Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

Real-Time Applications of Big Data- A Survey

Rubal
Department of Computer Science & Engineering
Guru Nanak Dev University, Regional Campus
Jalandhar, India

Sheetal Kalra
Department of Computer Science & Engineering
Guru Nanak Dev University, Regional Campus
Jalandhar, India

***Abstract*--With the rapid growth of technologies, large amount of data is produced from different sources that can either be structured or unstructured. Such type of data is very difficult to process and manage that contains millions records of information that includes social media, web sales, audios, videos, images etc. Timely analytics of this data is a key factor in success in many business and service domains. Examples of these domains are intelligent transport, healthcare, security, finance and military. To improve quality of information and decision making it is important to effectively analyze this large volume of data to answer new challenges.**

Keywords--Big data, Hadoop, HDFS, MapReduce, NoSql, Real-time data analytics.

I. INTRODUCTION

Big data is collection of massive and complex data sets whose size is beyond the ability of a typical database software tool. Traditional database software tools are unable to store, manage and analyze huge volume of data. Big data can be classified into two categories:

1) Data from the *physical world* which is obtained from sensors, scientific experiments, biological data, and remote sensing data etc.
2) Data that comes from the *human society* such as data from social sites, internet, government, health department, finance, and transportation [1]. Big data is characterized by 4V's, namely, volume, velocity, veracity, value and variety.

1) Variety means data comes from various sources that can be structured, semi-structured and unstructured type [2]. Different variety of data includes text files, audios, videos, sensor data etc.
2) Volume represents the size of data which is represented in terabytes and petabytes.
3) Velocity defines the motion of data and the analysis of streaming of data.
4) Value is the most important aspect of big data. It is used for business and IT infrastructure systems to store large amount of values in database. At the end, if we cannot extract value from our data, then it is useless in building the capability to store and manage it.
5) Veracity signifies the uncertainty and impreciseness in data.

For real-time processing data processing , data capturing and data exportation must be combine together. The main aim of big data real time processing is to realise the mesh of data and process it in short time. Various techniques and algorithms have achieved the acceptable performance, but not yet have proposed an entire feasible real time processing system for big data.

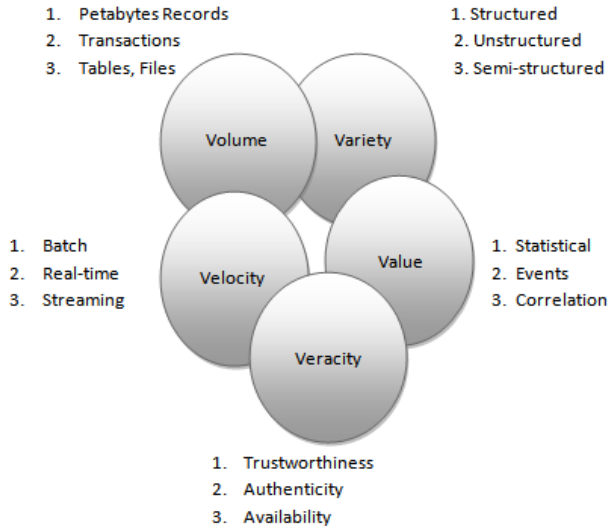

Fig 1. 5 V's of Big Data

Today business and organizations are learning more about their operations, products and goals using big data analysis. And they can use this knowledge into improved decision making so that they can achieve better performance and gain strategic advantage over their competitors.

II. RELATED WORK

More than 54 million results are shown on Google while searching about the big data application platforms, which provides the trends of big data and application platform used in industries. In 2011, many big companies are planned to use data on social media. Because of the increasing usage of Facebook, Twitter, snap-chat, Instagram and other applications of Web 2.0 for business. This is also known as an Internet-driven transformation of enterprises [3]. According to the prediction of McKinsey &Company, many marketers are expecting to spend their 19.5% of budgets on social media, which is nearly 3 times the current level of spending.
Undoubtedly, big data means big opportunities. It brings grand challenges to the digitization of industries. Therefore, In 2014 , According to survey of IDC, 70% of enterprises have either started or are planning to start big data related programs and projects, and the expected expense of which is 8 million US dollars in the coming year [3]. Nowadays, it can be seen that big data moves into enterprises. By the year of 2020, modern enterprises will need to manage 50 times more information, which is really a big challenge for medium-sized and small enterprises [3].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

