International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 9, Issue 10, October 2020)

Enhance the Overall performance of Big Data Processing in Cloud Computing

Sana Parveen[1], Prof. Sarwesh site[2]
[1]Research Scholar, [2]Assistant Professor, CSE, All Saints College of Technology
Email- sanabintidrees@gmail.com, Email: er.sarwesh@gmail.com

***Abstract*--The generation of technology and requirement fulfill the demand of digital universe data. Day to day the digital universe facts are exploded in phrases of megabyte and petabyte. The exploding price of facts demands the new technology of technology such as huge data processing. In this paper Optimized the performance of map limit programming model for the enhancement of data processing. The modified model of programming used clustering technique. The clustering approach comprises the procedure of map records in phrases of challenge group. The venture group of map facts correlated with exceptional index of facts for the processing of data node. The proposed model applied in Hadoop framework and programmed in java. For the contrast of overall performance used three preferred datasets and measure the processing time and matter cost of file. Today Big Data draws a lot of attention in the IT world. The rapid rise of the Internet and the digital economy has fuelled an exponential growth in demand for data storage and analytics, 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. This paper primarily focuses on discussing the various technologies that work together as a Big Data Analytics system that can help predict future volumes, gain insights, take proactive actions, and give way to better strategic decision-making. Further this paper analyzes the adoption, usage and impact of big data analytics to the business value of an enterprise to improve its competitive advantage using a set of data algorithms for large data sets such as Hadoop and MapReduce.**

***Keywords*-Big Data, Hadoop, MapReduce, Clustering, Optimization**

I. INTRODUCTION

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data.

It provides evolutionary breakthroughs in many fields with collection of large datasets. In general, it refers to the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing applications. These are available in structured, semi-structured, and unstructured format in petabytes and beyond. Formally, it is defined from 3Vs to 4Vs. 3Vs refers to volume, velocity, and variety. Volume refers to the huge amount of data that are being generated everyday whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semistructured etc. The fourth V refers to veracity that includes availability and accountability. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques [1]. Some of these extraction methods for obtaining helpful information was discussed by Gandomi and Haider [2]. The following Figure 1 refers to the definition of big data. However exact definition for big data is not defined and there is a believe that it is problem specific. This will help us in obtaining enhanced decision making, insight discovery and optimization while being innovative and cost-effective. It is expected that the growth of big data is estimated to reach 25 billion by 2015 [3]. From the perspective of the information and communication technology, big data is a robust impetus to the next generation of information technology industries [4], which are broadly built on the third platform, mainly referring to big data, cloud computing, internet of things, and social business. Generally, Data warehouses have been used to manage the large dataset. In this case extracting the precise knowledge from the available big data is a foremost issue. Most of the presented approaches in data mining are not usually able to handle the large datasets successfully. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such as data mining and statistical analysis. These challenges generally arise when we wish to perform knowledge discovery and representation for its practical applications.

A fundamental problem is how to quantitatively describe the essential characteristics of big data. There is a need for epistemological implications in describing data revolution [5]. Additionally, the study on complexity theory of big data will help understand essential characteristics and formation of complex patterns in big data, simplify its representation, gets better knowledge abstraction, and guide the design of computing models and algorithms on big data [4]. Much research was carried out by various researchers on big data and its trends [6], [7], [8].

II. Review Of Literature

Big Data is a data analysis methodology enabled by recent advances in technologies that support high- velocity data capture, storage and analysis. Data sources extend beyond the traditional corporate database to include emails, mobile device outputs, and sensor-generated data where data is no longer restricted to structured database records but rather unstructured data having no standard formatting [30]. Since Big Data and Analytics is a relatively new and evolving phrase, there is no uniform definition; various stakeholders have provided diverse and sometimes contradictory definitions. One of the first widely quoted definitions of Big Data resulted from the Gartner report of 2001. Gartner proposed that, Big Data is defined by three V's volume, velocity, and variety. Gartner expanded its definition in 2012 to include veracity, representing requirements about trust and uncertainty pertaining to data and the outcome of data analysis. In a 2012 report, IDC defined the 4th V as value—highlighting that Big Data applications need to bring incremental value to businesses. Big Data Analytics is all about processing unstructured information from call logs, mobile-banking transactions, online user generated content such as blog posts and tweets, online searches, and images which can be transformed into valuable business information using computational techniques to unveil trends and patterns between datasets. Another dimension of the Big Data definition involves technology. Big Data is not only large and complex, but it requires innovative technology to analyze and process. In 2013, the National Institute of Standard and Technology (NIST) Big Data workgroup proposed the following definition of Big Data that emphasizes application of new technology; Big Data exceed the capacity or capability of current or conventional methods and systems, and enable novel approaches to frontier questions previously inaccessible or impractical using current or conventional methods. Business challenges rarely show up in the appearance of a perfect data problem, and even when data are abundant, practitioners have difficulties to incorporate it into their complex decision-making that adds business value. In 2012, McKinsey & Company conducted a survey of 1,469 executives across various regions, industries and company sizes, in which 49 percent of respondents said that their companies are focusing big data efforts on customer insights, segmentation and targeting to improve overall performance [10] An even higher number of respondents 60 percent said their companies should focus efforts on using data and analytics to generate these insights. Yet, just one-fifth said that their organizations have fully deployed data and analytics to generate insights in one business unit or function, and only 13 percent use data to generate insights across the company. As these survey results show, the question is no longer whether big data can help business, but how can business derive maximum results from big data.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 9, Issue 10, October 2020)

