Big Data and Big Data Analytics: Concepts, Types and Technologies

Article · November 2018

DOI: 10.21276/ijre.2018.5.9.5

International Journal of Research and Engineering
ISSN: 2348-7860 (O) | 2348-7852 (P) | Vol. 5 No. 9 | September-October 2018 | PP. 524-528
Digital Object Identifier DOI® http://dx.doi.org/10.21276/ijre.2018.5.9.5

Big Data and Big Data Analytics: Concepts, Types and Technologies

Author(s) : [1]Youssra Riahi, [2]Sara Riahi
Affiliation(s): [1]Faculty of Informatics,
International University of Rabat, Technopolis parc, Morocco
[2] Department of Mathematics and Computer Science, Faculty of Sciences,
University of Chouaib Doukkali, Jabran Khalil Jabran Avenu , Morocco
Corresponding Author: riahiyoussra3@gmail.com

ORIGINAL ARTICLE

Abstract - Nowadays, companies are starting to realize the importance of data availability in large amounts in order to make the right decisions and support their strategies. With the development of new technologies, the Internet and social networks, the production of digital data is constantly growing. The term "Big Data" refers to the heterogeneous mass of digital data produced by companies and individuals whose characteristics (large volume, different forms, speed of processing) require specific and increasingly sophisticated computer storage and analysis tools. This article intends to define the concept of Big Data, its concepts, challenges and applications, as well as the importance of Big Data Analytics

Keywords: *Big Data; Big Data Analytics; Hadoop; Internet; Security*

I. INTRODUCTION

The digital data produced is partly the result of the use of devices connected to the Internet. Thus, smartphones, tablets and computers transmit data about their users. Connected smart objects convey information about consumer's use of everyday objects.

Apart from the connected devices, data come from a wide range of sources: demographic data, climate data, scientific and medical data, energy consumption data, etc. All these data provide information about the location of users of the devices, their travel, their interests, their consumption habits, their leisure activities, and their projects and so on. But also information on how the infrastructure, machinery and apparatus are used. With the ever-increasing number of Internet and mobile phone users, the volume of digital data is growing rapidly. Today we are living in an Informational Society and we are moving towards a Knowledge Based Society. In order to extract better knowledge we need a bigger amount of data. The Society of Information is a society where information plays a major role in the economical, cultural and political stage. [1]

II. WHAT IS BIG DATA ?

A. *Definition*

The term "Big Data" refers to the evolution and use of technologies that provide the right user at the right time with the right information from a mass of data that has been growing exponentially for a long time in our society. The challenge is not only to deal with rapidly increasing volumes of data but also the difficulty of managing increasingly heterogeneous formats as well as increasingly complex and interconnected data.

Being a complex polymorphic object, its definition varies according to the communities that are interested in it as a user or provider of services. Invented by the giants of the web, the Big Data presents itself as a solution designed to provide everyone a real-time access to giant databases.

Big Data is a very difficult concept to define precisely, since the very notion of big in terms of volume of data varies from one area to another. It is not defined by a set of technologies, on the contrary, it defines a category of techniques and technologies. This is an emerging field, and as we seek to learn how to implement this new paradigm and harness the value, the definition is changing. [2]

1) *Characteristics of Big Data*

The term Big Data refers to gigantic larger datasets (volume); more diversified, including structured, semi-structured, and unstructured (variety) data, and arriving faster (velocity) than before. These are the 3V.

a) 3V:

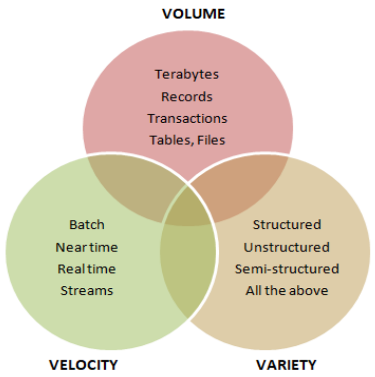

Fig. 1. *3V Concept*

-Volume: represents the amount of data generated, stored and operated within the system. The increase in volume is explained by the increase in the amount of data generated and stored, but also by the need to exploit it.

-Variety: represents the multiplication of the types of data managed by an information system. This multiplication leads to a complexity of links and link types between these data. The variety also relates to the possible uses associated with a raw data.

