A Review: Big Data Technologies with Hadoop Distributed Filesystem and Implementing M/R

[1] Renas Rajab Asaad, [2] Hawar Bahzad Ahmad, [3] Rasan Ismail Ali
[1,2,3] Department of Computer Science, Nawroz University, Kurdistan Region - Iraq

ABSTRACT

Today Big Data, is any set of data that is larger than the capacity to be processed using traditional database tools to capture, share, transfer, store, manage and analyze within an acceptable time frame; from the point of view of service providers, Organizations need to deal with a large amount of data for the purpose of analysis. And IT department are facing tremendous challenge in protecting and analyzing these increased volumes of information. The reason organizations are collecting and storing more data than ever before is because their business depends on it. The type of information being created is no more traditional database-driven data referred to as structured data rather it is data that include documents, images, audio, video, and social media contents known as unstructured data or Big Data. Big Data Analytics is a way of extracting value from these huge volumes of information, and it drives new market opportunities and maximizes customer retention. Moreover, this paper focuses on discussing and understanding Big Data technologies and Analytics system with Hadoop distributed filesystem (HDFS). This can help predict future, obtain information, take proactive actions and make way for better strategic decision making.

Keywords: Big Data, Hadoop Ecosystem, Hadoop Distributed File System, NameNode, DataNode.

1. Introduction

[1] There is a lot of work about the sale of data, or called "Big Data", which means the amounts of very large personal and professional information that can be analyzed to detect patterns, trends, groups and situations relating to human behavior and interactions. These data are a daily cumulative product of what Internet users and social media users have access to information about, such as photos and personal data, comments about their lives, ideas and affiliations, their diet, travel, health, sport, income, sex, entertainment and cultural interests. Four companies account for about 90 percent of this "treasure trove" - Google, Facebook, Apple and Amazon - according to a report by Data International. According to specialized reports, the digital world will reach 180 Zettabit in 2025, due to the strong demand for the use of the Internet in various areas of our lives. While Facebook and Google, for example, initially used data collected from users to better target their targeted ads, both companies and others have discovered in recent years that this data could be transformed into a

Academic Journal of Nawroz University
(AJNU) Volume 9, No 1 (2020).
Regular research paper : Published 17 Feb 2020
Corresponding author's e-mail : renas.rekany@nawroz.edu.krd

doi : 10.25007/ajnu.v9n1a530

number of "cognitive" [1]. And contribute to artificial intelligence feeding, some of which generate new sources of income from assessing user profiles by screening their writings and visual identification for purposes that can be sold to other companies for use in their own products or even for use by governments for legitimate security purposes. Values Market for technology companies and the Internet, and this explains the giant bypassed traditional companies in various other sectors. For example, Amazon's market value doubled 21 times in 10 years, while the market value of US retail giant Wal-Mart rose only 1.5 times in the same period [2].

HDFS is an abbreviation for Hadoop Distributed File System, a distributed, portable, and scalable system written in Java for the Hadoop framework where the Hadoop Distributed File System Block is formed by a cluster of data nodes. HDFS distributes data stored across servers in the cluster and stores multiple copies of data on different servers to ensure that no data is lost if a server fails (Hadoop file system provides high availability capabilities). The Hadoop file system is a fast-paced storage system where it is reclaimed at the end of the cluster. HDFS is a useful file system for caching and immediate results during MapReduce processing or as the basis for long-running clusters [3].

1.1 Objectives

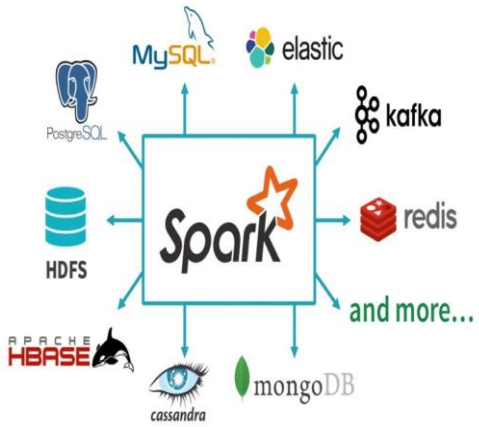
i. Understanding and Targeting Customers
ii. Healthcare Providers
iii. Education Sectors
iv. Improving Science and Research

2. Literature Review and Big Data Technologies

2.1 The Hadoop Ecosystem

The concept of Hadoop ecosystem includes the use of various parts of the core-hadoop, for example, MapReduce (a framework for dealing with a large amount of data), and Hadoop Distributed File System (HDFS), an advanced system for handling File System [4].

