iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VII **Month of publication:** July 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

Big Data Techniques for Efficient Storage and Processing of Weather Data

K.Anusha[1], K.Usha Rani[2]
[1] *Research Scholar,* [2] *Professor, Department of Computer Science*
Sri Padmavati Mahila University, Tirupati, India

Abstract*: *Big Data is mainly used to describe the exponential growth and availability of extremely large data sets that may be analysed computationally to reveal patterns, trends and associations. Applications of data analysis arise in many fields perhaps most importantly in Weather Forecasting. Weather Forecasting is one of the applications of Science and Technology to predict the state of atmosphere for upcoming time at a given location. Now-a-days as size of the weather data is increasing tremendously, Weather Forecasting seems to be built on Big Data Analytics. The accuracy of Weather Forecasting plays a vital role in daily routine of Business and their decisions. For accurate weather prediction, there is a need to store and process huge amounts of weather data. In general, a framework Apache Hadoop is most popularly used for storing and processing of large data sets. In this study, Spark and Cassandra Integration is experimented to evaluate the time taken to store and process large data sets for further processing. And the result is evaluated with Hadoop Map Reduce.

Keywords*: *Big Data, Hadoop, Map Reduce, Spark, Cassandra, and Weather Forecasting.

I. INTRODUCTION

Big Data is a collection of large volumes of heterogeneous data that is being generated often at high speeds. These data sets cannot be managed and processed using Traditional Data Management Tools. The size of the data will range from Terabytes to many Petabytes. Big data is an umbrella term that not only refers to huge amount of data available today but also the complete process of gathering, storing, retrieving, and analysing the available data. The task of capturing, managing and processing the huge amount of various types of data is very difficult in Traditional Databases. Big Data exceeds the processing capability of Traditional Databases. The Big Data from various sources available now have increased attention among various researchers in all fields with every attempt to exploit the value of knowledge resulting from its processing power and analysis [1] [2].

The SQL Databases are not the best to handle big data mainly because of their limited capacity and cost of scaling to accommodate large data. NoSQL databases try to solve some of the problems posed by Relational Database Management Systems (RDBMS).

Weather Forecasting is one such source which contributes huge data. The Weather dataset is collected from National Climatic Data Centre (NCDC). In weather forecasting, the raw data is received through the satellite and then delivered over to the different servers of the weather channels and this raw data is stored in the cluster.

A meteorologist is trying to predict how the weather will change during a quantified period and what the weather conditions will be during the period of the forecast. Weather affects everyone almost every day [3].

A. Weather Forecasts are Issued

1) to save lives
2) reduce property damage
3) reduce crop damage
4) to let the general public know what to expect.

Apache Hadoop, which is the common framework for storing Big Data is time consuming. In Hadoop, data is stored in HDFS and processed using Map Reduce.

Apache Spark is an important framework for performing general data analytics on distributed computing cluster like Hadoop. It provides an in- memory computation to increase speed of data processing over map reduce. It runs on top of existing Hadoop cluster and access Hadoop data store i.e. Hadoop Distributed File System (HDFS) [12].

Apache Cassandra was developed by Facebook. It's an open source distributed database management system. Cassandra is a scalable non-relational database that offers continuous availability, linear scale performance, operational simplicity and easy data

distribution across multiple data centres and cloud availability zones [18]. The scalability is very high compared to all other databases. Just to add nodes to the cluster in Cassandra achieves the goal of scalability. It's a peer to peer architecture and it uses consistent hashing to partition the data across the cluster of nodes and hence there's no single point of failure. Cassandra's architecture has the ability to scale, perform and offer continuous uptime. Cassandra provides a number of key features and benefits and is used as the essential database for modern online applications [19].

Due to the limitations in Hadoop Map Reduce framework, it is necessary to implement an efficient framework which reduces the processing time. Hence, Spark-Cassandra integration is experimented to produce better results.

The rest of the paper is organized as follows: Section 2 describes Literature Review; Section 3 represents the Methodology; Section 4 represents the Proposed System; Section 5 represents the experimental analysis; Section 6 represents the conclusion.

