Journal of Theoretical and Applied Information Technology
30th June 2019. Vol.97. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HADOOP AND BIG DATA CHALLENGES

[1]ABEDALLAH ZAID ABUALKISHIK

[1]American University in the Emirates, College of Computer and Information technology,
Dubai, United Arab Emirates.
E-mail: [1]abedallah.abualkishik@gmail.com

ABSTRACT

Today's technologies and advancements have led to eruption and floods of daily generated data. Raw data has no value if it's not analyzed to extract the hidden insight for business organization. Big data is heterogeneous, unstructured, and enormous. Collecting, storing, manipulating, interpreting, analyzing and visualizing Big data shape the dimensions of Big Data life cycle. Big data deals most of time with unstructured data that require real time and batch processing. The goal of any big data platform is to extract correlations, hidden sentiments, patterns, values, and insights of these raw data. However, Big data analytics pipeline is end-to-end challenging.

The paper objectives are of three-folds: Revisit the big data concept, dimensions and it is characteristics. Second, it aims to introducing Hadoop open source big data platform and the supportive utilities. Third, the paper aims to study the underlying challenges that surround Big data pipeline end to end.

Keywords: *Big Data, Big Data Pipeline, Big Data V's. Hadoop Platform, Challenges*

1. INTRODUCTION

While we are living in the era of Industry 4.0 that floods data in unprecedented way. 90% of today's world data volume was created in the last two years at average of 2.5 quintillion bytes of data daily [1]. The current exponential growth of data imposes the necessity for a new mechanism and tools to handle them. Therefore, the term Big Data was coined to handle and process massive and complex amount of structured, unstructured and heterogenous data at rest and data in motion that inundate a business daily, which usually cannot be handled using traditional data management processing applications.

The British Mathematician Clive Humby in 2006 said "Data is the new oil" embodied the fact that data need processing just like oil needs refining. The World Economic Forum [2] considered data as an assets and source of power just like oil. Ginni Rometty, IBM CEO addressed the Council of Foreign Relations in 2013 "I want you to think about data as the next natural resource.". Therefore, Big data gained lots of interest due to the expected high ROI exemplified by local governmental plans [3]. Tens of articles were published on Forbes, Economist, and New York Times to address this term and its expected value. Gartner, Inc. defined Big data as: "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" [4]. Utilizing Big data for any business organization offers cost reduction, competitive advantages, and allow for new business opportunities.

The new growth of data volume rises and shapes a new economy that is totally reliant on big data processing on real time or near real time. For example, autonomous car, delivery by drone, Twitter, Google, Netflix, Uber, Facebook. The Economist [5] pointed out that "Data are to this century what oil was to the last one" Data has changed the mechanism of economy via creating new infrastructure, new business model, new politics and monopolies, and ultimately new economy. Data is unlike any another resource in nature in the way of extracting, refining, processing, buying and selling. Thus, it requires special regulations, processing and treatments.

Designing a comprehensive and efficient Big data platform requires deep understanding of three components: Data, user needs, and the underlying technologies. Big data analyses data to discover

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

pattern and insights that would ultimately improve decision making, strategic planning, business model, and reduce time, effort and cost. When Big data is powered with analytical tools then it is possible to determine causes of failure and defects in real time and detecting and predicting defects and probably preventing them. Big data has been implemented successfully at several areas as: Tourism [6], [7], health care [8], support decision making [9].

The notion of supporting Big data started whit Google when they first designed Google File system that is also known as BigTable. Chang et al. [10] defined BigTable as "Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers.".

Hu et al. [11] used a systems-engineering approach to described Big data analytics into four phases as follows: Data generation: concerns with the mechanism of generating data, i.e. click streams and sensors. Data acquisition: refer to the process of obtaining and collecting data from raw various sources including Relational Data Base Management Systems (RDBMS), Not only SQL (NoSQL), document store, web scraping, data lakes etc. Then, data transmission mechanism to transmit raw data into a suitable storage. After that, perform some data pre-processing as cleansing and filtering data, and dealing with sparse data. Data storage concerns with insistently storing data at the hardware level and managing data at the software level. Finally, data analysis concerns with extracting insights and values out of the data via various methods and tools. Analysis can be classified into several categories: Structured data analysis, text analysis, web analysis.

This paper is organized as follows. Section two describe Big data analytics, section three introduces Hadoop Big data platform and its limitations, section 4 presents Big data pipeline and its surrounding challenges, section five ends the paper with a summary of conclusion.

This guide provides details to assist authors in preparing a paper for publication in JATIT so that there is a consistency among papers. These instructions give guidance on layout, style, illustrations and references and serve as a model for authors to emulate. Please follow these specifications closely as papers which do not meet the standards laid down, will not be published.

2. BIG DATA ANALYTICS

Humans started to store information since the very early civilization when they attempted to write scripts or stamped symbols on rock and clay disc to represent some knowledge. Storing and representing data become a trivial task with computers especially in last decade whereas data processing become easy and cheap.

Datafication is the process of collecting, storing, and analyzing data that model certain activity or process and turn it into measures to make it more visible and help to make clearer decisions. Datafication is tightly coupled with sensors industry that is proliferating due to the low cost of sensors, thus more activities, processes and things to be datified. Datafication can be consider the backbone of Big Data revolution. Consequently, data can be easily generated, collected, stored, modified, duplicated, analyzed, represented, and visualized.

Big data is a complex process that refers to the set of technologies and tools that collects, store, and process large, complex and varied types of structured, semi-structured and unstructured datasets in batches or real-time to extract correlations, trends, patterns, insights and knowledge. Data analytics refers to the methods and technologies that are used to analyze the data, extract hidden pattern and explore insights out of data. These methods are varied between simple statistical methods as mean and standard deviation to machine learning and deep learning.

