REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE

N° :……..

DOMAINE : MATHEMATIQUES ET DE L'INFORMATIQUE

FILIERE : INFORMATIQUE

OPTION : INFORMATIQUE DECISIONELLE ET OPTIMISATTION

Mémoire présenté pour l'obtention

Du diplôme de Master Académique

Par : Chemini Mounira

Intitulé

Big Data and Data mining

Soutenu devant le jury composé de :

……	Université de M'sila	Président
Dr.Hamani Said	Université de M'sila	Rapporteur
……	Université de M'sila	Examinateur

Année universitaire : 2017 /2018

THANKS

All my thanks to my parents ,

They have been with me every step of my life through good and bad times .

I thank sincerely Dr.hamani for his patience, support and advice that helped me achieve louse to work.

My thanks also go to the teachers who provided me with knowledge, allowed me to extend my greetings and my sincere thanks to all my colleagues who encouraged me during the course.

Thanks

DEDICATION

I dedicate my note work to my family .A special feeling of gratitude to my lovely parents Ali and Aicha who have raise me to be a good person .

My brother Djalal and my sister sara have never left me side and are very kind .

I also dedicate this note to many friends who have supported me.

Bibliography :

[1] https://www.kdnuggets.com/2017/02/origins-big-data.html.

[2] https://www.kdnuggets.com/2017/02/what-is-big-data.html.

[3] Wikipedia. Big data, 2014. http://en.wikipedia.org/wiki/Big_data, accessed April 2014.

[4] M. Stonebreaker, P. Brown, and D. Moore. Object-relational DBMSs, tracking the next great wave. Morgan Kauffman Publishers, Inc., San Francisco, California, 2 edition ,1998.

[5] http://searchcloudprovider.techtarget.com/feature/Big-data-analysis-in-the- cloud-Storage-network-and-server-challenges.

[6] For Big Data Analytics There's No Such Thing as Too Big The Compelling Economics and Technology of Big Data Computing, White Paper, March 2012.

[7]http://www.infoworld.com/d/cloud-computing/amazons-redshift-big-data-analytics-the-pros-and-cons-213049.

[8] https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics.

[9] https://www.whizlabs.com/blog/big-data-analytics-importance/. April 2014

[10] Everitt B S Cluster Analysis Heinemann ,1974 .

[11] Hartigan, J. A.. Clustering algorithms. New York: John Wiley and Sons, 1975.

[12] Krzanowski W J principles of Multivariate Analysis Oxford University Press, 1990.

[13] Hartigan J A and Wong M A Algorithm AS136: A K-means clustering algorithm Appl. Statist. 28 100–108 , mai 1979.

[14] K. Kline, D. Kline, and B. Hunt. SQL in a nutshell, a desktop quick O'Reilly Media ,Sebastopol , California edition ,2008.

[15] Apache Hadoop. HDFS Architecture Guide, 2013.

[16] Apache Hadoop. MapReduce Tutorial, 2013.

[17] Hadoop Map Reduce CookBook - Srinath Perera ,2012.

[18] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html, accessed April 2014.

[19] Greenplum. A unified engine for RDBMS and Map Reduce, 2009.

[20] https://bigdataarchitecture.com/ April 2014.

[21] Hadoop: The Definitive Guide MapReduce for the Cloud,2009.

Summary

Table of contents :

Chapter II : Big data analytics

Chapter III : Hadoop HDFS

Chapter VII: Implémentation.

Introduction:

The Big data is larger, more complex data sets, are defining three 3Vs (volume, variety and velocity) Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing.

Big Data analytics is the process of collecting, organizing and analyzing large sets of data(called Big Data) to discover patterns and other useful information

the technologies information has been using tools to analyze and process big data, which are retrieving them from different systems in many fields and in the industrial in particular as part of the business intelligence system, Data recovery, processing and use is undertaken for the purpose of developing products or developing a new product.

our work is to explore big data technology uses Hdfs (Hadoop distributed files system) and compare data analytics such as clustering distributed Hdfs and non distributed (windows) .

CHAPTER I

The big data

1.Introduction :

In this chapter we will present the big data and what is it but before we will have an overview the history of big data with use this latter.

Figure I The big data [1].

2.History :

In the 2013 Oxford English Dictionary introduced the term “Big Data” for the first time in Its dictionary. That of course begs the question ‘When was the term Big Data first used and Why?’. My curiosity led me to lot of research material but I relied mostly on Mr. Gil Press’s “A Very Short History of Big data” from Forbes, Mr. Steve Lohr’s “The Origins of ‘Big Data’: An Etymological Detective Story“ from The New York Times, Mr. Mark van Rijmenam’s “A Short History Of Big Data” from DataFloq.com.[1]

The story of how data became big starts many years before the current buzz around big data. Already seventy years ago we encounter the first attempts to quantify the growth rate in the volume of data or what has popularly been known as the “information explosion” (a term first used in 1941, according to the Oxford English Dictionary). [1]

Following info graphic is a quick summary and the story follows right below see *Figure II*.