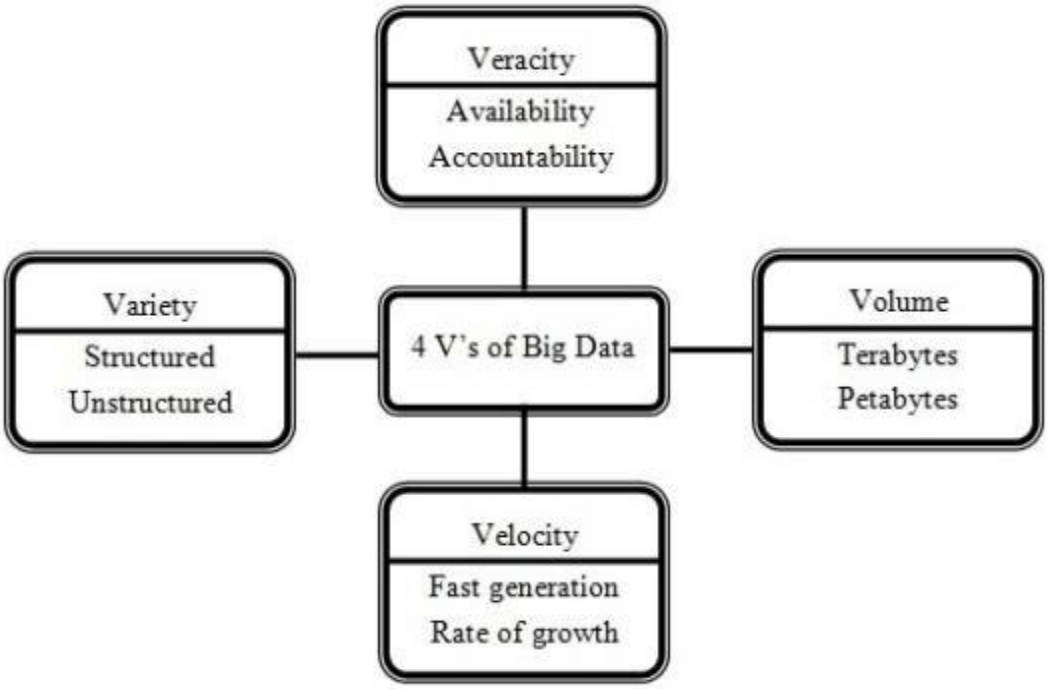

Figure.1-Big Data Characteristics

III. Challenges In Big Data Analytics

Recent year's big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, and Thomson Reuters etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However oppotunities always follow some challenges. To handle the challenges we need to know various computational complexities, information security, and computational method, to analyze big data. For example, many statistical methods that perform well for small data size do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenges in analyzing big data. Various challenges that the health sector face was being researched by much researchers [9], [10].

Here the challenges of big data analytics are classified into four broad categories namely data storage and analysis; knowledge discovery and computational complexities; scalability and visualization of data; and information security.

IV. Predictive Analytics

Predictive Analytics is the use of historical data to forecast on consumer behavior and trends [18]. It is the use of past/historical data to predict future trends. This analysis makes use of the statistical models and machine learning algorithms to identify patterns and learn from historical data [25]. Predictive Analysis can also be defined as a process that uses machine learning to analyze data and make predictions [22]. Sixty seven percent of businesses aim at using predictive analytics to create more strategic marketing campaign in future, and 68% sight competitive advantage as the prime benefit of predictive analysis [17]. Broadly speaking, predictive analysis can be applied in ecommerce for product recommendation, price management, and predictive search. Typically a large e-commerce site offers thousands of product and services for sale.

