International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 12, December 2015

Analysis of Data Using Hadoop and Mapreduce

[1] Bharti Kalra, [2] Dr. Anuranjan Misra, [3] Dr. D. K. Chauhan
[1] P.hD. Scholar(CSE) ,[2] Prof. & Head,Dept of CSE & IT [3]Director Technical,
Noida International University, Greator Noida(India)
[1]kalra.bharti1@gmail.com,[2] amc290@gmail.com,[3]prof.dkchauhan@gmail.com

Abstract: — We are in the world of technology, where data is moving around all the time. This data is becoming bigdata when it comes in huge volume and data can be structured, unstructured or semi-structured. hadoop is the technology to analysis the big data. The objective of this paper is to analysis the data that is collected from the open source using the hadoop and mapreduce programming model.

I.INTRODUCTION

Data is the most important assets for any organization. This Data can be collected by individual, or by any team or by any organization for any objective. So Big data, when this term comes into mind, so many definition moving around. Many people and organizations describe bigdata as its own way, based on the parameters (5vs)i.e. volume, variety, velocity ,variability and complexity. According to {2] Big Data is defined as the very large amount of datasets that hold variety of data.Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it [1].

II.BIG DATA ANALYTICS

According to a report [3], Co-sponsored by IBM, Big data analytics is where advanced analytic techniques operate on big data sets. Hence, big data analytics is really about two things—big data and analytics—plus how the two have teamed up to create one of the most profound trends in business intelligence (BI) today. Let's start by defining advanced analytics, then move on to big data and the combination of the two.

Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyse huge volumes of data that conventional analytics and business intelligence solutions can't touch.[4]

III.HADOOP

Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users. It is licensed under the Apache License 2.0[5].

The Apache Hadoop framework is composed of the following modules

***A.Hadoop Common*:** Contains libraries and utilities needed by other Hadoop modules

B.Hadoop Distributed File System (HDFS): A distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster

C.Hadoop YARN: A resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications

D.Hadoop MapReduce: A programming model for large scale data processing in addition to the previous overall configurations, the following individual components are also included on HDInsight clusters.

***E.Ambari*:** Cluster provisioning, management, and monitoring.

F.Avro (Microsoft .NET Library for Avro): Data serialization for the Microsoft .NET environment.

G.Hive & HCatalog: Structured Query Language (SQL)-like querying, and a table and storage management layer.

H.Mahout: Machine learning.

I.Oozie: Workflow management.

J.Phoenix: Relational database layer over HBase.

K.Pig: Simpler scripting for MapReduce transformations.

L.Sqoop: Data import and export.

M.Tez: Allows data-intensive processes to run efficiently at scale.

O.ZooKeeper: Coordination of processes in distributed systems.

This Paper proposes an implementation of Analytics of Big Data using Hadoop and Mapreduce.

IV. SIMULATION BACKGROUND

The implementation done on windows 8 with the hadoop 2.3.0, Map reduce and the JAVA version 6. The data is to be analyse using the word count algorithm available with the hadoop 2.3.0, map reduce is the core component of hadoop which is available to analysis data in parallel fashion, in two phase 1) Map 2) Reduce

Each map and reduce phase has a key-value pairs for input and output, Where map phase runs a map per function that generates a set of intermediate key-value pairs. Reduce phase reduce the number of data using the key-value pairs.

The reducer class display the final output by file name following the part-r-00000.

IV.PROBLEM DEFINITION AND METHODOLOGY

Consider database created by the sean lahman, who make baseball statistics, that is collected from open sources. This database contains Pitching, hitting and fielding statistics for major league baseball from 1871 through 2011[7]. This database contains tables numbered from 2.1 through 2.27, where each table contain data about players and leagues. A Sample size of master table contains 1031 entries and the AwardPlayers contain 2736 entries. These dataset first stored individually on a local file system. The aim of this data analysis to find out the solution of some aspects

A.Finding the number of individuals who served as managers and who are in the base ball hall of fame.
B.Award Won by the different number of players
C.Total no of who bats through left, right or both
D.Total no of throws through left or right.