Big data is characterized with a set of dimensions that describe what big data is. The four pillars for any big data system are: Volume, Variety, Velocity, and Veracity. These four V's are like pillars lending stability to the giant structure of Big data accompanied by the final objective of any Big data system, that is, Value which complements the whole purpose of big data system. The following is a brief explanation of Big Data V's:

Volume: how much data to process are there? This is the main characteristic of Big data that deals with the volume of collected data from various sources. Current Big data systems deal with Zettabyte (10^{21}) Bytes. The Large Hadron Collider generating about 1 Petabyte of data per second that no current computing platform can record and hence its filtered by experiments to keep only the most interesting data. The filtered data then aggregated into the European Organization for Nuclear Research (CERN) Data Center that process 1Petabyte of data a day [12]. Data is growing

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

exponentially in a daily base. Think about the data generated from social media, Wireless Sensors Network, emails, mobile usage, network analytics, and stock market. As the speed of data volume increases exponentially, there is a need for efficient distributed big data system that can store the massive volume of data. Currently, the most well-known Big data systems are: Hadoop, MongoDB, Amazon Web Service, Google BigQuery, Microsoft Azure, IBM Big data and many other.

Variety: Refers to the various data formats and sources that includes structured as RDBMS formatted data and unstructured data as text, images, multimedia, social media, voice, sensor data, graph-based data, and streaming data. The issue with unstructured data is that it has no defined metadata and does not follow certain pattern or format. Many studies pointed out that the percentage of unstructured data forms more than 90% of Big data, and this percentage is increasing. This characteristic requires the availability of a powerful data integration, transformation, and storage, and it impose several challenges through the Big data pipeline.

Velocity: Denotes the speed of collecting and processing various formats of data that ranges between batching system to real-time system. Velocity imposes certain requirements to design an efficient Big data system as storage capacity, constrains on network bandwidth. Accordingly, this requires an efficient collection, storing, and data processing algorithms.

Veracity: Deals with the data uncertainty or spars data. Dealing with sparse data needs filtration of clean and relevant data to extract meaningful and valued insights. This requires the big data system to have tolerant data processing algorithms.

Validity: Concerns with the acceptability degree of data for analysis. In particular, it aims to identify whether the data is up to date or not, collected as needed or not, conform to the standard or not. Therefore, any big data systems must be equipped with a robust data quality assessment tool to discard the data or adopt it.

Variability: States the frequency of data collection speed, and so, it is possible to detect anomalies and variances of collected data. Identifying anomalies and outliers usually leads to error prevention and will prevent error accumulation. Indeed, Big data characterized with high degree of uncertainty. Thus, preventing error at early stage would enhance the final analysis results.

Volatility: Refers to the speed of periodic update of data. Think about data coming from stock market, tweets, streaming of sensors signals...etc. Truly, it's a transactional data that forms a streaming or real-time data that need a powerful analytical big data computing platform. Big data system treats volatility cautiously especially in Big data decision making to issue a proper decision at certain time.

Vulnerability: Concerns with data security and data breach especially for cloud-based big data system and consequently the privacy and confidentiality of data. Security and privacy are the two major important aspects not only in Big data technology but in any other technology. At the end, a data breach for a big data is a big breach.

Visualization: Visualizing big data is a real issue since modeling millions of data points is very challenging and cannot be performed using traditional ways. Proper visualization requires enough in-memory technology to handle data scalability and computation response time.

Viability: Aims to identify whether the collected data is relevant to the use case at hand or not.

Value: Probably is the most significant characteristic for any big data system. Value can be a correlation, descriptive statistics, patterns, sentiments…etc. The aforementioned V's are useless if there is no value out of the big data. Indeed, Value is tightly coupled with veracity and velocity. The faster you derive meaningful insights out of data, the higher the value.

It Is worth mentioning that the Big data dimensions Vs are highly correlated to each other and not independent. The likelihood of change in one V is correlated with another change of another V [13].

3. HADOOP BIG DATA PLATFORM

Hortonworks believes that half of the world's data will be stored in Apache Hadoop within the next 5year. The importance of Hadoop efficiency is yielded from the big data strategy of parallel reading of large data files that are stored in internode network in a cluster. The bottleneck of any big data platform is the reading data from storage. i.e. disk latency. Increasing the computation power of CPU or the number of cores is not the answer but scaling out cluster of computers is a better approach along with parallel data processing, and this is exactly what Hadoop performed.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Apache Hadoop is a framework that perform distributed processing of massive datasets across clusters of computers that scale up from a single server to thousands. Apache Hadoop is an open source framework for reliable, scalable and distributed computing over a massive amount of data developed in Java and consist of four main subprojects: MapReduce, Hadoop Distributed File System (HDFS), YARN, and common Hadoop utilities like Hbase, Zookeeper, Avro and some other [14]. Hadoop is entirely dependent on Google's MapReduce and Google File System (GFS) technologies as its foundation. It is designed to process batches of enormous different formats amount of data using commodity hardware. Since the processing is performed in batches the response time is not immediate but acceptable. Hadoop replicates any stored file three times to guarantee reliability, availability, and fault tolerance. Figure 1 shows Apache Hadoop Ecosystem architecture.

