Middle-East Journal of Scientific Research 24 (Techniques and Algorithms in Emerging Technologies): 201-206, 2016
ISSN 1990-9233;

DOI: 10.5829/idosi.mejsr.2016.24.TAET23428

# An Effective Data Storage Management Using Advances in Hadoop Environment

*[1]D. Vivek, [1]K.S. Giriprasath and [2]P. Balasubramanie*

[1]Department of Information Technology, Nandha Engineering College (Autonomous),
Erode-52, Anna University, Chennai, India
[2]Department of Computer Science and Engineering, Kongu Engineering College, Erode-52

**Abstract:** The innovations in world increase with new forms of data, which is collected and stored in the device of retrieval. These data are outsourced for future analytics and management on existing performance of new technologies. Herewith, the advances in research issues shows that the replication of data is increased in every storage devices , further leads to the diversity of data storage and loss of data. In this paper, the effective storage of unstructured data is processed by sharing the data in the distributed management system. Different data are collected and distributed to various nodes of clusters which are monitored by the single node named as master node. Hence, the experimental results show that data in the shared node is easily collected by the users without data loss and it also explains the historical data analytics by different tools in big data environment.

**Key words:** Storage · Hadoop · Distributed · Clusters · Node

## INTRODUCTION

Big data is a relative term which describes a situation where the volume, velocity and varietyof data exceed an organization's storage or compute capacity for accurate and timely decision making. The data stored during transactional processing is the by-product of fast-growing online activity. Internets of things, such as measuring, call detail reports, environmental sensing and RFID systems; produce their own tidal influences of data.

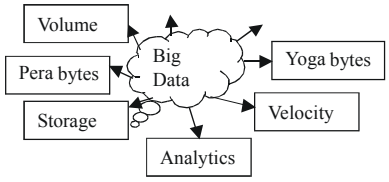


Fig. 1: Determining relevant data is key to delivering value from massive amounts of data.

All these forms of data are expanding and that is coupled with fast-growing flow of unstructured and semi structured data from social media. That's a lot of data, but it is the reality for many establishments. By some evaluations, organizations in all sectors have at least 100 terabytes of data, many with more than a petabyte.

The big data has less volume is the definition in which target move in a constant rate than by ever-increasing variety, velocity, variability and complexity.

**Variety:** Nearly 85% of an organization's data is unstructured, nonnumeric value but it still must be gathered into quantitative analysis and decision making. Text, video, audio and other unstructured data require changed architecture and technologies for analysis.

**Velocity:** RFID tags and smart measuring are driving an ever greater need to deal with the torrent of data in real time. This, joined with the need and drive to be more agile and deliver insight faster, is putting tremendous force on organizations to shape the necessary structure and skill base to react quickly enough.

**Variability:** In adding to the haste, at which data comes your way, the data movements can be highly mutable with daily, periodic and event-triggered highest loads that can be interesting to manage.

**Complexity:** Complications distributing with data increase with the escalating universe of data sources and are compounded by the need to link, match and convert data

**Corresponding Author:** D. Vivek, Department of Information Technology,
Nandha Engineering College(Autonomous), Erode-52, Anna University, Chennai, India.

across business objects and systems. Organizations need to appreciate relationships, such as complex hierarchies and data linkages, among all data.

A data environment can become extreme along any of the above dimensions or with a combination of two or all of them at once. However, Fig 1 explains the importance in understanding that all data are relevant or useful. Organization must capable of separating wheat from the chaff by focusing on information that counts but not with information overload.

Three technologies help to handle a big data and extracts a meaningful business values from it.

- Information management for big data. The data can be managed as a strategic, core asset, with ongoing process control for analysis of big data.
- High-performance analytics for big data. From big data there is a Gain rapid and the ability to solve complex problems with more data.
- Flexible deployment options for big data. Options are chosen from on- premises or hosted, software-as-a-service (SaaS) approaches for big data and big data analytics.

**Analytics Process Model:** In order to start doing analytics, some basic vocabulary needs to be defined. A first important concept here concerns the basics unit of analysis [1,2]. Customers from various perspectives can be considered. Customer Lifetime Value (CLV) measurement can be either for individual customers or household level. For example, in insurance fraud detection, one usually performs the analysis at insurance claim level. Also in web analytics, the basic unit of analysis is usually a web visit or session.