II. LITERATURE REVIEW

Extensive studies have been carried out on Big Data Techniques related to Weather Forecasting. In this section few studies related to Apache Hadoop, Apache Spark and Apache Cassandra are presented.

Dhanashri V. Sahasrabuddhe et.al [3] provides a short introduction of data structures used for representing Big Data for weather forecasting.

Khalid Adam Ismail Hammad et.al [4] provides an introduction about Big Data frameworks, platforms, Databases for Big Data, data storage and Big Data Management and storage. They also presented a Big Data analysis and Management including Big Data with Data Mining, Big Data over Cloud Computing and Hadoop Distributed File System and Map Reduce.

Hossein Hassani [5] believes that Big Data will soon be predicting our every move. Big Data is most commonly required after for building the predictive models in a world where forecasting persist to remain as an important statistical problem. The traditional forecasting tools are not able to handle the size, speed and complexity that is inherent in Big Data.

Timothy D. Rey [6] believes that there is significant value in the interdisciplinary notion of data mining for forecasting when used to solve time series problems.

Sam Madden [7] explains that existing tools do not lend themselves to sophisticated data analysis at the scale many users would like.

Arribas-Bel [8] was of the view that current statistical software is not able to tackle Big Data forecasting and also believe that this is an outstanding lack of a structure in these data sets and the size. Therefore, forecasting Big Data possess a challenge to organizations.

Rey and Wells [9] believe that Data Mining techniques can be imposed to help forecasting with Big Data. However, it should be noted that in the past, Data Mining techniques have been used mainly on static data as contrasted to time series. The opportunities for benefits through forecasting with Big Data are very different. At present, there is an increased research into using Big Data to obtain accurate weather forecasts and the initial results suggest that Big Data will benefit weather forecasts extremely. In fact, weather forecasting has been one of the main beneficiaries of Big Data.

Pedja Bogdanovich et.al [10] proposed one of the Data Structures known as ATree for Large datasets. The Atree data structure is used to store abstracts of data, which is less in size and is capable of satisfying most of the queries of user as well as the original data. The Atree data structure uses B-tree as its underlying data structure. B-tree is disk backed up and is slower than in-memory data structures used to provide faster access to data.

Veershetty Dagade et.al [11] proposed "Big Data Weather Analytics Using Hadoop". In this paper, the authors suggests that huge amount of weather data is shifted to HDFS system and Hive programming proves to be better to analyse data for huge volumes.

Mike Olson [14] proposed Hadoop and Map Reduce which provides more storage at lower cost than legacy systems.

Bhalchandra Bhutkar et.al [15] proposed that although traditional databases are useful for performing complex analytical queries in wide variety of applications they have several issues while handling the huge amount of data. NoSQL technologies like HBase, Cassandra, and Mongo DB have gained significance over the years because of their ability to handle big data in distributed environment. These technologies are mostly open source and provide a means of handling a significantly large volume of data with lower cost and easier management than traditional RDBMS .

Lekha R. Nair et.al [21] addresses the issue of real time analysing and filtering those numerous job advertisements from among the millions of other streaming tweets and classify them into various job categories to facilitate effective job search, utilizing Spark.

Basvanth Reddy et.al [22] addresses the issue of weather data analysis using Hadoop Map Reduce Technique.

Abdul Ghaffar Shoro et.al [23] explore the concept of Big Data Analysis and recognize some meaningful information from some sample big data source, such as Twitter twits, using one of industries emerging tools, known as Spark by Apache.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

III. METHODOLOGY

In general, Hadoop Map Reduce is used for Weather Forecasting [14], [22]. In this study, we experimented Spark for processing and Cassandra for data storage to compare with Hadoop Map Reduce. Brief description about Hadoop Map Reduce, Spark and Cassandra are presented in this section.

A. Hadoop Map Reduce Implementation

Hadoop is a parallel data processing framework that has conventionally been used to run map/reduce jobs. These are long running jobs that might take minutes or hours to complete [11], [14]. The flowchart of a Map Reduce execution is shown in figure 1 [24].

