IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-6, Special Issue-2, March 2018 E-ISSN: 2347-2693

An Efficient Survey on Predictive Analytics in Big Data

S. Chitra[1*], P. Srivaramanga[2]

[1]Department of Computer Science, Srimad Andavan Arts and Science College,Trichy
[2]Marudupandiyar College , Thanjavur

Available online at: www.ijcseonline.org

Abstract- Big Data has gained much attention from the academia and the IT industry. In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the boundary range. Currently, over 2 billion people worldwide are connected to the Internet, and over 5 billion individuals own mobile phones. By 2020, 50 billion devices are expected to be connected to the Internet. In the recent times the amount of data are generated and stored by various industries are rapidly increasing on the internet thus data scientists are facing a lot of challenges for maintaining a huge amount of data as the fast growing industries require the significant information for enhancing the business and for predictive analysis of the information. This paper focuses on the various states of art studies towards Big Data Analytics techniques.

Keywords Big Data Analytics, Hadoop, Map Reduce, Data Center, Hadoop Distributed File System (HDFS)

Introduction

Imagine a world without data storage; a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity. Data is the building block upon which any organization thrives.

Imagine a world without data storage; a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity. Data is the building block upon which any organization thrives.

The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.

1.1. Big Data Analytics

The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time [12]. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before [17]. Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage [17]. Naturally, business benefit can commonly be derived from analyzing larger and more complex data sets that require real time or near-real time capabilities; however, this leads to a need for new data architectures, analytical methods, and tools. Therefore the successive section will elaborate the big data analytics tools and methods, in particular, starting with the big data storage and management, then moving on to the big data analytic processing. It then concludes with some of the various big data analyses which have grown in usage with big data.

1.2 Characteristics of Big Data

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize big data: volume, variety, and velocity, or the three V's. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the

different kinds of uses and ways of analyzing the data [9].

Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as eXtensible Markup Language (XML) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume.

Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites [17]. Some researchers and organizations have discussed the addition of a fourth V, or veracity. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations [22].

2. LITERATURE REVIEW:

The following table 1 represents the literature review on the Big Data Analytics.

Table 1: Literature Review on Big Data Analytics

Author Name	Paper Title	Description of the Paper
S.Vikram Phaneendra & E.Madhusudhan Reddy	BigData- solutions for RDBMS problems- A survey	Big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. The authors focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc
Kiran kumara Reddi & Dnvsl Indira [R]	Different Technique to Transfer Big Data: survey	The author suggested using nice model to handle transfer of huge amount of data over the network. The Nice model uses a store –and-forward approach by utilizing staging servers. The model is able to accommodate differences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer big data and to solve issues like security, compression, routing algorithms.
T. J. Hacker and K. Mahadik [R]	Flexible resource allocation for reliable virtual cluster computing systems	The authors proposed scheduling polices for virtual high performance computing clusters. They presented a resource prediction model for each policy to estimate the resources needed within a cloud, the queue wait time for requests, and the size of the pool of spare resources needed.
Jimmy Lin [R]	MapReduce Is Good Enough	The authors used Hadoop which is currently the large –scale data analysis "hammer" of choice, but there exists classes of algorithms that aren't "nails" in the sense that they are not particularly amenable to the MapReduce programming model.
M. Durairaj and T. S. Poornappriya [R]	A Review on Big Data Analytics Tools for Telecommunication Industry	The authors in the papers presented the various Big Data Analytics Techniques that to be used for telecommunication industry for churn customer prediction as well as for the improve the revenue of the company.
M. A. Salehi, P. Radha Krishna, K. S. Deepak, and R. Buyya [R]	Preemption-aware energy management in virtualized data centers	The authors proposed a new MapReduce cloud service model for production jobs. Their method creates cluster configurations for the jobs using MapReduce profiling and leverages deadline-awareness, allowing the cloud provider to optimize its global resource allocation and reduce the cost of resource provisioning
Umasri.M.L, Shyamalagowri.D ,Suresh Kumar.S [R]	Mining Big Data:- Current status and forecast to the future	The author also started that there are certain controversy about Big Data. There certain tools for processes. Big Data as such hadoop, strom, apache S4. Specific tools for big graph mining were PEGASUS & Graph. There are certain Challenges that need to death with as such compression, visualization etc.
J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G.	Twister: a runtime for iterative mapreduce	The authors proposed a programming model and architecture to enhance MapReduce runtime that