Analytics is a term that is often used interchangeably with data science, data mining, knowledge discovery and others. The distinction between those is not clear cut. Different techniques can be used for this purpose such as:

- Statistics (eg: linear and logistic regression)
- Machine learning (eg: decision tree)
- Biology (eg: neural networks, genetic algorithms, swarm intelligence)
- Kernel methods ( eg: support vector machine)

Figure 2 explains the decision tree in a classification algorithm setting for predictive analysis, where the minimum equity and provisions a financial institution holds are directly determined by credit risk analysis, fraud analysis and insurance analysis etc.

Analytic S/w

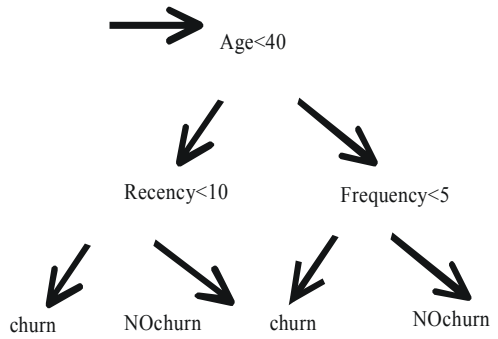


Fig. 2: Example for Classification Analysis

It is important that analytical models are developed in the most optimal way, taking into account various requirements [3]. In this paper the flow goes as following, section II explains the collection and pre-processing of data, section III explains the predictive analysis, Section IV some tools and technologies used in big data world.

**Data Collection:** Data are key ingredients for any list of data sources that are of potential interest before starting the analysis. It is important that every data pre-processing step is carefully justified, carried out, validated and documented before proceeding with further analysis.

Big data is a term terrified around in a lot of objects and understand the meaning of big data, but for those struggling to understand exactly what big data is, it can get annoying. Several definitions are available for big data as it is frequently used as an inclusive term for everything from actual data sets to big data technology and big data analytics. However, this object will focus on the actual types of data that are causative to the ever growing collection of data referred to as big data. Precisely we focus on the data created outside of an organization, which can be grouped into two extensive categories: structured and unstructured.

**Structured Data:**

**Created:** Twisted data is just that, data industries purposely create, generally for arcade research. This may consist of customer surveys or focus groups. Modern methods are included such as creating a constancy program that collects consumer information or asking users to create an account and login while they are shopping online.

The CREATE TABLE statement creates a table in a database.

Tables are organized into rows and columns; and each table must have a name.

```
CREATE TABLE table_name
(
column_name1 data_type(size),
column_name2 data_type(size),
column_name3 data_type(size),
....
);
```

**Provoked:** Giving people the opportunity to express their views is defined by the provoked data. Every time a customer degrees a restaurant, an employee, an acquiring involvement or a product they are creating provoked data. Rating sites, such as Help, also produce this type of data.

**Transacted:** Transactional data is also fairly self-explanatory. On every completion of transaction the data is collected on bases of business services, whether the purchase is completed through an online shopping cart or in-store at the cash register. Businesses also purchase online collect data on the steps that lead to transactional services. For example, a customer may click on a banner advertisement that leads them to the product pages which then spurs a purchase.

"Transacted data is a powerful way to understand exactly what was bought is explained by Forbes article. Matching this type of data with other information, such as weather, can yield even more visions.

**Compiled:** On every U.S. household a compiled data can be collected. Companies like Acxiom collect information on things like credit scores, location, demographics, purchases and registered cars that marketing companies can then access for supplemental consumer data.

**Experimental:** Experimental data is created when businesses experiment with different marketing pieces and messages to see which are most active with consumers. The experimental data is looked as a combination of created and transactional data.

**Unstructured Data:** People in the business world are generally very familiar with the types of structured data mentioned above. Unstructured data is not familiar till NOSQL and HADOOP arrived in the field. In fact, most data being created today is unstructured.

**Captured:** This type of data creation is passive by person's behavior. Every time someone enters a search term on Google that is data that can be captured for future benefit. By using big data technologies images are captured by the use of GPS devices.

**User-Generated:** User-generated data contains all of the data individuals are putting on the Internet every day. From tweets, Facebook posts, to comments on news stories, to videos put up on YouTube, individuals are creating a huge amount of data that businesses can achieve the utility amount of profit for each products.