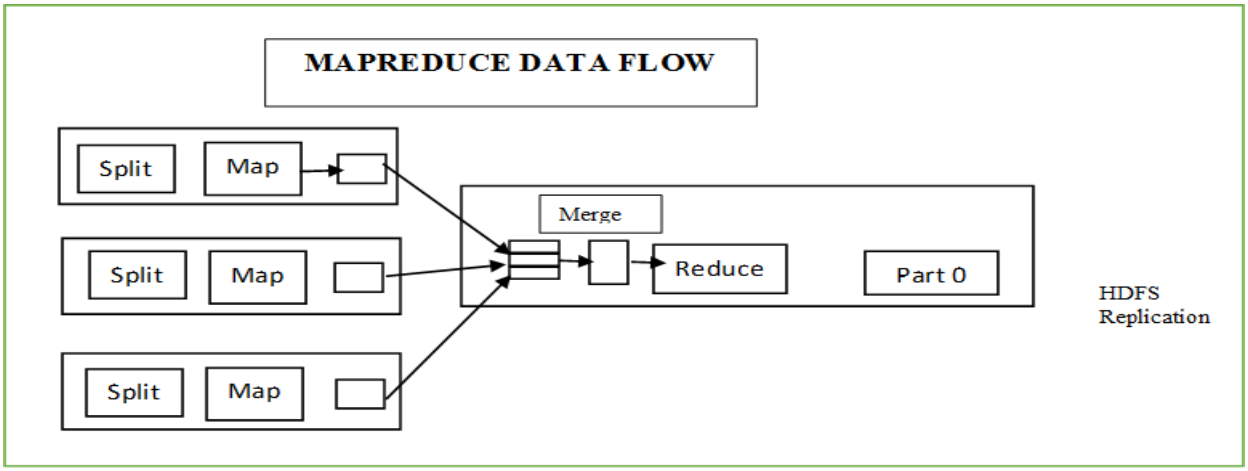


Fig 1: Flow chart of a Map Reduce Execution

In Hadoop Map Reduce implementation, the processing of data may take a long time. The input data is divided into splits and those splits are passed to mapper and then the output from maps is given as input to reducer. Hence, in this study, we experimented on Spark to process weather data in an efficient manner compared to Hadoop Map Reduce.

B. Spark Implementation

Spark is a faster execution engine which can provide up to 10x performance over to Map Reduce. Spark gets most of its speed by the construction of Directed Acyclic Graph (DAG) out of the job operations and uses memory to save an intermediate data, thus making the reads extremely efficient operations [17].

Spark consists of Master/Worker architecture. There is a driver node that talks to a single coordinator known as master that manages all the workers in which executors run [12], [13]. The architecture of Spark is represented in Figure 2 [13].

