Abstract— With the increase in size and complexity if modern datasets, the world is faced with new challenges in the automation and scalability of the very large data sets. Scheduling has been an active area of research in computing systems since their inception. The main objective is to study MapReduce framework, MapReduce model, scheduling in hadoop, various scheduling algorithms and various optimization techniques in job scheduling. Scheduling algorithms of MapReduce model using hadoop vary with design and behaviour, and are used for handling many issues like data locality, awareness with resource, energy and time.

Keywords— Big Data; MapReduce; Hadoop framework; MapReduce model

1. Introduction

The data generated by modern applications is huge in volume, velocity and variety, and to manage, analyze and mine this vast data for information we can’t use traditional data processing methods. This data is known as Big Data. Big data for the most part incorporates informational indexes with sizes past the capacity of regularly utilized programming apparatuses to catch, oversee, and process information within a practical amount of time. It incorporates almost all types of data including structured and unstructured data. Although the main focus of big data is to deal with highly unstructured enormous data.

Big data is important in order to reduce time and cost and and develop new products with a better insight. Traditional tools and techniques are not sufficient to deal with this type of data. Big data is important in order to reduce time and cost and and develop new products with a better insight. Hadoop is an open source appropriated handling structure that oversees information preparing and capacity for big data applications running in grouped frameworks. It is at the center of a developing community of big data advances that are basically used to help progressed examination activities, including prescient investigation, information mining and machine learning applications. Hadoop can deal with different types of information, giving clients greater adaptability for gathering, preparing and breaking down information than social databases and information distribution centers give.

Hadoop has the capability of storing and processing big data with a huge computing power depending upon the the number of computing nodes being utilized. It makes sure that data is protected and failure or loss of information does not occur. To achieve this, hadoop uses a distributed system where in case of failure of one node, the system delegates the work to some other node. It is also possible to increase the size of the system without affecting the processing speed in hadoop file systems. We may still face some challenges in memory utilization and scalability in hadoop system.

Some other issues include data security and compatibility. Due to the rapid increase in volumes of data in the warehouses, the performance and speed of big data technologies has declined in the past few years. To deal with such issues, the concept of data scaling was introduced into the field of big data. A scalable data platform is the one which is capable of handling drastic changes in the volumes and traffic of data.

This approach needs additional software and hardware to provide bigger storage and produce more efficient outputs. Map Reduce framework has become a manageable and scalable technology for a fault free processing of big data. It is a Java based framework which comprises of two phases - Map phase and the reduce phase.

1. Assistant Professor, Dept. of CSE, SRM Institute of Science and Technology, Chennai
2,3 UG Scholar, Dept. of CSE, SRM Institute of Science and Technology, Chennai.
lavanya27382@gmail.com
Map phase handles the sorting and filtering process and generates an intermediate data and the reduce phase undertakes the data generated by the map tasks and performs a summary operation to process the results. The number of map tasks can be determined by the number of data blocks. It is often tricky to calculate the number of reduce tasks. This shortcoming of map-reduce causes problems in judicious optimization of memory and thereby produces less scalable systems. This in turn causes improper utilization of CPU and IO resources.

2. RELATED WORK

A. Big Scholarly Data: A Survey

With rapid growth in scholarly information and sources and introduction of digital publishing, it is becoming more challenging to analyze, manage and harvest the vast information. The paper provides various data analysis methods, such as statistical analysis, social network analysis, and content analysis for dealing with big scholarly data. Statistical Analysis consists of distribution of quantity and continuous behavior. Network analysis can be done by finding Average Path Length, Clustering Coefficient and Degree Centrality to find out the interdependence of scholarly sources and authors. Textual Pattern Analysis and Topical Analysis can be done for Scholarly Text Mining which focuses on knowledge discovery and text mining.

B. A Distributed Computing Platform for fMRI Big Data Analytics

Interactions within Human brain are very complex and to understand behaviour of brain and analyze underlying data poses a challenge to neuroscience community. The paper proposes usage of a distributed scalable system for storing and visualizing multi-model neuroimage datasets and to use fMRI technique to perform imaging of brain. The process for analyzing has been divided into data management system, data processing pipeline, computing platform, cloud storage, data visualization and processing engines.

The implementation for the above method is done by using Amazon Simple Storage for parallel processing combined with local hard drives and cloud storage, secured with Postgresql and rank-1 dictionary algorithm to decompose input matrix and implement distributed algorithm to utilize computational power and memory capacity across many machines.

Due to emergence of multiple data sources and diverse fields which generate heterogenous data which interact by relational databases and NoSQL databases. The paper proposes use of elasticity property of cloud computing to deal with heterogeneous queries by using Virtual Disk Store System. It makes use of an algebraic tree to form queries, annotates the tree for information, generates operates execution plan using Dynamic Programming approach and evaluates cost of queries.