Big data is made up of many dissimilar types of data. The seven types of external data are included in the big data spectrum. Many types of internal data are available that contribute to big data as well, but hopefully breaking down the types of data helps you to better combine all of this data into big data is so powerful for business.

**Literature Survey:** The performance of Web services is highly related to the service status and network environments which are variable over time; it is an important task to predict the performance of service-oriented systems at run-time. To address this critical challenge, Y. Zhang [4] proposes an online performance prediction framework, called OPred, to provide personalized service-oriented system performance prediction efficiently. Hence, OPred outperforms the state-of-the-art performance prediction approaches in terms of prediction accuracy.

A parallel data mining system for analysing big graph data generated on a Bulk Synchronous Parallel (BSP) computing model named BSP-based Parallel Graph Mining (BPGM). This system has four sets of parallel graph mining algorithms programmed in the BSP parallel model and a well-designed workflow engine optimized for cloud computing to invoke these algorithms [5]. This Parallel based mining has the ability to analyze big graph data and achieved a better performance than the Hadoop-based data mining tools BC PDM and BSP based parallel platform BC-BSP.

To generate a large amount of intermediate data, data-aware cache frameworks for big-data applications are introduced by Y. Zhao and J. Wu [1]. In this framework a task queries the cache manager before executing the actual computing work. A novel cache description scheme and a cache request and reply protocol is designed. Hence, Testbed experiment results demonstrate that Dache significantly improves the completion time of MapReduce jobs. By the utilization of the research achievement in the sociology, Y. Yu [2] designed the Partner First algorithm based on the concept of Partner Circle. The Partner First algorithm achieves the purpose of path planning through the improved A* algorithm, which can significant reduce the search space by preferentially searching the services in the Partner Circle. There is a linear relationship between efficiency and service quantity

of the Partner First algorithm. It proves that using social network in dynamic service composition is efficient and effective.

Big data is a emerging learning that requires distributed computing for the DBNs. A distributed learning paradigm for the RBMs and the back propagation algorithm using MapReduce, a popular parallel programming model [6] was introduced. Thus, the DBNs can be trained in a distributed way by stacking a series of distributed RBMs for pre training and a distributed back propagation for fine-tuning. Concerning the communication cost, only data-level parallelism is performed in the developed distributed algorithm since a fully connected multi-layer network is considered. The ordinal optimization using rough models and fast simulation is introduced to obtain suboptimal solutions [3] in a much shorter timeframe. While the scheduling solution for each period may not be the best, ordinal optimization can be processed fast in an iterative and evolutionary way to capture the details of big-data workload dynamism.

A computational dynamic trust model [7] for user authorization is designed and rooted in findings from social science. This model distinguishes trusting belief in integrity from that in competence in different contexts and accounts for subjectivity in the evaluation of a particular trustee by different trusters. Simulation studies were conducted to compare the performance of the proposed integrity belief model with other trust models from the literature for different user behavior patterns. The semantic link network and multimedia resources are merged for provide a new outlook for organizing them with their semantics. The tags and the surrounding texts of multimedia resources are used to measure their semantic association. The hierarchical semantic of multimedia resources is defined by their annotated tags and surrounding texts. The semantics of tags and surrounding texts in different framework.

The rapid increase number of multimedia resources has brought an urgent need to develop intelligent methods to organize and process the multimedia resources [8]. Y. Liu et al , 2014 designed simantic link network to establish the relationship between the various multimedia resources. The similarity between tags and surrounding texts are implemented using 100 thousands of data like images from flicker where the clustering and searching data mining tasks can be used for the future understanding with the help of semantic link network. Large iterative multitier ensemble classifiers specifically tailored for big data is managed for investigation of huge data classification [9].

**Tools and Techniques for Big Data:** Big data [10,11] is a term alarmed around in a lot of items and understand the meaning of big data, but for those struggling to understand exactly what big data is, it can get annoying. There are several definitions of big data as it is frequently used as an inclusive term for everything from actual data sets to big data technology and big data analytics. However, this object will focus on the actual types of data that are causative to the ever growing collection of data referred to as big data.

The various tools of big data world is analyzed and tabulated in Table 1.

