DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

Ecole Nationale Polytechnique

Electronic Department
Laboratory of Communication and Photovoltaic Conversion

In partial fulfillment of the requirement for master's Degree

---

Big Data , Trends and challenges

Youcef BOUKHELKHAL

Supervised by :

PhD. Mourad ADNANE

Presented in public on : 12 th October 2017

Jury members:

President	Dr. R. SADOUN	Professor	ENP
Examiner	Mr. M. BELOUCHRANI	Professor	ENP
Supervisors	Mr. M. ADNANE	PhD	ENP

ENP 2017

DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

Ecole Nationale Polytechnique

Electronic Department
Laboratory of Communication and Photovoltaic Conversion

In partial fulfillment of the requirement for master's Degree

---

Big Data , Trends and Challenges

Youcef BOUKHELKHAL

Supervised by :

PhD. Mourad ADNANE

Presented in public on : 12 th October 2017

Jury members:

President	Dr. R. SADOUN	Professor	ENP
Examiner	Mr. M. BELOUCHRANI	Professor	ENP
Supervisors	Mr. M. ADNANE	PhD	ENP

ENP 2017

Dedication :

This work is for everyone that supported me, beginning with my mother and my father: who took care of me since my childhood till this day.

This work is for my little sisters: may they accomplish much more than this.

I would like to dedicate this work to my grand-parents, and all of my family. also to my cousin Oussama, who I consider as my only brother.

And finally, I would like to thank all my friends whom I met at every stage of my education, and every stage of my life.

Sincerely, Youcef

Acknowledgment:

I would like to thank Dr Mourad ADNANE for his availability, for his valuable help, his availability throughout this thesis, his precious advices during the thesis, we learned a great a deal from him, not to forget his good humor and kindness.

BOUKHELKHAL Youcef
ALGIERS - 2017

Abstract :

العربية :

أصبحت البيانات الضخمة مجالا مثيرا للاهتمام، وتغطي مجموعة كاملة من التطبيقات في مختلف المجالات، نظرا لمزاياها، والقيمة التي يمكن أن تضيفها للعالم

الهدف من هاذا المشروع هو اشتكشاف أسرارالبيانات الضخمة ، وفوائدها على العالم، والأهم من ذلك التحديات التي تواجهها مع بعض الحلول الموجودة، وأخيرا التهديدات الأكثر شيوعا.

الكلمات المفتاحية : البيانات ، البيانات الضخمة ، تحليلات البيانات، الحوسبة السحابية، إنترنت الأشياء.

Resume :

Big Data est devenue un domaine intéressant, il couvrent toute une gamme d'applications dans divers domaines, en raison de ses avantages et de la valeur qu'elle peut ajouter.

Ce projet consiste à explorer les mégadonnées (Big Data), ses avantages pour le monde , et plus important encore, les défis qu'il rencontre avec certaines solutions existantes, et enfin ses menaces les plus courantes.

Les mots clés : données, Big Data, L'analyse des données, le cloud computing, l'Internet des objets.

Abstract:

Big data became an interesting field, and it covers a whole spectrum of applications in various domains, due to its advantages, and the value that it can add.

This project consists of exploring the big data, its benefits to the world, and more importantly the challenges that is facing with some existed solution, and finally its most common threats.

Key words: Data, Big data, Data analytics, cloud computing, Internet of Things.

Contents

List of figures

List of figures

Chapter 1 : INTRODUCTION:

1.1 Data history:

data starts with one progression , historically data was being generated and accumulated by workers , in other words by employees of companies were entering data into the computer systems , but then things evolved to the internet and now users could generate their own data , for example websites . This level is larger than the first by orders of magnitude , now that we are talking about scalability , so here it scaled up from just employees entering the data, to users entering their own data , so all of sudden the amount of data being accumulated was way higher than it was historically , and now there is even a third level in this progression, because now machines and systems are generating data, for examples the buildings in the cities are full of monitors, and smart meters that being entering the data, the plains and satellites taking pictures and accumulating data , and that is orders of magnitude is even way higher than users[1].

1.2 data explosion:

now the data generated on the earth is growing exponentially for many reasons some of them are:

- retailer databases which are recording customers activities.
- organizations working on logistics, financial services, health care and other services are also storing more data.
- websites and public social media is creating of digital materials.
- after recognition technics has been improved computer now are extracting any kind of information from images and videos ...
- because smart objects are increasing, big data is being generated by expanding internet of things.
- and finally, scientific researches are generating a vast quantity of data.

the graph (Figure1) shows the world data growth, data is growing at a 40% compound annual rate, reaching nearly 45 ZB by 2020

$(1ZB(ZettaByte) = 10^9 TB(TeraByte))$

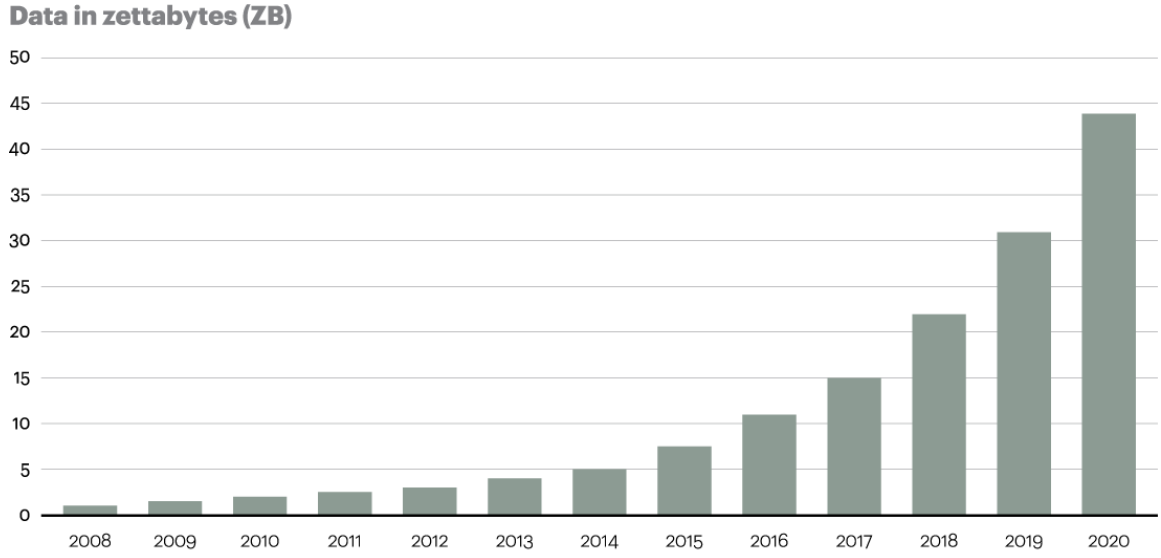

Figure 1:Data growth

So, because of the colossal amounts of data being generated, that requires new technologies to store and process it, and here where the term big data appeared[1].