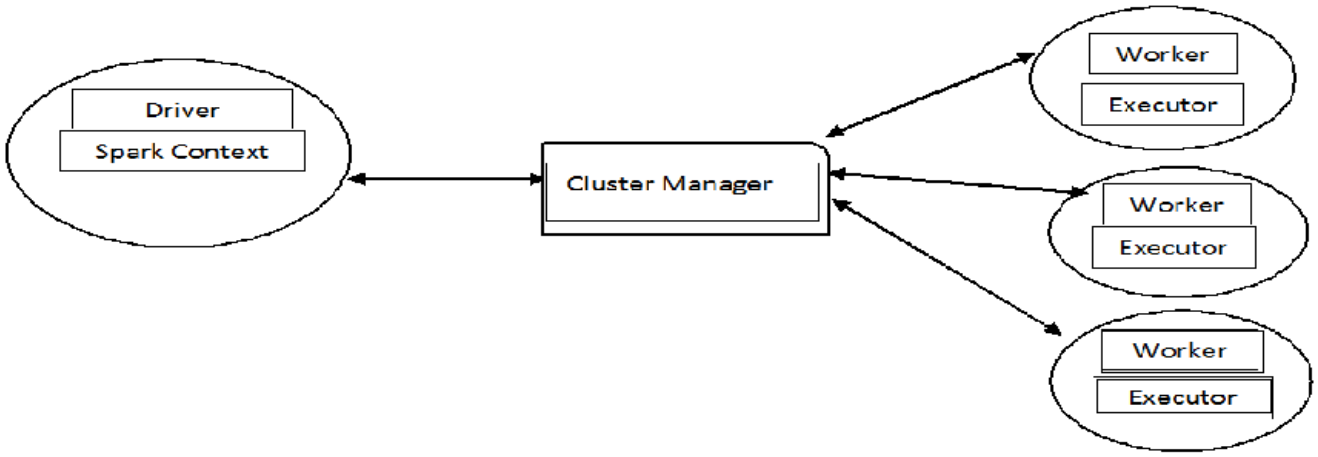


Fig 2: Spark Architecture

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

Spark has considered to run on top of the Hadoop and it is one of the best alternative to the traditional batch map/reduce model that can be used for real-time stream data processing and for fast interactive queries that will finish within seconds. So, Hadoop supports both traditional map/reduce and Spark. The Spark platform which runs on top of Hadoop is represented in Figure 3 [17].

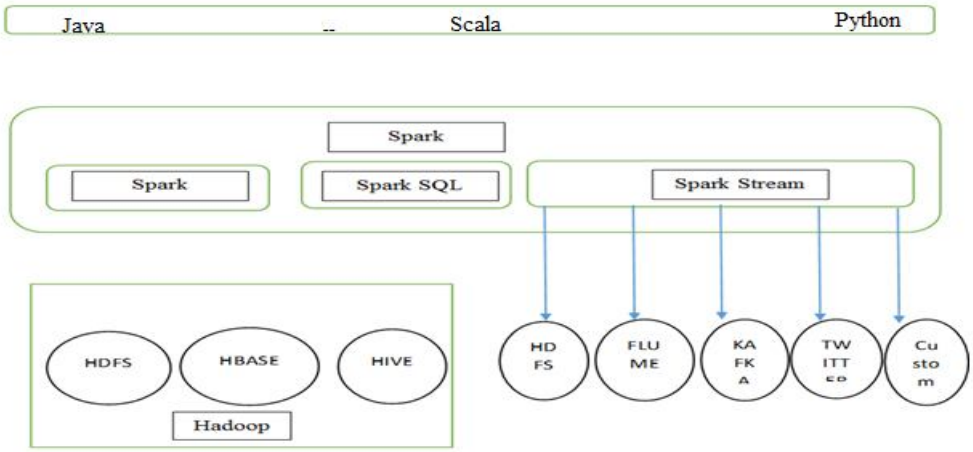


Fig 3: Spark Platform

Spark has capabilities like, in-memory data storage and real time processing. Spark uses more RAM instead of network and disk I/O and is quite fast as compared to Hadoop. The performance of spark can be several times faster than other Big Data technologies. Spark holds the intermediate results in memory instead of writing them to disk which is very useful. Spark works both with an in-memory and on disk. Spark will attempt to store as much as data in memory and then will spill to disk. Core concept in Spark is the Resilient Distributed Datasets (RDD) [18]. RDD is just like a table in a database. It can hold any type of data and stores data on different Partitions.

The Resilient Distributed Dataset is an important concept and is at the heart of Spark. It is designed to support in-memory data storage distributed across a cluster in a manner that it is both fault-tolerant and efficient. Fault-tolerance is achieved, by tracking the transformations applied to coarse-grained sets of data. Efficiency is achieved using the parallelization of processing across multiple nodes in the cluster, and minimization of data duplication between these nodes. Once data is loaded into an RDD, two simple operations can be carried out:

1) Transformations create a new RDD by changing the original through the processes such as mapping, filtering, and more.
2) Actions, such as counts, which measures but do not alter the original data.

The original RDD remains unchanged. The sequence of transformations from RDD1 to RDDn is logged and can be repetitive in the event of data loss or the failure of a cluster node.

Transformations are lazily evaluated, meaning that they are not executed until there is a need of subsequent action for the result. This will generally improve performance, as it can avoid the need to process the data unnecessarily. It can also, in certain conditions, introduce processing bottlenecks that cause applications to stand while waiting for a processing action to conclude [22].