Table 1: Tools used for Big data

| Tools | Developer | Operating System |
|---|---|---|
| Hadoop | Apache | Windows, Linux |
| Map Reduce | Google | Independent |
| Grid Gain | GitHub | Windows, Linux |
| HPCC | Lexis Nexis Risk Solution | Linux |
| Storm | Twitter | Linux |
| Cassandra | Facebook | Independent |
| Hbase | Apache | Independent |

The functionalities are managed and found that there significant advantages to build our own storage solution. Our substantial amount of flexibility from designing various tools to model big table in addition with big data implementation [12].

**Hadoop Methodology:** Open-source software named Hadoop framework for storing and processing big data in a clustered environment on large clusters of commodity hardware. Two different tasks are accomplished such as: massive data storage and faster processing. For starters, let's take a fast look at some of those terms and what they mean.

- Open-source software. Open source software differs from commercial software due to the broad and open network of developers that create and manage the programs. Traditionally, it's free to download, use and contribute to, though more and more commercial versions of Hadoop are becoming available.
- Framework. It specifies everything needed by one can be developed and run with own software applications is provided – programs, tool sets, connections, etc.
- Distributed. Data is divided and stored across multiple computers and computations can be run in parallel across multiple connected machines.

- Massive storage. The Hadoop framework can store huge amounts of data by breaking the data into blocks and storing it on clusters of lower-cost commodity hardware.
- Faster processing. Hadoop processes large amounts of data in parallel across clusters of tightly connected low-cost computers for quick results.

**Data into Hadoop:** To get data from Hadoop there are numerous forms of techniques. Here are just a few:

- Using JAVA commands you can load the files in the file system and HDFS takes care of making multiple copies of data blocks and allocating those blocks over multiple nodes in Hadoop.
- You don't have to write MapReduce code.
- If a huge number of files are presented, a shell script that will run multiple "put" commands in parallel will speed up the process.
- Create a cron job to scan a directory for new files and "put" them in HDFS as they show up. This is useful for things like downloading email at regular intervals.
- Mount HDFS as a file system and simply copy or write files there.
- Structured data are imported by Sqoop from a relational database to HDFS, Hive and HBase. It can also extract data from Hadoop and export it to relational databases and data warehouses.
- Use Flume to continuously load data from logs into Hadoop.
- Use third-party vendor connectors

The following is the hadoop management for word count in the database to manage the storage process improvement.

```
public class WordCount {
public static class NewMapper extends Mapper<Object, Text, Text, IntWritable> {
private final static IntWritable one = new IntWritable(1);
private Text word = new Text();
public void map(Object key, Text value, Context context)
throwsIOException, InterruptedException {
StringTokenizeritr = new StringTokenizer(value.toString());
while (itr.hasMoreTokens()) {
word.set(itr.nextToken());
context.write(word, one);
}
}
}
public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Job job = new Job(conf, "wordcount");
job.setJarByClass(WordCountNew.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
job.setMapperClass(NewMapper.class);
job.setCombinerClass(NewReducer.class);
job.setReducerClass(NewReducer.class);
FileInputFormat.addInputPath(job, new Path (args[0]));
FileOutputFormat.setOutputPath(job, new Path (args[1]));
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
```

The data in the business environment is organized by more variety and volume of data. So the data are reduced by map reduction algorithm.The map and reduce processes can be adapted to a large degree to conform with the preferences of the user and thus they have potential of performing a wide range of computational jobs.

**Map Reduce:** A typical example of a Map Reduce job is the Word Count problem. This consists of counting the number of times a word occurs in an input text. To do this the input file is split into a number of blocks. Each map task will have such a block as input. For each word in the block an output will be created by the mapper, having the word as key and the number 1 as value. The sorting function of the Hadoop framework makes sure that each output of a certain key, in this case word, end up in a specific reduce task. In the reduce phase the values are simply added and the total number of times a word occurs is obtained.

**Experimental Results and Discussions:** In order to determine the reliability of the assumption that adding machines to a computer cluster will reduce computational time, tests with varying number of machines have been performed. The computational time is displayed as a function of number of processors in the computer cluster.

