and real-time requests, the response time becomes a critical factor. The third requirement for big data processing is the highly efficient utilization of storage space. Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing. Finally, the fourth requirement is the strong adaptivity to highly dynamic workload patterns. Key data analytics providers include: Splunk, Clickfox, Rainstor, MapR, Cloudera, Hadoop, Greenplum, and Progress.

2.2 The V-s of Big Data

The Big Data concept is intrinsically related to the predictive analytics. The later deals with the process of extracting information from data and using it to predict future trends and behavior patterns. What stands in the foundations of the predictive analytics is capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting it to predict future outcomes [An15].

2.2.1 Volume

This dimension refers to the quantity of data, as big data is frequently defined in terms of massive data sets with measures such as petabytes and zetta-bytes commonly referenced. And these vast amounts of data are generated every second. Today big data is generated by machines, networks and human interaction on systems like social media, and the volume of data to be analyzed is massive.

2.2.2 Variety

Variety refers to the increasingly diversified sources and types of data requiring management and analysis. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. Imagine how different an EKG signal is from a tweet on social media. So it becomes emergent to integrate all such complex and

multiple data types: structured, semi-structured and unstructured.

2.2.3 Velocity

This feature refers to the increasing speed at which big data is created and the increasing speed at which the data needs to be stored and analyzed. The data sources can be the business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. This has been further accelerated thanks to the increase of connected devices (which will soon become the largest source of data) and the worldwide success of the social platforms.

The flow of data is massive and continuous. Sampling data can help deal with issues like volume and velocity.

2.2.4 Veracity

This dimension refers to the noise and abnormality in data being generated. Is the data that is being stored and mined relevant to the problem being analyzed? Does it consider the context? Given the increasing volume of data being generated at an unprecedented rate there is an urge to manage the uncertainty associated with particular types of data.

Besides these 4 traditional V's, there are two additional dimensions that are keys to operationalizing big data, and they are **validity** and **volatility**.

2.2.5 Validity

Like big data veracity, validity means the correct and accurate data for the intended use. The validity of big data sources and subsequent analysis must be accurate, if you are to use the results for decision making.

2.2.6 Volatility

Big data volatility refers to how long the data is valid and how long it should be stored. In this world of real-time data, you need to determine at what point the data is no longer relevant to the current analysis.

3. Big Data in Telecom.

Big data is about the use of existing data, integration of new sources of data, and the use of new tools in a more timely way to increase efficiency or to enable new business models. Today, big data is becoming a business imperative because it enables organizations to accomplish several objectives:

- apply analytics beyond the traditional analytics use cases to support real-time decisions, anytime and anywhere
- deep dive into information that can be used in data-driven decision making
- optimize all types of decisions, whether they are made by individuals or are embedded in automated systems by using insights that are based on analytics
- provide insights from all perspectives, from historic reporting to real-time analysis, to predictive modeling

The rapid growth of emerging economies and the importance of information technology throughout a product lifecycle has led to the customization of the products. The success or the failure of the company depends on the customer. Additionally, in order to have a tailored product portfolio [Ma08] to cater to different markets, organizations need to have a clear understanding of customer requirements in different segments and must design products which meet these expectations. And the process of knowledge exchange is bidirectional: it is also the service provider or the manufacturing company that by analyzing the customer behavior, his location etc. should be able to understand his needs and have it 'time to market' ready. This is valid especially in the Telecom Industry.

Furthermore, the spread of mobile-phone technology worldwide is one of the most significant changes that has affected our society. And this is valid not only in the industrialized countries but also in the developing economies. While in the industrialized countries the mobile network has been used for sophisticated purposes (like online shopping, book flights, taxi-s or vacations), there are countries in the world, including Albania, where mobile technology has been used as a substitute for usually weak telecommunication and transport infrastructure as well as underdeveloped financial and banking systems. For example, popular mobile services such as Cell Bazaar in Bangladesh allow customers to buy and sell products, SoukTel in the Middle East offers an SMS-based job-matching service, and the M-PESA mobile-banking service in Albania allows individuals to make payments to banks, or to individuals.