C. Databases

1) *Hadoop Distributed File System:* In general, HDFS is used to store data in Hadoop. Hadoop and HDFS use a Master-Slave architecture. HDFS has number of advantages. But a number of benefits are derived from using Cassandra over HDFS. One of the important benefit of Cassandra is rather than using a master-slave design, Cassandra has a peer-to-peer distributed "ring" architecture which is very easy to setup, and maintain. In Cassandra, all nodes are the same. There is no idea of a master node, with all nodes communicating with each other via a gossip protocol [26]. Hence, in this study, we used a NoSQL database Cassandra to store large datasets of weather.
2) *NoSQL Database – Cassandra*: A NoSQL database provides a mechanism for storage and retrieval of data other than the tabular relations used in relational databases. NOSQL databases are increasingly used in Big Data and real-time web applications. Apache Cassandra is a NOSQL database best for high speed and online transactional data [19] [20]. Apache

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

Cassandra has a master less ring architecture which is efficient, easy to set and maintain. The architecture of Cassandra is represented in Figure 4 [20].

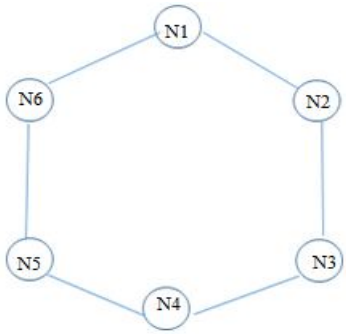


Fig 4: Cassandra's maser less ring architecture [20]

3) Input Dataset: In this study, the weather data is collected from National Climatic Data Centre (NCDC) website. NCDC has data for every month with hourly basis. The weather dataset contains fields like date, location id and observations for each weather parameters like temperature, humidity, pressure, etc. [25].

IV. PROPOSED SYSTEM

In this study, we proposed a method using both Spark and Cassandra for processing of weather data with reduced time compared to Hadoop Map Reduce.

A. Spark-Cassandra Connector

The architecture for the proposed system is represented in the following figure 5.

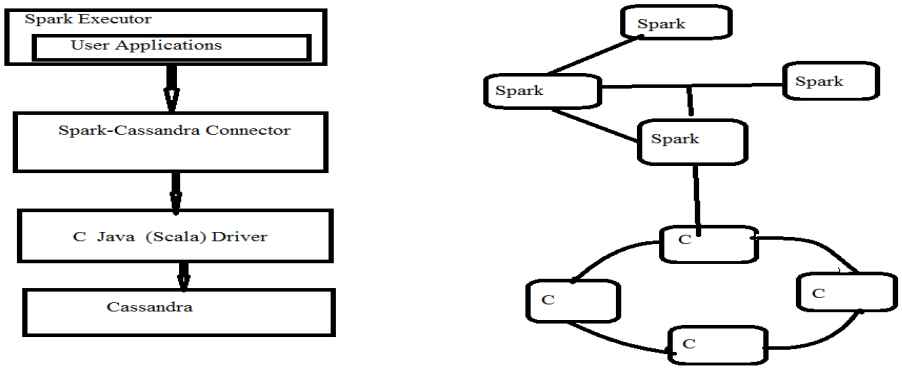


Fig 5: Architecture for proposed system (Spark – Cassandra Connector)

Apache Cassandra is a NoSQL database platform particularly suited for these types of Big Data challenges. Cassandra's data model is an outstanding fit for handling data in sequence regardless of data type or size. When writing data to Cassandra, data is sorted and written sequentially to disk. When retrieving data by row key and then by range, there is a fast and efficient access pattern due to minimal disk seeks. Using Spark with Cassandra lessens the number of Spark transformations required on our data because Cassandra does the work in its cluster.

B. Proposed Algorithm

1) Collect weather data from weather stations.
2) Run the Spark-Cassandra Connector.
3) Load weather data into Cassandra.
4) Extract weather data from Cassandra through Spark RDD's.
5) Evaluate the time taken for processing.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

The General Flow Chart for storing and processing of Weather Data using Spark and Cassandra is shown in the following figure 6.