1.3 Big Data definition:

Big data is an abstract concept, basically it refers to huge volume of data that cannot be stored and processed, using the traditional approach in a given speed, and big data doesn't mean any data that is in order of giga byte or tera byte, even a small amount of data can be referred as big data, depending on the context being used.

As an example, let say an email service cannot handle an attachment of 100 MB in the emails, in this scenario we say 100 MB is big data.

Another example, 10 Tera bytes of images needs to be resized, and using a traditional system will not be sufficient to complete this task in given time, so this data can be referred as big data[2].

1.4 why big data is useful:

Google processes data of hundreds of PB, and Facebook generates log data of over 10 Petabyte (PB) per month, Baidu; a Chinese company, processes data of tens of PB and

Taobao; a subsidiary of Alibaba, generates data of tens of Terabyte (TB) on online trading per day, Now the question may accrue, if the big data is complex to store and process, why these companies are investing a lot of money in this field and trying to gather data as much as possible?

The answer is that because they found that data is incredibly valuable, analyzing all data and finding pattern in it allows them to extract powerful information which can bring more benefits,

Big data is a disruptive force that will affect organizations across industries, sectors and economies. And almost every department within a company will undergo adjustments to allow big data to inform and reveal. Data analysis will change, becoming part of a business process instead of a distinct function performed only by trained specialists. Big data productivity will come because of giving users across the organization the power to work with diverse data sets through self-service tools.

to make the image clearer here is some examples:

when a user goes to a website, like Facebook, google, amazon … they provide him with some advertisement, recommendations and products chosen specifically for this user. To achieve that, these kind of web sites they make sure to store and analyses every detailed data properly that the user generates (like searching, the mouse clicks, even the time spent on looking at a web page ….), so according to these data they can estimate his preferences.

this is another example from the real life : there is this time when the hurricane sandy was about to hit on New Jersey in the United States , so the retail biggest company in United states called Walmart used big data to profit from it , so they studied the purchase patterns do different customers when hurricane is about to strike or any kind of other calamity on a particular area , and when they did an analysis of it they found that people tend to buy emergency stuff like flash light , lifejackets …, and interestingly people also buy a lot of strawberry pop-tarts , so they used this information and make benefits of it . and they didn't do any theory studies, but only analyzing the stored data they came with this information[2].

1.5 Conclusion:

Once companies begin leveraging big data for insight, the action they take based on that insight has the potential to revamp business as it is known today, marketing department can gain immediate feedback on a new branding campaign by analyzing blog comments and social-networking conversations.

New companies that understand the value of big data will not only challenge existing competitors, but may also begin defining the way business is done in their industries. Customer relationships will experience transformation as companies attempt to quickly understand concepts, that previously could not be captured.

Achieving the vast potential of big data calls for a thoughtful, holistic approach to data management, analysis and information intelligence. Across industries, organizations that get ahead of big data, will create new operational efficiencies, new revenue streams, differentiated competitive advantage and entirely new business models. Business leaders should begin thinking strategically about how to prepare their organizations for big data and big opportunities.

Chapter 2 :
Big Data challenges:

2.1 Introduction:

in this chapter we talk about the difficulties and the challenges that the big data has been facing, and the solutions that has been developed to handle those obstacles, and also its interaction with different fields. The big data is criticized by the 5Vs which will be more explored in the next lines[3]:

2.2 Big Data 5Vs:

Volume:

starting with the first V which is the volume data, the challenge posed by data volume is most noticeable, since the total volume of the stored data is increasing exponentially, in different fields:

- In science, such as biology, meteorology, astronomy, etc…., scientists encounter computing limitation constantly due to the increasing data volume.
- On the Web, applications such as Google and Facebook are dealing with the numbers of customers that have never been considered by local applications. The sizes of the data sets consumed by today's Web applications can be extraordinarily big.
- Such big volume issues can also be found in the areas of finance, communication and business informatics, due to the wide application of information technology and the increasing intensity of online transactions.

Variety:

Because of all this enormous is coming from multiple sources, that leads to the second V which is variety, so because of the variety sources that generating data with different formats, like audios, videos, pictures and txt files.

This data is classified into three forms, structed, instructed and semi-structed as it is shown in figure 2.