Navigating and searching for a product out of thousands on a website could be a major setback to consumers. However, with the invention of recommender system, an E-Commerce site/application can quickly identify/predict products that closely suit the consumer's taste [24].

Using a technology called Collaborative Filtering a database of historical user preferences is created. When a new customer access the ecommerce site, the customer is matched with the database of preferences, in order to discover a preference class that closely matches the customer taste. These products are then recommended to the customer [24]. Another technology that is used in ecommerce is the clustering algorithm. Clustering algorithm works by identifying groups of users that have similar preferences. These users are then clustered into a single group and are given a unique identifier.

V. Cloud Computing For Big Data Analytics

The development of virtualization technologies have made supercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been one of the most robust big data technique. Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data. Cloud computing harmonize massive data by ondemand access to configurable computing resources through virtualization techniques. The benefits of utilizing the Cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction.

Open challenges and research issues of big data and cloud computing are discussed in detail by many researchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, and resource management [29], [30]. So Cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools. Big data application using cloud computing should support data analytic and development. The cloud environment should provide tools that allow data scientists and business analysts to interactively and collaboratively explore knowledge acquisition data for further processing and extracting fruitful results. This can help to solve large applications that may arise in various domains. In addition to this, cloud computing should also enable scaling of tools from virtual technologies into new technologies like spark, R, and other types of big data processing techniques. Big data forms a framework for discussing cloud computing options. Depending on special need, user can go to the marketplace and buy infrastructure services from cloud service providers such as Google, Amazon, IBM, software as a service (SaaS) from a whole crew of companies such as NetSuite, Cloud9, Jobscience etc. Another advantage of cloud computing is cloud storage which provides a possible way for storing big data. The obvious one is the time and cost that are needed to upload and download big data in the cloud environment. Else, it becomes difficult to control the distribution of computation and the underlying hardware. But, the major issues are privacy concerns relating to the hosting of data on public servers, and the storage of data from human studies. All these issues will take big data and cloud computing to a high level of development.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 9, Issue 10, October 2020)

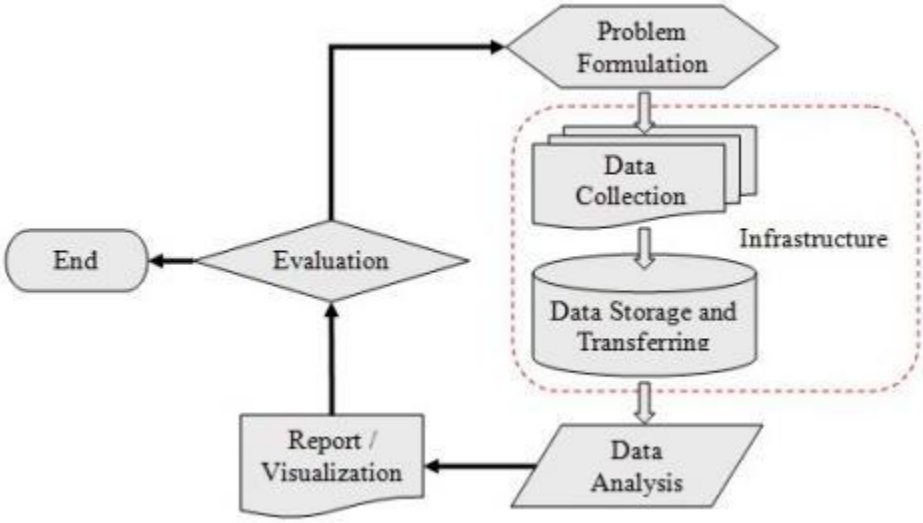

Figure 2-Big Data Flow

VI. Big Data Technologies

Apache

Flume Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store. Flume deploys as one or more agents, each contained within its own instance of the Java Virtual Machine (JVM). Agents consist of three pluggable components: sources, sinks, and channels. Flume agents ingest incoming streaming data from one or more sources. Data ingested by a Flume agent is passed to a sink, which is most commonly a distributed file system like Hadoop. Multiple Flume agents can be connected together for more complex workflows by configuring the source of one agent to be the sink of another. Flume sources listen and consume events. Events can range from newline-terminated strings in stdout to HTTP POSTs and RPC calls — it all depends on what sources the agent is configured to use. Flume agents may have more than one source, but at the minimum they require one. Sources require a name and a type; the type then dictates additional configuration parameters. Channels are the mechanism by which Flume agents transfer events from their sources to their sinks. Events written to the channel by a source are not removed from the channel until a sink removes that event in a transaction. This allows Flume sinks to retry writes in the event of a failure in the external repository (such as HDFS or an outgoing network connection).