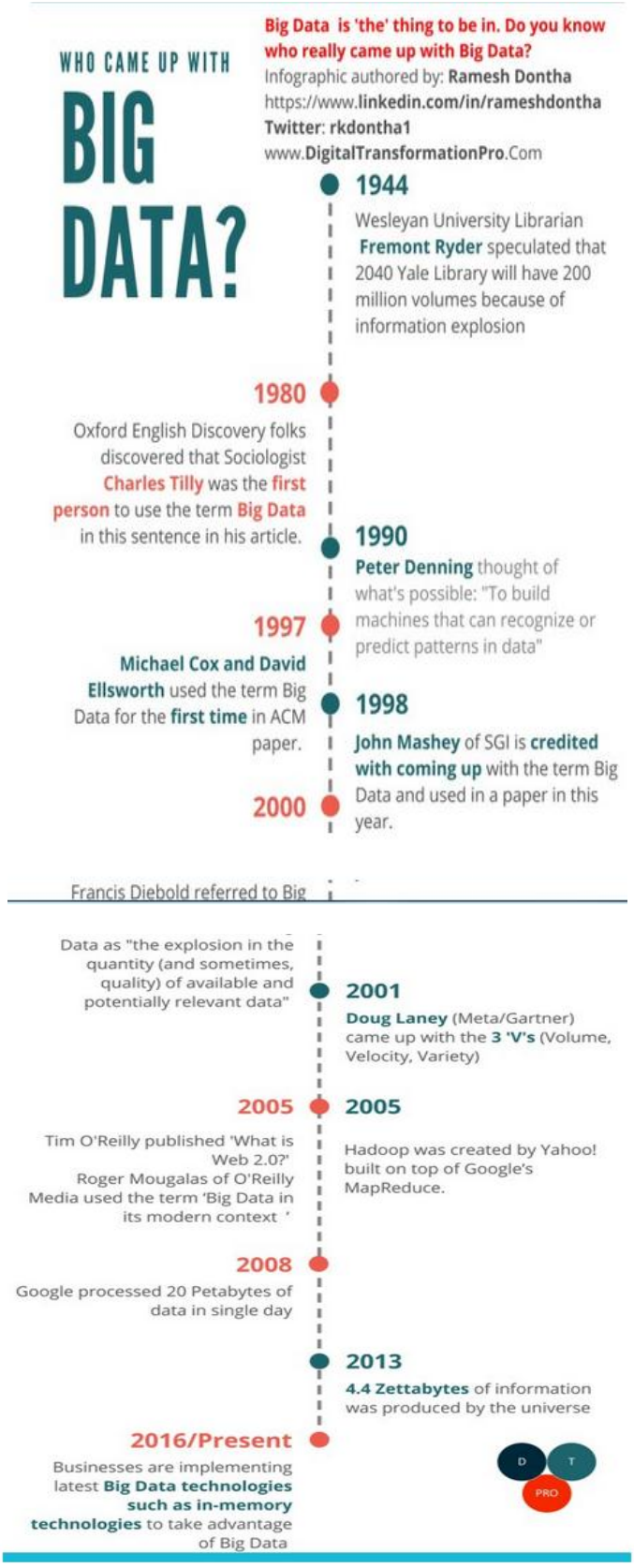

Figure II The story of big data[1].

3. What is the big data :

3.1 Definition:

Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing application. So Wikipedia's definition is focusing on 'volume of data' and 'complexity of processing that Data'. Good start but doesn't clearly answer the question of what is the volume threshold of data that makes it Big Data. Is it 100 GB , A Peta Byte [2].

As per O'Reilly media:

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it.[3]

3.2 Sources of big data :

3.2.1 Media :

Media is the most popular source of big data, as it provides valuable insights on consumer preferences and changing trends. Since it is self-broadcasted and crosses all physical and demographical barriers, it is the fastest way for businesses to get an in-depth overview of their target audience, draw patterns and conclusions, and enhance their decision-making.

Media includes social media and interactive platforms, like Google, Face book, Twitter, YouTube, Instagram, as well as generic media like images, videos, audios, and podcasts that provide quantitative and qualitative insights on every aspect of user interaction[3].

3.2.2 Cloud :

Today, companies have moved ahead of traditional data sources by shifting their data on the cloud. Cloud storage accommodates structured and unstructured data and provides business with real-time information and on-demand insights. The main attribute of cloud computing is its flexibility and scalability. As big data can be stored and sourced on public or private

clouds, via networks and servers, cloud makes for an efficient and economical data source[3].

3.2.3 The web :

The public web constitutes big data that is widespread and easily accessible.

Data on the Web or 'Internet' is commonly available to individuals and companies alike. Moreover, web services such as Wikipedia provide free and quick informational insights to everyone. The enormity of the Web ensures for its diverse usability and is especially beneficial to start-ups and SME's, as they don't have to wait to develop their own big data infrastructure and repositories before they can leverage big data[3].

3.2.4 IOT :

Machine-generated content or data created from Iot constitute a valuable source of big data. This data is usually generated from the sensors that are connected to electronic devices. The sourcing capacity depends on the ability of the sensors to provide real-time accurate Information[2].

IoT is now gaining momentum and includes big data generated, not only from computers and Smart phones, but also possibly from every device that can emit data. With IoT, data can now be sourced from medical devices, vehicular processes, video games, meters, cameras, Household appliances, and the like [3].

3.2.5 Databases:

Businesses today prefer to use an amalgamation of traditional and modern databases to acquire relevant big data. This integration paves the way for a hybrid data model and requires low investment and IT infrastructural costs. Furthermore, these databases are deployed for several business intelligence purposes as well. These databases can then provide for the extraction of insights that are used to drive business profits. Popular databases include a variety of data sources, such as MS Access, DB2, Oracle, SQL, and Amazon Simple, among others[1].

3.3 Types of data:

3.3.1 Structured data:

Structured Data is used to refer to the data which is already stored in databases, in an ordered manner. It accounts for about 20% of the total existing data, and is used the most in programming and computer-related activities[2].

There are two sources of structured data- machines and humans. All the data received from sensors, web logs and financial systems are classified under machine-generated data.

These include medical devices, GPS data, data of usage statistics captured by servers and applications and the huge amount of data that usually move through trading platforms, to name a few[2].

Human-generated structured data mainly includes all the data a human input into a computer, such as his name and other personal details. When a person clicks a link on the internet, or even makes a move in a game, data is created- this can be used by companies to figure out their customer behavior and make the appropriate decisions and modifications[2].

3.3.2 Unstructured data:

While structured data resides in the traditional row-column databases, unstructured data is the opposite- they have no clear format in storage. The rest of the data created, about 80% of the total account for unstructured big data. Most of the data a person encounters belongs to this category- and until recently, there was not much to do to it except storing it or analysing it manually[2].

Unstructured data is also classified based on its source, into machine-generated or human-generated. Machine-generated data accounts for all the satellite images, the scientific data from various experiments and radar data captured by various facets of technology[2].

Human-generated unstructured data is found in abundance across the internet, since it includes social media data, mobile data and website content. This means that the pictures we

upload to out Face book or Instagram handles, the videos we watch on YouTube and even the text messages we send all contribute to the gigantic heap that is unstructured data[2].

3.3.3 Semi-structured data**:**

The line between unstructured data and semi-structured data has always been unclear, since most of the semi-structured data appear to be unstructured at a glance. Information that is not in the traditional database format as structured data, but contain some organizational properties which make it easier to process, are included in semi-structured data. For example, NoSQL documents are considered to be semi-structured, since they contain keywords that can be used to process the document easily[2].

4. Big data attributes:

4.1 Volume:

Volume is the most challenging aspect of Big Data since it imposes a need for scalable storage and a distributed approach to querying. Big enterprises already have a large amount of data accumulated and archived over the years. It could be in the form of system logs , record keeping...etc[1].

The amount of this data easily gets to the point where conventional database management systems may not be able to handle it. Data warehouse based solutions many not necessarily have the ability to process and analyze this data due to lack of parallel processing architecture A lot can be derived from text data, locations or log files. For example, email, communication patterns, consumer preferences and trends in transaction-based data, security investigation[2]. Spatial and temporal (time-stamped) data absorb storage space quickly.

Big Data technologies offer a solution to create value from this massive and previously Unused/ difficult to process data[2].

4.2 Velocity:

Data is flowing into organizations at a large speed. Web and mobile technologies have

enabled generating a data flow back to the providers. Online shopping has revolutionzed Consumer and provider interactions. Online retailers can now keep log of and have access to customers every interaction and can maintain the history and want to quickly utilize this information in recommending products and put the organization on a leading edge[2] . Online marketing organizations are deriving lot of advantage with the ability to gain insights instantaneously. With the invention of the smart phone era there is even further location based data generated and its becoming important to be able to take advantage of this huge amount of data[1].

4.3 Variety:

All this data generated with social and digital media is rarely structured data. Unstructured text documents, video, audio data, images, financial transactions, interactions on social websites are examples of unstructured data. Conventional databases support 'large objects' (LOB's), but have their limitations if not distributed. This data is hard to fit in conventional neat relational database management structures and is not very integration friendly data and needs a lot of massaging before applications can manage it. And this leads to loss of information. If the data is lost then it's a loss that cannot be recovered. Big Data on the other hand tends to keep all the data since most of this is write once and read many times type of data. Big Data believes that there could be insights hidden in every bit of data[1].