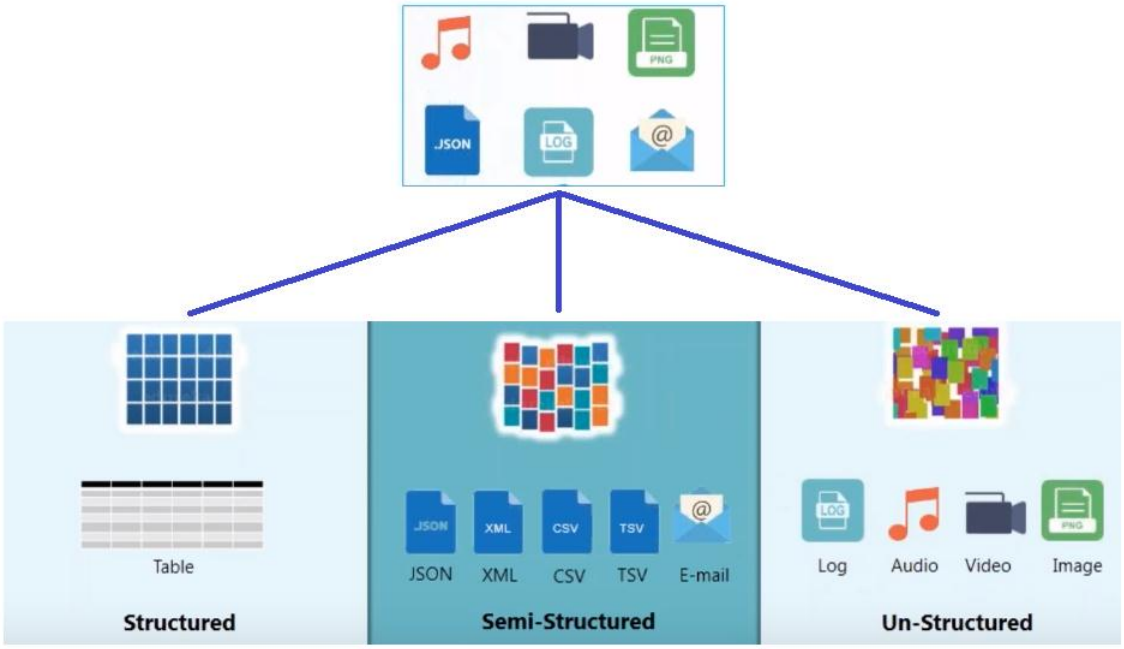

Figure 2:data variety

structured format: also called tabulated format, that the system has a proper schema for the data, so every data is organized and identified in the system.

Unstructured format: is essentially the opposite of structured data and that include all kind of formats as videos, images and txt files … which makes managing this data more complex.

Semi- structured: it is data that doesn't have a fixed rigid schema and no separation between the data and the schema, but every file in there is tagged and marked in another file, and that called self-describing, and that kind of data is most used in the web servers.

[redacted]Velocity:

This V refers to the speed of the accumulation of this data, increasing rate at which data are accumulated, and it is not just about input data. The velocity of a system's outputs matters too. such as Facebook's recommendations, or into dashboards used to drive decision-making.

So, the velocity of the data means how frequently the data arrives and is stored, and how quickly it can be retrieved.

The term velocity refers to the data in motion, the speed at which the data is moving. Data such as financial market, movies, and ad agencies should travel very fast for proper rendering

Value:

Where the data has perceived or quantifiable benefit to the enterprise or organization using it. Every organization wants to convert big data into business values, and that the whole importance of the big data. so, the challenge is not only to collect and manage vast volume and different type of data, but also to extract meaningful value from it, and that require a lot of technologies and efforts.

Veracity:

Uncertainty, quality or the data validity, and it refers to how much we trust in this data and the value that it can give us, how accurate is that data in predicting business value.

2.3 Sources of Big Data:

Those 5 V s challenges vary in different domains, in this part we will classify the big data sources and see the most common fields that we find big data as a powerful tool[3].

Earth sciences:

The advancement of sensing and computing simulation technologies enabled collection and generation of massive data sets every second at different spatiotemporal scales for monitoring, understanding and presenting complex earth systems.

For example, the IPCC (Intergovernmental Panel on Climate Change) alone produced 10,000 TB of climate data.

Internet of Things:

the interconnection via the Internet of objects and enabling them to send and receive data , becomes a huge source , since this devices are connected and contain an advanced sensors ,they are streaming and generating data across the globe , about time

and location of humans, movement of automobiles, vibration of machine, temperature, precipitation, humidity and chemical changes in the atmosphere … , and geographical footprints from interconnected mobile devices, personal computers, sensors, RFID tags and cameras .. ,

Big Data generated from the various sensors of IoT contains a valuable information which is useful in many applications, including better product-line management, more effective and timely criminal investigation, boosting agriculture productivity, and accelerating the development of smart cities and more.

Social sciences:

Economists, political scientists, sociologists and other social scholars use Big Data mining methods to analyze social interactions, health records, phone logs, government records and other digital traces, and using social networks and their activities in the web such as browsing, blogging or buying.

While it is still challenging to quickly extract spatiotemporal patterns from big social data too, for example, help predict criminal observe emerging public health threats and provide more effective intervention.

Astronomy:

By observing the sky using advanced sky survey technologies. Astronomy is producing a spatiotemporal map of the universe, this tach generates vast amounts of data each moment, besides observational data, for example the Large Hadron Collider is investigating how the universe originated and operates at the atomic level and that produces 60 TB of experimental data per day.

The astronomy big data not only records information on how universe evolves, but also can be used to understand how earth evolves and to protect earth from outer space impact.

the challenge in this astronomical data is managing and making sense of the information efficiently.

Business:

Business fields take a big portion of big data, which is used to optimize product placement, analyze customer transaction and market structure, develop personalized product recommendation systems, manage risks and support timely business decisions.

This Business intelligence generate large volume, high velocity and highly unstructured data such as what, where and when a transition occurred and other actions, analyzing this data has enhanced for decisions on strategy, managing optimization and competition.

[redacted]Industry:

This industrial revolution systems are characterized by self-control and self-optimization and Big Data poses a host of challenges to industry including the following:

- seamless integration of energy and production.
- centralization of data correlations from all production levels.
- optimization of performance of scheduling.
- storage of Big Data in a semi-structured data model to enable real-time queries and random access without time-consuming operations.
- realization of on-the-fly analysis to help organizations react quickly to unanticipated events and detect hidden patterns that compromise production efficiency.

So big data analytics could be leveraged to tackle these challenges in Industry.

- In addition to the reviewed six sources, Big Data challenges may also come from other relevant domains such as medical research, public health, smart cities, security management, emergency response and disaster recovery.
- data resources fields and how big data is used proves that the 5Vs vary in different digital earth relevant domains

2.4 Big Data technology challenges:

this section reviews the technological challenges posed by Big Data, because of its 5V features in many different sectors of industry, government and the sciences[3].

[redacted]Data storage:

Storing data is not only related to the volume of the big data, but also to its velocity and its variety, So Storing Big Data on traditional physical storage is problematic as hard disk drives (HDDs) and traditional data protection mechanisms (e.g. RAID: redundant array of independent disks) are not efficient.

And, the Big Data requires the storage systems to be able to scale up quickly which is difficult to achieve with traditional storage systems.

