BIG DATA AS A SOURCE OF STATISTICS

Aamir Siddiqui,
Data Scientist,
Mumbai, India

Abstract:

Over the last two decades the internet has grown exponentially and generating huge amount of data in peta bytes to zetta bytes which comes mostly in unstructured form. The IoT is generating huge amount of data every day. These massive data can be stored and analysed using big data tools like Hadoop, Cassandra, Kafka, MongoDB, Spark. The objective of this paper is to show how to store and process the big data that will give better statistics and understanding about the population in consideration.

Keywords: Big Data, Zetta Bytes, IoT, Hadoop, MongoDB, Spark.

I Introduction

The demand for rapid statistical services could be met by leveraging the emerging sources of Big Data, such as those relating to sensors, transactional and social media data [1]. The term Big Data has no formal definition but is usually defined as the data sets that are too large that cannot be process by traditional relational database to analyse and draw meaningful conclusion. It comes in structured, semi-structured and unstructured form from different sources in different sizes from peta bytes to zetta bytes. Big data come from social media, online transactions, sensors used in Internet of things (IoT), audio, video, pictures, internet click stream logs in a real time. There are five V's in Big Data which are described as follows:Volume: It comes in huge volume. This can be data of Twitter feeds, clickstreams or IoT.

Velocity: It is the rate at which data is received and processed.

Variety: It refers to different types of data that are available on the internet.

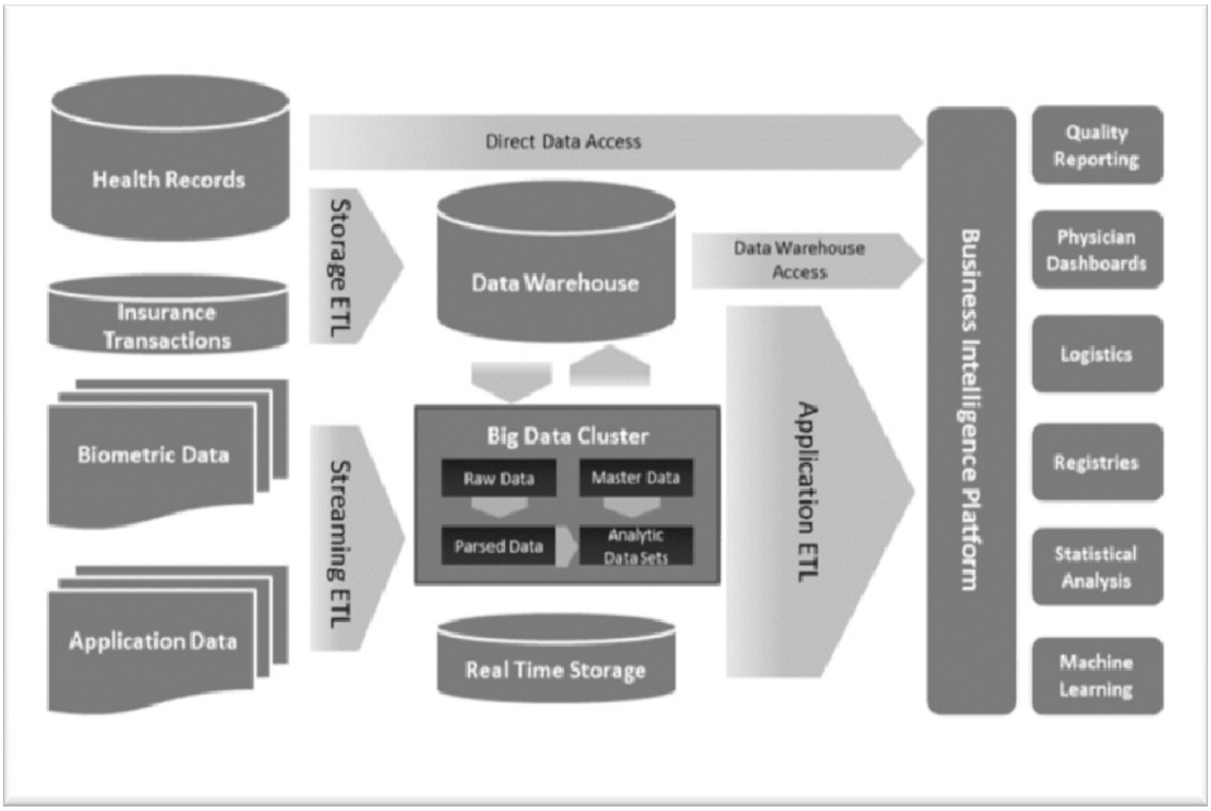
Value: How to extract useful data from the available data and Veracity: It is the quality of the data.

The global big data and business analytics market was valued at 169 billion U.S. dollars in 2018 and is expected to grow to 274 billion U.S. dollars in 2022[2].

II Big Data Storage and Processing Tools

In last few years the size of the data has grown exponentially. The challenge is to store the growing data for processing and analysis. Technologies such has Apache Hadoop, MapReduce, MongoDB, Spark, Kafka can be used to collect and process data. These technologies can process structured semi-structured and unstructured data. There are three types of data processing techniques batch processing stream processing and interactive analysis. In batch processing data is collected over a defined time and return the result when computation is completed for example Apache Hadoop. In streaming model data is processed as and when it is generated and is done in real time for example Apache Storm and Splunk. The interactive analysis allows to directly interact in real time for example Apache Drill.

The architecture of the big data is (Figure 1)



Source: [4]

Apache Hadoop

Apache Hadoop is an open-source framework designed for distributed storage and processing of large data sets across clusters of computers. It has a Hadoop Distributed File System (HDFS) for storage. HDFS breaks up files into chunks and distributes them across the cluster. It has Yarn for job scheduling and cluster resource management.

MapReduce

It is used for parallel processing. MapReduce refers to two separate and distinct tasks that Hadoop programs perform. The first is to map the job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples. The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job[3].

Apache Mahout

It is meant for machine learning. It produces scalable machine learning algorithms, extracts relationship from the data sets. It runs on Apache Hadoop using the

MapReduce application. Its algorithms are clustering, classification, pattern mining, collaborative filtering, dimensionality reduction and regression.

Apache Spark

Apache Spark is an open-source distributed cluster-computing framework. Spark is a data processing engine that provides faster analytics than MapReduce. It's in memory processing is faster when compared to Hadoop. Spark is faster as everything is done here in memory. It provides live data streaming processing. It has a Resilient Distributed Dataset (RDD). This is where Spark does most of the operations such as data transformation. Spark core is the basic building block of Spark. It performs job scheduling, memory operations, fault tolerance. Spark SQL allows querying data via SQL. It has separate library for Machine Learning MLlib. It has a library for visualization GraphX.

Apache Storm

It is an open source distributed real-time computational system for processing live data. It can process over a million jobs in a second. It is easy to integrate with any programming language. It has two nodes master node and the worker node. The master node runs "Nimbus" for distributing codes and assigning tasks to machines. Worker nodes run Supervisor that handle one or more worker processes on their nodes. It cannot manage its cluster so it depends on Apache Zookeeper. Zookeeper facilitates communication between Nimbus and Supervisors.

III Big Data Challenges

Big data is so big that organizations are struggling to store the data in the existing environment. Every year some new technologies are coming up to meet the challenges of storing big data like Apache Spark, MongoDB to name a few. This helps the organization to the store huge volume of data which is generating every second. Another big challenge is data mining to separate useful data from the entire dataset. A lot of time goes into data cleaning and data wrangling. Data Scientist from Data Science department of organization helps to clean and filter data and apply Artificial Intelligence and Machine Learning algorithms to develop Deep Learning and Predictive Models.

IV Big Data Analytics and Statistics

The more information we have the better statistics we will get to deliver the best service. The Big Data analytics are used in industries such as Banking and Finance, Media and Entertainment, Education, Government, Administration, Healthcare, Retail, Airlines, Manufacturing, Agriculture, Telecom, Marketing. It can be used to discover hidden patterns, correlation, trends and other information. Data Scientists can use the big data to make machine learning model to make more informed decision. Let's take an example of Banking Industry to understand the potential of Big Data. Millions of dollars are lost to online theft which affects bank monetarily and also damage its reputation as a risky bank. Transactional data can be used to detect and prevent fraudulent transaction in a bank. By preventing fraud huge amount of bank's money can be saved. Demographic data can be used to better understand customers and recommend different banking products to different customers based on the customer segmentation algorithm. It can be used for loan recommendation and identify which customer can repay loan and which customer cannot repay this will bring down NPA of the bank. Big data will also help bank to identify Internal Fraud which is usually done by the staff of the bank. Big data can help to prevent fraud in many departments of the bank like CBS, ATM, Credit Card, Debit Card, Loan and Advances, Internet Banking, Mobile Banking, Cheques, E-Wallet, Internal Employee Fraud, Treasury, Trade Finance.

V Conclusion

Data is generating at a very high velocity in huge volume. This is the right time to make use of big data using statistics and machine learning algorithms to uncover patterns and trends in the data sets to help organizations and governments to deliver better services. With proper tools and architecture in place big data can be analysed and turned into gold for the organization.

References

1. BigData UN Global Working Group (2019) https://unstats.un.org/unsd/bigdata/taskteams/si-gsd/default.asp

2. Shanhong Liu, Aug 9, 2019, Statista, https://www.statista.com/statistics/254266/global-big-data-market-forecast/

3. IBM(2019): https://www.ibm.com/analytics/hadoop/mapreduce

4. Austin, Christopher & Kusumoto, Fred. (2016). The application of Big Data in medicine: current implications and future directions. Journal of interventional cardiac electrophysiology: an international journal of arrhythmias and pacing. 47. 10.1007/s10840-016-0104-y.)

СТАТИСТИЧНА ЦІННІСТЬ СИМБІОЗУ ШТУЧНОГО ІНТЕЛЕКТУ ТА BIG DATA

Антоненко Ярослав Олексійович,
аспірант,
Національна академія статистики, обліку та аудиту

У час надшвидкого розвитку інформаційно-комунікаційних технологій та темпів зростання обсягів інформаційних ресурсів, актуальність і цінність використання Big Data (великих даних) та штучного інтелекту (ШІ) неможливо переоцінити. Всі інформаційні гіганти наших часів вкладають величезні кошти в використання і розвиток Великих даних та ШІ. Про це свідчать опитування IT директорів різних компаній, проведені компанією Gartner. Усі респонденти повідомили, що в організаціях, де вони працюють, вже використовується або буде використовуватися ШІ. Також, згідно з цим