-Velocity: represents the frequency at which data is generated, captured, and shared. The data arrive by stream and must be analyzed in real time.

b) 5V:

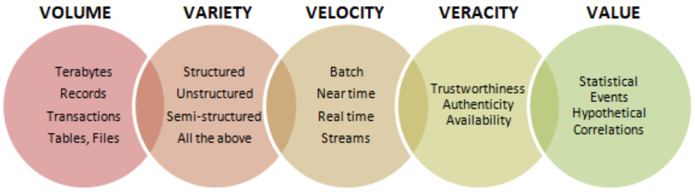

Fig. 2. *5V Concept*

To this classical characterization, two other "V"s are important:
-Veracity: level of quality, accuracy and uncertainty of data and data sources.
-Value: the value and potential derived from data.

III. WHAT IS BIG DATA ANALYTICS ?

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data. [3]
The analysis of structured data evolves due to the variety and velocity of the data manipulated. Therefore, it is no longer enough to analyze data and produce reports, the wide variety of data means that the systems in place must be capable of assisting in the analysis of data. The analysis consists of automatically determining, within a variety of rapidly changing data, the correlations between the data in order to help in the exploitation of it.

Big Data Anlytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways. [4]

A. Types of Big Data Analytics

a) Descriptive Analytics

It consists of asking the question: What is happening?
It is a preliminary stage of data processing that creates a set of historical data. Data mining methods organize data and help uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.

b) Diagnostic Analytics

It consists of asking the question: Why did it happen?
Diagnostic analytics looks for the root cause of a problem. It is used to determine why something happened. This type attempts to find and understand the causes of events and behaviors.

c) Predictive Analytics

It consists of asking the question: What is likely to happen?
It uses past data in order to predict the future. It is all about forecasting. Predictive analytics uses many techniques like data mining and artificial intelligence to analyze current data and make scenarios of what might happen.

d) Prescriptive Analytics

It consists of asking the question: What should be done?
It is dedicated to finding the right action to be taken. Descriptive analytics provides a historical data, and predictive analytics helps forecast what might happen. Prescriptive analytics uses these parameters to find the best solution.

IV. HADOOP FOR BIG DATA APPLICATIONS

Big Data are collections of information that would have been considered gigantic, impossible to store and process, a decade ago.
The processing of such large quantities of data imposes particular methods. A classic database management system is unable to process as much information. Hadoop is an open

source software product (or, more accurately, ‘software library framework’) that is collaboratively produced and freely distributed by the Apache Foundation – effectively, it is a developer’s toolkit designed to simplify the building of Big Data solutions. [5]

Hadoop is used by companies with very large volumes of data to process. Among them are web giants such as Facebook, Twitter, LinkedIn, eBay and Amazon. Hadoop is a distributed data processing and management system. It contains many components, including: HDFS, YARN,Map Reduce. HDFS is a distributed file system that provides high-performance access to data across Hadoop clusters. [6] MapReduce is a core component of the Apache Hadoop software framework. Hadoop enables resilient, distributed processing of massive unstructured data sets across commodity computer clusters, in which each node of the cluster includes its own storage. MapReduce serves two essential functions: It parcels out work to various nodes within the cluster or map, and it organizes and reduces the results from each node into a cohesive answer to a query.[7]

Hadoop relies on two servers:

JobTracker: there is only one JobTracker per Hadoop cluster. It receives Map/Reduce tasks to run and organizes their execution on the cluster.When you submit your code to be executed on the Hadoop cluster, it is the JobTracker’s responsibility to build an execution plan. This execution plan includes determining the nodes that contain data to operate on, arranging nodes to correspond with data, monitoring running tasks, and relaunching tasks if they fail.[8]

•**TaskTracker**: several per cluster. Executes the Map/Reduce work itself (as a Map and Reduce task with the associated input data).
The JobTracker server is in communication with HDFS; it knows where the Map/Reduce program input data is and where the output data must be stored. It can thus optimize the distribution of tasks according to the associated data.

To run a Map/Reduce program, we must:
•Write input data in HDFS
•Submit the program to the cluster's JobTracker.
•Retrieve output data from HDFS.

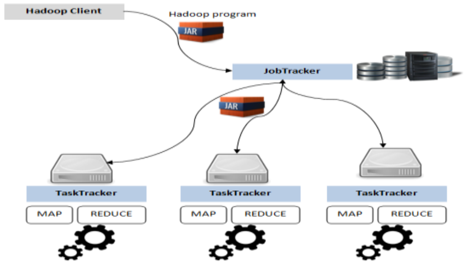

Fig. 3. Hadoop Architecture

All TaskTrackers report their status continuously through heartbeat packages. If a TaskTracker fails (missing heartbeat or failed task), the JobTracker notifies the redistribution of the task to another node.