Inderpal Bhandar, Chief Data Officer at Express Scripts noted in his presentation at the Big Data Innovation Summit in Boston that there are additional Vs that IT, business and data scientists need to be concerned with, most notably big data Veracity. Other big data V's getting attention at the summit are: Veracity ,validity and volatility[1].

4.5 Veracity:

Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed[2].

inderpal feel veracity in data analysis is the biggest challenge when compares to things like volume and velocity. In scoping out your big data strategy you need to have your team and partners work to help keep your data clean and processes to keep 'dirty data' from accumulating in your systems[1].

4.5 Validity:

Like big data veracity is the issue of validity meaning is the data correct and accurate for the intended use. Clearly valid data is key to making the right decisions. Phil Francisco, VP of Product Management from IBM spoke about IBM's big data strategy and tools they offer to help with data veracity and validity[2].

4.6 Volatility

Big data volatility refers to how long is data valid and how long should it be stored. In this world of real time data you need to determine at what point is data no longer relevant to the current analysis[1].

Big data clearly deals with issues beyond volume, variety and velocity to other concerns like veracity, validity and volatility. To hear about other big data trends and presentation follow the Big Data Innovation Summit on twitter[1] .

5. Need of big data analytics :

With the above-mentioned attributes of big data, data is massive, comes at a speed and highly unstructured that it doesn't fit conventional relational database structures.

with so much insight hidden in this data, an alternative way to process this enormous data is necessary. Big corporations could be well resourced to handle this task but the amount of data being generated every day easily outgrows this capacity[6].

cheaper hardware , cloud computing and open source technologies have enabled processing big data at a much cheaper cost[6].

Lot of data means lot of hidden insights. The ability to quickly analyze big data means the possibility to learn about customers, market trends, marketing and advertising drives equipment monitoring and performance analysis and much more, and his is an important reason that many big enterprises are in a need of robust big data analytics tools and technologies[6].

Big data tools mainly make use of in-memory data query principle, Queries are performed where the data is stored, unlike conventional business intelligence (BI) software that runs queries against data stored on server hard drive[6].

in memory data analytics has significantly improved data query performance, Big data analytics not just helps enterprises make better decisions and gain an edge into real-time processing , it has also inspired businesses to derive new metrics and gain new sources of revenue out of insights gained, note that temporal data naturally leads to big data as does spatial data., early attempts to deal with large warehouses ,including non-scalar data, used so called ORDBMS[4], i.e. object relations databases. Big Data outperforms ORDBMS in various ways, including the need for more complicated backups, recovery and faster search algorithms, benefits of using big data technologies may come at a downside of a loss of privacy of the data , in terms of privacy , some companies sell customer data to other companies and this can be a problem [4].

6. Data Analytics in Cloud Computing:

Cloud computing is built around a series of hardware and software that can be remotely accessed through any web browser, usually files and software is shared and worked on by multiple users and all data is remotely centralized instead of being stored on user's hard

drives, Analytics in cloud computing, such as tracking social media engagment and statistics is simply applying the principles of analytics to information housed on cloud drives rather than on individual servers or drives[5].

Much of the benefit from data analysis comes from its ability to recognize patterns in a set and make predictions regarding past experience ,usually the process is referred to as data mining, which simply means discovering patterns in data sets to better understand trends. With all the benefits data analysis and big data offer, much of their potential is missed because employees lack quick, reliable access to said information. Gartner estimates 85% of Fortune 500 companies do not reap the full benefit of their big data analytics because of lack of accessibility to data, causing them to miss potential opportunities to better connect with and meet clients' needs. As analysis moves towards cloud drives, data analysis gains accessibility as company employees can access company information remotely from any location, freeing them from being chained to local networks and thus making data more accessible. Recently, Time Warner unveiled its data analytics cloud system, which allows their 4,000 employees to better utilize sales data in hopes of equipping them to increase profit margins[7].

6. Conclusion:

We introduced this chapter with an overview of the big data , Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making, in addition we present the source of it ,types and Consequently we present the cloud computing in big data analytics.

CHAPTER II

Big data analytics

1.Introduction:

After this we have present what it is big data .in this chapter we will define the Big Data analytics, In addition we will stress the importance of Big Data Analysis and show how the analysis of Big Data will improve decisions in the future.

Figure I Big data analytics[1].

2.Big data analytics:

2.1 Definition :

Big Data analytics is the process of collecting, organizing and analyzing large sets of data (called Big Data) to discover patterns and other useful information, Big Data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with Big Data typically want the knowledge that comes from analyzing the Data[8].

2.2 The importance of big data analytics:

The Big Data analytics is a revolution in the field of Information Technology, The use of Data analytics by the companies is enhancing every year. The primary focus of the companies is on customers. The analytics divide into different types as per the nature of the environment : Prescriptive Analytics, Predictive Analytics, and Descriptive Analytics See *Figure II. 2* [9].

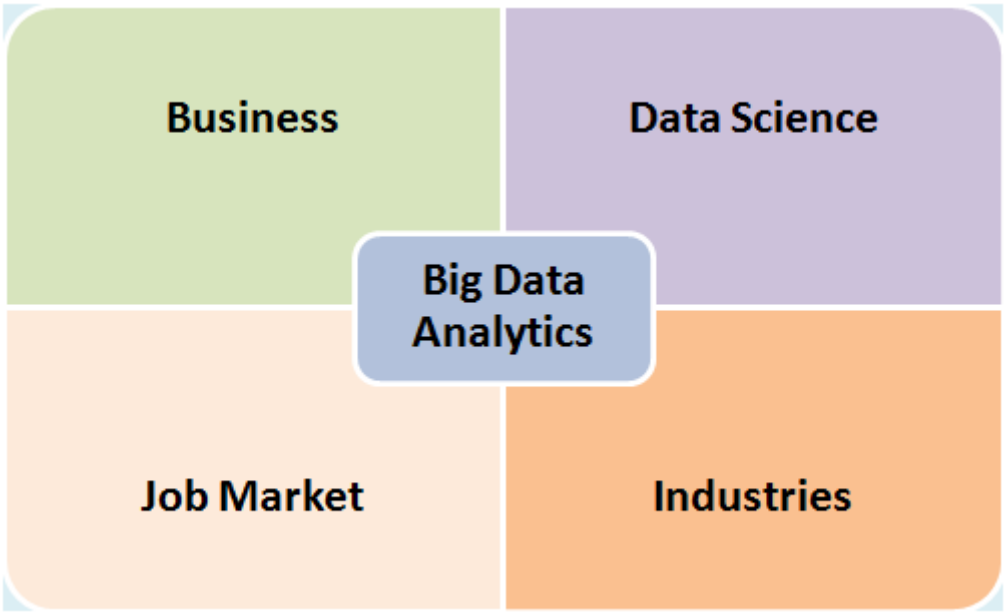

Figure II. 2 Types of big data analytics[9].