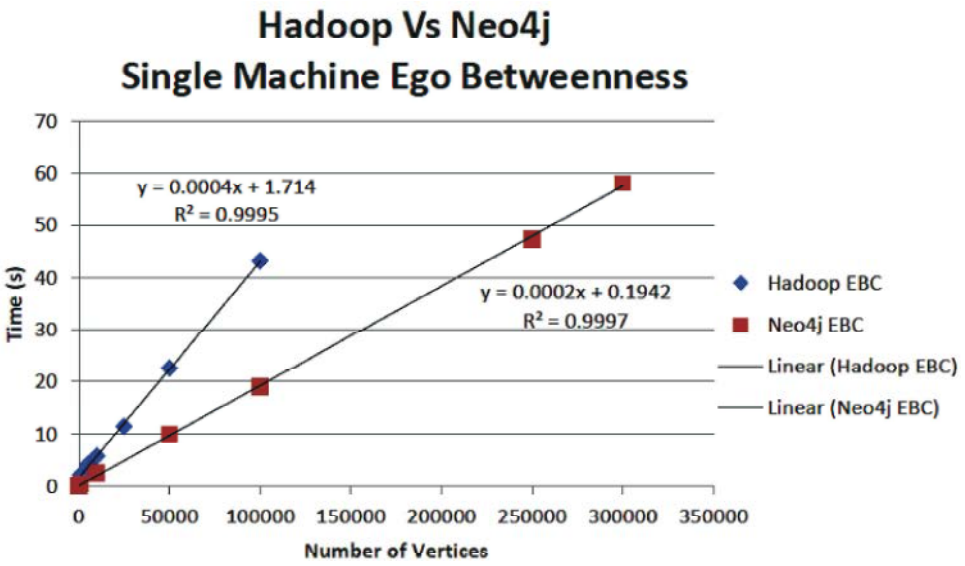


Fig. 3: Comparison and word count analysis

## CONCLUSION

Present days, data volume and variety applications are increasing, the storage management for data are referred as insufficient. Big data is ranking at the top directed at any database-driven application written for the real time applications. This paper results the approach for handling different big data areas with storage techniques and enlists the uses of those stored data to different users, who can manage the data for their usage. The HADOOP environment will manage the data for distributed storage management with the help of map reduce techniques.

## REFERENCES

1. Zhao, Y. and J. Wu, 2013. "Dache: A data aware caching for big-data applications using the Map Reduce framework," Proc. IEEE Infocom, 19(1): 35-39.
2. Yu, Y., J. Chen, S. Lin and Y. Wang, 2015. "A Dynamic QoS-Aware Logistics Service Composition Algorithm Based on Social Network," 2(4).
3. Cao, J., F. Zhang, W. Tan and C. Wu, 2014. "Evolutionary Scheduling of Dynamic Multitasking Workloads for Elastic Cloud Computing," Stuff. Mit. Edu, 2(3).
4. Zhang, Y., S. MembLiuber, Z. Zheng and M.R. Lyu, 2014. "An Online Performance Prediction Framework for Service-Oriented Systems," 44(9): 1169-1181.
5. Liu, Y., B. Wu, H. Wang and P. Ma, 2014. "BPGM?: A Big Graph Mining Tool, 19(1).
6. Zhang, K. and X.W. Chen, 2014. "Large-Scale Deep Belief Nets With MapReduce," IEEE Access, 2: 395-403.
7. Zhong, Y., B. Bhargava, Y. Lu and P. Angin, 2015. "A Computational Dynamic Trust Model for User Authorization," IEEE Trans. Dependable Secur. Comput., 12(1): 1-15.
8 Liu, Y., L. Chen, X. Luo, L. Mei, C. Hu and Z. Xu, 2014. "Semantic link network based model for organizing multimedia big data," IEEE Trans. Emerg. Top. Comput., 2(3): 1-1.
9 Abawajy, J., A. Kelarev and M. Chowdhury, 2014. "Large Iterative Multitier Ensemble Classifiers for Security of Big Data," IEEE Trans. Emerg. Top. Comput., 2(2): 1-1.
10. Hu, H., Y. Wen, T.S. Chua and X. Li, 2014. "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," IEEE Access, 2: 652-687.
11. Dong, X., R. Li, H. He, W. Zhou, Z. Xue and H. Wu, 2015. "Secure Sensitive Data Sharing on a Big Data Platform," 20(1).
12. Xu, L.E.I., C. Jiang and J. Wang, 2014. "Information Security in Big Data?: Privacy and Data Mining," pp: 1149-1176.