Cloud storage services offer virtually unlimited storage, However, transferring to and hosting big Data on the cloud is expensive given the size of data volume, it need to have an advanced algorithm to address all the data considering the data usage, its transmission and accumulation.

Data transmission:

In general, the life cycle of data is as follows:

- data collection from devices and sensors to storage.
- data integration from multiple data centers (ex: storage servers).
- data management for transferring the integrated data to processing platforms
- data analysis for moving data from storage to analyzing host

Each stage requires a fast bus to transfer large amount of data, Therefore, smart preprocessing techniques and data compression algorithms are needed to effectively reduce the data size before transferring the data, which is considered as a big challenge.

Data management:

It is challenging to efficiently manage, analyze and visualize big, unstructured and structured data. The variety and veracity of Big Data are redefining the data management paradigm.

Now days Hadoop and NoSQL are the most common technologies to clean, store, and organize data (especially unstructured data).

When we talk about data management we need to mention metadata, so metadata is about files that contain information (as addresses and indexes) about data, which is essential for big data managing. the challenge remains to automatically generate these files to describe Big Data and relevant processes,

And Big Data also poses challenges to database management systems (DBMSs) because traditional RBDMSs lack scalability for managing and storing unstructured Big Data. NoSQL databases are designed for Big Data and developing efficient indexing and querying algorithms is still a challenging issue.

Data processing

Data processing may involve various processes, including:

- Validation: Ensuring that supplied data is clean, correct and useful.
- Sorting: arranging items in some sequence and/or in different sets.
- Summarization: reducing detail data to its main points.
- Aggregation: combining multiple pieces of data.
- Analysis: the collection, organization, analysis, interpretation and presentation of data.
- Reporting: list detail or summary data or computed information.
- Classification: separates data into various categories.

Processing large volumes of data in real time or not, requires dedicated computing resources, and this is partially handled by the increasing speed of CPU, network and storage, However, the computing resources required for processing Big Data far exceed the processing power, offered by traditional computing paradigms, so the new technique accrued called cloud computing, which is based on the principle of bringing the CUP to the data instead data to the CUP. this technique will be more explored in the solutions section.

Cloud computing offers virtually unlimited and on-demand processing power as a partial solution; however, it has many new issues:

- First is the limitation of cloud computing's network bandwidth which impacts the computation efficiency over large data volumes.
- Secondly it a challenging task to track and ensure data locality and to support data processing involving intensive data exchange and communication.
  (data locality: is the process of moving the computation close to where the actual data resides on the node, instead of moving large data to computation).

Big Data requires preprocessing before conducting data analysis and mining, for better quality. Large, high-dimensional spatiotemporal data cannot be managed by existing data reduction algorithms, within a tolerable time frame and acceptable quality.

For example, traditional algorithms are not able to preprocess the massive volumes of continuously incoming intelligence and surveillance sensor data in real time.

Highly efficient and scalable data reduction algorithms are required for removing the potentially irrelevant, redundant, noisy and misleading data, and this is one of the most important tasks in Big Data research.

Data analysis:

Data analysis is an important phase to extract useful information extraction and predictions from data. Big Data analysis requires sophisticated scalable and effective algorithms which include parallel processing platforms (e.g. Hadoop) to harness the power of distributed processing.

Furthermore, most existing analytical algorithms require structured homogeneous data and have difficulties in processing the heterogeneity of Big Data. This gap requires either new algorithms that cope with heterogeneous data or new tools for preprocessing data to make them structured to fit existing algorithms.

Data visualization:

Big Data visualization uncovers hidden patterns and discovers unknown correlations to improve decision-making, Since Big Data is often heterogeneous in type, visualization is critical to make sense of Big Data, but it is difficult to provide real-time visualization and human interaction for visually exploring and analyzing Big Data.

The summarized key functionalities for Big Data visualization as follows:

- highly interactive graphics incorporating data visualization best practices.
- integrated, intuitive and approachable visual analytics;
- web-based interactive interfaces to preview, filter or sample data;

Designing and developing these functionalities is challenging because of the many features of Big Data.

Data integration:

It is basically about integrating big data in any field, and it is critical for achieving the Big data value through integrative data analysis and cross-domain collaborations,

Data architecture:

big data architectures outline the hardware and software components that are necessary to a full big data solution. Big data architecture documents may also describe protocols for data sharing, application integrations and information security.

So, if your business has big plans for big data, a strong big data architecture is required to executing those plans.

And An ideal architecture would seamlessly synthesize and share data, computing resources, network, tools, models and, most importantly, people.

Data security:

The increasing dependence on computers and Internet over the past decades makes businesses and individuals vulnerable to data breach and abuse. Big Data poses new security challenges for traditional data encryption standards, methodologies and. Previous studies of data encryption focused on small-to-medium-size data, which does not work well for Big Data due to issues of the performance and scalability.

Thus, effective policies for data access control and safety management need to be investigated in Big Data and these need to incorporate new data management systems and storage structures.

In the cloud era, since data owners have limited control on virtualized storage, ensuring data confidentiality, integrity and availability becomes a fundamental concern.

Data privacy:

The unprecedented networking among smart devices and computing platforms contributes to Big Data but poses privacy concerns where an individual's location, behavior and transactions are digitally recorded. For example, social media and individual medical records contain personal health information raising privacy concerns. Another example is that companies are using Big Data to monitor workforce performance by tracking the employees' movement and productivity. These privacy issues expose a gap between the convention policies/regulations and Big Data and call for new policies to address comprehensively privacy concerns.

Data quality:

Data quality is defined by four aspects and the nature of Big Data turns this aspects challenges:

- accuracy: For example, social media data are highly skewed in space, time and demographics, and location accuracy varies from meters to hundreds of kilometers.
- completeness: incomplete data increases the risk of false discoveries.

- redundancy: filtering should be conducted at the point of data collection in real-time to reduce the data redundancies.
- consistency: ensuring data consistency is challenging with Big Data especially when the data change frequently and are shared with multiple collaborators.

2.5 Big Data tools and technologies:

The last section reviews the technology challenges posed by Big Data from 11 different aspects. While some of these challenges (such as analysis, visualization and quality) exist before Big Data era. The 5Vs of Big Data bring the challenges to a new level as discussed above. Big Data poses unique challenges from several aspects including analysis, visualization, integration and architecture.