Today, many organizations are applying big data analytic for various activities like fraud detection, traffic management, decision making etc. All of these activities require a platform that can process the data and provide useful insights. As Hadoop's ecosystem evolves, new open source tools and capabilities are emerging that helps in generating more capable and business relevant big data analytics environment. There are various studies on benchmarking and enhancing the MapReduce paradigm and NoSQL technologies separately. They analyze new distributed storage systems such as Cassandra and Hbase. Some of the features they study include data models, consistency, storage mechanisms, availability and query support [4].

MapReduce and NoSQL technologies have been used together in the industry for web appliactions like social media, online gaming and e-commerce. DataStax has introduced DataStax Enterprise which is big data platform built on top of Cassandra and includes support for Apache Hadoop and side products like Hive and Pig. HDFS is not the only platform for performing big data tasks. DataStax uses CFS on Cassandra data store which offers replacement of HDFS [12]. Datastax use DataStax OpsCenter to manage CFS and database clusters. DataStax OpsCenter is a visual management and monitoring solution for Cassandra database clusters. It allows a developer to easily manage and monitor all aspects of database from any desktop without installing any client software.

Due to great significance and value of big data, many countries have launched their initiatives on big data related research and applications. In March 2012, the Obama administration launched officially the big data research and development initiative with the investment of more than US$200 million [1].

III. PROCESS MODEL FOR BIG DATA APPLICATION

1. *Data Collection*: It is the process of collecting data from various sources and then storing it in file system known as Hadoop distributed file system (HDFS). Hadoop platform is used for processing large amount of unstructured data. HDFS is a base component of Hadoop and is used as the primary storage system. RDBMS and NoSQL are two types of databases used for big data. In a normal relational database, data is found and analyzed using queries, based on structured query language. Non-relational databases also use queries. They are just not constrained to use only SQL, but can also be used as other query languages to extract information out of data stores. Hence, the term is NoSQL (Not only SQL). It is not replacement of SQL instead it is compliment to SQL. It provides an easy way to store unstructured data from various sources [4,16]. There are various issues that relational model cannot address, so in that case non-relational databases are used to provide more scalability and superior performance.

2. *Data Cleaning:* It is a process of checking and correcting the corrupt data and inaccurate records in database [5]. If corrupt data is found then such files need to be restructured in more processed form. Data cleaning requires descriptive analysis and after cleansing process, all the data sets should be consistent. This form of analytics allows you to condense big data into smaller and more useful bits of information or summary of what happened in past. It helps to understand the relationship between the customers and products and also gain advantage from the past behavior and understanding what approach to take in future. For example, Netflix uses descriptive analysis to find the correlation among the movies and provide recommendations to improve their recommendation engine. Descriptive analysis is basically about the past and to determine what to do next.

3. *Data Classification:* Data classification involves filtering of data based on their structure. Structured, unstructured and semi-structured data need to be classified in order to perform meaningful analysis.

4. *Predictive and Prescriptive Analytics:* The predictive analytics is about the future and provides the companies with actionable insights based on the data. It cannot predict the future but only forecasts what might happen in the future. For predictive analysis it's important to have as much data as possible. More data means better predictions. Then prescriptive analysis provides advice based on prediction that is "what to do". Prescriptive analytics tries to know what the effect of future decisions will be, in order to alter the decisions before they are actually made. This will enhance decision-making a lot as future outcomes are taken into consideration while doing the prediction. It requires two main components: one is actionable data and other is feedback system to track the outcome produced by action taken. Prescriptive analytics has a large impact on the business and also on decision makers of an organization for taking effective decisions. It can optimize the scheduling, production and delivery of the right products to right customers within right amount. One well-known company 'Ayata' is using and taking the advantage of prescriptive analysis [9].