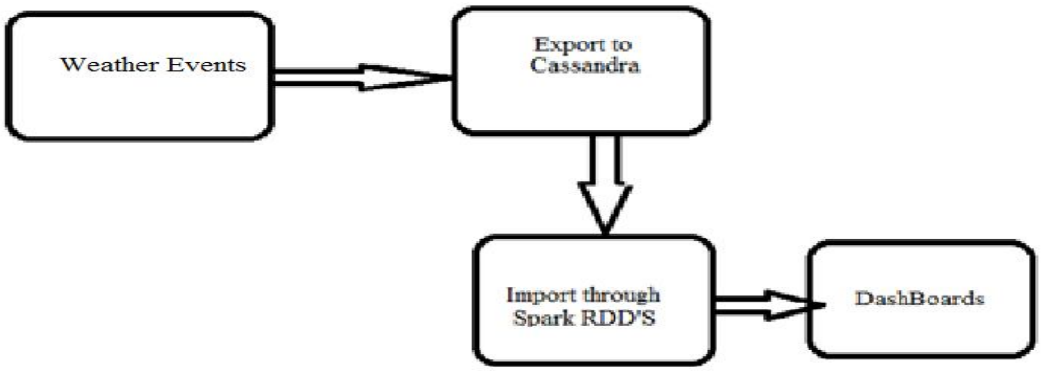


Fig 6: Flow Chart for Weather Data Analysis

V. EXPERIMENTAL ANALYSIS

The benchmarking was done based on the comparison between Hadoop Map Reduce implementation and the proposed method i.e., Spark-Cassandra connector implementation.

In this experiment, it is observed that the time taken to process the data by using the proposed method is less compared to the time taken in Hadoop Map Reduce implementation.

The results are as shown in Table 1

TABLE I

COMPARISION OF TIME (SEC)

Data Size	Time taken to import (sec)	
	Hadoop Map Reduce	Proposed Method
512 MB	30	1.2
1 GB	45	1.7

Hadoop Map Reduce is better only for full scan of data. Hence, in the proposed system, Spark-Cassandra integration is experimented to filter the weather data and to extract only few columns using Spark SQL. The following figure 7 shows the graphical representation of analysis.