Fig 1. Hadoop Ecosystem



2.2 Spark

Apache Spark is an engine for processing Big Data within Hadoop, and it's a part of the Hadoop ecosystem, but it's more widespread. Spark is faster than the standard Hadoop engine a hundred times, MapReduce. A survey in the AtScale about Big Data said, twenty-five percent of the results had already used Spark in production, and thirty-three percent they are developed Spark projects. and many dealers offering Spark-based products [5].

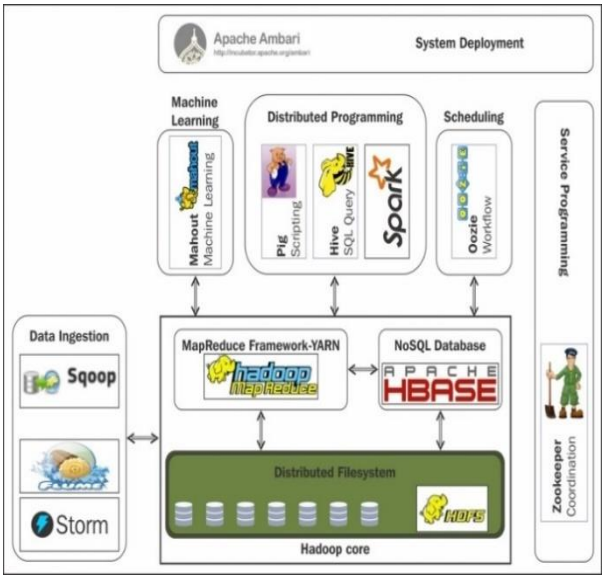


Fig 2. Apache Spark

2.3 R

R is a language and environment for statistical

doi : 10.25007/ajnu.v9n1a530

information processing and graphics. It is a GNU project similar to S language and environment developed by John Chambers and colleagues at Bell Laboratories. R can be considered as a different implementation of S. There are some important differences, but many codes written for S do not change under R [6].

2.4 Data Lakes

An easy way to access huge stores of data, many enterprises are setting up data lakes. These are vast data repositories that collect data from many different sources and store it in its natural state. The warehouse also collects data from disparate sources like data lake, but processes it and structures it for storage. In this case, the lake and warehouse metaphors are fairly accurate. If data is like water, a data lake is natural and unfiltered like a body of water, while a data warehouse is more like a collection of water bottles stored on shelves [7].

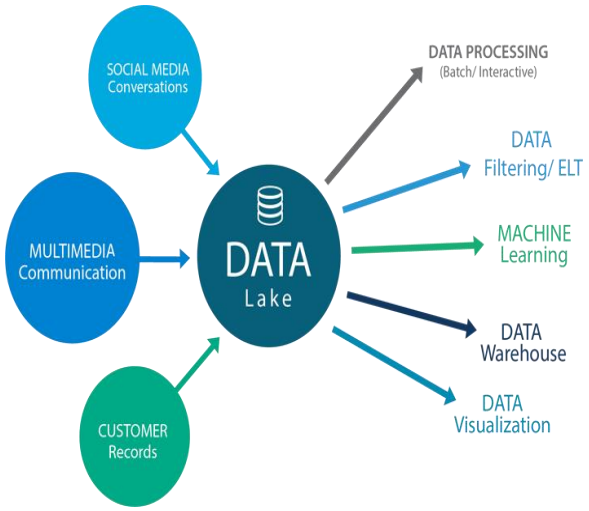


Fig 3. Data Lakes

2.5 NoSQL Databases

Storing information in a traditional way as structured in rows and columns with the database management system (DBMS). Developers and DB administrators query, manipulate and manage the data in that relational DBMS or (RDMS) using a special language known as SQL.