To address Big Data challenges a variety of methodologies, techniques and tools are identified to facilitate the transformation of data into value. Computing infrastructure, especially cloud computing, plays a significant role in information and knowledge extraction.

Efficient handling of Big Data often requires specific technologies, such as massive parallel processing, distributed databases, scalable storage systems, and advanced computing architectures, platforms, infrastructures and frameworks, this section introduces some of these methodologies and technologies that underpin Big Data handling[4].

Big data and traditional systems:

In traditional system the data is generated at a steady rate and it is structured in nature, which is simple for traditional system to process it, now the data is heterogenous and it is being generated at an alarming rate by multiple sources so the velocity is high and they are all unstructured data, and our traditional systems is not capable of handling that.

Multiple processing units:

Let's say we add more processing units for the same data processing, by this we increased the processing power but again there is a problem, because all the processing units are accessing data from a single point, so bringing data to processing generates a network overhead, and that causes a network congestion. And sometimes there might be of situations, like processing unit downloading data from the data warehouse, and

other units must wait in queue in order to access that data, and this fails for real-time processing, so we use the distributed data.

[redacted] Distributed and parallel approach:

The idea of the distributed and parallel systems is to use multiple data warehouses, with multiple processing units, and each one has its own task, then it comes a higher level of processing to combine each of the processing results coming from the units, now each processing unit contains its own data warehouse, and this is known by data locality in Hadoop terms, which means the data is locally available into the processing[4].

[redacted]framework (Hadoop):

so now after we saw the distributed and parallel approach, now we need a framework that manage all of it aspects and can deals with the problems of storing and processing. the most common framework is called Hadoop[4].

Hadoop is an open-source framework that allows us to store and process large data sets in parallel and distributed fashion based on master/slave architecture (figure 3).

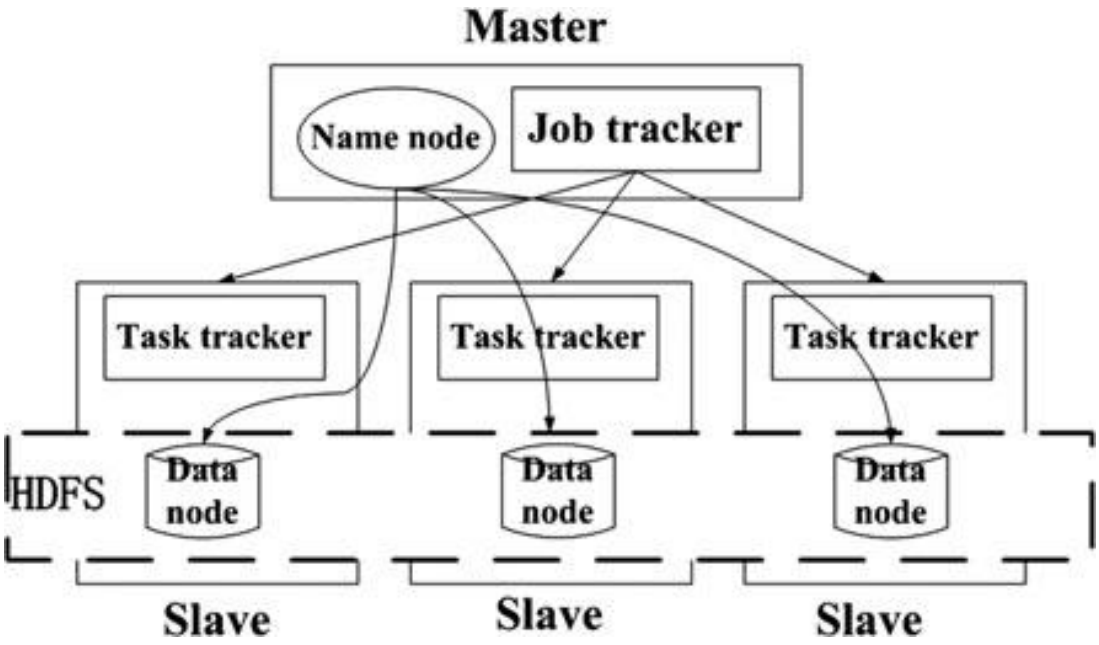

Figure 3:Hadoop cluster topology

Hadoop framework is based on two layers to handle the storing and processing problems:

- To solve the storage problem, we have HDFS (Hadoop distributed file system), which takes the mission of distributing all the data over different interconnected machines, the group of this machines called a Hadoop cluster.

- to process big data, we have something called MapReduce, and this is the programming unit of Hadoop, it allows a parallel and distributed processing of data, that is stored in the Hadoop cluster, and this is known as map, finally when the intermediate output is combined to provide the final output, this is called reduce and hence MapReduce.

The two important Hadoop core components are:

HDFS:

as it is mentioned before, HDFS is the responsible of storing the data, it also uses master/slave architecture as it is shown in the diagram (figure 4).