For example, if the network between a Flume agent and a Hadoop cluster goes down, the channel will keep all events queued until the sink can correctly write to the cluster and close its transactions with the channel. Sink is an interface implementation that can remove events from a channel and transmit them to the next agent in the flow, or to the event's final destination and also sinks can remove events from the channel in transactions and write them to output. Transactions close when the event is successfully written, ensuring that all events are committed to their final destination.

Apache Sqoop

Apache Sqoop is a CLI tool designed to transfer data between Hadoop and relational databases. Sqoop can import data from an RDBMS such as MySQL or Oracle Database into HDFS and then export the data back after data has been transformed using MapReduce. Sqoop also has the ability to import data into HBase and Hive. Sqoop connects to an RDBMS through its JDBC connector and relies on the RDBMS to describe the database schema for data to be imported. Both import and export utilize MapReduce, which provides parallel operation as well as fault tolerance. During import, Sqoop reads the table, row by row, into HDFS. Because import is performed in parallel, the output in HDFS is multiple files.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 9, Issue 10, October 2020)

Apache Pig

Apache's Pig is a major project, which is lying on top of Hadoop, and provides higher-level language to use Hadoop's MapReduce library. Pig provides the scripting language to describe operations like the reading, filtering and transforming, joining, and writing data which are exactly the same operations that MapReduce was originally designed for. Instead of expressing these operations in thousands of lines of Java code which uses MapReduce directly, Apache Pig lets the users express them in a language that is not unlike a bash or Perl script. Pig was initially developed at Yahoo Research around 2006 but moved into the Apache Software Foundation in 2007. Unlike SQL, Pig does not require that the data must have a schema, so it is well suited to process the unstructured data. But, Pig can still leverage the value of a schema if you want to supply one. PigLatin is relationally complete like SQL, which means it is at least as powerful as a relational algebra. Turing completeness requires conditional constructs, an infinite memory model, and looping constructs. Issues in Information

Apache Hive

Hive is a technology developed by Facebook that turns Hadoop into a data warehouse complete with a dialect of SQL for querying. Being a SQL dialect, HIVEQL is a declarative language. In PigLatin, you specify the data flow, but in Hive we describe the result we want and hive figures out how to build a data flow to achieve that result. Unlike Pig, in Hive a schema is required, but you are not limited to only one schema. Like PigLatin and SQL, HiveQL itself is a relationally complete language but it is not a Turing complete language.

Apache ZooKeeper

Apache Zoo Keeper is an effort to develop and maintain an open-source server, which enables highly reliable distributed coordination. It provides a distributed configuration service, a synchronization service and a naming registry for distributed systems. Distributed applications use ZooKeeper to store and mediate updates to import configuration information. ZooKeeper is especially fast with workloads where reads to the data are more common than writes. The ideal read/write ratio is about 10:1. ZooKeeper is replicated over a set of hosts (called an ensemble) and the servers are aware of each other and there is no single point of failure.