HDFS relies on two servers:
•NameNode: unique on the cluster. It stores information about file names and their characteristics. It is the master of the HDFS that controls slave DataNode.
•Secondary NameNode: The Secondary NameNode monitors the state of the HDFS cluster and takes “snapshots” of the data contained in the NameNode. If the NameNode fails, then the Secondary NameNode can be used in place of the NameNode.[9]
•DataNode: multiple by cluster. Stores the contents of the files themselves, fragmented into blocks (64KB by default)

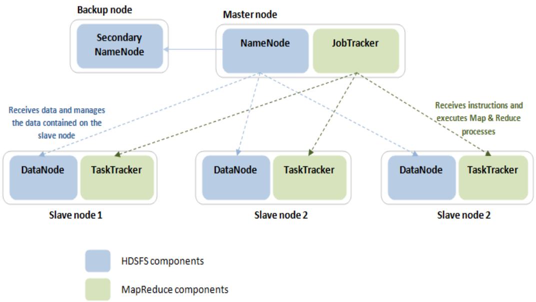

Fig. 4. General Architecture

V. MAP REDUCE CONCEPT

MapReduce is a Java environment for writing programs intended for YARN. Java is not the simplest language for this, there are packages to import and class paths to provide. The data exchanged between Map and Reduce, in the entire job are pairs (key, value):
-Key: it is any type of data: integer, text. . .
-Value: it is any type of data.
The two functions Map and Reduce receive and send such pairs.

A. Map

The Map function receives an input pair and can produce any number of pairs in output: none, one or more. The types of inputs and outputs are as desired. This very little constraint specification allows so many things. In general, the pairs Map are constituted as follows:
• The text value is one of the rows or one of the n-tuples of the file to be processed
• The key of type integer is the position of this line in the file.
YARN launches a Map instance for each row of each file in the data to be processed. Each instance processes the row it has been assigned and produces output pairs.

B. Reduce

The Reduce function receives a list of input pairs. These are the pairs produced by the instances of Map. Reduce can produce any number of output pairs, but most of the time it is one. On the other hand, the crucial point is that the input pairs processed by an instance of Reduce all have the same key.

YARN launches a Reduce instance for each different key that Map instances have produced, and provides only the pairs with the same key. This is what makes it possible to aggregate values. Generally, Reduce must do a processing on the values, such as sum all values between them, or determine the largest of the values. . .

When designing a MapReduce treatment, we must think about the keys and values necessary so it can works. Reduce tasks receive a list of pairs with the same key and produce a pair that contains the expected result. This output pair can have the same key as the input.

C. Steps for a MapRedcue job

1.Preprocessing of input data, eg: decompression of files

2.Split: Separate data into separately processable blocks and formatted (key, value),

eg in rows or tuples

3.Map: application of the map function on all the pairs (key, value) formed from the input data, this produces other pairs (key, value) output

4.Shuffle & Sort: redistribution of data so that the pairs produced by Map having the same keys are on the same machines

5. Reduce: Aggregation of pairs with the same key to get the final result.

a) Schema explanation

1. At the beginning, YARN will inquire about the location of the data using the Name node and have them decompress if necessary by the Data nodes concerned.

2. The Split phase consists of constructing pairs (n° of n-tuple, n-tuple) to be provided to the Map tasks.

3. YARN creates Map processes on each machine containing part of the data and provides them with the pairs of their machines successively.

4. Each Map task analyzes its data and provides or not a pair. It can consist of converting strings into numbers, making calculations, and so on.

5. YARN sorts the pairs leaving Map according to their key and sends them to the machine that runs the Reduce task concerned by this key.

6. The Reduce tasks receive a list of pairs and perform the reduction of the values (max, sum, avg…). They emit only the final value.

b) Schema

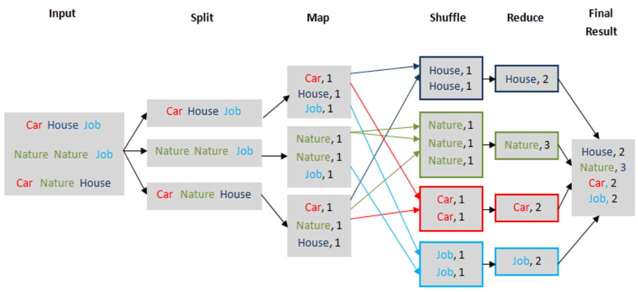

Fig. 5. MapReduce word count process

VI. MAP REDUCE CONCEPT

YARN (Yet Another Resource Negociator) - also called MapReduce 2.0 (MRv2) - which takes the place of MapReduce. YARN is placed on top of HDFS (Hadoop Destributed File System) to provide operating system capabilities for Big Data analytics applications. This oranization allows simultaneous execution of multiple applications while providing better tracking of the data throughout its life cycle. It also allows to mix workloads in batch, interactive and in real time.