5. *Data Delivery:* In this last step, reports are generated based on the analysis of big data then these reports are used for meaningful decision making in any organization or department. Raw data is difficult to understand and boring as it include noise data which affect the precision of data analysis and also confuse officials while making decisions. So in this case, data visualization is the effective way to extract the valuable information and make it a visual representation, which is the basis for decision making in an organization. More and more is not an obstacle instead it promote the organizational development [5,6]. The only thing is the way to manage this huge amount of data effectively. Various tools are used for data visualization for example: dygraphs, datawrapper, googlecharts, leaflet, graphs, pie charts, fusion charts, chartkick and many more.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

IV. REAL-TIME APPLICATIONS OF BIG DATA

1. Financial market trading: In today's era huge amount of financial data are generated in every second for stock from multiple markets, currency exchange rates and commodity prices. Companies use this data which is big as well as dynamic to detect opportunities and threats so that they can react quickly to them. For example, predicting increases or decreases in prices of security matters before a change actually occurs. Earlier such decisions are helpful in making higher chances of more profit.

2. Intelligent transportation: It is one of the most important applications of real-time big data analytics. These days different types of sensing techniques are used to monitor traffic conditions in big and crowded cites. Two types of sensors are used: road sensors and vehicle sensors. Road sensors include road monitoring cameras and vehicle sensors include GPS systems, on-board cameras and speedometer etc. [7]. These sensors generate huge amount of data which is utilized to enhance the transportation services in big cities.

3. Military Decision making: The Military wars are very dynamic. The key to win a war is not strength but also ability to collect the proper information about the current situation and make the right decision quickly [7]. And also information about the enemy resources and movements need to be collected and carefully analyzed.

4. Early warnings for natural disasters: Early warnings can save thousand lives. These systems need to analyze huge distributed data in real time from ground sensors, remote sensors; weather information from satellites etc. example of such system is Tsunami early warning system which involves sensors for earthquake detection, satellites to associate weather information and geographic area maps [8].

V. BIG DATA IN BIG COMPANIES

The need of big data comes from the big companies like Google, Facebook, yahoo etc. for the purpose of analysis of huge amount of data which is in unstructured form. Companies combine data from all sources i.e web browsing patterns, industry forecasts, social media, existing customer records, etc. to predict trends, pinpoint customers, prepare for demand and monitor real-time analytics and results [9]. Today top most retailers are using big data to gain competitive advantage.

Use cases:

1. Big data analytics allow organization to learn more about their customer's characteristics as well as purchasing habits in order to improve the marketing efforts and increase profits.

For example:

- Nestle company created 24/7 monitoring center to listen all the conversation about the company and its products on social media and keep track of customer's sentiment. The company is actively engage in this in order to satisfy the customer needs and build the customer loyalty.
- Starbucks coffee company collects data on the purchasing habits of customers in order to send special ads and coupon offers to the customers' mobile phones. It also identify trends indicating that whether customers are losing interests in their products and then is direct special offers to them in order to regenerate their interests [9].

2. Companies analyze the data about the consumer from various sources and after checking the transaction history they start to make personalized offers to those consumers.

For example:

- Spotify is a company for entertainment uses data from users' playlists and profiles to provide recommendation to each user. By combining data from million users it is able to make recommendations.
- Walmart a retail industry combines data from social media, public data and internal data to monitor about what the customer and his friends are saying about a specific product and use this data to send targeted messages about the product to the users and also share discount offers [9,10].

3. By putting the data together companies do market basket analysis using which retailers can better optimize the product selection and pricing, after which they decide where to target advertisements.

For example:

- Etihad airlines uses big data analysis to determine which destinations and routes should be added in order to increase revenue.
- P&G uses simulation models in order to create best design for their products. Thousands of iterations are done to develop best design.

4. Financial firms use big data analysis to identify fraud schemes by combining data at various points.

For example:

- Zions bank uses data analytics to detect anomalies across channels to indicate potential fraud. The fraud team receives data from different sources to detect the threats and anomalies [9].