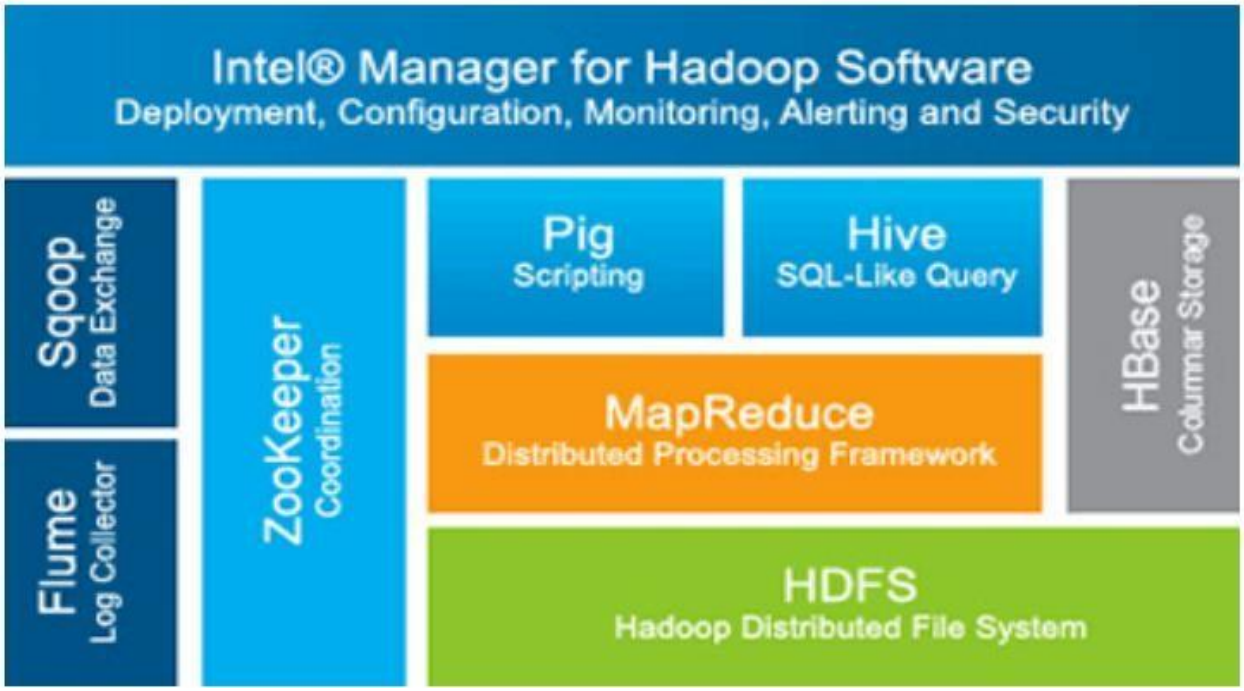

Figure.3 Hadoop Architecture

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 9, Issue 10, October 2020)

MongoDB

MongoDB is an open source, document-oriented NoSQL database that has lately attained some space in the data industry. It is considered as one of the most popular NoSQL databases, competing today and favors master-slave replication. The role of master is to perform reads and writes whereas the slave confines to copy the data received from master, to perform the read operation, and backup the data. The slaves do not participate in write operations but may select an alternate master in case of the current master failure. MongoDB uses binary format of JSON-like documents underneath and believes in dynamic schemas, unlike the traditional relational databases. The query system of MongoDB can return particular fields and query set compass search by fields, range queries, regular expression search, etc. and may include the user- defined complex JavaScript functions. As hinted already, MongoDB practice flexible schema and the document structure in a grouping, called Collection, may vary and common fields of various documents in a collection can have disparate types of the data. The MongoDB is equipped with the suitable drivers for most of the programming languages, which are used to develop the customized systems that use MongoDB as their backend player.

There is an increasingly demand of using MongoDB as pure in-memory database; in such cases, the application dataset will always be small.

Apache Cassandra

Apache Cassandra is the yet another open source NoSQL database solution that has gained industrial reputation which is able to handle big data requirements. It is a highly scalable and high-performance distributed database management system that can handle real-time big data applications that drive key systems for modern and successful businesses. It has a built-for-scale architecture that can handle petabytes of information and thousands of concurrent users/operations per second as easily as it can manage much smaller amount of data and user traffic. It has a peer to peer design that offers no single point of failure for any database process or function, in addition to the location independence capabilities that equate to a true network-independent method of storing and accessing data, data can be read and written anywhere. Apache Cassandra is also equipped with flexible/dynamic schema design that accommodates all formats of big data applications, including structured, semi-structured, and unstructured data.