NoSQL databases is storing unstructured data and having quick performance, although they don't have the same level of consistency as relational DBMS. Common NoSQL databases contain MongoDB, Redis, Cassandra, Couchbase and many others; Nowadays, most of RDBMS users offer NoSQL databases [8].

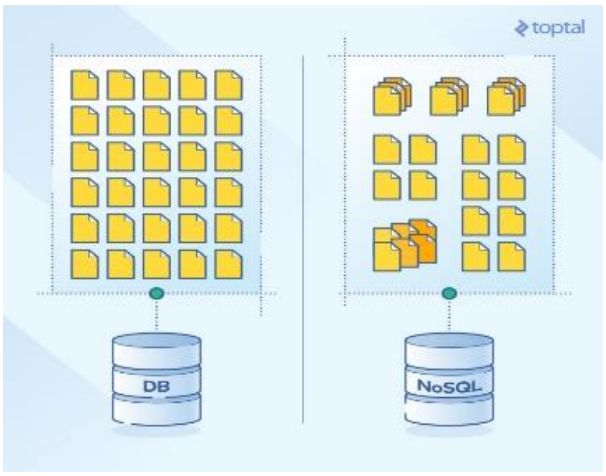


Fig 4. NoSQL Database

2.6 Predictive Analytics

According to the past data, predictive analytics is forecasting next events or behavior and its a subset of big data analytics. It depends on some techniques such as modeling, data mining and machine learning to predict future events. The purpose of using predictive analytics is for marketing, business analytics, forgery detection, finance and credit scoring.

doi : 10.25007/ajnu.v9n1a530

Zion Market Research says the Predictive Analytics market produced $3.5 billion in income in 2016, could be reached at 2022 in $11. Means the past ten years, advances in A.I. have enabled wide improvements in the abilities of solutions by predictive analytics. As a result, companies start to invest more in big data solutions with predictive abilities. Vendors like IBM, Microsoft, SAS, Statistica, SAP, RapidMiner and others, offer predictive analytics solutions [9].

Fig 5. Predictive Analytics

2.7 Artificial Intelligence

The amount of information we produce every year is doubled. Since in 2016 we produced a quantity of information equivalent to all that mankind produced until 2015. Every minute we do hundreds of thousands of searches through (Google) and publications through (Facebook). This information reflects our way of thinking and feeling. Soon, everything around us maybe our clothes too, will be connected to the Internet which means more information available about our lives. It is estimated that within ten years there will be one hundred and fifty billion sensors connected to the Internet, twenty times more than the number of humans on earth, and then the amount of information will double every twelve hours.

What is new is that this technology has become famous in the world of politics. Through "push" - a modern form of patriarchy - and on a large scale, governments are trying to direct citizens towards more harmonious behavior with the environment. So the government is not only interested in what we do but also wants us to do what seems to be acceptable. "Big push" is a product of payment from big data. Some people seem to have some kind of digital power that allows governments to control the masses efficiently without having to be involved in a democratic process. This can be judged by a wise ruler - or this is what we hope - supported by the data through which he can make decisions that serve the economy and society as if he had a magic digital wand [10].

2.8 Blockchain

Blockchain is a technology for storing, verifying and licensing digital transactions in the Internet by distributing databases to customers in a highly encrypted, decentralized, high-security and cryptographic format that would be impossible to break under the technologies available today.

Many researchers and experts assert that the technology of Blockchain will be the gateway to a world of innovation in the Internet space and the destabilization and change of business methods in a way that may disappear with many companies around the world as money transfer companies unless installed wave and adapt their work with the latest technologies.

This technique is a new type of database, which is unconventional or classical, decentralized, and provides an ever-increasing list of records that go beyond blocks or blocks.