Most regular big data applications are implemented using open loop approach. In this approach big data is analyzed and new opportunities and challenges are obtained from the information which is further used to enhance the profitability of that domain. Unlike this, real-time big data applications use closed-loop approach in which actions are based on the current and previous situations. Real-time analytical processing involves single or multiple analytical services. These services need to deploy fast algorithms that provide various alternative options within a bounded time [1]. The successful design, implementation and operations of such real-time big data applications are very beneficial in reducing risks and enhancing profitability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

Today various large events like sports, parades and concerts etc. are organized, for which proper management of crowd is required. For this, two types of traffic sensing and tracking technologies are used. Traffic monitoring through traffic sensors and vehicle tracking allow for better information about the crowded area and which in turn is used to better control on where the traffic is directed. Using this real time tracking application an accurate view about where the person is moving can be obtained and it also helps to identify the overcrowded sections. And further this information is used to distribute the police forces for optimal control over the area. Hence, to make effective decisions this huge amount of data about the location of people has to be collected, organized and analyzed on the spot [6]. Also, such big events are prone to accidents, with proper analysis of the gathered data from the sensors and tracking systems, the authorities can handle and identify the location of the accident and take remedial action within time.

VI. TOOLS USED FOR BIG DATA PROCESSING

A. Hadoop

Hadoop is a java based open source that is used for processing large data sets over thousands of distributed nodes [11]. It utilizes a scale-out architecture in which commodity servers are configured as cluster, where each server has inexpensive internal disk drives. Then the data is broken down into blocks and spread throughout a cluster. After this in the next step Mapreduce tasks are carried out on smaller subsets of data [12]. Two main components of Hadoop are HDFS and MapReduce.

1. HDFS: It is the base component of Hadoop framework that manages the data storage. It stores data in the form of data blocks in the hard disk. The default size of data block is 64MB [13]. Keeping large block size means small number of blocks can be stored so, it reduce the memory requirement on the master node which is also known as NameNode. It is used to store metadata information. Block size also has an impact on the job execution time and also on the cluster performance. A file can be stored in HDFS during upload process in different size. For file sizes which are smaller than the block size, the Hadoop cluster may not perform optimally.

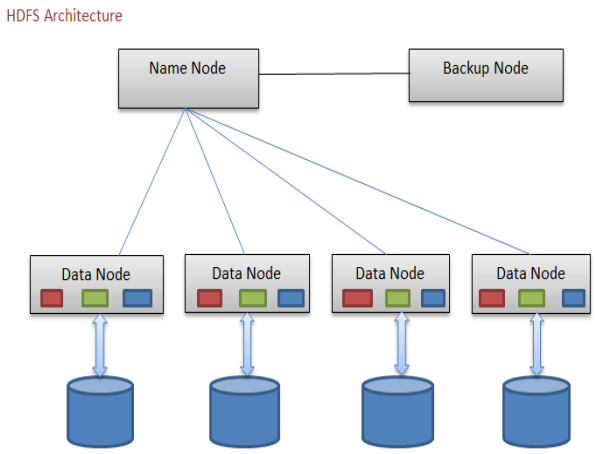

Fig 2. Hadoop deployment with HDFS

HDFS uses master-slave architecture as shown in (Fig. 2). It consists of a primary NameNode and a secondary NameNode. Primary NameNode is a master server that manages the file system name space and also handles access to data by clients. And secondary NameNode is used for failover purposes. In this there are number of DataNodes which manages the storage attached to the boxes that it runs on [14]. HDFS uses file system namespace which enables data to be stored in files. Each file is divided into more no. of blocks, which is further divide into set of DataNodes. The NameNode is used for tasks such as opening, closing and renaming files. It also handles job of mapping blocks to DataNodes. And DataNodes are used to handle block replication, creation and removal of data when it is instructed by NameNode.

2. *The HadoopMapReduce paradigm*: It has evolved for processing of big data. MapReduce is the second main component of Hadoopframework [14,15]. It is used for data processing part on the Hadoop cluster. MapReduce jobs are written in Java or any other language like (Python, Ruby, etc).