2.2.1 Big Data Analytics and Data Sciences:

The analytics involves the use of advanced techniques and tools of analytics on the data obtained from different sources in different sizes.

Big Data analytics involves the use of analytics techniques like machine learning, data mining, natural language processing, and statistics. The data is extracted, prepared and blended to provide analysis for the businesses. Large enterprises and multinational

Organizations use these techniques widely these days in different ways [9].

2.2.1 Businesses and Big Data Analytics:

Big Data analytics tools and techniques are rising in demand due to the use of Big Data in Businesses. Organizations can find new opportunities and gain new insights to run their business efficiently. These tools help in providing meaningful information for making better business decisions.

The companies can improve their strategies by keeping in mind the customer focus, big data analytics efficiently helps operations to become more effective. This helps in improving the profits of the company[7].

Big data analytics tools like Hadoop helps in reducing the cost of storage, This further increases the efficiency of the business. With latest analytics tools, analysis of data becomes easier and quicker, This, in turn, leads to faster decision making saving time and energy[8].

2.3 Companies in the analysis of big data :

To turn big data into a business advantage, businesses have to review the way they manage data within data centre. The data is taken from a multitude of sources, both from within and without the organization. It can include content from videos, social data, documents and machine-generated data, from a variety of applications and platforms. Businesses need a System that is optimized for acquiring, organizing and loading this unstructured data into their databases so that it can be effectively rendered and analyzed, Data analysis needs to be deep and it needs to be rapid and conducted with business goals in mind [10].

The scalability of big data solutions within data centers is an essential consideration. Data is vast today, and it is only going to get bigger. If a data centre can only cope with the levels of data expected in the short to medium term, businesses will quickly spend on system Refreshes and upgrades. Forward planning and scalability are therefore important[10].

In order to make every decision as desired there is the need to bring the results of knowledge discovery to the business process and at the same time track any impact in the various dashboards, reports and exception analysis being monitored. New knowledge discovered through analysis may also have a bearing on business strategy, CRM strategy and financial strategy going forward. See *Figure II1.1*

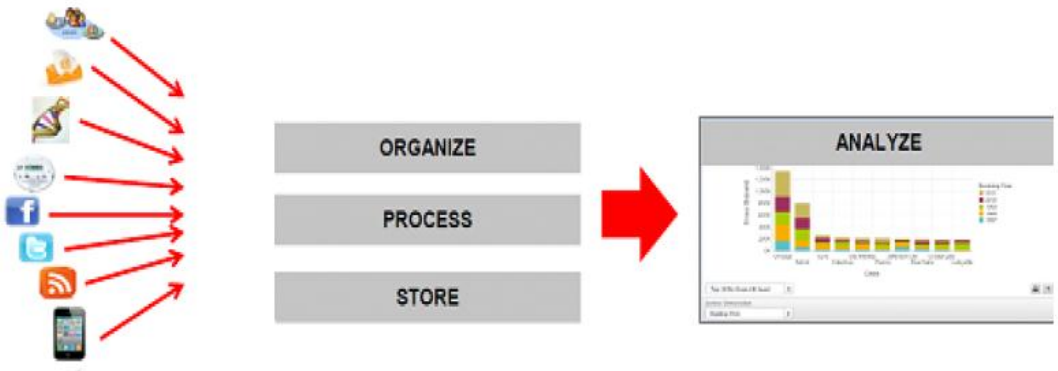

Figure II1.1 Big Data Management[10].

Big data analytics and the Apache Hadoop open source project are rapidly emerging as the preferred solution to business and technology trends that are disrupting the traditional data management and processing landscape. Enterprises can gain a competitive advantage by being early adopters of big data analytics. Even though big data analytics can be technically challenging, enterprises should not delay implementation[8].

As the Hadoop project mature and business intelligence (BI) tool support improves, big data analytics implementation complexity will reduce, but the early adopter competitive advantage will also wane, Technology implementation risk can be reduced by adapting existing architectural principles and patterns to the new technology and changing requirements rather than rejecting them[10].

4. Clustering Method :

4.1 Definition of Clustering :

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters[11].

4.2 Types of Clustering:

clustering algorithms can be classified as **:**

- Hard Clustering:

In hard clustering, each data point either belongs to a cluster completely or not.

- Soft Clustering:

In soft clustering, instead of putting each data point into a separate cluster, a probability of that data point to be in those clusters is assigned [11].

Clustering algorithms can be classified also as **:**

- Connectivity models:

As the name suggests, these models are based on the notion that the data points closer in data Space exhibit more similarity to each other than the data points laying farther away, These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance Increases, also, the choice of distance function is subjective, these models are very easy to interpret but lacks scalability for handling big datasets [13].

- Centroid models:

These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters, K-Means clustering algorithm is a

popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optimal[13].

- Distribution models:

These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian), These models often suffer from over fitting ,A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions[11].

- Density Models:

These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS[13].

Another type of classification algorithms :

- Hierarchical clustering:

requires: a distance (Euclidean distance, correlation similarity, Manhattan distance) and a data fusion approach.

we look for the 2 closest points according to the distance and we group them in a cluster. The points are replaced by their center, then we search again the nearest points (or clusters) to group them in a cluster, and this iteratively[13].

Each time we calculate the distance between 2 clusters, we use an aggregation method below. iterates until it has only one cluster [11].

In the following will present a clustering algorithm:

4.3 k-means Clustering:

K-means is one of the simplest unsupervised learning algorithms that solve the well

known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriority. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centrist as bray center of the clusters resulting from the previous step. After we have these k new centroid, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more[11].

4.3.1 Algorithmic steps for k-means clustering :

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using.

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3)[11]

As follow is the algorithm:

Input: K (the number of clusters),

D(a set of lift ratios)

Output: a set of K clusters

Method

Arbitrarily choose K objects from D as the initial cluster centers;

Repeat:

1.(re) assign each object to the cluster to which the objects is the most similar, based on the mean value of the objects in the cluster;

2.Update the cluster means i.e.. calculate the mean value of the objects for each cluster.

Until : no change ;

5. Conclusion:

In this chapter we present the data analytics, the importance of it , needs of this data in businesses , With the use of big data analytics becoming more and more important to businesses, it is even more vital for them to find a way to analyze the ever (faster) growing disparate data coursing through their environments, finally we presented a concept the clustering (algorithm k-means).

CHAPTRE III

Hadoop HDFS

1. Introduction:

The Hadoop is open source who solve the problem of big data, in this chapter we will present what is it and show map reduce example .

2. Hadoop presentation:

Formal definition of Hadoop by Apache: “The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models[15].

It is designed to scale up from single servers to thousands of machines, each offering local computation and storage, rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly available service on top of a cluster of computers, each of which may be prone to failures” [15].

Hadoop was initially inspired by papers published by Google, outlining its approach to handle an avalanche of data, and has since become the standard for storing, processing and analyzing hundreds of terabytes, and even peta bytes of data[15].

Hadoop frame word envelopment was started by Doug Cutting and the framework got its name From his son ‘s elephant toy [16].

Hadoop has drawn the inspiration from Google's File System (GFS).