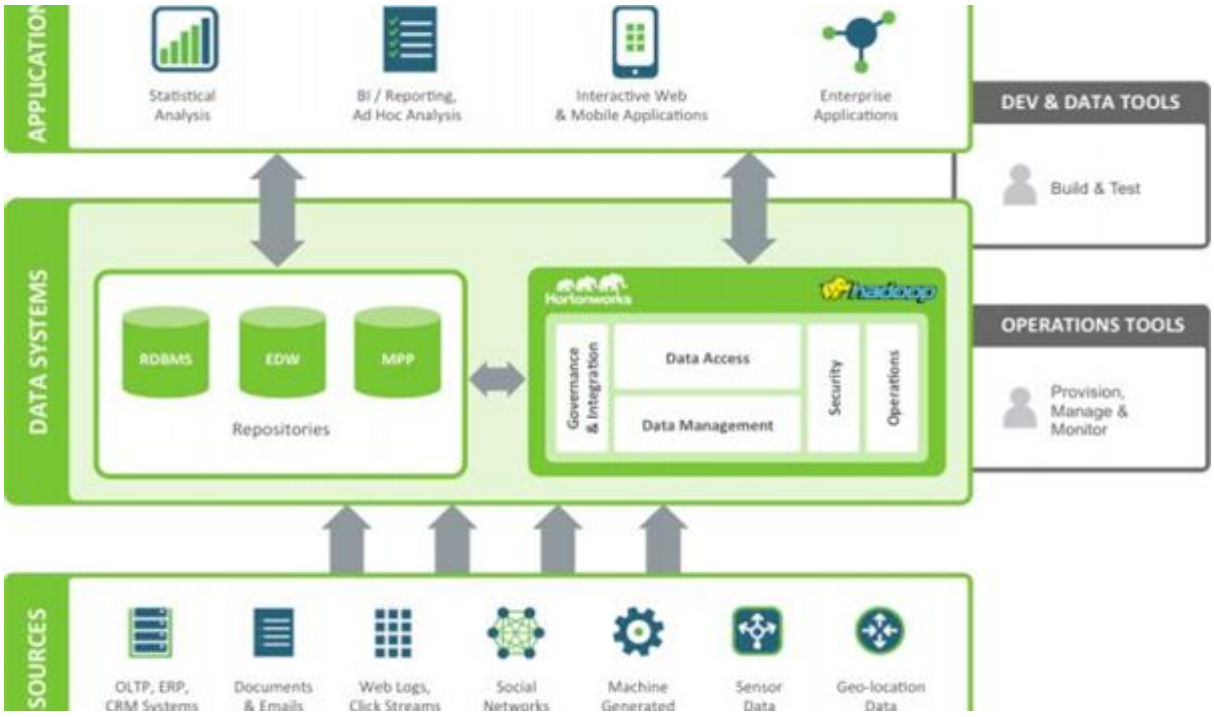

Figure 4. Data Architecture with Hadoop Integrated with existing data system

VII. Big Data Framework

Apache Spark

Apache Spark an open source big data processing framework built around speed, ease of use, and sophisticated analytics. It was originally developed in 2009 in UC Berkeley's AMP Lab, and open sourced in 2010 as an Apache project. Hadoop as a big data processing technology has been around for ten years and has proven to be the solution of choice for processing large data sets. MapReduce is a great solution for one-pass computations, but not very efficient for use cases that require multi-pass computations and algorithms. Each step in the data processing workflow has one Map phase and one Reduce phase and you'll need to convert any use case into MapReduce pattern to leverage this solution. Spark takes MapReduce to the next level with less expensive shuffles in the data processing. With capabilities like in-memory data storage and near real-time processing, the performance can be several times faster than other big data technologies. Spark also supports lazy evaluation of big data queries, which helps with optimization of the steps in data processing workflows. It provides a higher-level API to improve developer productivity and a consistent architect model for big data solutions.

Spark holds intermediate results in memory rather than writing them to disk, which is very useful especially when you need to work on the same dataset multiple times. It's designed to be an execution engine that works both in-memory and on-disk. Spark operators perform external operations when data does not fit in memory. Spark can be used for processing datasets that larger than the aggregate memory in a cluster. Spark will attempt to store as much as data in memory and then will spill to disk. It can store part of a data set in memory and the remaining data on the disk. You have to look at your data and use cases to assess the memory requirements. With this in-memory data storage, Spark comes with a great performance advantage. Spark is written in Scala Programing Language and runs on the Java Virtual machine. It currently supports programming languages like Scala, java, python, Clojure and R. Other than Spark Core API, there are additional libraries that are part of the Spark ecosystem and provide additional capabilities in Big Data analytics. Spark Streaming is one among the spark library that can be used for processing the real-time streaming data. This is based on micro based on micro batch style of computing and processing. Spark SQL provides the capabilities to expose the spark datasets over JDBC API and allow running the SQL like queries on Spark data using traditional BI and visualization tools. MLlib, GraphX are some other libraries from spark.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 9, Issue 10, October 2020)

VIII. COMPARISION

Provider	"On-Premises Option"	Development Technology	Portability	Reliability	Security
Amazon AWS	Bring Your Own	Whatever you can get to run in an AMI	Code is easy to move. The rest...?	Reliability by the spreading of wealth	We have ACL's and digital certificates, but please ignore all that Xen Dom0 stuff.
Google App Engine	Sure...for your dev environment	* Hip coders love Python, right?	Are you kidding?	Trust the Magic	As your new Dark Overlord, we invite you to not question our intent. We will ensure you "do no evil"
Microsoft Azure	Absolutely! The future is hybrid.	* If you love MSFT, you already know it. (.NET)	You mean between Microsoft products, right?	Promises, promises	Our SDL will protect you! That, and ForeFront and Morro!
Salesforce.com (force.com)	Absolutely not! The future is pure cloud	It's new, it's improved, it's APEX!	Heh. That's funny	We got magic, too.	We've invested millions in security, but please don't click on web links in email! You're the last line of defense!