Fox [R]		supports iterative MapReduce computations efficiently. They showed how their proposed model can be extended to more classes of applications for MapReduce
Albert Bifet [R]	Mining Big Data In Real Time	Stated that streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, allowing organizations to react quickly when problem appear or detect to improve performance. Huge amount of data is created everyday termed as " big data". The tools used for mining big data are apache hadoop, apache big, cascading, scribe, storm, apache hbase, apache mahout, MOA, R, etc
F. Tian and K. Chen [R]	Towards optimal resource provisioning for running mapreduce programs in public clouds	The authors proposed a cost function that models the relationshipbetween the amount of input data, Map and Reduce slots, and the complexity of the Reduce function for the MapReduce job. Their proposed cost function can be used to minimize the cost with a time deadline or minimize the time under certain budget
Bernice Purcell [R]	The emergence of "big data" technology and analytics	Started that Big Data is comprised of large data sets that can't be handle by traditional systems. Big data includes structured data, semi-structured and unstructured data. The data storage technique used for big data includes multiple clustered network attached storage (NAS) and object based storage. The Hadoop architecture is used to process unstructured and semi-structured using map reduce to locate all relevant data then select only the data directly answering the query. The advent of Big Data has posed opportunities as well challenges to business
J. Zhan, L. Wang, X. Li, W. Shi, C. Weng, W. Zhang, and X. Zang [R]	Cost-aware cooperative resource provisioning for heterogeneous workloads in data centers	The authors proposed a cooperative resource provisioning solution using statistical multiplexing to save the server cost
Sameer Agarwal, Barzan Mozafari [R]	BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data	Presents a BlinkDB, a approximate query engine for running interactive SQL queries on large volume of data which is massively parallel. BlinkDB uses two key ideas: (1) an adaptive optimization framework that builds and maintains a set of multi-dimensional stratified samples from original data over time, and (2) A dynamic sample selection strategy that selects an appropriately sized sample based on a query's accuracy or response time requirements.
Y. Song, Y. Sun, and W. Shi [R]	A two-tiered on-demand resource allocation mechanism for vm-based data centers	The authors proposed a two-tiered on-demand resource allocation mechanism consisting of the local and global resource allocation
M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica [R]	Improving mapreduce performance in heterogeneous environments	The authors studied the problem of speculative execution in MapReduce. They proposed a simple robust scheduling algorithm, Longest Approximate Time to End (LATE), which uses estimated finish times to speculatively execute the tasks that hurt the response time the most.
T. Sandholm and K. Lai [R]	Mapreduce optimization using regulated dynamic prioritization	The authors designed a system for allocating resources in shared data and compute clusters that improves MapReduce job scheduling. Their approach is based on isolating MapReduce clusters in VMs with a continuously adjustable performance.
X. Wang, D. Shen, G. Yu, T. Nie, and Y. Kou [R]	A throughput driven task scheduler for improving mapreduce performance in job-intensive environments	The authors proposed a task scheduling technique for MapReduce that improves the system throughput in job-intensive environments without considering the energy consumption.
Z. Ren, X. Xu, M. Zhou, J. Wan, and W. Shi [R]	Workload analysis, implications and optimization on a production hadoop	The authors proposed a job scheduling algorithm to optimize the completion time of small