Each block or block contains information and data,

doi : 10.25007/ajnu.v9n1a530

including time data and a link to the previous block, and the data is stored, not modified or manipulated, and if desired to do so there must be a decentralized consensus by the various elements of the network.

These databases store different data, from business transactions to business and medical records, customer IDs and any other type of data. They accept the verification of the source of transactions in a decentralized manner and without the intervention of any third party.

This technology provides the completion of transactions without a third medium, which means reducing the duration and costs, and can be programmed for a company or a group of institutions and the modification of the technology to be adapted for a specific purpose. [11].

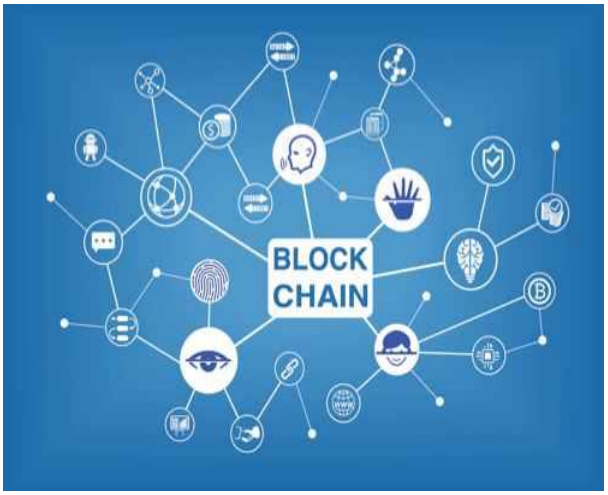


Fig 6. Blockchain

2.9 Prescriptive Analytics

Is the analysis by which statistics and modeling are used to determine future performance based on current and past data. Predictive analyzes can look for patterns that may appear or recur in the future, allowing companies and investors To modify their strategic plans and use their resources effectively to take advantage of possible future events. [12].

3. Hadoop Distributed File System

HDFS is the extremely significant component of Hadoop Ecosystem. Its the essential storage system of Hadoop. Based on a java file system that provides (scalable, fault tolerance, reliable and cost-efficient data storage) for Big data. Also, Its already configured with default configuration for many installations. Most of the time for large clusters configuration is needed. Hadoop interact directly with HDFS by shell-like commands [13].

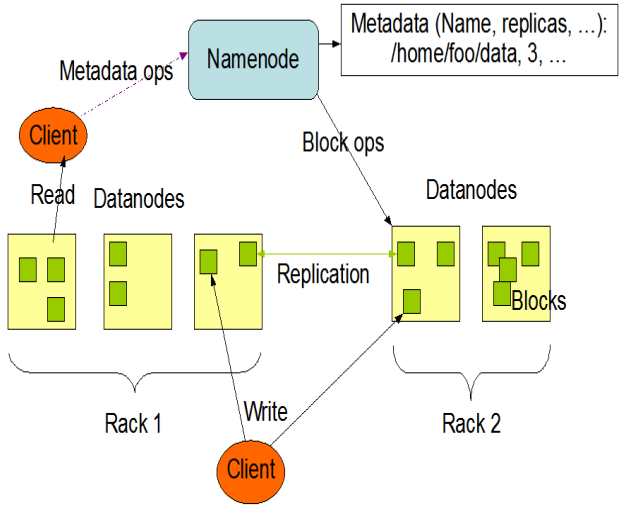


Fig 7. HDFS Architecture

The main components of Hadoop Distributed File System are:

3.1 NameNode and DataNodes

HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file

doi : 10.25007/ajnu.v9n1a530

system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode [13].