Figure 5- Comparative study of Different Provider

	AWS *monthly cost*	Azure *monthly cost*	Google *monthly cost*	IBM *monthly cost*
Name of Services	**EBS**	**Managed Disk**	**Persistent Disk**	**Block Storage**
Magnetic 500 GB	$22.50	$21.76	$20.00	N/A
SSD 500 GB, 1000 IOPs	$50.00	$66.56 *P20 – 2300 IOPs*	$85.00 *Includes 15,000 IOPs*	$100.00
SSD 500 GB, 2000 IOPs	$192.50 *PIOPs SSD*	$66.56 *P20 – 2300 IOPs*	$85.00 *Includes 15,000 IOPs*	$175.00
SSD 500 GB, 5000 IOPs	$650.00 *PIOPs SSD*	$122.88 *P30 – 5000 IOPs*	$85.00 *Includes 15,000 IOPs*	$290.00
Snapshots 500 GB	$25.00	$25.00	$13.00	*Info not available*

Figure 6- Block Storage Scenarios

IX. CONCLUSION

Today's technology landscape is changing fast. Organizations of all shapes and sizes are being pressured to be datadriven and to do more with less. Even though big data technologies are still in a nascent stage, relatively speaking, the impact of the 3V's of big data, which now is 5v's cannot be ignored. The time is now for organizations to begin planning for and building out their Hadoop-based data lake. Organizations with the right infrastructures, talent and vision in place are well equipped to take their big data strategies to the next level and transform their businesses. They can use big data to unveil new patterns and trends, gain additional insights and begin to find answers to pressing business issues. The deeper organizations dig into big data and the more equipped they are to act upon what's learned, the more likely they are to reveal answers that can add value to the top line of the business. This is where the returns on big data investments multiply and the transformation begins. Harnessing big data insight delivers more than cost cutting or productivity improvement but it definitely reveals new business opportunities. Data-driven decisions always tend to be better decisions

REFERENCES

[1] Apache Software Foundation. (2010). Apache ZooKeeper. Retrieved April 5, 2019 from https://zookeeper.apache.org

[2] Chae, B., Sheu, C., Yang, C. and Olson, D. (2018). The impact of advanced analytics and data accuracy on operational performance: A contingent resource based theory (RBT) perspective, Decision Support Systems, 59, 119-126.

[3] Chambers, C., Raniwala, A., Adams, S., Henry, R., Bradshaw, R., and Weizenbaum, N. (2018). Flume Java: Easy, Efficient Data-Parallel Pipelines. Google, Inc. Retrieved April 1, 2015 from http://pages.cs.wisc.edu/~akella/CS838/F12/838-CloudPapers/FlumeJava.pdf

[4] Cisco Systems. Cisco UCS Common Platform Architecture Version 2 (CPA v2) for Big Data with Comprehensive Data Protection using Intel Distribution for Apache Hadoop. Retrieved March 15, 2015, from http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/Cisco_UCS_CPA_for _Big_Data_wi th_Intel.html

[5] DATASTAX Corporation. (2013, October). Big Data: Beyond the Hype - Why Big data Matters to you [White paper]. Retrieved March 15, 2015 from https://www.datastax.com/wp-content/uploads/2011/10/WP-DataStaxBigData.pdf

[6] Davenport, T & Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review, 90, 70-76.