YARN also maintains compatibility with MapReduce's Application Programming Interface (API), requiring only a recompile of the applications already developed.

The main difference is the separation of resources management (ResourceManager - RM) and task or application control (ApplicationMaster - AM) into two daemons. AM is a framework with a specific library that negotiates RM resources with the NodeManager (NM) to run and monitor tasks.

YARN (Yet Another Resource Negotiator) is a mechanism for managing jobs on a cluster of machines. YARN allows users to launch Map-Reduce jobs on data in HDFS and monitor their progress, retrieve the messages (logs) displayed by the programs. Eventually, YARN can move a process from one machine to another in the event of a failure or of advancement judged too slow.

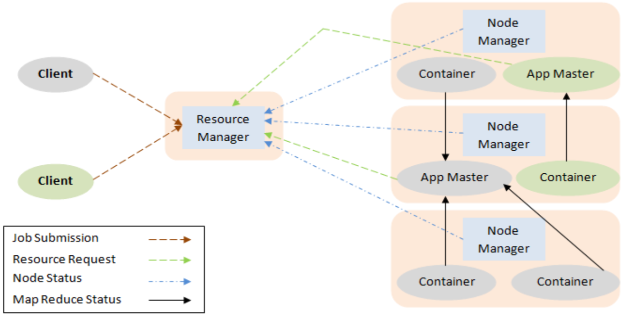

Fig. 6. General Architecture

•**RM (Resource Manager**): The central daemon of YARN. It manages resources assignments (CPU/Memory) when it comes to applications. It has two components: a scheduler which is in charge of resources allocation to the running application but it doesn't ensure restarting in case of task failure. The second component is the Application Manager which is in charge of App Masters management in the cluster. It ensures restarting of application masters on different nodes in case of failure.
•**NM (Node Manager):** The slave daemon of YARN. NM is responsible for containers monitoring their resource usage and reporting the same to the RM [10]. NM tracks the status of the node on which it is running.
•**AM (Application Master):** There is only one application master per application. It negotiates resources from the RM and works with the NM. It manages the application life cycle. The AM acquires containers from the RM's scheduler before contacting the corresponding NMs to start the application's individual tasks. [11]

YARN is an evolution of the architecture of Hadoop allowing to unload the JobTracker which tended to accumulate too many roles and thus became complex. This rethinking of the roles allowed also to decouple Hadoop from Map Reduce and, in so doing, to no longer remain bounded to MapReduce. This will allow Hadoop, in addition to better scalability, to be enriched by new frameworks covering needs with little or no coverage with Map Reduce.

VII. Conclusion

Big data refers to the set of numerical data produced by the use of new technologies for personal or professional purposes. Big Data analytics is the process of examining these data in order to uncover hidden patters, market trends, customer preferences and other useful information in order to make the right decisions. Big Data Analytics is a fast growing technology. It has been adopted by the most unexpected industries and became an industry on its own. But analysis of these data in the framework of the Big Data is a process that seems sometimes quite intrusive.

Analytics is a data science. BI takes care of the decision-making part while Data Analytics is the process of asking questions. Analytics tools are used when company needs to do a forecasting and wants to know what will happen in the future, while BI tools help to transform those forecasts into common language [12]. More often, Big Data is considered as the successor to Business Intelligence. This comparison will be discussed in a future work.

REFERENCES

[1] Perspectives on Big Data and Big Data Analytics-Database Systems Journal vol. III, no. 4/2012

[2] The Big Data Revolution, Issues and Applications, Azzeddine Riahi, Sara Riahi- IJARCSSE, Volume 5, Issue 8

[3] Deep learning applications and challenges in big data analytics-Najafabadi et al. Journal of Big Data (2015) 2:1 DOI 10.1186/s40537-014-0007-7

[4] BIG DATA ANALYTICS: CHALLENGES AND APPLICATIONS FOR TEXT, AUDIO, VIDEO, AND SOCIAL MEDIA DATA-International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1, February 2016

[5] Big Data- The definitive guide to the revolution in business analytics-Fujitsu

[6] Khttp://searchbusinessanalytics.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS

[7] http://searchcloudcomputing.techtarget.com/definition/MapReduce

[8] http://www.informit.com/articles/article.aspx?p=2008905

[9] http://www.informit.com/articles/article.aspx?p=2008905