	cluster: A case study on taobao	MapReduce jobs. Their approach extends job priorities to guarantee the rapid response for small jobs
H. Chang, M. S. Kodialam, [R]	Scheduling in mapreduce-like systems for fast completion time	The authors proposed various online and online algorithms for the MapReduce scheduling problem to minimize the overall job completion times. Their algorithms are based on solving a linear program (LP) relaxation.
B. Moseley, A. Dasgupta, R. Kumar, and T. Sarlos [R]	On scheduling in map-reduce and flow-shops	The authors proposed a dynamic program for minimizing the makespan when all MapReduce jobs arrive at the same time. They modeled the problem as a two-stage flow shop problem, and proved that the dynamic program yields a PTAS if there is a fixed number of job-types.
M. Pastorelli, A. Barbuzzi, D. Carra, M. Dell'Amico, and P. Michiardi [R]	Hfsp: sizebased scheduling for hadoop	The authors proposed a size-based approach to scheduling jobs in Hadoop to guarantee fairness and near-optimal system response times. Their scheduler requires a priori job size information, and thus, it builds such knowledge by estimating the sizes during job execution.
J. Wolf, D. Rajan, K. Hildrum, R. Khandekar, V. Kumar, S. Parekh, K.-L. Wu, and A. Balmin [R]	Flex: A slot allocation scheduling optimizer for mapreduce workloads	The authors Proposed a flexible scheduling allocation scheme, called Flex, to optimize a variety of standard scheduling metrics such as response time and makespan, while ensuring the same minimum job slot guarantees as in the case of Fair scheduler.
T. Sandholm and K. Lai [R]	Dynamic proportional share scheduling in hadoop	The authors proposed a dynamic priority parallel task scheduler for Hadoop that prioritizes jobs and users and gives users the tool to optimize and customize their allocations to fit the importance and requirements of their jobs such as deadline and budget.
A. Verma, L. Cherkasova, and R. H. Campbell [R]	Aria: automatic resource inference and allocation for mapreduce environments	The authors proposed a job scheduler for MapReduce environments that allocates the resources to production jobs. Their method can profile a job that runs routinely and then uses its profile in the designed MapReduce model to estimate the amount of resources required for meeting the deadline
A. Verma, L. Cherkasova, and R. H. Campbell [R]	Two sides of a coin: Optimizing the schedule of mapreduce jobs to minimize their makespan and improve cluster performance	The authors proposed a job scheduler that minimizes the makespan for MapReduce production jobs with no dependencies by utilizing the characteristics and properties of the jobs in a given workload.
R. Nanduri, N. Maheshwari, A. Reddyraja, and V. Varma [R]	Job aware scheduling algorithm for mapreduce framework	The authors proposed a heuristic scheduling algorithm to maintain a resource balance on a cluster, thereby reducing the overall runtime of the MapReduce jobs. Their job selection and assignment algorithms select the job that is best suitable on a particular node while avoiding node overloads.
J. Leverich and C. Kozyrakis [R]	On the energy (in) efficiency of hadoop clusters	The authors proposed a method for energy management of MapReduce jobs by selectively powering down nodes with low utilization. Their method uses a cover set strategy that exploits the replication to keep at least one copy of a data-block. As a result, in low utilization periods some of the nodes that are not in the cover set can be powered down

3. BIG DATA ANALYTIC TECHNIQUES

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like

Apache Hive, Base and Zookeeper. HDFS and MapReduce are explained in following points.

3.1 HDFS Architecture

Hadoop includes a fault-tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates *clusters* of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers.

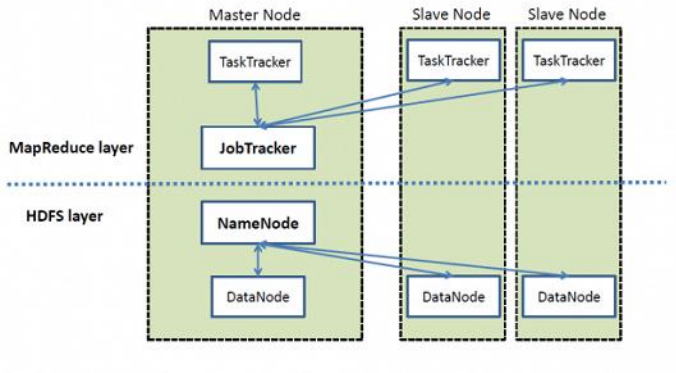

Figure 1: Architecture of Hadoop

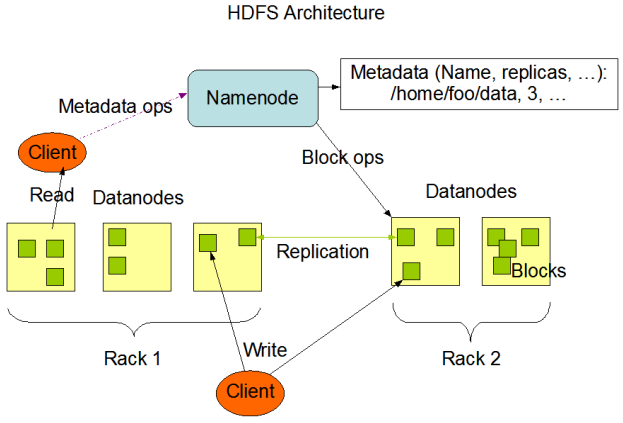

Figure 2: Architecture of Hadoop Distributed File System

3.2 MapReduce Architecture

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur on multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. In a traditional data warehousing scenario, this might entail applying an ETL operation on the data to produce something usable by the analyst. In Hadoop, these kinds of operations are written as MapReduce jobs in Java. There are a number of higher level languages like Hive and Pig that make writing these programs easier. The outputs of these jobs can be written back to either HDFS or placed in a traditional data warehouse. There are two functions in MapReduce as follows:

map – the function takes key/value pairs as input and generates an intermediate set of key/value pairs

reduce – the function which merges all the intermediate values associated with the same intermediate key.