V.ALGORITHM

We consider these steps to find out the solution of the above aspects by the mapreduce Programming model

a. Take the Data set
b. Load this data set to a directory
c. Writing the map, Reduce and Driver code for analysis the data.
d. Creating the jar file
e. Executing the jar file
f. Loading the output file

VI.EXPERIMENT RESULT

Fig A show the result of first query that is How many player are individual managers and players are base ball hall of fame, Fig B shows the result of second query, that is the name of award won by the different number of players.Fig C shows the result of third aspect that is how many players bat through left right or both hand.

VII. FUTURE WORK

Here is the very short analysis of the data. We found the result of very short aspects.In future we will publish some more results with some new aspects.

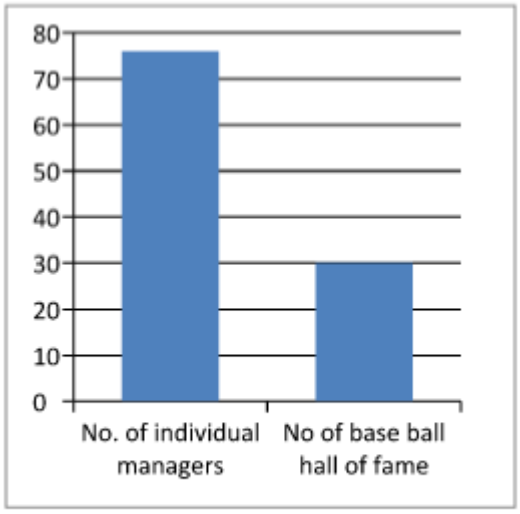

Fig A

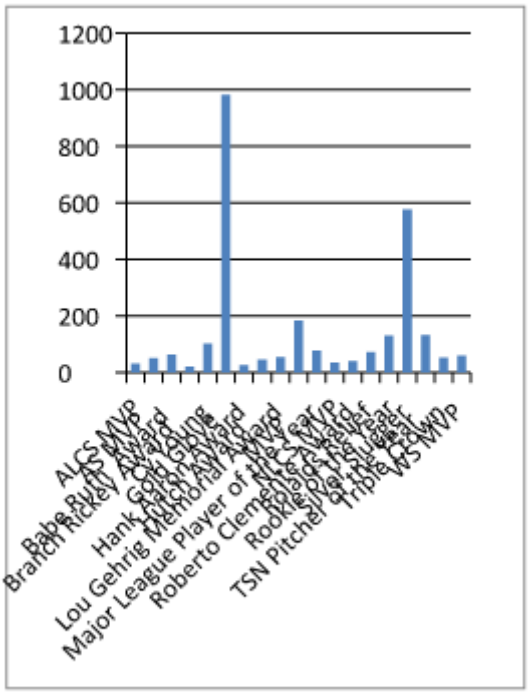

Fig B

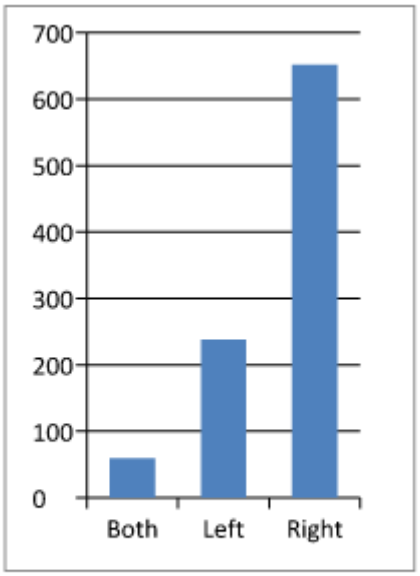

Fig C

VI. CONCLUSION

Our aim in this paper to analysis the research data through the mapreduce Programming model and summarizes the result of some aspects of multiples columns.

REFERENCES

[1] https://www.oreilly.com/ideas/what-is-big-data
[2] using *privacy by design* to achieve big data innovation without compromising privacy, by deloitte, june 10, 2014.
[3] big data analytics fourth quarter 2011 by philip russomftp://ftp.software.ibm.com/software/tw/defining_big_data_through_3v_v.pdf
[4]https://www.umb.edu/academics/caps/corporate/big_data_analytics
[5] http://opensource.com/life/14/8/intro-apache-hadoop-big-data
[6]https://azure.microsoft.com/en in/documentation/articles/hdinsight-hadoop-introduction/
[7]http://www.stat.ncsu.edu/people/monahan/courses/st590g/baseball/lahman591-csv/readme59.txt