Big data has become a universal part of telecom industry because of the massive amount of data being generated every minute through a connected world. The explosion of smart devices and the always more powerful networks (FTTH, 3G, 4G, LTE, Wi-Fi etc.) has enabled the telecom operators to have access to information about their customers' behavior, preferences, movement, etc. Not only human-to-human communication but also human-to-machine and machine-to-machine (M2M) communication generate huge amount of data which could be helpful for all industries including telecom. During the last years, the M2M communication is expected to surpass human generated data in the near future. Most operators conduct analytics programs that enable them to use their internal data to boost the efficiency of their networks, segment customers, and drive profitability with some success. The spread of smart devices, diversity of networks (FTTH, 3G, 4G, Wi-Fi...) and the multitude of new services (mobile advertising, e-government, m-payments, streaming music, location services, etc.) create a wealth of information about

consumers' usage and purchase behavior. All of which generates very large volumes of data - in the billions of call records monthly, driving the Big Data initiatives that so many companies are currently undertaking. And with these efforts the value of location is beginning to become ever more apparent. Each of these call records generate X and Y coordinates (longitude and latitude) which are the markers that when associated to socio-economic or demographics data.

Telecom companies are among the pioneers for Big Data adoption. This sector is among those where data explosion - driven by data intensive applications such as call data records, network traffic monitoring and digital content, - is most likely to trigger action in terms of organizational change and product/service offering diversity. Current adoption of Big Data technologies in telecom neared 40% in October 2012, with close to 40% of respondents planning to adopt them within the next three years.

Big data promises to promote growth of efficiency and profitability across the entire telecom value chain, including the below areas:

- Using insights into customer behavior and usage to develop new products and services
- Optimizing the quality of service by analyzing network traffic in real time
- Analyzing call data records in real time to identify fraudulent behavior immediately
- Allowing call center reps to flexibly and profitably modify subscriber calling plans immediately
- Tailoring marketing campaigns to individual customers using location-based and social networking technologies
- Improve customer retention by creating ad hoc products and pricing almost for each customer (or per customer segment). This is called dynamic profiling and customer segmentation

All the above, if done appropriately, would immediately increase average revenue per user (ARPU), can improve the customer experience within the telecom industry, reduce churn, reduce revenue loss etc.

4. Hadoop Architecture and Map-Reduce paradigm

Hadoop[2] [Hadoop] runs MapReduce tasks over Big Data. It provides also Hadoop Distributed File System (HDFS) for supporting file-oriented, distributed data management operations efficiently. It has been highlighted that Hadoop is of type MAD (Magnetism, Agility, and Depth) system meaning that:

- it is able to capture all data sources
- it is able to adapt its engines to changes that may occur in big data sources
- it is able of supporting depth analytics over big data sources much more beyond the possibilities of traditional SQL-based analysis tools.

In these terms, Hadoop can be considered as the evolution of next-generation Data Warehousing systems, with particular regards to the ETL phase of such systems. MapReduce is the core of Hadoop. MapReduce is a programming model with the associated computational framework that is inspired to the primitives Map and Reduce of functional languages. More specifically, Map splits computational tasks into smaller computational tasks (this involves in the split of the target data domain as well) and assigns to then appropriate {Key, Value} pairs. These smaller computational tasks are executed very efficiently, even by exploiting parallelism. Thus, unstructured data, such as text, can be mapped to a structured key/value pair, where, for example, the key could be the word in the text and the value is the number of occurrences of the word. This output is then the input to the "Reduce" function. Reduce then performs the collection and combination of this output, by combining all values which share the same key value, to provide the final result of the computational task.

The MapReduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results. The MapReduce job starts by the JobTracker assigning a portion of an input file on the HDFS to a map task, running on a node. On the other hand, the Task Tracker nodes actually run the jobs and communicate results back to the Job Tracker. That communication between nodes is often through files and directories in HDFS, so inter-node communication is minimized.

MapReduce nodes and the HDFS work together. At the very first step, there are all various types of data including log files, sensors etc. The Hadoop File System stores a replica of all such data (in blue, beige, yellow and pink) across the Data Nodes. In the second step, the client executes the map job, and reduce a job on a particular data set, and sends both of them to the Job Tracker. The later, distributes the jobs across the Job Tracker in Step 3. The task tracker runs the mapper, and the mapper produces the output that is then stored in HDFS. In the last step, step 4, the reduce job runs across the mapped data in order to put together the results. As previously explained, Hadoop is a real MAD system. This because the data are stored first in HDFS. Map-reduce interprets the data at the processing

[2] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

time and not in the loading time. While the magnetism and agility is provided by the fact that data are stored as files in HDFS.