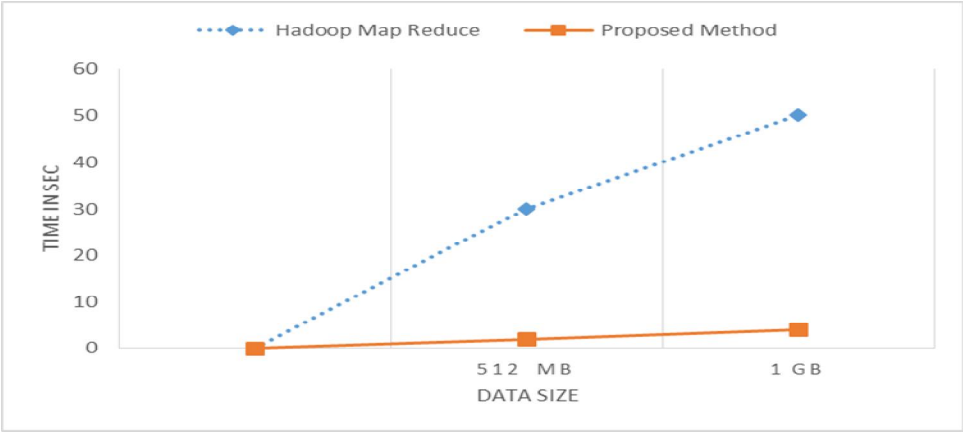


Fig 7: Hadoop Map Reduce vs Spark-Cassandra Benchmarking

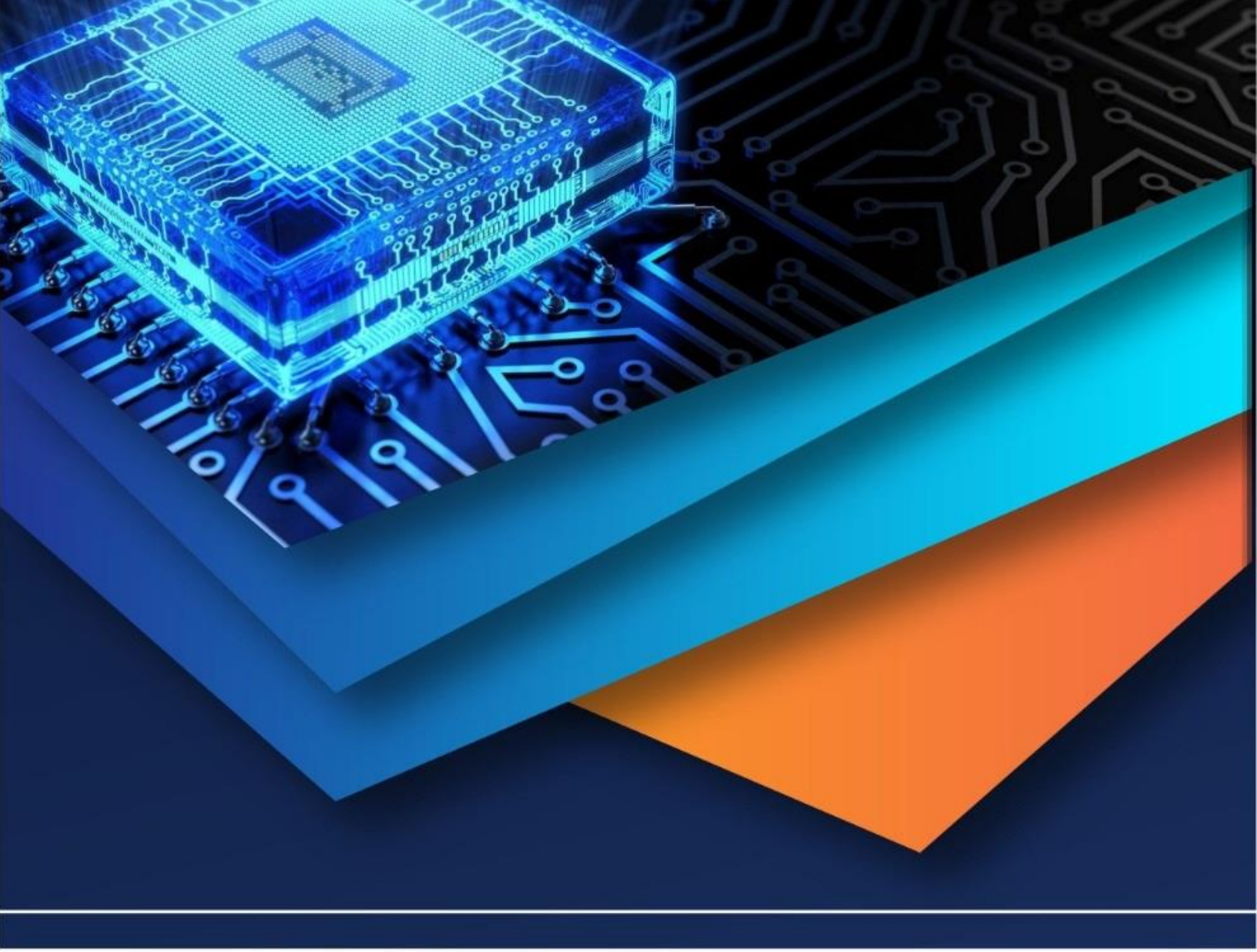
From the chart it can be seen that the performance of the proposed method's implementation is much better than the corresponding Hadoop Map Reduce implementation.

VI. CONCLUSIONS

In this study, we have proposed a method for storage and processing of weather data using Big Data techniques. The method experimented is an integration of Spark and Cassandra and the result is compared with Hadoop Map Reduce implementation. It is observed that the performance of proposed method has better performance. Hence it could be concluded that the proposed system is better than the Hadoop Map Reduce.