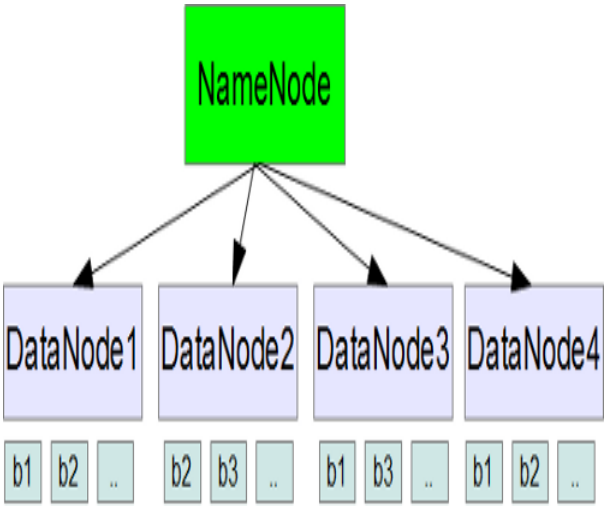


Fig 8. HDFS Components

4. YARN

YARN (Yet Another Resource Negotiator), is a JavaScript Package Manager used for processing [14].

4.1 ResourceManager

It is a cluster level component and runs on the main machine. It manages resources and schedule applications running on top of YARN [14].

4.1.1 NodeManager

It is a node level component (one on each node) and runs on each slave machine. It is responsible for managing containers and monitoring resource utilization in each container. It also keeps track of node health and log management. It continuously communicates with Resource Manager to remain up-to-date [14].

So, you can perform parallel processing on HDFS using MapReduce.

4.2 MapReduce

MapReduce engine, which consists of a function tracker, and MapReduce functions are pushed by client applications. In turn, the job tracker drives the work out to a task-tracking contract in the cluster, trying to keep work as close as possible to the data. With the conscious file system, the function tracker is aware of the node that contains the data, and any other machines located nearby. If the work can not be hosted on the actual node where the data exists, the priority is given to the nodes on the same shelf. This reduces the amount of data passing on its main spine. If the task tracker fails or takes time, this part of the task is rescheduled. The task tracker on each node generates a separate virtual Java machine to prevent the same task tracker from failing if the task it performs disables the default Java machine. The heartbeat of the task tracker is sent to the tracker function every few minutes to confirm its condition. Job Status Tracker and Job Tracker reports are reviewed through a web browser by a Jetty server.[14] [15].

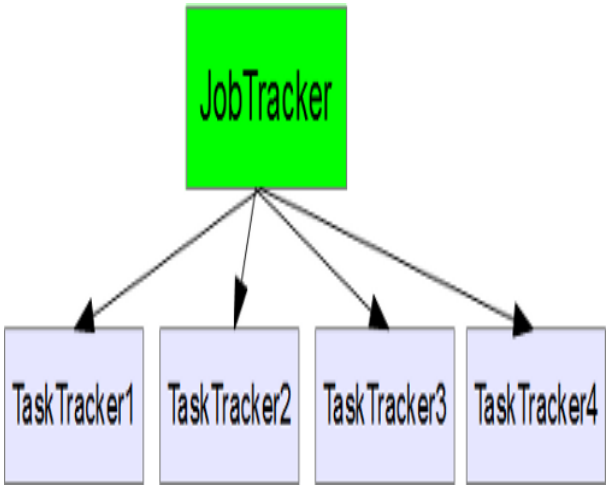


Fig 9. YARN Components

5. M/R Implementation and results

This is a simple linear Regression study computed in R programming language on MapReduce. Steps:

doi : 10.25007/ajnu.v9n1a530

i. Problem Statement.

ii. Implementation in R.

Code by R Studio:

```
score = read.table(file.choose(), header = TRUE)
attach(score)
plot(Day,Score)
reg = lm(Score~Day)
summary(reg)
abline(reg)
plot(reg)
```

Residuals:

1 2 3 4 5

4 7 -20 3 6

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	69.000	13.675	5.046	0.0150 *
Day	17.000	4.123	4.123	0.0259 *