[7] Dhawan, S & Rathee, S. (2013). Big Data Analytics using Hadoop Components like Pig and Hive. American International Journal of Research in Science, Technology, Engineering & Mathematics, 88, 13-131. Retrieved from http://iasir.net/AIJRSTEMpapers/AIJRSTEM13-131.pdf

[8] Edwards, P., Peters, M. and Sharman, G. (2001). The Effectiveness of Information Systems in Supporting the Extended Supply Chain, Journal of Business Logistics 22 (1), 1-27

[9] EMC Corporation. (2013, January). EMC Accelerates Journey to Big Data with Business Analytics-as-aService [White paper]. Retrieved from http://www.emc.com/collateral/white-papers/h11259-emc-acceleratesjourney-big-data-ba-wp.pdf

[10] EMC Corporation. Big Data, Big Transformations [White paper]. Retrieved from http://www.emc.com/collateral/white-papers/idg-bigdata-umbrella-wp.pdf

[11] EMC Solutions Group. (2012, July). Big Data-as-a-Service [White paper]. Retrieved from https://www.emc.com/collateral/software/white-papers/h10839-big-data-as-a-service-perspt.pdf

[12] Enterprise Hadoop: The Ecosystem of Projects. Retrieved from http://hortonworks.com/hadoop/

[13] George, L. (2014, September). Getting Started with Big Data Architecture. Retrieved April 5, 2015, from http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/

[14] IBM Corporation. IBM Big Data Platform. Retrieved from http://www01.ibm.com/software/in/data/bigdata/enterprise.html

[15] Intel Corporation. Big Data Analytics - Extract, Transform, and Load Big data with Apache Hadoop [White paper]. Retrieved April 3, 2015 from https://software.intel.com/sites/default/files/article/402274/etl-big-datawith-hadoop.pdf

[16] McClary, D. (2013, June). Acquiring Big Data Using Apache Flume. Retrieved March 3, 2015 from http://www.drdobbs.com/database/acquiring-big-data-using-apache-flume/240155029

[17] Millard, S. (2013). Big Data Brewing Value in Human Capital Management – Ventana Research. Retrieved April 2, 2015 from http://stephanmillard.ventanaresearch.com/2013/08/28/big-data-brewing-value-in-humancapital-management

[18] Mosavi, A. and Vaezipour, A. (2013). Developing Effective Tools for Predictive Analytics and Informed Decisions. Technical Report. University of Tallinn.

[19] Oracle Corporation. (2013, March). Big Data Analytics - Advanced Analytics in Oracle Database [White paper]. Retrieved March 5, 2015 from http://www.oracle.com/technetwork/database/options/advancedanalytics/advanced-analytics-wp- 12c-1896138.pdf?ssSourceSiteId=ocomen

[20] Oracle Enterprise Architecture. (2015, April). An Enterprise Architect's Guide to Big Data - Reference Architecture Overview [White paper]. Retrieved from http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

[21] Penchikala, S. (2015, January). Big Data Processing with Apache Spark - Part 1: Introduction. Retrieved from http://www.infoq.com/articles/apache-spark-introduction

[22] Puri, R. (2013). How Online Retailers Use Predictive Analytics To Improve Your Shopping Experience. Retrieved April5, 2015 from http://blogs.sap.com/innovation/analytics/how-online- retailers-use-predictiveanalytics-to-improve-your-shopping-experience-0108060

[23] Sanders, N.R. (2014). Big Data Driven Supply Chain Management: A Framework for Implementing Analytics and Tuning Information into Intelligence, 1st Edition, Pearson, NJ

[24] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2002). Recommendation systems for large e- commerce: Scalable neighborhood formation using clustering. In Proceedings of the fifth international conference on computer and information technology, 1.

[25] Shmueli, G. & Koppius, O. (2011). Predictive Analytics in Information Systems Research. MIS Quarterly, 35(3), pp. 553-72.

[26] Sorkin, S. (2011). Splunk Technical Paper: Large-Scale, Unstructured Data Retrieval and Analysis Using Splunk. Retrieved April 15, 2015 from https://www.splunk.com/content/dam/splunk2/pdfs/technicalbriefs/splunk-and-mapreduce.pdf

[27] The Bloor Group. IBM and the Big Data Information Architecture. Retrieved April 3, 2015 from http://insideanalysis.com/wp-content/uploads/2014/08/BDIAVendor-IBMv01.pdf

[28] Tiwari, S. (2011). Using Oracle Berkeley DB as a NoSQL Data Store. Retrieved April 5 2015 from http://www.oracle.com/technetwork/articles/cloudcomp/berkeleydb-nosql-323570.htm

[29] Transparency Market Report. (May, 2015).Big Data Applications in Healthcare likely to Propel Market to US$48.3 Bn by 2018. Retrieved June 26, 2015, from http://www.transparencymarketresearch.com/pressrelease/big-data-market.htm

[30] Villars, R. L., Olofson, C. W., & Eastwood, M. (2011, June). Big data: What it is and why you should care. IDC White Paper. Framingham, MA: IDC.