Cloudera[3] delivers the modern platform for data management and analytics. It offers fastest, easiest and most secure data platform built on Apache Hadoop. Cloudera allows to access to all your data in your possession in a single platform. With Cloudera you can efficiently capture, store, process and analyze vast amounts of data in order to solve your most challenging business problems quickly and securely, at a low cost.

Below you can find some print screens of the environments set up, which has been create by using VBox, inside of it two nodes of Hadoop have been created. Additionally also Cloudera has been installed inside the same server node.

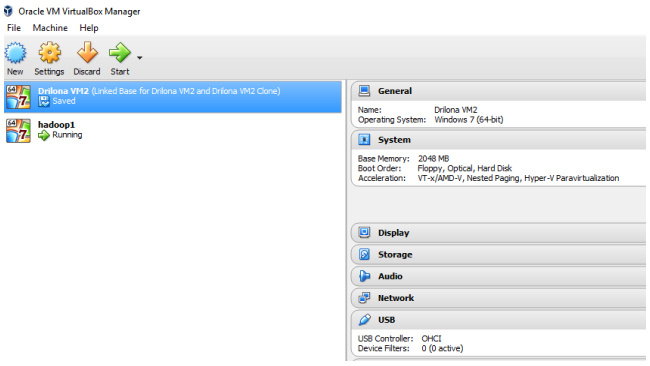

Figure 2. VirtualBox with two server nodes

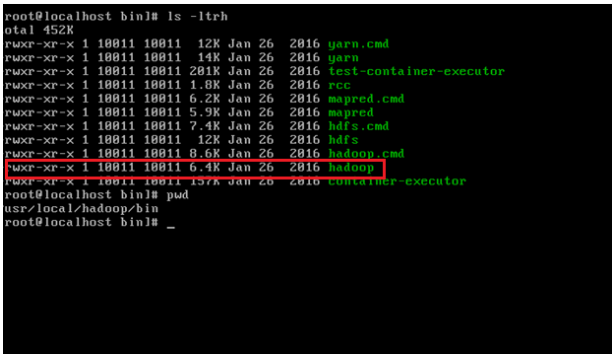

Figure 3: Hadoop

[3] https://www.cloudera.com/content/dam/cloudera/Resources/PDF/solution-briefs/Industry-Brief-Big-Data-Use-Cases-for-Telcos.pdf

5. Conclusion

Big data has become a universal part of telecom industry because of the massive amount of data being generated every minute though connected world. Big data promises to promote growth of efficiency and profitability across the entire telecom value chain, including the below areas:

- Using insights into customer behavior and usage to develop new products and services
- Optimizing the quality of service by analyzing network traffic in real time
- Analyzing call data records in real time to identify fraudulent behavior immediately
- Allowing call center reps to flexibly and profitably modify subscriber calling plans immediately
- Tailoring marketing campaigns to individual customers using location-based and social networking technologies

As a result, the average revenue per user (ARPU) would immediately increase, the customer experience within the telecom industry can be improved, churn can be reduced, revenue loss can be reduced etc. The experimental environment is set up by using VBox. Cloudera delivers the modern platform for data management and analytics. It offers fastest, easiest and most secure data platform built on Apache Hadoop. Cloudera allows us to access to all sample data in possession by telco operator in a single platform. Analytics programs will be used to enable telco operator to use its internal data to boost the efficiency of its networks, segment customers, and drive profitability with some success.

References

[CuSoDa15]Analytics over Large-Scale Multidimensional Data: The Big Data Revolution, Alfredo Cuzzocrea, Il Yeol Song, Karen C. Davis.

[An15] Abrief introduction on Big Data 5 Vs characteristics and Hadoop Technology – ICCC2015, Ishwarappa Anuradha

[Ma08] Behavioral Segmentation of Telecomunication Customers, Emilia Mattila, Master of Science Thesis Stockholm, Sweden 2008