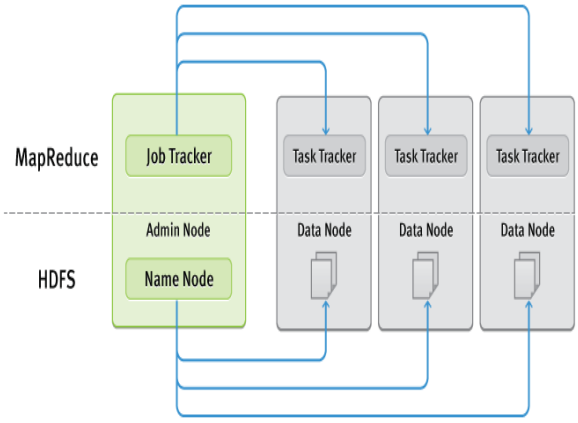

Fig 3. Interaction among the service components [13]

A single job consists of multiple tasks. And the number of tasks runs on the data nodes depend on the amount of memory installed. More memory is installed on each data node for better cluster performance. The job execution process is handled by JobTracker and TaskTracker services (see Fig.3). The selection of job scheduler for a cluster depends on the application workload. If the jobs are small then FCFS scheduler can use to serve the purpose. For more balanced cluster Fair scheduler is used [15]. If the node which runs the job tracker service fails the all the jobs submitted in the cluster will be discontinued and must be submitted again once the node is recover or another node is made available for the same purpose. The secondary NameNode plays important role in check pointing. It has same configuration as that of NameNode so that in case of failure this node can be promoted as NameNode [14,15].

B. Cassandra

Cassandra is peer-to-peer architecture developed by Facebook. It is a non-relational and column oriented distributed database.CFS was designed by DataStax corporation. Cassandra is tolerant against the single point of failure and also provides horizontal scalability. Cassandra is capable of handling petabyte of information and thousands of user operations per second across many data centers.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

Cassandra is a real time NoSQL database that supplies high performance at massive scale. It uses peer-to-peer architecture rather than a master-slave that is why it is easy to setup and maintain. Cassandra uses gossip protocol for handling all the nodes. In this all nodes are same and communicating with each other using this protocol there is no concept of master node. If one node fails then cluster detect the failure and automatically routes the user request from other side. Once the node is recovered again it rejoins the cluster. Cassandra has no single point failure that is it is capable of offering continuous availability.

1. CFS (Cassandra File System):

The Cassandra file system was designed by DataStax enterprise to run analytics on Cassandra data stores. CFS provides the storage that run Hadoop-styled analytics on Cassandra data. In contrast to HDFS, implementation of CFS is peer-to-peer or masterless based on cassandra.

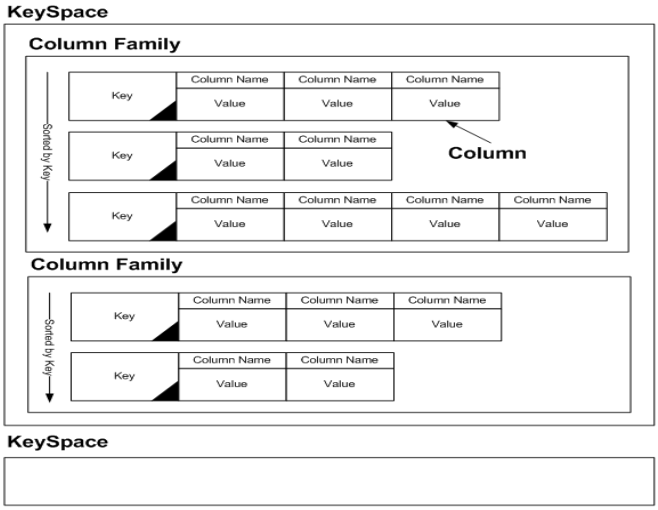

Fig.4. Hierarchy of Cassandra data model [12].

2. Benefits of CFS:

There are various benefits of using CFS over HDFS. CFS performs extremely well under different workloads.

1. *Better Availability*: It provides continuous availability for analytics in a database cluster.

Replication and redundancy capability of Cassandra provides complete customization that is how many copies of data should be maintained in a cluster, which in turn ensures data availability and no chance of data loss[12].

2. *Simpler Deployment*: CFS use peer-to-peer architecture, there is no need of any master-slave configurations, and no need of complex storage requirements for storage area networks. A cluster can be set up in a few minutes. Users can easily create cluster they desire.

3. *Multi-Data Center support*: CFS supports multi-data center, cloud and hybrid environments. Today many modern businesses need to run analytic operations across more than one data centers. It supports running a single database cluster across many data centers, in which any node in the cluster is being able to service reads and writes requests. A job Tracker is configured for each data center so that each location has its own for handling MapReduce and analytic processing jobs.