Hadoop was spun from notch in 2006 to become a sub-project of Lucerne and was renamed to.

Hadoop Yahoo has been a key contributor to Hadoop evolution, By 2008 yahoo web search engine index was being generated by a 10,000 core Hadoop cluster.

Hadoop is an open source framework by Apache, and has invented a new way of storing and processing data[15].

Hadoop does not rely on expensive, high efficiency hardware instead it leverages on benefits from distributed parallel processing of huge amount of data across commodity, low-cost servers.

This infrastructure stores as well as processes the data and can easily scale to changing needs.

Hadoop is supposed to have limitless scale up ability and theoretically no data is too big to handle with distributed architecture [16].

Hadoop is designed to run on commodity hardware and can scale up or down without system interruption, It consists of three main functions: storage, processing and resource management .

It is presently used by big corporations like Yahoo, eBay, linkdln and face book Conventional data storage and analytics systems were not built keeping in mind the needs of big data. And hence no longer easily and cost-effectively support today's large data sets [16].

3. Hadoop attributes:

• Fault tolerant - Fault tolerance is the ability of the system to stay functional without interruption and without losing data even if any of the system components fail [17].

One of the main goals of Hadoop is to be fault tolerant, since Hadoop cluster can use thousands of nodes running on commodity hardware, it becomes highly susceptible to failures .

Hadoop achieves fault tolerance by data redundancy/replication.

And also provides ability to monitor running tasks and auto restart the task if it fails.

• Built in redundancy - Hadoop essentially duplicates data in blocks across data nodes ,and for every block there is assured to be a back-up block of same data existing somewhere across the data nodes. Master node keeps track of these node and data mapping ,and in case of any of the node fails the other node where back-up data block resides, takes over making the infrastructure failsafe[16].

A conventional RDBMS has the same concerns and uses terms like: persistence, backup and

recovery, These concerns scale upwards with Big Data[16].

• Automatic scale up/ down - Hadoop heavily relies on distributed file system and hence it comes with a capability of easily adding or deleting the number of nodes needed in the cluster.

• Move computation to data - Any computational queries are performed where the data resides . This avoid overhead required to bring the data to the computational environment[17] .

4.Hadoop components:

Let's take a look at two most important components that are the foundation to Hadoop framework.

4.1 Hadoop Distributed File System – HDFS:

HDFS is a distributed file system designed to run on commodity hardware. HDFS has a master/ slave architecture. See Figure 3.1. It's a write-once and read multiple times approach.

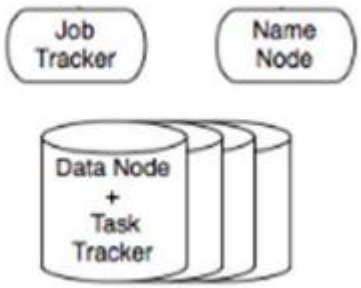

Figure II1.1 Hadoop cluster simplified visualization[18].

An HDFS cluster consists of a single Name Node (latest version 2.3.0 has redundant name node to avoid single point of failure), a master server machine that manages the file system and regulates access to the file system by the clients. There are multiple data nodes per cluster .

As shown in Figure 3.2 [19], data is split into blocks and stored on these data nodes .

Name Node maintains the map of data distribution. Data Nodes are responsible for data read and write operations during execution of data analysis.

Hadoop also follows concept of Rack Awareness., What this means is a Hadoop Administrator user

Can define which data chunks to save on which racks., This is to prevent loss of all the data if an Entire rack fails and also for better network performance by avoiding having to move big chunks of bulky data across the racks. This can be achieved by spreading replicated data blocks on the the machines on different racks[19].

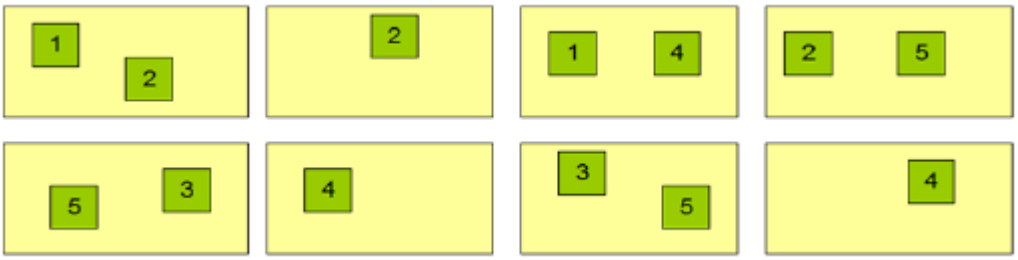

Figure III.2 Hadoop data replication on data nodes [19].

Refer Figure 3.3 [19], the Name Node and Data Node are commodity servers, typically Linux Machine Hadoop runs different software on these machines to make it a Name Node or a Data Node HDFS is built using the Java language, Any machine that Java can be run on can be Converted To act as the Name Node or the Data Node. A typical cluster has a dedicated machine That run only the Name Node software, Each of the other machines in the cluster runs one instance Of Data Node software, The Name Node manages all HDFS metadata[21].

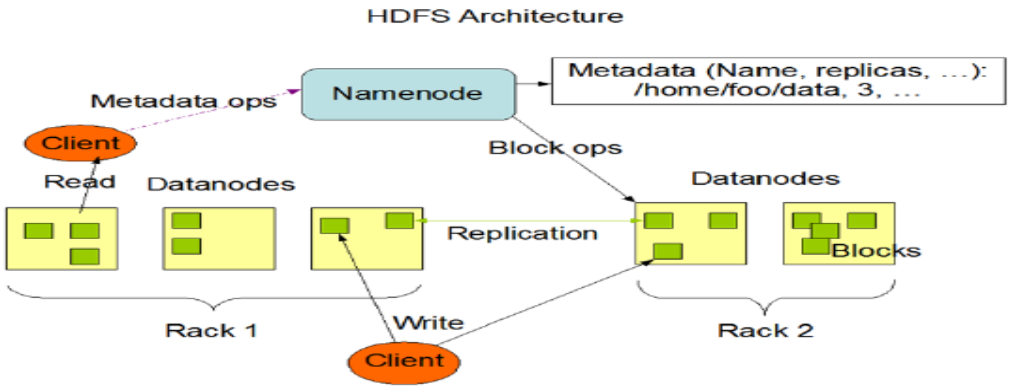

Figure III.3 Hadoop detailed architecture[21].

4.2 Map Reduce:

Map Reduce is a software framework introduced by Google to perform parallel processing on large datasets...assuming that large dataset storage is distributed over a large number of machines. Each machine computes data stored locally, which in turn contributes to distributed and parallel processing. There are two parts to such a computation - Map and reduce. Data nodes assigned to the Map phase, take raw input data and based on the type of computation required produce intermediate data that is stored locally, reduce nodes take these intermediate outputs and combine them to derive final output which is then stored in HDFS[19].

Hadoop tries to collocate data and the computation, name node with its knowledge of how the data is distributed, tries to assign the task to the node in which the data locally resides .