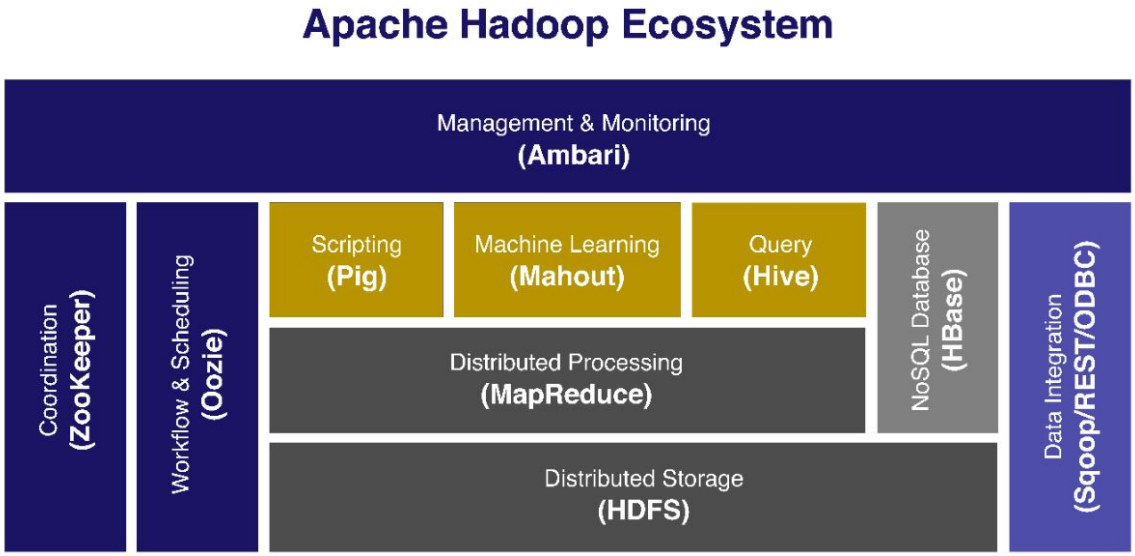

Figure 1: Apache Hadoop Ecosystem architecture

The following is a brief of Apache Hadoop 2.9.1 version framework:

HDFS: Is a distributed file system that aims to provide high throughput access to large spanned dataset files across multiples clusters and make them look as a single file. HDFS works in conjunction with MapReduce framework where programs are brought to data not the opposite as what usually happens in traditional programming to make the entire cluster available for reading and processing data, and to eliminate the delay of network transmission due to transfer enormous size of data. HDFS works collaboratively with MapReduce to access data on multiples clusters designed for streaming reads. HDFS replicates any data file into 3 replicas by default, two of them must be on the same rack and the remaining one on another rack to reduce transmission time. Hadoop splits large files into blocks of size 128MB by default. HDFS applies Master/Slave architecture where the maser called NameNode that manages file system namespace and metadata. The Slave is DataNode that stores data file and periodically reports status to NameNode [15].

Hadoop MapReduce: Is software framework introduced by Google [16] to facilitate distributed computing on large datasets of clusters of computers. MapReduce programming model is a representation model of the divide and concur processing model. The input data is split into small pieces (Map) over cluster's nodes, and then the results are filtered and aggregated in the (Reduce) step down to a single answer. Nevertheless, MapReduce1 has several drawbacks as: Centralized handling of job control flow, scalability, resource utilization, tight coupling of resource management infrastructure with certain programming language, and the over usage of Hadoop beyond the initial design. MapReduce has been evolved from MapReduce1 to MapReduce2 or the so-called YARN which concerns with scheduling and cluster resource management and covers the limitations of the first release.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

YARN: framework for job scheduling and cluster resource management. YARN was developed to overhauls the limitations of MapReduce. YARN has split the MapReduce1 main functionalities into two components concerns with resource management, and job scheduling and monitoring via two separate daemons: Resource Manager and Application Master. This new architecture has solved the limitations of MapReduce 1.

Ambari: A web-based tool for provisioning, managing, and monitoring Hadoop clusters. It also provides a dashboard for viewing cluster health and ability to view MapReduce, Pig and Hive applications visually. Apache Ambari provides easy to use RESTful APIs which allows application developers to easily integrate Ambari with their own applications.

HBase™: Is an implementation of Google BigTable. It is a scalable and distributed database dedicated for structured data storage for large table that can scale to petabyte. Hbase is highly scalable with new added nodes, low latency and support random read/write, highly available, strong consistency, and a very good choice for sparse data, in other words, Hbase is considered Hadoop database and can leverage the HDFS via adding new nodes to cluster. Hbase support a flexible and dynamic data model, and does not require an explicit schema definition up front which makes Hbase a natural choice for several big data applications. [17].

Hive™: Apache Hive is a data warehouse system built on top of Hadoop and was originally developed at Facebook. Hive facilitates data summarization, analysis, querying and managing large datasets, and ad-hoc queries. It's a tool that enable easy data extract/transform/load (ETL). Hive provides an SQL interface to data stored in Hadoop that is known as HiveQL which will be compiled later by MapReduce. It allows for abstraction of non-relational and semi-structured data. Hive works smoothly with other applications in Hadoop ecosystem due to HCatalog which is a table and metadata management layer that enable users with different data processing tools as Pig and MapReduce to easily read/write data on the grid [18].

Pig™: Is a tool for analyzing large dataset via Pig Latin scripting language which was designed to simplify MapReduce programming. Pig's consist of a compiler that produces sequences of MapReduce programs translated from Pig Latin script. It's a high-level data flow programming language where it defines a data stream and a series of Extraction, Transformation, and Load (ETL) [19].

Mahout™: A Scalable machine learning and data mining library focused mainly on clustering, classification and batch based collaborative filtering (recommendation). The aim of Mahout is to find insights out from Big data stored in Hadoop. Mahout works very well with distributed environment and can scale effectively in the cloud. It includes several MapReduce enabled clustering implementation like K-means and Canopy [20].

Apache Spark™: Is a memory-centric computational engine framework that is fast and scalable. Since its memory based, it is taking over the role of MapReduce as the major framework for working with Big Data. Spark includes build-in libraries that support ETL, machine learning, stream processing, and graph computation, SQL and DataFrames. Spark was designed to extend MapReduce model. Spark is characterized with high speed due to in-memory computation, and faster than MapReduce for complex application on disk. Besides, Spark can be used for a wide range of workloads (batch application, iterative algorithms, interactive queries, streaming data), and it is easy to use since it supports wide range of APIs that for Scala, Python, Java, and R [21].

Avro™: A data serialization efficient binary format that facilitates interoperability with applications of different programming languages due to binary encoding that can be used for long term storage in Hadoop. Avro uses the human readable text file format JSON to define data types and protocols since it supports transmitting data objects of the form: attribute-value. Avro supports versioning for MapReduce application, that is, handling field addition and deletion forward and backward compatibility [22].

ZooKeeper™: A centralized coordinator for reliable distributed applications coordination in Hadoop ecosystem. ZooKeeper concerns with maintaining configuration information, naming, distributed synchronization, partial failures handling that is inevitable in distributed environment, and generic group service [23].

Oozie: Is a Java based workflow server-based system that coordinates, manages, and executes Hadoop jobs. Oozie represents workflows as control flows and action as nodes via Directed Acyclic Graphs (DAG) [24].

Hadoop is an efficient open source Big data platform that is characterized with high scalability

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of data storage via adding more nodes as needed in a cost-effective manner. Hadoop supports several types of data storage (structured to unstructured) processing. Also, Hadoop is fault tolerant platform. In addition to the open source Hadoop Ecosystem, many distributors are adding their own proprietary add-ons to enhance the Ecosystem: Cloudera uses Cloudera Manager, and has Impala for SQL queries against Big Data; IBM has Big SQL, an ISO compliant version of SQL-2011 to access NoSQL datastores.

Despite the aforesaid advantages, Hadoop as any emerging technology has its own disadvantages. For example, Hadoop sustain at least 3 copies of a single data file, two at the same rack and one copy on another rack to sustain and assure data availability and fault tolerance. Big data is already large and voluminous, following this approach will magnify it further and will degrade the performance. In addition, sparse data and small data files i.e. file smaller than 128MB, have shown a significant degradation of Hadoop performance as Hadoop lacks the ability to effectively support random reading due to its high capacity design. Hadoop work effectively only with large dataset.

Hadoop is suitable to deal with batching processing not online streaming data. Hence, it is not suitable for real time processing data as processing data is time consuming. Additionally, Hadoop does not fully support cyclic data flow (iterative processing) i.e. output of current phase is input for next phase. Moreover, Hadoop does not support memory caching, and so, there is latency and late response. Besides, Hadoop does not support query optimization. Therefore, there is no efficient cost-based query execution.

Security is a real concern in Hadoop due to large stored size of data where there is a possibility of malicious code within data. Additionally, Hadoop is vulnerable by nature since it was developed by Java that was heavily exploited by cybercriminals. Cohen and Acharya [25] pointed out that sensitive data stored within Hadoop system is a an attractive target for exfiltration, corruption, unauthorized access and modification. Moreover, Hadoop missing the data protection via encryption at the levels of storage and network.

Most of Hadoop limitations were fixed via new software releases. For example, Kerberos and Knox fix security limitations. Apache Storm, Spark, and Flink solve the caching and other Hadoop limitations. However, another main challenge in this domain is the lack of highly skilled technician. Big data platform is very complex and rapidly evolve. Thus, acquire all the needed knowledge for Big data and keep updated is challenging.

Hadoop ecosystem is almost 10 years now and it keep on evolving rapidly. Its observed that the Big data technology is rapidly changing and evolving every almost two years. Hadoop 3 uses Docker to boots agility and package isolation to faster build apps, and better resource isolation for disk and network. Hadoop 3 applies Erasure Coding (EC), that is, less storage overhead from 3x replica to 1.5 replica in which the 0.5 replica is used for parity checking while assuring same level of data recoverability. EC reduces the storage cost by ~50% compared with 3x replication. Furthermore, Hadoop 3 supports multiple standby NameNode that assures scalability and availability. A significant change is enabling new resources as disks and Graphical Processor Units (GPUs) that boost integration and container.

4. CHALLENGES IN BIG DATA

Big data is a giant concept that includes several unique utilities (software and hardware) to manipulate large, complex, and heterogenous amount of data. The challenges that surround Big Data could be derived from both perspectives, i.e. The data and the utilities. Indeed, each discussed Big Data's V includes some challenges, for instance, handling the rapid growth of data size is problematic and requires efficient scalable storage system, data file formats for several types of data, i.e. columnar data files, compression, and duplication. Another challenge is related to the velocity of data whereas generating insights from the stored data is needed in a timely manner and efficient cost via new generation ETL and analytical tools.

Herewith, we address the challenges related to each phase of Big Data Analytical Pipeline according to [26] in Figure 2. The first three phases concern data management, analysis/modelling and interpretation concern data analytics. This section aims to address some of these challenges in respect to each phase in this pipeline:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

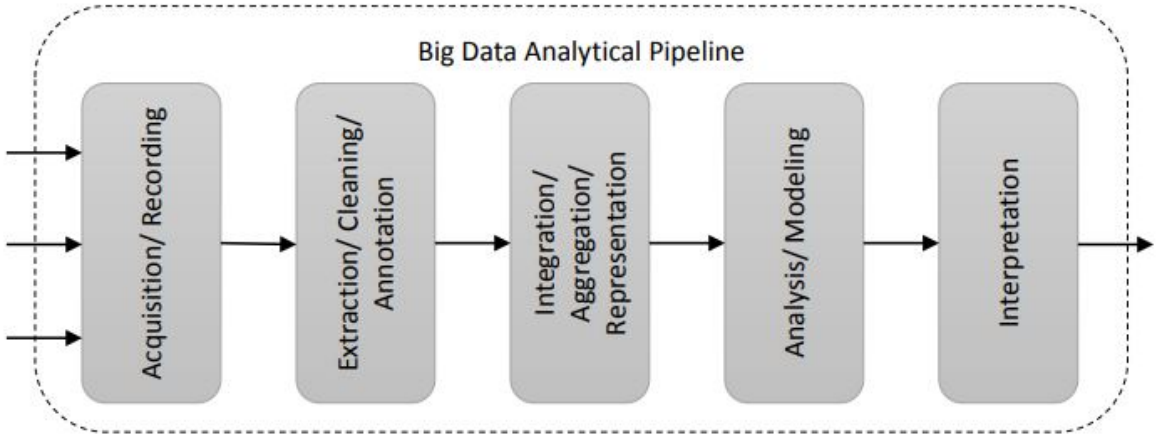

Figure 2: Big Data Analytics Pipeline.

Data Acquisition/ Recording: Nowadays, data is generated everywhere. Think about data of interest generated by a smart phone, sensors, website logs files …etc. Big data sources produce massive amount of data that can be cleaned, filtered and compressed to increase the efficiency of the Big data platform. Think about the format of data, what data to collect and what data to discard? The first apparent challenge at this stage is data storage that can process the enormous amount of heterogenous data with high input and output speed. The main challenge that can degrade the performance of any Big data platform is the disk latency. Thus, scaling up more powerful computers with powerful CPU will not help, instead scaling out to clusters of disks is the answer. The later problem that is associate with disk latency is the network latency in which moving enormous amount of data on the network will be very costly for performance [27], [28], [29]. Thus, the computation is brought to data not data brought to computation which is one of the driving principles of HDFS and MapReduce.

Another challenge lies in designing an effective online filter that doesn't affect the potential value of any discarded data, tolerate and process missing, incomplete, redundant and overlapped data, especially data collected from various sources which is very common in Big data environment i.e., sensory networks. Therefore, in order to develop an efficient acquisition process, there is a need for an efficient data reduction, filtering, and compression algorithms that fit the collected data.

One of the challenges of designing Big data system is that it must be scalable to accommodate the changing of data volume over time intervals. Therefore, this variation of data volume requires additional computational processing power via using multiple cores CPU [30]. Many platforms utilized the advancement of processing power as Apache Spark, Apache Storm whereas the computations are performed in memory.

Data extraction and cleaning: Usually, data preparation for analysis is costly in terms of money and time. Acquiring and collecting data in the previous phase does not mean the data is in the needed and suitable format that is ready for effective analysis. Still there is a need to extract the needed data for analysis out of the acquired data. For example, the acquiring streaming videos from CCTV that requires face detection. In this scenario, an efficient extraction process is required to extract the relevant data of interest in order to reduce the overall processing time. This would require extracting updated data only that require an efficient change detection mechanism, for example, log sniffing [31] that detects and change in the log file comparing to the last scan. The technical challenge of the extraction process is entirely dependent on the Big data application, i.e. Tweets, sensory data, CCTV images or video… etc. Data cleaning concerns with extracting correct and reliable data for next phase, that is, data analysis. For example, sensors can be inaccurate or faulty, inconsistent, irrelevant, incomplete data are all examples about real cases where data need cleansing before valid analysis.

Data Integration, aggregation, representation: After understanding and cleansing data, it must be served to Big data warehouse in a suitable format and in a timely fashion either in real time or in batches. Data integration handles easy access to multiple data sources [32].

Effective Big data system requires acquiring data from multiple heterogeneous sources. Thus,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

efficient tools of integration and transformation are needed to tackle the heterogeneity problems and to standardize data. The large volume of proliferated data raises a concern about tracking data provenance and the degree of its reliability. Big data provenance concerns with tracking the root of origin to its original sources, the creation process, and the propagation process too [33], [34]. Wu et al. [35] defined data provenance as a lineage process of data. Nunes et al. [36] defined it as the process of data derivation. Problems associated to data provenance are inherent from the nature of Big data process and mainly due to the size [37].

Interoperability is a real challenge in Big data platform that arises after solving the integration problems [38]. Interoperability is dependent on the application at hand. For example, think about a Big data system that aggregates heterogeneous types of data like surveys results, voice calls, and certain related images. Aggregating, merging, and corroborating these data together is challenging and problematic.

Big data is mainly unstructured, complex, and heterogenous data that are aggregated from various resources like: sensors, social media streams, emails, documents, images, voice, videos, etc. The challenges of data integration are inherited from Big Data's Vs in order to handle the data and process it to obtain values. For example, the volume challenge requires scalability to accommodate large data volume growth that requires additional storage to fulfill analytics demands. The veracity requires predefined protocols, transformation and integration to handle various formatted data and structures of API. Dong and Srivastava [39] pointed out that Big Data Integration (BDI) differs from traditional data integration in several aspects in which the number of data sources of BDI is higher, very dynamic due to newly added data sources, heterogenous data sources, varied data quality of data sources i.e. coverage, accuracy, timeliness. Handling schema mapping, record linkage, and data fusion are challenging and still evolving to get best results.

Data governance is highly needed to maximize data quality, and security via following data compliance procedures and respecting data user/business ownership. Data governance aims to guarantee data accuracy and data consistency that are real concern for decision maker to issue a decision based on reliable data. Achieving data governance requires a combination of robust data policy and procedure, central data control management, and security. The growing volume and veracity of data increases the complexity of data governance and management to generate business values.

Data aggregation has several streaming data challenges like: Streaming data requires real time application that handles, filters, and analyzes continuous generated data on-the-fly to. Streaming data is generated usually in small sizes that include various types like: log files, status and activity of certain sub-component in a system, information related to social network, i.e. tweets, financial and stock trading…etc. This type of data needs periodic, sequential and incremental record-by-record processing to update related metrics, statistics, and reports over a sliding time window. Challenges are inherited from the nature of such data.

Heterogeneous schema of different data sources is problematic and differ in connection methods and protocols, data rates and frame size, and processing requirements that form an integration challenge at earlier stage at certain level: across multiple data sources and multiple streaming systems. Later, it is very eager to optimize the aggregated data size to utilize the computing resources like the cache memories hierarchy to achieve smooth data transition within the system.

Most of the algorithms that handles streaming data assume that data is complete, immediately available, and free of cost of which is not the case most of the time. Incomplete data is very common in streaming environment. Many open challenges are still there: handling unpredictable missing values in a stream! Automatically select the best-fit imputation technique [40].

Fast processing is needed to process the flooded continuous data with a tolerable latency in millisecond whereas input adaptors may become a bottleneck.

Processing streaming data requires two main layers: Storage and processing. The storage layer must support high speed read/write, ordering, and filtering to enable fast and inexpensive computation. The processing layer digests data from storage layer for processing and notifying the storage layer back to delete the data if it's no longer needed. Furthermore, both layers must be scalable and fault tolerant.

Handling several heterogeneous types of data is problematic for data representation purposes. To this end, transforming unstructured and semi-structured data into structured data that is ready for analysis is a need that requires full understanding of the business use case and objectives that would help

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

you to decide what data to includes, discard, and analyze.

Sparse data or missing data enforces a major challenge on any Big data system. Missing data need to be substituted, regenerated, or skipped from the pipeline since its rebel effect on the pipeline will lead to wrong interpretation of results. Moreover, it may depreciate the scalability, efficiency, and performance of the big data system. Efficient processing of data with high throughput is a real concern in processing Big data.

Modelling and Analysis: The challenges of modeling and analyzing Big data are mainly related to scalability issue. The nature of Big data (heterogeneity, volume, noisy, streaming, dynamic, inter-related, un-trustworthy) impose a fundamental significant difference on modelling and analysis comparing to classical data. The good news about analyzing and modelling Big data is that data big enough to skip and ignore missing, incomplete and noisy data. The remaining volume would be enough for approximate analysis. Still there is a need to handle this missing data.

Representing Big data is not as simple as representing small data. Several methods can be used to represent Big data as fuzzy set, rough set, and formal analysis [41]. These methods may not work well with large voluminous, and scalable data. Thus, a hybrid of these methods would be needed to meet the representation requirements. Analyzing Big data requires super computation power to deliver the desired performance. To do so, parallelism is needed. Systematic and formal computation methods are needed as machine learning, deep learning, and reinforcement learning that are less dependent on memory computation and minimizes computational cost processing and complexities [42], [43].

"A picture is worth a thousand words". Visualizing data aims to represent data in a systematic form using graph theory techniques. Visualizing data is another challenge in Big data. Visualizing millions of data points or records in real time or even static time is a challenging task. Basically, it needs special hardware that is based on either parallel computing, cloud computing, or distributed processing. Scalability and dynamics are two major challenges in visualizing big data. These challenges are related to the Big data's V's. For example, volume: aims to smoothly represent large voluminous of data. Variety: several heterogenous data are there to handle, i.e. structured, semi-structured and unstructured data. Velocity: concern with visualizing Big data in real time or in batches. Value: To obtain comprehensive insights out of the represented data [44].

GPUs has improved the speed and scalability of computers with intensive and massive parallel processing and enormous number of cores whereas some GPUs containing around 6000 cores, that is 200 times more than the most powerful CPUs in Today's technology. GPUs can handle the volume and velocity of Big data for the main purpose of rendering images and graphs. GPUs capabilities have extended the main purpose to other areas as financial modelling, AI, Deep Learning, and Big data. GPUs have extended the scalability of data mining algorithms of handling large data set size in many applications domain, significantly [45].

Due to voluminous size of data, most data visualizing tools perform its functions at data centers rather than via fetching data to the tools to avoid data trafficking and bandwidth. The best solution for tackling visualization processing is via decomposing the task into sub-tasks to enhance parallel processing [46]. Representing and visualizing Big data usually yield visual noise and data loss as its limited within a narrow representation area whereas data points are over written over itself and can't be recognized. Furthermore, plotting every data point will cause over-plotting that will affect perceptual and cognitive capability of reader, and can results in high latency that cause poor response time [47]. In order to tackle some of these challenges, there is a need to fully understand the data at hand, apply proper data reduction techniques or approximation techniques like sampling, filtering, and aggregation, and make sure it is clean, so it is possible to alleviate the over load of processing.

The aim of all visual tools is to support the analytical reasoning via interactive visual interfaces. There are many tools that visualize big data as Google chart, FusionCharts, D3, Tibco Spotfire, Watson Analytics and many others, most of them characterized with deficient performance, response time and scalability.

Interpretation: Results interpretation is the main objective for this pipeline. Decision maker must scrutinize the assumptions that were used in each phase in the pipeline to guarantee proper and valid decision. Think about the assumptions of data skipping procedures, data collection and recording, and statistical assumptions. All these accumulated assumptions would accumulate errors, affect the interpretation of final data results, and would push

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to repeat the analysis. In addition, erroneous data would be there due to several reasons as error in measurement, fault sensor, outdated data, bugs, and so on. The complexity of all these conditions together make it difficult to interpret the data properly.

Predictive analysis and the interpretation of Big data results that are based on statistical analysis must not be treated the same as interpretation of classical data interpretation mainly due to the size and the accumulated erroneous that would happen in the pipeline. Interpretation of Big data results must be conducted cautiously by taking into consideration the underlining data arguments. Fan et al. [48] pointed out that the notion of statistical significance is not relevant to Big data, many conventional methods are not applicable to Big data due to scalability limitation, and several inherent distinctive features of Big data as: noise accumulation, heterogeneity, spurious correlation, and incidental endogeneity

Interpreter must consider spurious correlation into account. The later infers a falsified correlation between two variables due to unseen factor or coincidence. However, in Big data it would happen due to voluminous size of data that would declare a falsified correlation. Fan and Lv [49] found in an experiment that the correlation coefficient between random variables in the tested dataset is increase against the dataset size. Moreover, the incidental endogeneity whereas the underlying assumptions of regression analysis validity might not be satisfied in regression analysis or other statistical methods. For example, the assumption of dependence between residual and independent variable.

5. SECURITY AND PRIVACY

Security and privacy are most likely the most important two challenges and aspects in Big data [50], [51]. The security and privacy challenges are correlated to Big data analytics since the implementation of Big data analytics is performed either on the cloud or at data repositories [52], [53]. Security in Big data mainly concerns about data itself. Data security concerns with integrity, availability, confidentiality, accountability, and data ownership protection. To do so, securing data can be done at two levels: Algorithmic approaches that stresses the privacy and security requirements directly on the code [54], [55], and data model-oriented approaches that concerns with hashing and encryption data models [56], [57].

Big data platform forms a seductive source for hackers and abusers to take over the data, leak data, and abuse it, i.e. Facebook and Cambridge Analytica incident. One challenge is derived from the fast evolving of Big data platforms that make privacy and security pace slower. Moreover, the privacy and security aspects are subjected to governmental regulations, discovering new bugs, i.e. iPhone and FaceTime. Moreover, social media giants (Facebook and Twitter) keep on changing the terms of security and privacy following certain events, triggers, and governmental regulations.

Cyberattacks occurred recently caused several data breach. Facebook announced that almost 30 million account were affected in late September 2018. A vulnerability in Google+ social network API's affects the personal data of hundreds of thousands of records according to Wall Street Journal between 2015-2018. Thus, organizations must protect their data assets. Indeed, protecting data in Today's technology that is proliferated with massive numbers of connected devices, and different technologies as cloud and IoT outperforming Today's technology. For example, the duration of unnoticed data breach could be very long and undiscovered. It's difficult to recognize that there is a malicious resident virus that is hided into your system like the infamous case of Yahoo data breach that almost affect 3billions users in 2013. Attackers can hide lie in an organization's platform without being noticed and having access to their records, accounts, emails. A very well know example about hidden threats is the cyberattack that hits the Iranian nuclear enrichment (Stuxnet). On the other, the attack duration could be minutes.

Security is essentials and a must in any recent technology, and so in Hadoop. Thus, following a strict security policy is not optional and can be challenged for several reasons: The inherent nature of Hadoop distributed framework system requires that not all the interactions follow client-server model. The distributed and partitioned Hadoop file system requires authorization and authentication at several levels, for example, a submitted job by certain client for execution via node x could be executed by unknown nodes to that client. Apache Hadoop ecosystem includes large number of technologies whereas each single component is subject to several factors as versionality and cross-component integration.

There are several security issues surround Big data pipeline end-to-end. The threats to Big data security start with the end-point input devices that might provide unauthentic data. Mobile Device Management (MDM) platform can be used to secure the authenticity of data input. Moreover,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

there is a need to secure the Big data platform internally which requires a review of user data access permissions. The later issues can be complemented via a comprehensive monitoring policy for user behavior that can detect any anomalies of data access and would help to build a behavioral model and usage pattern for each user.

Securing unstructured data is not as simple as it looks since there could be some vulnerabilities of malicious and malignant code is embedded within the data. Thus, it requires some forms of encryption that will affect the computational complexity exponentially. There are several security measures as hashing, masking sensitive data, or two-ways encryption algorithms.

Securing Big data in cloud is a challenge. The attacker can tamper the data exchanged in the cloud CCSS and the Big data server as HDFS. For example, the attacker can carry out DOS attack and spoof the responses and shut down the server [58].

Other challenges face Big data technology is to communicate Big data technology, methods, numbers, assumptions, data, results, and outcomes to top management and executives for decision making purposes due to lack of deep understanding and the unalignment between Big data experts and decision makers [59], [60].

The lack of skillful and talented professionals in Big data is another challenge. Davenport and Dyché [61] pointed out that traditional quantitative analysis is not enough to handle Big data. Douglas [62] pointed out that according to the US department of Labor, the shortage number of skillful Big data personal in the US only will be 120,000 to 190,000 by 2018. McKinsey forecasts that there will be a shortage of almost 140,000–190,000 workers with "deep analytical" experience will be needed in the U.S, and 1.5 million managers will need to become data-literate. Moreover, there is a real shortage of professional security personnel estimated by 2 million in 2019 according to the non-profit information security advocacy group ISACA.

6. CONCLUSION:

Big data is a complicated technology end-to-end. The Big data pipeline starts with acquiring data, recording, extracting, cleaning, integrating. aggregating, representing, analyzing and modelling, visualizing and interpreting results. Big data are heterogenous data aggregated from multiple resources divided between unstructured, semi-structured, and structured data. At each phase of Big data there is certain challenge and difficulty that affect the reliability the use case solution at hand.

Big data characterized with several characteristics whereas volume and variety are the most influencing one that create several challenges later. Any change in Big data V will affect the other Vs. Value is the final objective of Big data platform that can be a correlation, descriptive statistics, inferences, hidden pattern.

Hadoop is an open source Big data platform that can handle Big data processing efficiently and effectively. Mastering Hadoop requires lots of knowledge and experience to drive and operate several utilities at the hardware and software level. Despite this, Hadoop has its pros and cons.

Apart from the challenges exist at each phase in the Big data pipeline. Privacy and security are the most challenging and critical aspects. Due to the complicated inherent nature of Big data, it's difficult to explain the results to the top management too. The lack of skilled professionals in data science in general, and Big data, in particular, is another challenge. Future work will be studying the future trends of Big data.

Acknowledgment: The author would like to thank Dr. Glen Mules for his valuable feedback on the first draft of this paper.

References

[1] I. mareketing Cloud, "10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations." 2017.

[2] W. E. Forum, Personal Data: The Emergence of a New Asset Class. 2011.

[3] W. House, "Fact Sheet: Big Data Across the Federal Government," 2012.

[4] "Gartner IT Glossary (n.d.). Retrieved from." [Online]. Available: http://www.gartner.com/it-glossary/big-data/.

[5] The Economist, "Fuel of the future- Data is giving rise to a new economy."

[6] M. Fuchs, W. Höpken, and M. Lexhagen, "Big data analytics for knowledge generation in tourism destinations - A case from Sweden," J. Destin. Mark. Manag., 2014.

[7] S. Sinha and A. Bansal, "A Framework for Effective Data Analytics for Tourism Sector: Big Data Approach," Int. J. Grid High Perform. Comput., vol. 9, no. 4, 2017.

[8] R. Priyadarshini, R. K. Barik, C. Panigrahi, H. Dubey, and B. K. Mishra, "An Investigation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Into the Efficacy of Deep Learning Tools for Big Data Analysis in Health Care," Int. J. Grid High Perform. Comput., vol. 10, no. 3, 2018.

[9] N. Elgendy and A. Elragal, "Big Data Analytics in Support of the Decision Making Process," in Procedia Computer Science, 2016.

[10] F. Chang et al., "Bigtable: A distributed storage system for structured data," Trans. Comput. Syst., 2008.

[11] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," IEEE Access, 2014.

[12] E. O. for N. R. (CERN), "CERN Data Centre passes the 200-petabyte milestone." .

[13] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," Int. J. Inf. Manage., 2015.

[14] A. Hadoop, "Apache Hadoop." 2018.