4. *Full Data Integration*: CFS provides another benefit to have big data platform that handles real-time, analytic and search workloads in one cluster. Also Full mixed workload support is built and transparently handled by DataStax enterprise. This benefit results in the eliminations of data heaps in organization and there is no need to create and maintain separate routines for handling and moving purpose. Any data written to Cassandra is replicated to search and analytics nodes and vice versa[17].

5. *Commodity Hardware support*: CFS support commodity Hardware that is it runs well on commodity Hardware and requires no other special server or any equipment.

VII. COMPARISON BETWEEN HADOOP AND CASSANDRA

Three major differences are:

1. Cassandra excels at real-time transaction processing whereas Hadoop excels batch-oriented analytical solutions.

2. Cassandra use peer-to-peer or masterless architecture on the other hand side Hadoop implements master-slave architecture.

3. Cassandra has robust fault tolerance capability that is it is not simply protect data against data loss which is done by HDFS in Hadoop when it copies data blocks across the Hadoop cluster. Instead, it also replicates data blocks to multiple nodes and supports replication between geographically distributed nodes.

VIII. PERFORMANCE ANALYSIS

Analysis of performance can be done by using Hadoop MapReduce and Cassandra by running Hadoop with three different configurations:

a) Hadoop-native: In this configuration HDFS is used for inout and output placements.
b) Hadoop-Cassandra-FS: in this the input reads from Cassandra and writes output to the shared file system.
c) Hadoop-Cassandra-Cassandra: reads input from Cassandra and writes output back to Cassandra.

The three Hadoop setups work differently under different works loads when input size is larger than output size (read>>write), input size is nearly same to output (read=write) and when input size is smaller than output (read<<write) [12].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

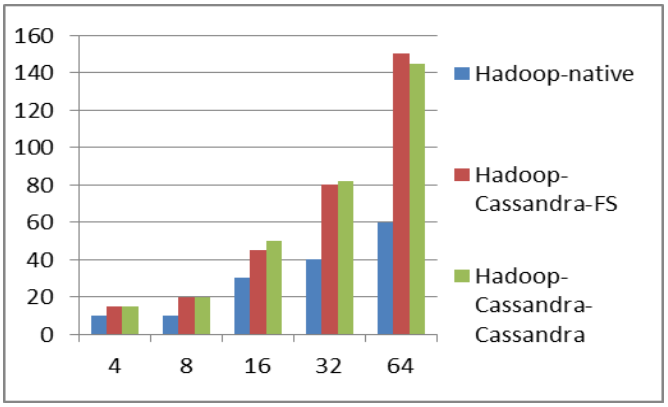

Fig.5(a). Read>> Write

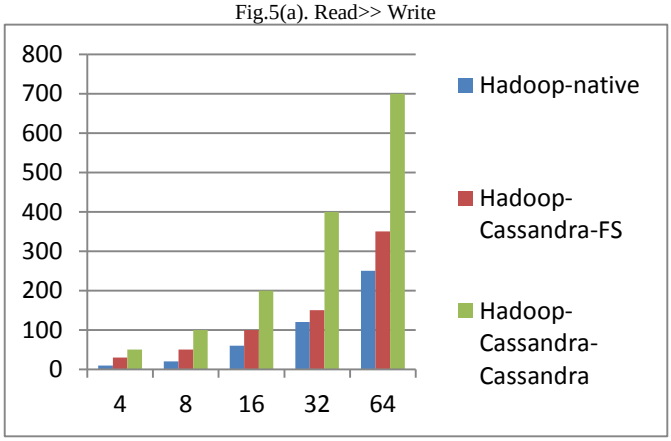

Fig.5(b) Read= Write

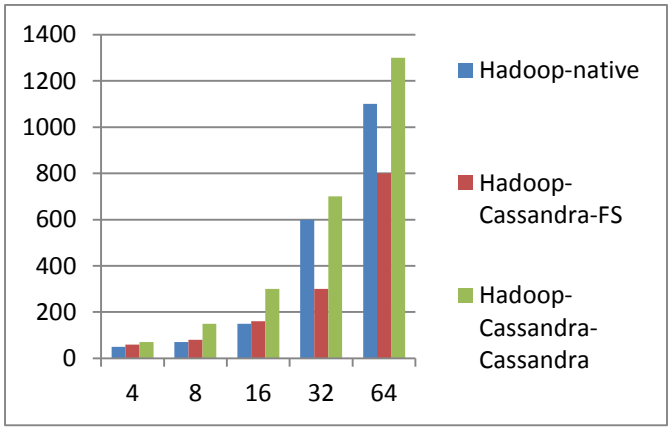

Fig.5(c). Read << Write

From the above graphs it is clear that HDFS performs poor under heavy workloads.