Programmers can write custom map and reduce functions and Map Reduce function automatically takes care of distributing and parallelizing the tasks across an array of commodity machines in the cluster underneath, It as well manages inter-machine communication leaving programmers to to focus on actual map-reduce functions Hadoop uses this fault tolerant, reliable ,distributed , parallel computing framework to analyze large datasets distributed over HDFS[19].

Both Map and Reduce functions operate on data conceptualized as key - value pairs.

- Map Phase - In the Map phase, each mapped reads raw input, record by record, and converts it into Key/Value pair and feeds it to the map function, depending upon how the user has defined the Map function, map function produces intermediate output in the form of new key/value pairs.

A number of such mappers located on the cluster parallel process raw data to produce a set of intermediate key/value pairs which are locally stored on each map per. Input data is split into Map tasks[20].

- map (k1, v1) -> k2, v2 (shuffle and sort - gathers all pairs with same key).
- Reduce Phase - merges all intermediate values associated with intermediate keys.
- reduce (k2, list(v2)) -> v3 (merge - combines together values for same keys – in case of queries

used for thesis - reduce will sum the values).

4.2 .1 Map Reduce jobs:

Map Reduce illustration with word count example [22]

○ Here we try to derive word frequency with Map Reduce program

○Assume two file inputs

■ file 1: “apple banana guava watermelon mango apple”
■ file 2: “mango kiwi guava cantaloupe mango”
○ We will illustrate the following operations using the Map Reduce algorithm
■ Map
■ Combine
■ Reduce
○ With a two-node cluster we would have two task nodes and that means two mappers available to distribute the first mapping task.
○ Map Phase I - split
■ mapper 1 takes file 1 as input
■ mapper 1 would produce following output in <key, value> format
<apple, 1>
<banana, 1>
<guava, 1>
<watermelon, 1>
<mango, 1>
<apple, 1>
■ mapper 2 takes file 2 as input
■ mapper 2 would produce following output
<mango, 1>
<kiwi, 1>
<guava, 1>
<cantaloupe, 1>
<mango, 1>
○ Map Phase II - combine
■ mapper 1 output
<apple, 2>
<banana, 1>
<guava, 1>
<watermelon, 1>
<mango, 1>
■ mapper 2 output
<mango, 2>
<kiwi, 1>
<guava, 1>
<cantaloupe, 1>
○ Reduce phase
■ Reducer will produce following final result
<apple, 2>

<banana, 1>
<guava, 2>
<watermelon, 1>
<mango, 3>
<kiwi, 1>
<cantaloupe, 1>
○ main method for Map Reduce Java program

■ This is in line with Hadoop version 1.2.1

4.2 .2 Map Reduce code:

```
public static void main(String[] args) throws Exception
{
JobConf conf = new JobConf(WordCount.class); //
Create a new job with the given configuration
conf.setJobName("wordcount");
conf.setOutputKeyClass(Text.class); //Set the key
class for the job output data.
conf.setOutputValueClass(IntWritable.class); //
Set the value class for job outputs.
conf.setMapperClass(Map.class);
conf.setCombinerClass(Reduce.class);
conf.setReducerClass(Reduce.class);
conf.setInputFormat(TextInputFormat.class);
conf.setOutputFormat(TextOutputFormat.class);
FileInputFormat.setInputPaths(conf, new
Path(args[0]));
FileOutputFormat.setOutputPath(conf, new
Path(args[1]));
JobClient.runJob(conf);
}
```

■ Map function

```
public static class Map extends MapReduceBase implements
Mapper<LongWritable, Text, Text, IntWritable> {
private final static IntWritable one = new
IntWritable(1);
private Text word = new Text();
public void map(LongWritable key, Text value,
OutputCollector<Text, IntWritable> output, Reporter
reporter) throws IOException {
String line = value.toString();
StringTokenizer tokenizer = new
StringTokenizer(line);
while (tokenizer.hasMoreTokens()) {
word.set(tokenizer.nextToken());
output.collect(word, one);
}
}
}
```

■ Reduce function

```
public static class Reduce extends MapReduceBase
```

```
implements Reducer<Text, IntWritable, Text, IntWritable> {
public void reduce(Text key, Iterator<IntWritable> values,
OutputCollector<Text, IntWritable> output, Reporter reporter)
throws IOException {
int sum = 0;
while (values.hasNext()) {
sum += values.next().get();
}
output.collect(key, new IntWritable(sum));
}
```

5. Conclusion:

In this chapter we present what is it Hadoop, Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. HADOOP is a framework with all the sub components like"MAP REDUCE" and "Hadoop Distributed File System".

.

CHAPTER VI :

IMPLEMENTATION

1.introduction :

After the description of big data analytics, At this chapter we will present the installation of the different components of our testing environment and perform a comparison based on the size f datasets between Hadoop Hdfs (K-means Map reduce algo) and windows

(K-means Weka).

2. System configuration:

Data sets description :

The 4 files (data) : Data1=300 lines.

Data 2=800 lines.

Data3 =1500 lines.

Data4=3000 lines.

Software components:

We installing Weka 3.7.12.JAR and Hadoop 2.3.

Machine hardware :

System windows 7 edition Integral .

Processeur: intel(r) core ™ i3 CPU 2.10GHz 2.10GHz.

Système d'expoltation : 64 bits .

3.Hadoop Installation steps:

To start install the Hadoop in windows , we must install the tools and understand the basic of Apache Hadoop. requires three essential steps:

Let's talk about the step number one: Installing and configuration Cygwin ,we need install Cygwin (open source software) because it provides optional packages which lets run and execute the age-old Unix-like commands right in our Windows system.

We install Cygwin from official site of Cygwin and download the executable depending on your system preference (32-bit or 64-bit) see *Figure II1. 1*

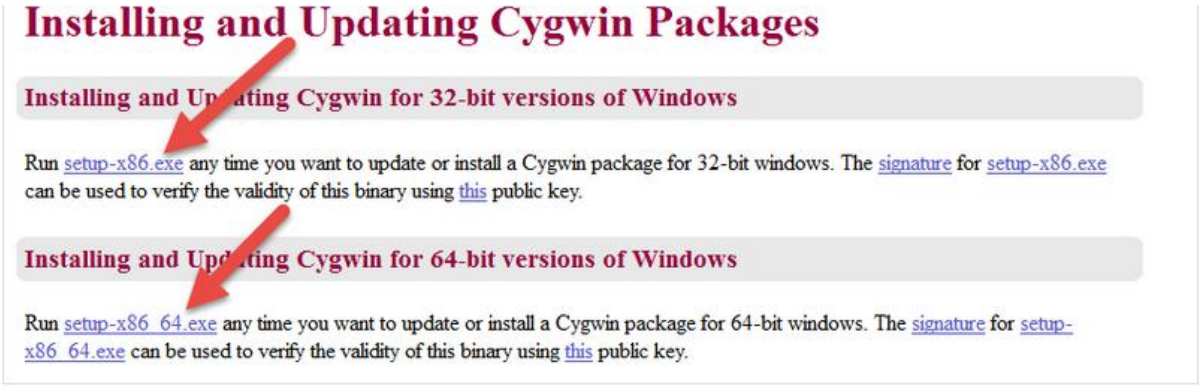

Figure II1. 1 website updating Cygwin[22].

Once the installation is complete, we can start using Cygwin by launching it using the desktop shortcut or from the start menu .see *Figure II1. 2*

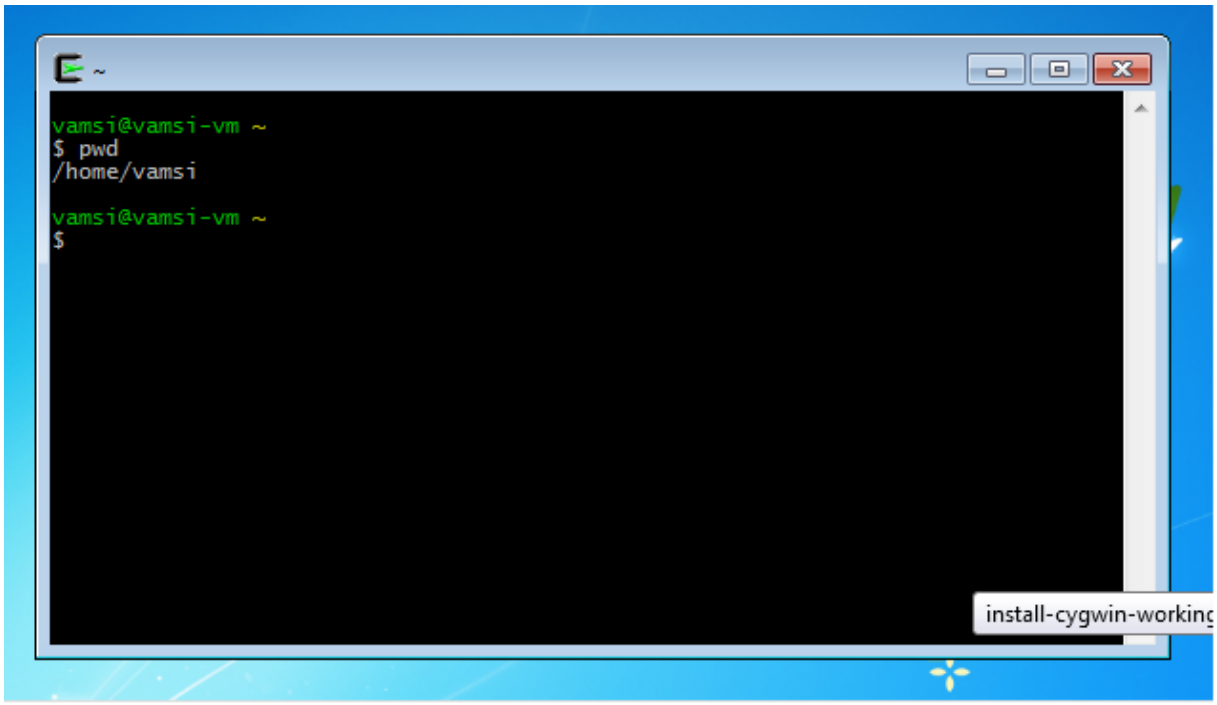

Figure II1. 2 Cygwin window.

The Step number two : Installing java JDK (Java Development kit) because it provides tools such as the java compiler used by IDEs for developing java programs. The JDK also contains a java Runtime Environment (JRE) which en able java programs, such as Eclipse to run our system.

you can find the JDK at http://www.technetwork/java/javase/download the correct version (automatically detect your system and offer to download correct version).

we install the java environment development (Eclipse IDE) from http:// www.eclipse.org/downlods. Eclipse is a general purpose technology platform[]

Open the Eclipse and start the application project .

But there are a few steps we need to pay attention, I would like to talk, I have done this steps after installing Cygwin and java jdk

Setting JAVA_HOME in Windows

Setting JAVA_HOME in Cygwin

Create server sshd in cygwin.

The step number three: Installing and configuration apache Hadoop 2.3.0.

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment.

- we Download Hadoop 2.3 for Windows (112.5 MB) from my box account - https://app.box.com/s/11fwozokqmc1ohttt117
- Also we Download the configuration file from my box account https://github.com/prabaprakash/Hadoop-2.3-Config/archive/master.zip see *Figure II1. 3*

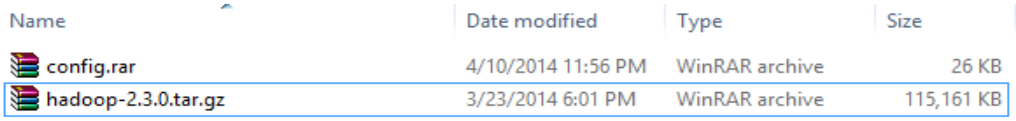

Figure II1. 3 Hadoop files.

After installing Hadoop we will create a HDFS in our system by format this command : see *Figure II1. 3*

```
c:\hadoop-2.3.0\bin>cd..

c:\hadoop-2.3.0>cd sbin

c:\hadoop-2.3.0\sbin>start-dfs.cmd
c:\hadoop-2.3.0\sbin>start-yarn.cmd
starting yarn daemons see figure
```

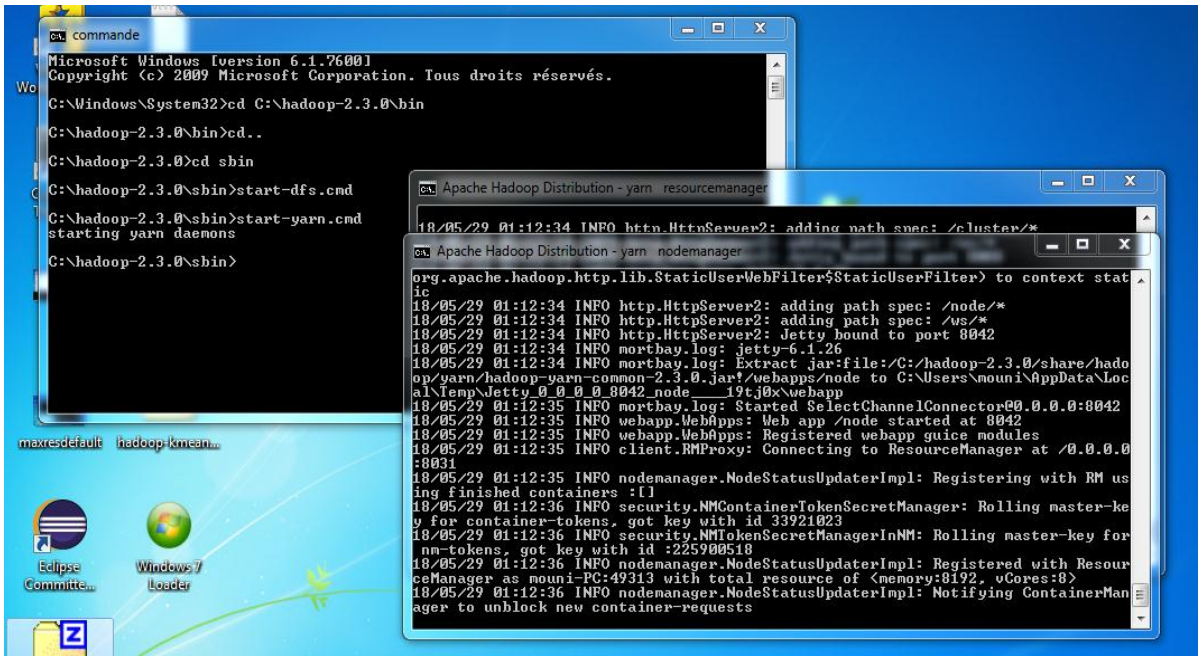

Figure II1.3 Hdfs screen.

4.Testing results:

K_means using Mapreduce:

The main function of the Mapreduce k- means program is :

MAP : put every point (data sets is about points have x and y) near to her centroids

Reduce: update centroids.

after that Compares between old centroids with the new centroids if they same ,he wil be

return again and the results of the experiment so verify ;This experiment is to compare times execution of 4 files(data) in Hadopp and weka .

We used Map reduce k-means implementation downloaded from link: https://github.com/jgalilee/hadoop-kmeans.

With k=3 and data sets as described in section2 .

The testing result is as follow

---------- Hadoop part ----------

data-sets---------- Execution time (second) ---- nbr cluster ---- nbr of points

data1 ----------> 0,087 ------ 3 ---- 300

data2----------> 0,048 ------ 3 ---- 800

data3----------> 0,569 ------ 3 ---- 1500

data4----------> 0,542 ------ 3 ---- 3000

We used k-means weka implementation downloaded from link:

https://spirceforge.net/projects/weka/files/weka3-7/3.7.12/

With k=3 and data sets as described in section2 .

The testing result is as follow

----------Weka part ----------

data-set----------Temps d'exécution (seconde) ---- nbr cluster ---- nbr de points

data1 ----------> 0.266 ------ 3 ---- 300

data2----------> 0.337 ------ 3 ---- 800

data3----------> 0.730 ------ 3 ---- 1500

data4----------> 0,500 ------ 3 ---- 3000

4.1 Comparative curve:

Test the files of 300, 800,1500 and 3000 respectively, and the result is shown in table

Figure IV.1

the test results show that the execution time of the data sets ,weka part : increases with the increase of data size, When the file is relativel big, the execution time increases ,

Hadoop part :with tha same of the file size the execution time faster than weka.

Data set	Nombre points	Nombre clusters	Exécution time K_means	
			Weka	Hadoop
Data 1	300	3	**0.266**	**0,087**
Data 2	800	3	**0.337**	**0,048**
Data 3	1500	3	**0.730**	**0,569**
Data 4	3000	3	**0,500**	**0,542**

Figure IV.1 Execution time of data sets.

Figure IV.2 Comparative graph .

Figure IV.2 test results show that the execution of files from 1 to 4, the execution time increased faster while the size file big but at weka is slowly ,then Hadoop results.

5. Conclusion:

Hadoop, as an open source project for the Apache foundation, is a distributed computing framework that deals with large amounts of data and has been widely used in the Internet industry. The purpose of this experiment is to study the method of building Hadoop platform and to study the performance of test platform.

for materiel reason that not getting the result that Hadoop is the best and fast in execution of big data.

Conclusion:

Firstly In the chapter one we talk about the Big data is an evolving term that describes any voluminous amount of structured, semi structured and unstructured data that has the potential to be mined for information.

In the chapter two we present the Big Data Analytics is the process of examining large data sets containing a variety of data (types Big Data) to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful information, Companies and enterprises that implement Big Data Analytics often reap several business benefits.

In the chapter three we talk about Hadoop is an open source software framework for storing data and launching applications on standard machine clusters. This solution offers massive storage space for all types of data.

Finely in the chapter of implementation we compared the results between Hadoop Hdfs and not Hdfs (Weka), we found expected results because usually Hadoop is the fastest in analyzing huge data.

Bibliography :

[1] https://www.kdnuggets.com/2017/02/origins-big-data.html.

[2] https://www.kdnuggets.com/2017/02/what-is-big-data.html.

[3] Wikipedia. Big data, 2014. http://en.wikipedia.org/wiki/Big_data, accessed April 2014.

[4] M. Stonebreaker, P. Brown, and D. Moore. Object-relational DBMSs, tracking the next great wave. Morgan Kauffman Publishers, Inc., San Francisco, California, 2 edition ,1998.

[5] http://searchcloudprovider.techtarget.com/feature/Big-data-analysis-in-the- cloud-Storage-network-and-server-challenges.

[6] For Big Data Analytics There's No Such Thing as Too Big The Compelling Economics and Technology of Big Data Computing, White Paper, March 2012.

[7]http://www.infoworld.com/d/cloud-computing/amazons-redshift-big-data-analytics-the-pros-and-cons-213049.

[8] https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics.

[9] https://www.whizlabs.com/blog/big-data-analytics-importance/. April 2014

[10] Everitt B S Cluster Analysis Heinemann ,1974 .

[11] Hartigan, J. A.. Clustering algorithms. New York: John Wiley and Sons, 1975.

[12] Krzanowski W J principles of Multivariate Analysis Oxford University Press, 1990.

[13] Hartigan J A and Wong M A Algorithm AS136: A K-means clustering algorithm Appl. Statist. 28 100–108 , mai 1979.

[14] K. Kline, D. Kline, and B. Hunt. SQL in a nutshell, a desktop quick O'Reilly Media ,Sebastopol , California edition ,2008.

[15] Apache Hadoop. HDFS Architecture Guide, 2013.

[16] Apache Hadoop. MapReduce Tutorial, 2013.

[17] Hadoop Map Reduce CookBook - Srinath Perera ,2012.

[18] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html, accessed April 2014.

[19] Greenplum. A unified engine for RDBMS and Map Reduce, 2009.

[20] https://bigdataarchitecture.com/ April 2014.

[21] Hadoop: The Definitive Guide MapReduce for the Cloud,2009.

ملخص :

الهدف من عملنا استكشاف تقنية البيانات (Hadoop distribted نظام ملفات) Hdfs ومقارنة تحليلات البيانات مثل تجميع الملفات الضخمة الموزعة وغير الموزعة.

ولهذا الغرض قمنا باستخدام خوارزميةk_means clustering في نظام hadoop وايضا في Weka.

كلمات المفاتيح : Hadoop Hdfs ,Mapreduce , Cluster.

Abstract:

The objective is explore big data technology uses Hdfs (Hadoop distribted files system)

and compare data analytics such as clustering distributed Hdfs and non distributed

For this purpose we have used an algorithm k-means clustering in hadoop and wika .

Keys word : Hadoop Hdfs ,Mapreduce,Cluster.

Résumé :

L'objectif est d'explorer la technologie Big Data utilise Hdfs (système de fichiers Hadoop distribués) et comparer les données analytiques telles que la mise en cluster distribuée Hdfs et non distribuée.

A cet effet nous avons utilisé un algorithme k-means clustering dans hadoop et wika.

Mot clés: Hadoop Hdfs, Mapreduce, Cluster.