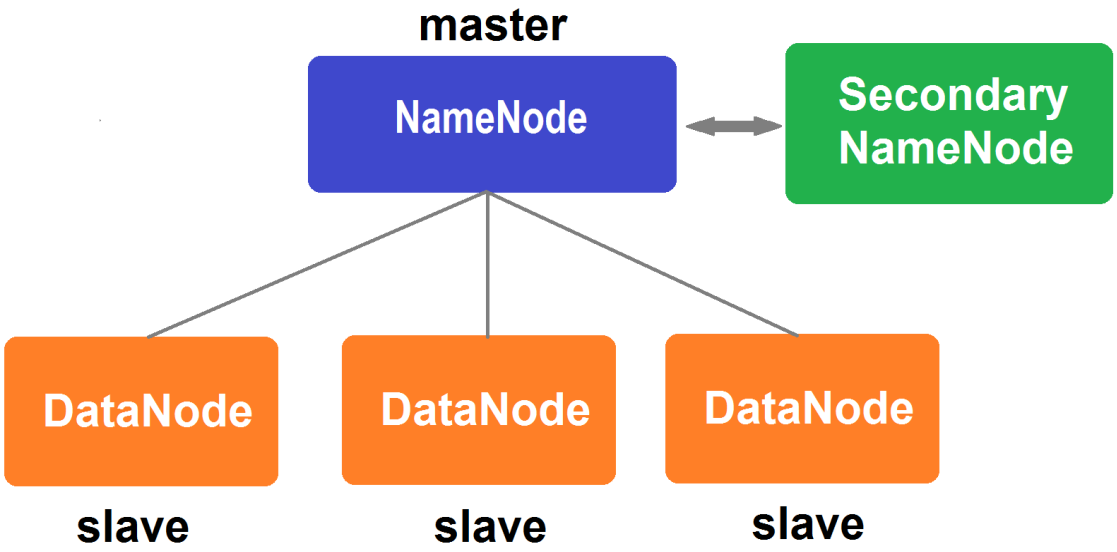

Figure 4:HDFS architecture

The master node is known as NameNode, and the slave node known as DataNode,

*NameNode:

- maintain and manages DataNodes.
- records metadata, information about data blocks, addresses, size permissions, hierarchy, etc.
- receives heartbeats and block reports from all the DataNodes.

- provides the clients nodes with addresses for reading and writing from the DataNodes.
- It also tackles the job of mapping blocks to DataNodes, which are then responsible for managing incoming I/O requests from clients.

*DataNode:

- store the actual data
- serves read and write from the clients

*Secondary NameNode:

the secondary name node is not a backup for the first name node but it has other purposes, we need to understand first the metadata mechanism.

metadata is group of files that contain information about our data, and all the modifications that have took place across The Hadoop cluster, and this metadata is maintained by HDFS, using two files, FsImage and EditLog. so FsImage contains all the modifications that have been made across the Hadoop cluster ever, since the NameNode was started, so let's say that the NameNode started 20 days ago, so FsImage will contain all the details of all the changes that happened in the past 20 days, so obviously you can imagine that there will be a lot of data contained in this file.

The EditLog also contains the most recent changes and modifications that took place in the past 1 hour, and this one is small and it resides in the ram of NameNode machine,

the second NameNode performs the tasks known as checkpointing, the checkpointing it is the process of combining the EditLog with the FsImage, it takes a copy of the EditLog from the NameNode, and then, it combines it with the old FsImage in order, to get the most recent FsImage .as it is shown in the diagram (Figure 5)

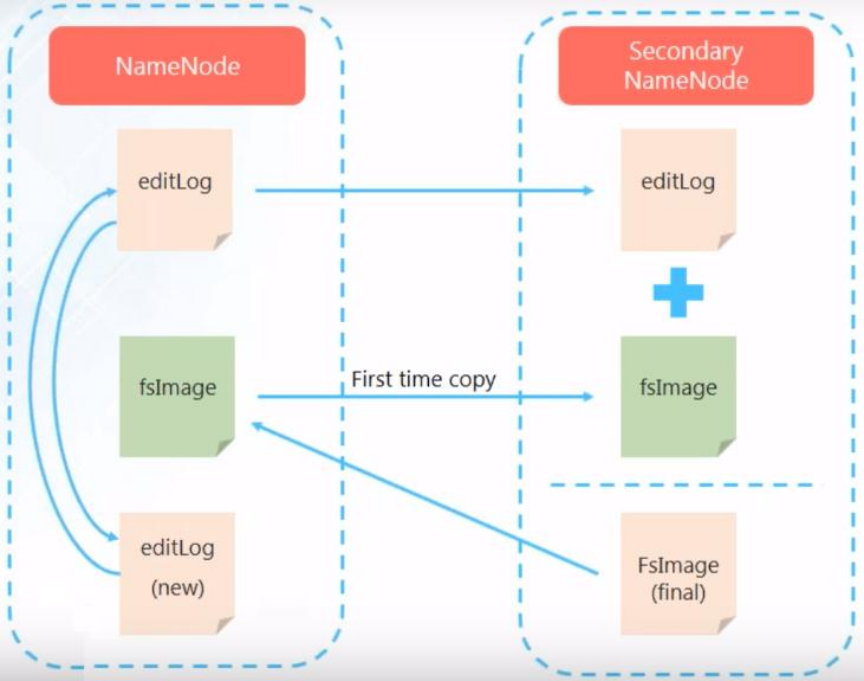

Figure 5:Secondary NameNode checkpointing

The HDFS system need to update the FsImage in order to incorporate all the recent changes regularly, in order to avoid the overloading, the EditLog which can affect the processing power of the NameNode, and also the FsImage stored in the secondary NameNode considered a backup in case of replacing the NameNode.

HDFS is a Block structured file system and each file is divided into a block of particular size (128MB by default).

As an example, we suppose a client wants to store a file of 380 MB size, the HDFS divides the file into three blocks, the first and the second block will occupy 128MB for each, and the 3 third block will be out the remaining size of the file, which is 124 MB, so after the file has been divided into data blocks, this data blocks will be distributed across all the data nodes in the Hadoop cluster.

In this scenario, HDFS allocate only 124 MB for the last part of the file, and not 128, in order to save the 4MB, and this helps HDFS to save the wasted space, by using only that much of space that is needed to store the last part of the file. And also, all that parts will be duplicated by a specific factor (3 by default) to prevent any kind of failure at the DataNode level as it shown in the image (Figure 6).