IX. CONCLUSION

Big data collection, cleaning, classification and analysis require long time to complete. And this is the main reason that real-time application cannot get immediate benefit from big data analytics. To build effective real time big data application several challenges need to be addressed which are real time event transfer, real time situation detection, real time analytics, real time decision making and real time responses. So to create the most efficient and effective application we need to identify the most suitable solutions to these challenges and build a real time application which is reliable and capable of meeting the time demands. In future big data will also become a new point of economic growth. With this companies will upgrade and transform to the mode of Analysis as a service(AaaS).

REFERENCES

[1] X. jin, B. W. Wah, X. Cheng, Y. Wang," *Significance and challenges of big data Research*" Big Data Research 2 (2015) : 59-64.
[2] N. Kshetri," Big Data's impact on privacy, security and consumer welfare", Telecommunications Policy 38 (2014) 1134-1145.
[3] B. Hu, Y. Ma, Liang- Jie Zhang, J. Shi, J. Zhong," *A key value based application platform for Enterprise Big Data*", IEEE International Congress on Big Data, 2014.
[4] R. P. Padhy, S. C. Satapathy and M. R. Patra. "*Rdbms to nosql: Reviewing some next-generation non-relational databases"*, International Journal of AdvancedEngeneering science and technologies (2011) : 15-30.
[5] J. Archenaa, E. A Mary Anita," *A Survey of Big data Analytics in Healthcare and Government*", 2nd International Symposium on big data and cloud computing (ISBCC'15), procedia Computer Science 50 (2015) : 408-413.
[6] J. Zhang, Y. Chen, T. Li, "*Opportunities of Innovation under Challenges of Big Data*", 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, 2013.
[7] N. Mohamed, J. Al-Jaroodi, "*Real-Time Big Data Analytics: Applications and Challenges*", IEEE, 2014.
[8] G. Chen, S. Wu, Y. Wang," *The Envolvement of Big Data Systems: From the perspective of an Information Security Application*", in Big Data Research 2 (2015) :65-73.
[9] Thomas H. Davenport, J. Dyche," *Big Data in Big Companies*", International Institute of Analytics, in May 2013.
[10] F. Tekiner and J. A. Keane," *Big Data Framework*", IEEE International Conference on Systems, Man and Cybernetics, 2013.
[11] B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvan," *Big Data and Hadoop- A Study in Security perspective*", 2nd International Symposium on big data and cloud computing (ISBCC'15), procedia Computer Science 50 (2015) 596- 601.
[12] E.Dede, B.Sendir, P. Kuzlu, J. Hartog, M. Govindaraju," *An Evaluation of Cassandra for Hadoop*",IEEE Sixth International Conference on Cloud Computing, USA, 2013.
[13] C. Jie, C. Dongjie, "Research on Big data information Retrieval based on hadoop architecture", IEEE Workshop on Electronics, Computers and Applications, 2014.
[14] A. B. Patel, ManashviBirlz, U. Nair," *Addressing Big Data Problems using Hadoop and MapReduce*", Nirma University International Conference on Engineering, 6-8 Dec 2012.
[15] K. Singh, R. Kaur ," *Hadoop- Addressing Challenges of Big Data*", IEEE International Advance Computing Conference (IACC), 2014.
[16] J. R. Lourenco, VeronikaAbrampva, Bruno Cabral, Jorgr Bernardino, Paulo Carreiro, Marco Vieira," *NoSQL in practice: a write-heavy enterprise application*", IEEE International congress on Big Data, 2015.
[17] A.chebotko, A.Kashlev, S. Lu,"A Big data Modeling Methodology for Apache Cassandra", IEEE International Congress on Big data, 2015.