Residual standard error: 13.04 on 3 degrees of freedom

Multiple R-squared: 0.85, Adjusted R-squared: 0.8

F-statistic: 17 on 1 and 3 DF, p-value: 0.02586

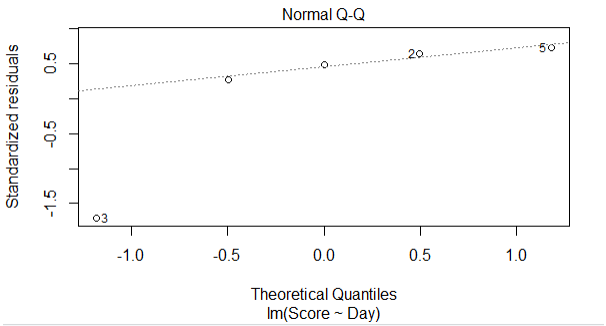


Fig 10. Normal Q-Q Result

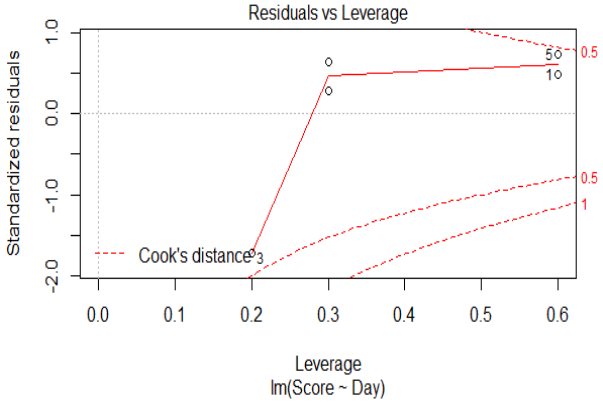


Fig 11. Residuals vs Leverage Result

iii. Implementation Map / Reduce.

Code by Java:

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class ProjectionMapper extends Mapper<LongWritable, Text, IntWritable, Text> {
 @Override
 public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
      String line = value.toString();
          String[] ScoreData = line.split("\\t");
        IntWritable userId = new IntWritable(Integer.parseInt(ScoreData[1]));
          Text intervalscore = new Text(ScoreData[0].toString() + "-" + ScoreData[2].toString());
          context.write(userId, intervalscore);}  }
```

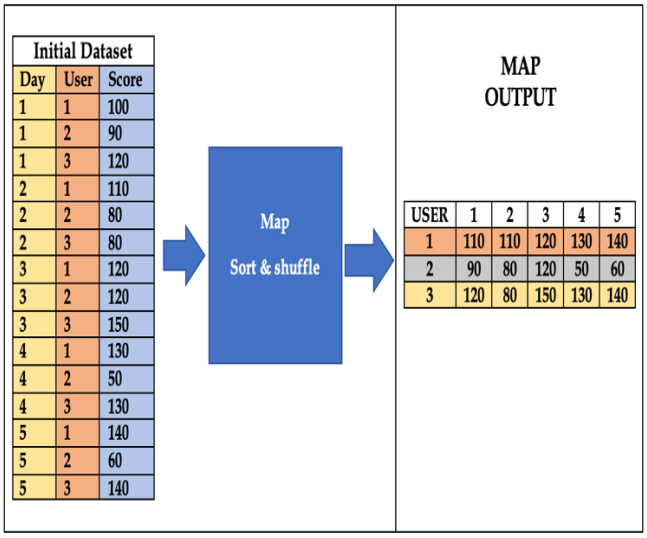
doi : 10.25007/ajnu.v9n1a530