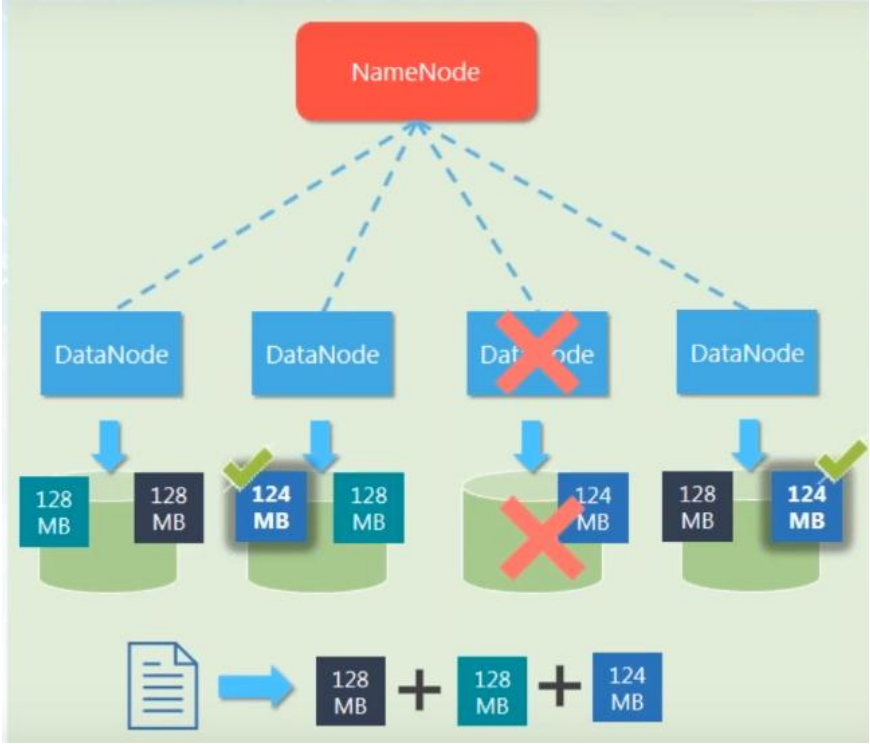

Figure 6:HDFS storing mechanism

The advantages of using the Hadoop distributed file system is as follow:

- First, we do not need to worry about the infrastructure of the hardware because HDFS is an upper layer that manage automatically all storing processes, which makes it easy to use.
- HDFS is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel, so it is easy to upgrade the storage capacity at the hardware layer without affecting the stored data.
- All the DataNodes are interconnected to use the pipelines mechanisms and that improves the speed of writing and reading.
- automatically duplicates the data that is stored in it and creates multiple copies. This is done to ensure that in case there is a failure, or data lost.

MapReduce:

MapReduce is a programming framework that allows us to perform distributed, and parallel processing on large data sets in a distributed environment, in order to facilitate,

and simplify the processing of vast amounts of data, on large clusters of commodity hardware in a reliable, fault-tolerant manner using the next three aspects:

- Partition a large problem into smaller sub-problems.
- Independent sub-problems executed in parallel.
- Combine intermediate results from each individual worker (processing unit), and the workers can be, parts of a processor core, Cores in a multi-core processor, group of processors.

The core idea behind MapReduce is mapping the dataset into a collection of pairs, and then reducing overall pairs with the same key. The overall concept is simple, but it is quite expressive when we consider that:

- All data can be mapped into pairs somehow.
- The keys and values may be of any type: strings, Tables, integers, structures, files …

The MapReduce algorithm contains two important tasks, namely Map and Reduce (Figure 7):

- The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).

- The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs), into a smaller set of tuples. The reduce task is always performed after the map job.

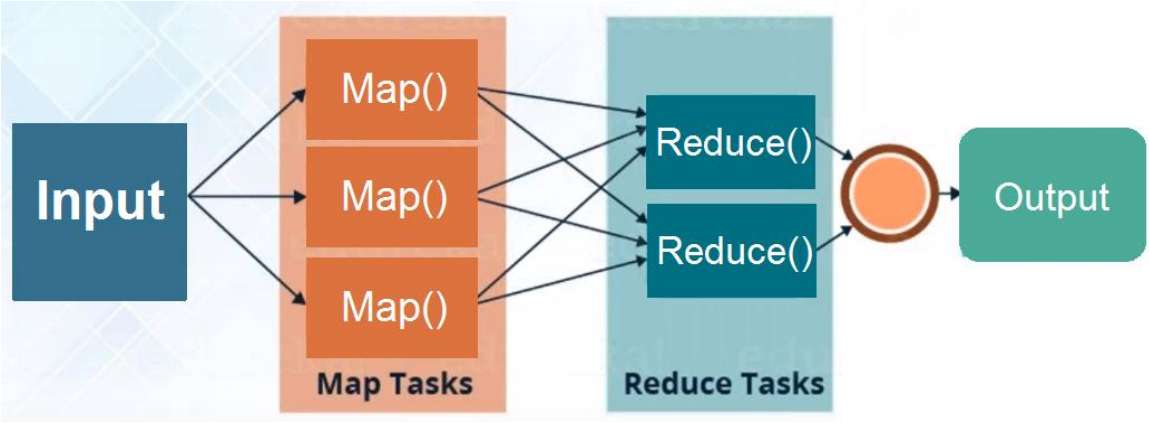

Figure 7:Map and Reduce

The Figure 8 illustrate an example, if we want to count the number of word occurrences, so that we can get frequencies. Thus, we want the reduce script to simply sum the values of the collection of pairs which have the same key.

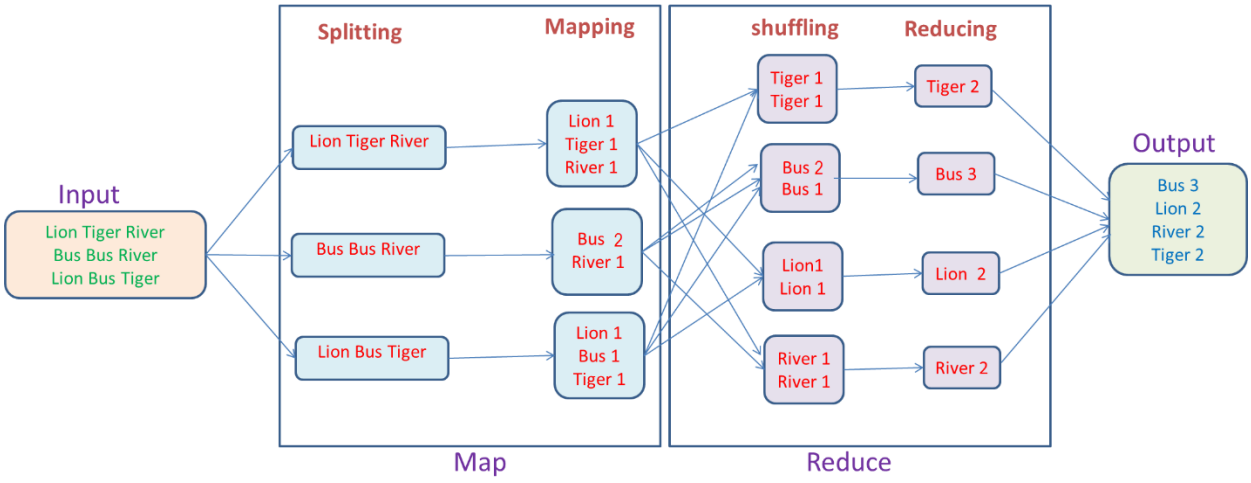

Figure 8: example: words counter

2.6 Cloud computing and Big Data:

Cloud Computing is highly used in big data analytics due to its cost-efficient computing paradigm, in which information and computer power can be accessed from anywhere by customers. Cloud Computing is the Internet-based development and use of computer technology.