REFERENCES

[1] (2013) website. [Online]. Available: http://www.cse.wustl.edu/~jain/cse570-13/ftp/m_10abd.pdf
[2] Torlone. [Online]. Available: http://torlone.dia.uniroma3.it/bigdata/L1-Introduzione.pdf
[3] Dhanashri V. Sahasrabuddhe, Pallavi P. Jamsandekar "Data Structure for representation of Big Data of Weather Forecasting: A Review" Volume 3 Issue 6, Nov-Dec 2015.
[4] Khalid Adam Ismail Hammad, Mohammed Adam Ibrahim Fakharaldien, Jasni Mohamed Zain "Big Data Analysis and Storage", September 10, 2015.
[5] Hossein Hassani and Emmanuel Sirimal Silva, "Forecasting with Big Data: A Review", Volume 2, Issue 1, pp 5–19, March 2015.
[6] Timothy D. Rey, Justin Kauhl "Using Data Mining in Forecasting Problems", in SAS GLOBAL FORUM 2013, paper 085-2013.
[7] Sam Madden, "From Databases to Big Data", IEEE std.2012.
[8] Arribas-Bel D (2014) Accidental, open and everywhere: emerging data sources for the understanding of cities, Volume 49, pp 45-53, May 2015.
[9] T Rey, C Wells "Integrating data mining and forecasting", in INFORUMS, December 2013, Volume 39.
[10] The ATree: A Data Structure to Support Large Scientific Databases. Pedja Bogdanovich, Hanan Samet. 1999, Springer-Verlag Berlin Heidelberg, pp. 237-248.
[11] Veershetty Dagade, Mahesh Lagali, Supriya Avadhani, Priya Kalekar "Big Data Weather Analytics Using Hadoop", Volume 14 Issue 2, April 2015.
[12] Jameel Mohammed (2015) Aptuz homepage [Online]. Available: http://aptuz.com/blog/is-apache-spark-going-to-replace-hadoop.
[13] The Gitbook website. [Online]. Available: https://jaceklaskowski.gitbooks.io/mastering-apache-spark/content/spark-architecture.html.
[14] Mike Olson (2012) Cloud era website. [Online].Available: https://blog.cloudera.com/wp- content/uploads/2010/05/Olson_IQT_Quarterly_Spring_2010.pdf.
[15] Bhalchandra Bhutkar, "Data Management using Apache Cassandra" SAS Research and Development (India) Pvt. Ltd.
[16] Apache Spark [Online]. Available: https://www.tutorialspoint.com/apache_spark/apache_spark_introduction.htm.
[17] Apache Spark [Online]. Available: https://www.infoq.com/articles/apache-spark-introduction.
[18] Apache Cassandra [online]. Available: https://academy.datastax.com/planet-cassandra/what-is-apache-cassandra.
[19] Apache Cassandra [Online]. Available: https://www.tutorialspoint.com/cassandra/cassandra_introduction.htm.
[20] Data Analytics [Online]. Available: https://www.safaribooksonline.com/library/view/data-analytics with/9781491913734/ch04.html.
[21] Lekha R. Nair, DR. Sujala D. Shetty "Streaming Twitter Data Analysis Using Spark For Effective Job Search" Journal of Theoretical and Applied Information Technology –Volume 80. No 2, October 2015.
[22] Basvanth Reddy, Prof. B.A Patil "Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique" Journal of Advanced Research in Computer and Communication Engineering-Volume 5, Issue 6, June 2016.
[23] Abdul Ghaffar Shoro & Tariq Rahim Soomro, "Big Data Analysis: A Spark Perspective" Global Journal of Computer Science and Technology: C Software & Data Engineering Volume 15 Issue 1 Version 1.0 Year 2015.
[24] Hadoop Map Reduce flow [Online]. Available: http://data-flair.training/blogs/hadoop-mapreduce-flow-how-data-flows-in-mapreduce/
[25] NCDC weather data [online]. Available: https://www.ncdc.noaa.gov/orders/qclcd/
[26] Apache Cassandra [Online]. Available: https://www.datastax.com/wp-content/uploads/2012/09/WP-DataStax-HDFSvsCFS.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY