**Chapter 5**

**Big Data and Health care**

**Chapter 5**

**Big Data and Health care**

5.1 Big Data Definition

Machines are currently joined as a common community in an Industry 4.0 factory. Such a technology includes the use of sophisticated simulation methods so that big data can be reliably analysed and make reasonable decisions in order to explain uncertainty [31]. Today’s **Big data** is a hot topic that directs ways to investigate analytically to extract information from data sets that are too large in size and it becomes complex to be processed by traditional data-processing application software within the tolerable acceptable time.Big data can be unstructured, semi-structured and structured data. However, the main focus is on unstructured data.There are several numbers of characteristics of Big Data. The most important features of big data are given below [32][33].

1. Volume

The amount of produced and stored data is too large in size. For example, Facebook and generate above 500TB and 20 PB data in day respectively and about 1TB of new trade data is generated per day in The New York Stock Exchange.

1. Variety

This describes the different types and natures of the data. Big data can be generated from text, images, audio, video.

1. Velocity

The velocity means the speed at which the data is generated and processed.

1. Veracity

It refers to the uncertainty of data quality and the data value.

5.2 Big Data Application and Health care

5.2.1 Big Data Application

Big data has augmented the mandate of information supervision experts so much so that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have paid more than $150 billion on data management and analytics. The following are the application fields of big data [34]. Some approaches explaining how control, smart clock, travel, smart safety, smart phone, intelligent television, security systems, etc. are designed using big data analytical techniques [ 35] [36].

1. Government
2. International development
3. Manufacturing
4. Healthcare
5. Education
6. Media
7. Insurance
8. Internet of Things (IoT)
9. Information technology
10. Sports

5.2.2 Big Data in Healthcare

Big data analytics has been facilitating healthcare improvement by providing personalized medicine, medical risk interference and predictive analytics, automatic reporting of patient data. Some fields of development are more ambitious than actually implemented. The volume of data produced within healthcare organizations is not small. With the additional implementation of m-Health, s-Health and e-Health the volume of data continuously increased. The electronic health record data, imaging data, patient created data, sensor data, and other systems generated data are the sources of data and difficult to process. The practice of big data in healthcare has elevated important moral challenges ranging from risks for distinct rights, confidentiality and independence, to transparency and faith.

Big data in health research is predominantly auspicious in terms of tentative biomedical research, as data-driven investigation can change more rapidly than hypothesis-driven research.

5.3 Big Data Analytics

Big data analytics is a complex process of investigating huge and diverse data sets and to reveal information like hidden patterns, unidentified relationships, marketplace trends and client preferences that can help making well-versed decisions.

Big data analytics allow big data analysts, data scientists, analytical modelers and other professional to scrutinise increasing bulks of structured transaction data, semi-structured and unstructured data. Internet search data, server logs, social media, survey and sensors data that are the examples of the mix of semi-structured and unstructured data should be dealt with big data analytics.

5.4 Apache Hadoop and Hadoop Architecture

5.4.1 Apache Hadoop

The most powerful tool for analysing big data is known as Apache Hadoop. Apache Hadoop is defined as an open source framework for storage and processing data [37]. Hadoop consists of four main parts.

* The Hadoop Distributed File System (HDFS) is a distributed file system designed to store data on commodity machines and it provides very high bandwidth across the cluster. Dananode and Namenode are the two components of HDFS.
* YARN (Yet Another Resource Negotiator) is a platform for the management of Hadoop's resources and job scheduling programs that will run on the Hadoop distributed processing framework. It has two basic components, those are the ResourceManager and NodeManager.
* Hadoop Common is a set libraries and utilities.
* Hadoop MapReduce is an implementation of the programming model. MapReduce performs two basic functions. The functions are mapping and reducing.

Hadoop has a master and slave topology. One master node and several nodes for slaves are the main components. The function of master node is to allocate a job to multiple slave nodes and to manage resources. the actual execution is performed by the slave nodes.

5.4.2 Hadoop Architecture

Hadoop Architecture encompasses three major layers. They are

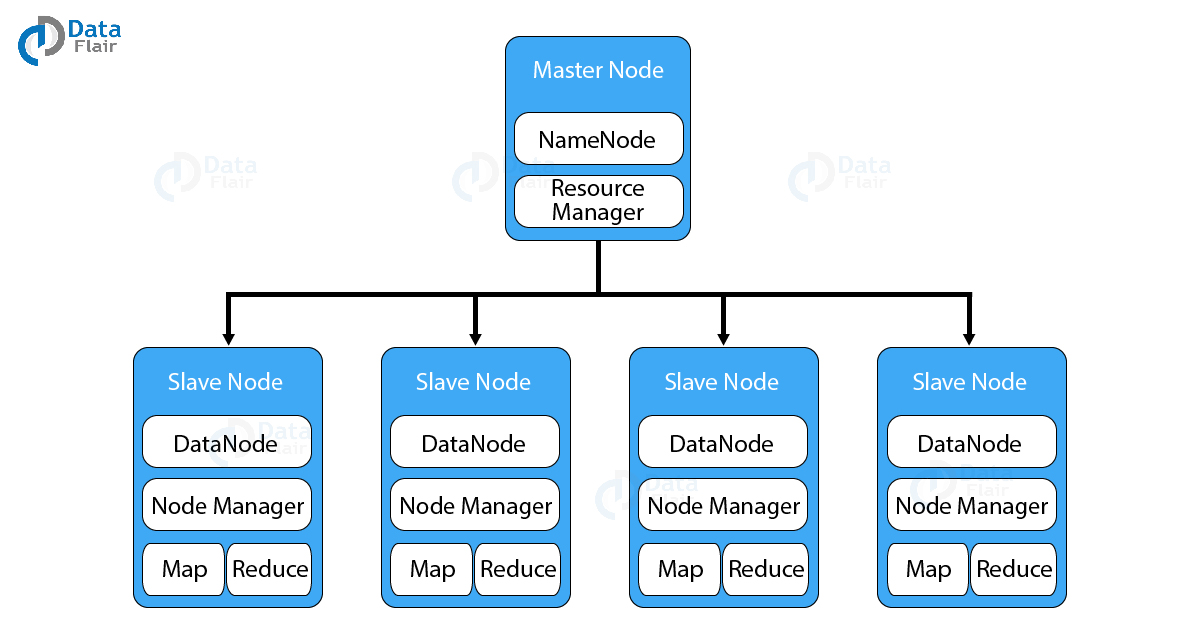
* HDFS
* Yarn
* MapReduce

HDFS runs for data storage of Hadoop. HDFS divide the data into smaller blocks units and stores blocks in a distributed manner. The default block size is 128MB or 256 MB. Two daemons running in HDFS are one for master node called NameNode and other for slave nodes called DataNode.

NameNode daemon controls the space management and file access by the client, opening, closing and renaming files, keeps track of mapping of modifications to file system and blocks to DataNodes or directories.

DataNode daemons are used for storing actual data. The smaller data unit are stored on a group of slave machines. This DataNode perform the read and write request from the file system’s client. DataNode is responsible for the creation, deletion and replication of blocks on demand from NameNo

Figure 5.1: Hadoop Architecture [38]



5.5 MapReduce framework

5.5.1 MapReduce

**MapReduce** is a software framework that allows to write applications for processing big data. These applications can be run in parallel on a cluster. It is so reliable and fault-tolerant. MapReduce job encompasses mapping tasks and reducing tasks. The data goes through the following phases and the figure 5.2 shows the dataflow diagram of MapReduce.

*Input splits:* A MapReduce function input is divided into a selectable part called the Input split. This is the first step towards introducing a system for map reduction.

*Mapping:* Mapping function requires a set of key-value pairs, each process and generates null or more output key / value pairs, and the map's types of input and outputs may different.

*Shuffling:* Each process includes the mapping phase output. The purpose of the mapping process production is to compile the appropriate data. The same terms are paired with their respective frequencies.

Zahed is graduate

Zahed is graduate from ICE RU. ICE is the best dept in RU.

from ICE RU

ICE is the

best department in

RU

Zahed,1 is ,1 graduate, 1

from ,1 ICE,1 RU,1

department,1

best,1

is ,1 is ,1

from, 1

graduate,1

department,1

ICE,1 ICE, 1

graduate,1

RU, 1 RU, 1

in, 1

in, 1

ICE,2

best,1 department,1 from,1 graduate, 1 ICE, 2 in, 1 is, 2 RU, 2 the, 1

Zahed, 1

ICE,1 is ,1 the,1

best,1 department,1 in, 1

RU, 1

best,1

from, 1

the, 1

Zahed, 1

is ,2

RU,2

the,1

Zahed,1

Input data

splitting

mapping

shuffling

reducing

Output data

Figure 5.2: MapReduce explanation with an example

*Reducing:* During this step the shuffling phase output values are applied. This process incorporates shuffling values and returns one output value. In a nutshell, the whole dataset is condensed. This phase in our example adds shuffled values, i.e. the total occurrence of every word is calculated.

5.6 Apache Spark, Hive, Hbase and Pig.

5.6.1 Apache Spark

Apache Spark is a general cluster computing system deployed in open-source applications. Spark offers an app with involved concurrent data and fault tolerance to configure whole clusters.

The Resilient Distributed Dataset (RDD), a read-only multiset of data objects spread over a network of computers that is managed tolerably by error, became Apache Spark's technical basis [39].

Apache Spark is a large-scale data processing and visualization system developed specifically for large-scale management Including Hadoop Distributed File System (HDFS), OpenStack Swift, Amazon S3, and Cassandra. Spark offers to access data across a variety of sources. Apache Spark is designed to increase analytics on Hadoop and offers a full set of additional tools, including a fully functional Machine Learning Library (MLlib), a GraphX and stream processing system. Apache Spark was created in 2009 at UC Berkeley's AMPLab and was donated to Apache Software Foundation in 2013, which was the most successful donation project.

5.6.2 Apache Hive

Apache Hive is an open source Apache Hadoop data warehouse program. Hive includes a Hive SQL-style language that allows for large datasets of data analyses and syntaxes contained in Hadoop file systems. Before becoming a sub-project of Hadoop, Hive originated as a Facebook initiative. Hive is a top-level initiative under the Apache Software Foundation, currently open access community [40].

Hive provides a Relational framework to access data stored in various databases and filesystems combined with Hadoop. Hive has a SQL-like design for the querying of data in different data bases and in file systems. To order to execute SQL programs and queries over distributed files, conventional SQL queries must be introduced to MapReduce Java API. With the abstraction SQL required to embed SQL-like queries (HiveQL) in underlying Java without allowing queries to be introduced in the low-level Java API, Hive enable the portability of SQL-based applications in Hadoop. Apache Hive is used and designed by other firms, such as Netflix and the Financial Industry Regulatory Authority (FINRA), though initially being created by Facebook [41].

5.6.3 Apache Hbase

The database with Hadoop is Apache HBase (HBase). It is a large data center, distributed, scalable. HBase is a software member of the Apache Hadoop community that helps to read and write Big Data in real-time. Apache HBase's primary purpose is to host massive tables (billions of rows of X million columns) on top of generic hardware clusters according to the Apache Software Foundation [42].

The HBases supports per-column encoding, in-memory and Bloom filters as illustrated in the initial HBase. Input and output tables for MapReduce tasks performed in Hadoop can be accessed via the Java API and via REST, Avro and Thrift Gateway API. Based on its similarities with Hadoop and HDFS, HBase is a column-driven data store of main meaning and is commonly used. HBase is operating on HDFS and is ideal for fast reading and writing operations with high-performance, low input / output latency on large datasets [43].

5.6.4 Apache Pig

In order to simplify queries with large data sets for Apache Hadoop and Map Remediation, Apache is a high-ranking procedural-speech tool. Apache Pig has a layer "Pig Latin," which can be used to access Database-like SQL data sets in Hadoop applications.

It was claimed that when data is loaded, DBMSs are considerably faster than the MapReduce method, but the initialization of data on database systems requires considerably longer. RDBMS provides column management help out of the box function for compact files, random data access indexes and fault tolerance at the transaction level. Pig Latin is proceedings and blends in very well with the pipeline model, whereas SQL is rather declarative [44].

Pig was developed as a research project by Yahoo in order to create and operate map reduction employment on very large data sets. In 2007 the Apache Software Foundation open source software Pig [45].