Cloud Computing is a style of computing paradigm in which typically real-time scalable resources such as files, data, programs, hardware and services.

These customers pay only for the used computer resources (in general those resources are virtualized) and services, having no knowledge of how a service provider uses an underlying computer technological infrastructure to support them. The service load in Cloud Computing is dynamically changed upon the customer requests. Cloud Computing shifts the computation from local, individual devices to distributed, virtual, and scalable resources, thereby enabling end-users to utilize the computation, storage, and other application resources, which forms the Cloud.

Due to cloud computing advantages, many companies now use cloud computing services to processes their own data.

Cloud computing is closely related to big data. The main objective of cloud computing is to use huge computing resources and computing capacities under concentrated management, to provide applications with resource sharing at a granularity and provide big data applications with computing capacity. The development of cloud computing provides solutions for the storage and processing of big data[4].

2.7 Conclusion:

This chapter shows the different challenges that big data is facing, and to address this challenge, there are a variety of methodologies, techniques and tools to facilitate the transformation of data into value. Computing infrastructure, especially cloud computing, plays a significant role in information and knowledge extraction. Efficient handling of Big Data often requires specific technologies, such as massive parallel processing, distributed databases, data-mining grids, scalable storage systems, and advanced computing architectures, platforms, infrastructures and frameworks.

Chapter 3:
The Present and Future threats of Big Data:

3.1 Introduction:

This chapter treats the most common threats in present and the predicted ones, those threats can technological or social.

3.2 Big Data Privacy:

In the big data era, data privacy includes two aspects:

- the protection of personal privacy, as the advances on data acquisition is made, personal interests, habits, and body properties, etc. of users may be more easily acquired, of which the user may not be aware. And even an unauthorized collection of information about the existence and characteristics of personal things can happen unintentionally.

- Personal privacy data may also be leaked during storage, transmission, and usage, even if acquired with the permission of users,

Organizations that own big data usually attempt to mine valuable information in the data with advanced algorithms. The privacy data protection technology is of great importance. Therefore, privacy protection in the big data era will become a new and challenging problem[5].

3.3 Data Quality:

There are a lot of factors that may restrict data quality, for example, generation, acquisition, transmission, and transmission may all influence data quality. Data quality is mainly manifested in its accuracy, completeness, redundancy, and consistency.

Even though a lot of measures have been taken to improve data quality, the quality related problems could not be completely solved. Therefore, effective methods to automatically detect data quality and repair some damaged data need to be investigated[3].

3.4 Big Data Safety Mechanism:

Big data brings challenges to data encryption due to its large scale and high variety. The performance of previous encryption methods on small and medium-scale data could not meet the demands of big data; and that makes big data cryptography approaches need to be developed for more security[3].

3.5 Big data technologies in the future:

Although technologies represented by Hadoop have achieved a great success, Data with a larger scale, more variety, and more complex structures, such technologies are definitely to fall behind and will be replaced given the rapid development of big data[3].

3.6 conclusion:

Throughout the history of human society, the demands and willingness of human beings are always the source powers to promote scientific and technological progress. In the big data era, big data may provide reference answers for human beings to make decisions through mining and analytical processing, but could not replace human thinking. It is human thinking that promotes the widespread utilizations of big data. Big data is more like an extendable and expandable human brain other than a substitute of human brain. With the emergence of Internet of Things, development of mobile sensing technology, and progress of data acquisition technology, people are not only the user and consumer of big data, but also its producer and participant. And all applications closely related to human activities based on big data will be increasingly concerned and will certainly cause enormous changes of social activities in the future society.

Conclusion:

The quantity of computer data generated on planet earth is growing exponentially from many sources, retailers, financial services, healthcare, public data, from social media, smart objects (IoT), and finally several areas of scientific researches.

It is obvious that big data is part of a wider innovation revolution in our domestic, social and business worlds. Due to its 5Vs (Volume, Variety, Velocity, Value, Veracity), transforming this big data to knowledge create a lot of challenges, not only to collect and manage vast volume and different type of data, but also to extract meaningful value from it.

Today the leading Big Data technology is Hadoop, this is an open source platform for reliable scalable distributed computing, and provides the first viable platform for big data analytics, Hadoop is already used by most big data pioneers, for example, LinkedIn currently uses it to generate over 100 billion personalized recommendations every week.

Now days cloud computing plays a big role in big data, for those companies who cannot afford an internal big data infrastructure, cloud-based big data solutions are already available, where public big datasets need to be utilized running everything in the cloud, and makes a lot of sense as data does not have to be downloaded, for example, amazon web services, already hosts many public data sets containing government and medical information.

looking further ahead, quantum computing may greatly improve big data processing, quantum computers will in theory excel at the massively parallel processing of unstructured data.

Bibliography

[1] NATHAN MARZ, JAMES WARREN. Big Data : PRINCIPLES AND BEST PRACTICES OF SCALABLE REAL - TIME DATA SYSTEMS, 2015

[2] O'Reilly Media. Big Data Now, 2012

[3] Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu & Fei Hu. Big Data and cloud computing: innovation opportunities and challenges, 2016

[4] Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland. Big Data Optimization: Recent Developments and Challenges, 2016

[5] Terence Craig, Mary E. Ludloff . Privacy and Big Data. 2011