HDFS is based on the principle of "Moving Computation is cheaper than Moving Data". Other Components of Hadoop [6]:

- **HBase**: it is open source, Non-relational, distributed database system written in Java. It runs on the top of HDFS. It can serve as the input and output for the MapReduce.
- **Pig**: Pig is high-level platform where the MapReduce programs are created which is used with Hadoop. It is a high level data processing system where the data sets are analyzed that occurs in high level language.
- **Hive:** it is Data warehousing application that provides the SQL interface and relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query and analysis.
- **Sqoop**: Sqoop is a command-line interface platform that is used for transferring data between relational databases and Hadoop.
- **Avro**: it is a data serialization system and data exchange service. It is basically used in Apache Hadoop. These services can be used together as well as independently.
- **Oozie**: Oozie is a java based web-application that runs in a java servlet. Oozie uses the database to store definition of Workflow that is a collection of actions. It manages the Hadoop jobs.
- **Chukwa**: Chukwa is a data collection and analysis framework which is used to process and analyze the large amount logs. It is built on the top of the HDFS and MapReduce framework.
- **Flume**: it is high level architecture which focused on streaming of data from multiple sources.
- **Zookeeper:** it is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information etc.

4. CONCLUSIONS

This paper describes the concepts of Big Data, characteristics of Big data and Big Data Processing tools. This paper provides the detailed literature survey on the performance of the Big Data Analytics. From the above literature survey, the scheduling of the MapReduce and the Energy Consumption is found to be better processing techniques to improve the performance of the Big Data Analytics.

REFERENCES

[1] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[2] Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348- 2355.

[3] T. J. Hacker and K. Mahadik. Flexible resource allocation for reliable virtual cluster computing systems. In Proc. ACM Conf. High Perf. Comp., Networking, Storage and Analysis, page 48, 2011.

[4] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).

[5] M. Durairaj, T.S. Poornappriya, "A Review on Big Data Analytics Tools for Telecommunication Industry", International Journal of Control Theory and Applications, Volume 9, Issue 27, pp.185-193, 2016.

[6] M. A. Salehi, P. Radha Krishna, K. S. Deepak, and R. Buyya. Preemption-aware energy management in virtualized data centers. In Proc. of the 5th IEEE Intl. Conf. on Cloud Computing, pages 844{851, 2012}.

[7] Umasri.M.L, Shyamalagowri.D ,Suresh Kumar.S "Mining Big Data:- Current status and forecast to the future" Volume 4, Issue 1, January 2014 ISSN: 2277 128X.

[8] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox. Twister: a runtime for iterative mapreduce. In Proc. 19th ACM Int'l Symp. High Performance Distr. Comp., pages 810-818, 2010.

[9] Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data" Copyright © 2013ì ACM 978-1-4503-1994 2/13/04.

[10] Y. Song, Y. Sun, and W. Shi. A two-tiered on-demand resource allocation mechanism for vm-based data centers. IEEE Transactions on Services Computing, 6(1):116{129, 2013.

[11] M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica. Improving mapreduce performance in heterogeneous environments. In Proc. of the 8th USENIX Conf. on Operating systems design and implementation, pages 29{42, 2008.

[12] Anitya Kumar Gupta, Srishti Gupta, "Security Issues in Big Data with Cloud Computing", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.6, pp.27-32, 2017

[13] S. Sathyamoorthy, "Data Mining and Information Security in Big Data", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.3, pp.86-91, 2017

[14] Z. Ren, X. Xu, M. Zhou, J. Wan, and W. Shi. Workload analysis, implications and optimization on a production hadoop cluster: A case study on taobao. IEEE Transactions on Services Computing, 7(2):307{321, 2014

[15] Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) 15–20 DEC 2012.

[16] F. Tian and K. Chen. Towards optimal resource provisioning for running mapreduce programs in public clouds. In Proc. IEEE Int'l Conf. on Cloud Computing, pages 155-162, 2011.

[17] V.K. Gujare, P. Malviya, "Big Data Clustering Using Data Mining Technique", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.2, pp.9-13, 2017

[18] J. Zhan, L. Wang, X. Li, W. Shi, C. Weng, W. Zhang, and X. Zang. Cost-aware cooperative resource provisioning for heterogeneous workloads in data centers. IEEE Transactions on Computers, 62(11):2155-2168, 2013.