```
import java.io.IOException;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class ProjectionReducer extends Reducer<IntWritable, Text, IntWritable, DoubleWritable> {
    protected void reduce(final IntWritable key, final Iterable<Text> values, final Context context) throws IOException, InterruptedException {
        int targetPeriod = 10;
        Participant participant = new Participant(values,targetPeriod);
        double projectedScoreAtTargetPeriod = participant.getProjectedScoreAtTargetPeriod();
        double projectedCumulativeScoreAtTargetPeriod = participant.getProjectedCumulativeScoreAtTargetPeriod();
System.out.println("For Participant: " + String.valueOf(key) + " the projected value at period " + String.valueOf(targetPeriod) + " is " + String.valueOf(projectedScoreAtTargetPeriod));
        System.out.println("For Participant: " + String.valueOf(key) + " the projected cumulative value at period " + String.valueOf(targetPeriod) + " is " + String.valueOf(projectedCumulativeScoreAtTargetPeriod));
context.write(key, new DoubleWritable(projectedCumulativeScoreAtTargetPer
iod));}}
```

Table 1: Hadoop Map Reduce Apply Regression Function

MAP OUTPUT							REDUCE Apply Regression Function	REDUSE OUTPUT	
User	1	2	3	4	5			User	Score
1	110	110	120	130	140			1	?
2	90	80	120	50	60			2	?
3	120	80	150	130	140			3	?

Table 2: Hadoop Map Reduce Sort and Shuffle



6. Conclusion

Recently, the big data technologies are changing very fast, the fields are changed such as in (Social Media, Medical Sector, Education, Internet in general, Weather Science, and in Astronomy). The big data is of great importance. It offers a high competitive advantage for companies if they can benefit from it and address it because it provides a deeper understanding of its customers and their requirements. This helps to make appropriate and appropriate decisions within the company in a more effective manner based on information extracted from customer databases.

Hadoob allows companies to store their data the way they are - organized or unorganized - so we do not have to spend money and time to create data for databases and rigid tables. Since Hadoop can easily handle inflation, it can be the ideal platform to capture all data from multiple sources simultaneously. The most popular adjective of Hadoob is its ability to store data at a much lower price than can be done with logical database management software. But this is only the first part of the story. The ability to keep this vast amount of data at a low price means companies can use all their data to make better decisions.

doi : 10.25007/ajnu.v9n1a530

7. References

1. A Day in Big Data. BIG DATA for smarter customer experiences. 2014. [ONLINE] Available at: http://adayinbigdata.com. [Accessed 03 November 15].
2. EMC Solutions Group. Big Data-as-a-Service. 2012, July. Retrieved from https://www.emc.com/collateral/software/white-papers/h10839-big-data-as-a-service-perspt.pdf
3. Dhawan, S & Rathee, S. Big Data Analytics using Hadoop Components like Pig and Hive. American International Journal of Research in Science, Technology, Engineering & Mathematics, 88. 2013 Retrieved from http://iasir.net/AIJRSTEMpapers/AIJRSTEM13-131.pdf
4. Enterprise Hadoop: The Ecosystem of Projects. Retrieved from http://hortonworks.com/hadoop/
5. Penchikala, S. Big Data Processing with Apache Spark - Part 1: Introduction. 2015, January Retrieved from http://www.infoq.com/articles/apache-spark-introduction
6. Grunsky, E. C. "R: a data analysis and statistical programming environment–an emerging tool for the geosciences." Computers & Geosciences. 28.10.2002.
7. Fang, Huang. "Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem." Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on. IEEE, 2015.
8. Tiwari, S. Using Oracle Berkeley DB as a NoSQL Data Store. 2011. Retrieved April 5 2015 from
9. Waller, Matthew A., and Stanley E. Fawcett. "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management." Journal of Business Logistics 34.2, pp.77-84. (2013).
10. O'Leary, Daniel E. "Artificial intelligence and big data." IEEE Intelligent Systems pp.96-99. 28.2 (2013).
11. MICHAEL, JW, ALAN COHN, and JARED R. BUTCHER. "BlockChain technology." The Journal (2018).
12. Deka, Ganesh Chandra. "Big data predictive and prescriptive analytics." Handbook of Research on Cloud Infrastructures for Big Data Analytics. IGI Global, Pp.370-391. 2014.
13. Shvachko, Konstantin, et al. "The hadoop distributed file system." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. Ieee, 2010.
14. Vavilapalli, Vinod Kumar, et al. "Apache hadoop yarn: Yet another resource negotiator." Proceedings of the 4th annual Symposium on Cloud Computing. ACM, 2013.
15. IBM. 2015. IBM - What is MapReduce. from: https://www.01.ibm.com/software/data/infosphere/hadoop/mapreduce/.

doi : 10.25007/ajnu.v9n1a530