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

[16] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. OSDI - Symp. Oper. Syst. Des. Implement., 2004.

[17] A. Hbase, "Apache Hbase." .

[18] A. Hive, "Apache Hive." Apache Hive.

[19] A. Pig, "Apache Pig."

[20] A. Mahout, "Apache Mahout." .

[21] A. Spark, "Apache Spark." .

[22] A. Avro, "Apache Avro." .

[23] A. ZooKeeper, "Apache ZooKeeper." .

[24] A. Oozie, "Apache Oozie." .

[25] J. Cohen and S. Acharya, "Towards a More Secure Apache Hadoop HDFS Infrastructure," Netw. Syst. Secur., 2013.

[26] C. Baru, M. Bhandarkar, R. Nambiar, M. Poess, and T. Rabl, "Benchmarking Big Data Systems and the BigData Top100 List," Big Data, 2013.

[27] M. Merabet, S. mohamed Benslimane, M. Barhamgi, C. B. Lyon, and C. Bonnet, "A Predictive Map Task Scheduler for Optimizing Data Locality in MapReduce Clusters," Int. J. Grid High Perform. Comput., vol. 10, no. 4, 2018.

[28] S. Carlin and K. Curran, "Cloud Computing Technologies," Int. J. Cloud Comput. Serv. Sci., 2012.

[29] P. C. Wong, H. W. Shen, C. R. Johnson, C. Chen, and R. B. Ross, "The top 10 challenges in extreme-scale visual analytics," IEEE Comput. Graph. Appl., 2012.

[30] A. Jacobs, "The Pathologies of Big Data," Queue, 2009.

[31] T. Jörg and S. Dessloch, "Near real-time data warehousing using state-of-the-art ETL tools," in Lecture Notes in Business Information Processing, 2010.

[32] D. AnHai, H. Alon, and I. Zachary, Principles of Data Integration. 2012.

[33] B. Glavic, K. S. Esmaili, P. M. Fischer, and N. Tatbul, "Efficient Stream Provenance via Operator Instrumentation," ACM Trans. Internet Technol., 2014.

[34] R. Hammad and C. S. Wu, "Provenance as a service: A data-centric approach for real-time monitoring," in Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014, 2014.

[35] E. Wu, S. Madden, and M. Stonebraker, "SubZero: A fine-grained lineage system for scientific databases," in Proceedings - International Conference on Data Engineering, 2013.

[36] I. Nunes, Y. Chen, S. Miles, M. Luck, and C. Lucena, "Transparent provenance derivation for user decisions," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012.

[37] Y. W. Cheah, R. Canon, B. Plale, and L. Ramakrishnan, "Milieu: Lightweight and configurable big data provenance for science," in Proceedings - 2013 IEEE International Congress on Big Data, BigData 2013, 2013.

[38] A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq, "Challenges of data integration and interoperability in big data," in Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, 2014.

[39] X. L. Dong and D. Srivastava, "Big data integration," in Proceedings - International Conference on Data Engineering, 2013.

[40] G. Krempl et al., "Open challenges for data stream mining research," ACM SIGKDD Explor. Newsl., 2014.

[41] L. A. Zadeh, "Fuzzy Sets-Information and Control-1965," Inf. Control, 1965.

[42] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient Machine Learning for Big Data: A Review," Big Data Research. 2015.

[43] C. Yoo, L. Ramirez, and J. Liuzzi, "Big data analysis using modern statistical and machine learning methods in medicine," International Neurourology Journal. 2014.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[44] L. Wang, G. Wang, and C. A. Alexander, "Big Data and Visualization: Methods, Challenges and Technology Progress," Digit. Technol., 2015.

[45] A. Cano, "A survey on graphic processing unit computing for large-scale data mining," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 2018.

[46] H. Childs et al., "Research challenges for visualization software," Computer (Long. Beach. Calif)., 2013.

[47] Z. Liu, B. Jiang, and J. Heer, "ImMens: Real-time visual querying of big data," Comput. Graph. Forum, 2013.

[48] J. Fan, F. Han, and H. Liu, "Challenges of Big Data analysis," National Science Review. 2014.

[49] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," J. R. Stat. Soc. Ser. B Stat. Methodol., 2008.

[50] C. Wu and Y. Guo, "Enhanced user data privacy with pay-by-data model," in Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013, 2013.

[51] M. Jensen, "Challenges of privacy protection in big data analytics," in Proceedings - 2013 IEEE International Congress on Big Data, BigData 2013, 2013.

[52] A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, and A. Rasin, "HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads," VLDB Int. Conf. Very Large Data Bases, 2009.

[53] H. Herodotou et al., "Starfish: A Self-tuning System for Big Data Analytics.," in CIDR. Vol. 11., 2011.

[54] M. M. Alani, "Security threats in cloud computing," in SpringerBriefs in Computer Science, 2016.

[55] M. Jang, M. Yoon, and J. W. Chang, "A privacy-aware query authentication index for database outsourcing," in 2014 International Conference on Big Data and Smart Computing, BIGCOMP 2014, 2014.

[56] A. Arasu, K. Eguro, R. Kaushik, and R. Ramamurthy, "Querying encrypted data," in Proceedings - International Conference on Data Engineering, 2013.

[57] A. Boldyreva, N. Chenette, and A. O'Neill, "Order-preserving encryption revisited: Improved security analysis and alternative solutions," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011.

[58] I. Muttik and C. Barton, "Cloud security technologies," Inf. Secur. Tech. Rep., 2009.

[59] M. Pacino, "Analytics, Big Data and Data Science," Bus. Intell. J., vol. 22, no. 2, pp. 55–56, 2017.

[60] A. Alharthi, V. Krotov, and M. Bowman, "Addressing barriers to big data," Bus. Horiz., 2017.

[61] T. H. Davenport and J. Dyché, "Big data in big companies," SAS Inst. Inc., 2013.

[62] M. Douglas, "Big data raises big questions," Gov. Technol., vol. 26, no. 4, pp. 12–16, 2013.