C. HDM: A Composable Framework for Big Data Processing

For handling big data, application and tasks are divided or pipelined into a sequence of jobs but those are not optimized. In the paper a Hierarchically Distributed Data Matrix is presented solution for a functional, strongly-typed data representation for writing composable big data applications. Due to data flow process and multiple applications pipeline is required and integration, composition and management with big data is not natively supported. It comprises of function fusion, local aggregation, operation reordering and caching.

D. Toward Efficient and Flexible Metadata Indexing of Big Data Systems

Although various techniques to manage data have been generated but still these is not much development in storing metadata or data description. The paper suggests a distributed indexing technique which uses hash table and distributes hashing for fast access to data. It uses hierarchical index layers which are distributed across multiple nodes to avoid a single point of failure. To provide high resilience they are logically stacked and allows for flexible queries, by vertical stacking which also permits for higher granularity. To reduce performance overhead and latency distributed hashing, hierarchy and composite identification is used.

E. Fuzzy Based Scalable Clustering Algorithms for Handling Big data using Apache Spark

In the field of big data and knowledge discovery, it is important that the data generated is used for knowledge and an effective way to do this is by clustering. The paper discusses using of c-Means
Clustering algorithm which partitions the points in a set of C fuzzy clusters. Apache Spark is used for implementation as in iterative clustering algorithms it is faster than others. Parallel processing of data points within a chunk is done and it achieves significant reduction in runtime without compromising clustering quality.

F. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial
Due to large scale technological advancements and exponential growth in terms of data generated a need for data acquisition and management has emerged. The paper discusses lifecycle of such generated Big Data and its impact. The lifecycle can be represented in a chain view comprising of data generation, data acquisition, data storage and data analytics. An alternate strategy is to use layered map which comprises of application layer, computing layer and infrastructure layer. AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data in which actually Cataloging and searching through newly published articles is difficult especially in cases of algorithms which are sometimes not accompanied by pseudo code captions but contain algorithmic procedures which are descriptive information and difficult to search through. The paper proposes a rule based method which searches for captions and then looks for keywords. A second method is a machine learning approach which looks for content of pseudo code instead of captions. It creates tuples of detected line number. A hybrid model of the above two rules is used for efficiency.

G. Non-parametric Distributed learning architecture
There are many issues associated with the computation which further lead to problems associated with scalability and efficiency of big data. Scalable nearest neighbour sparse graph approximation by exploiting graph structure where a nearest neighbor sparse graph can be used to give spatial optimization. Memory space needs to be preserved without compromising the performance of a sparse graph. To tackle this problem a graph can be partitioned into inter and intra graphs and the result of this partitioning can be further used for approximation. This kind of an approach makes sure that the similar attributes within and among each partitioned group is intact. The increase in the data produced by enterprise has grown rapidly in the past few years due to which there is a need of processing data. It is an open source platform which provides high availability and scalability of map reduce. Warehousing and protecting big data state of the art analysis, methodologies, future challenges some of the issues of data warehousing have become a very serious challenge in field of big data research. There are aspects that can be considered to make the management of this data more efficient and flexible. Inefficiencies caused due to factors such as size and design can be tackled by working on the OLAP cubes to introduce a more solid structure and an inventive design to make the process smooth.

H. Scalable Uncertainty-Aware Truth Discovery in Big Data Social Sensing Applications for Cyber-Physical Systems
In social sensing the use of cyber physical systems is growing but the correctness of data sources and reliability is questionable. The paper discusses use of a scalable uncertainty truth discovery scheme. It primarily focuses on disaster affected areas for live updates. A truth discovery scheme is implemented which provides confidence to the data sources. Multiple GPUs running thousands of cores run in few orders of magnitude faster than current schemes which can be of high impact in life threatening situations.

3. RESEARCH CHALLENGES
This section presents numerous significant problems that have not been fully addressed in the current studies. In this event, considerable research effort is still required to improve the efficiency of MapReduce framework and is discussed below.

➢ Energy efficiency
➢ Dynamic Resource allocation
➢ Processing big data in cloud computing
➢ Load balancing
4. CONCLUSION

MapReduce has proven to be a useful programming model framework for large scale data processing. This is because of its remarkable flexibility, which allows automatic parallelization and execution on a large-scale cluster with more than thousands of nodes. Various research papers have been reviewed for the survey mentioned in the literature. Each scheduler considers the resources like CPU, Memory, Job Deadlines etc. To overcome the issue of big data storage and processing the open source framework Hadoop can be used. Starvation caused by non preemption scheduling is reduced by pre-emption to the deadline constraint scheduling model. Real Time MapReduce scheduler achieves good cluster utilization. Dynamic Task Splitting Scheduler improves the performance and fairness of users through dynamically splitting the tasks. Future work will consider enhancement in scheduling using particular scheduler.

5. REFERENCES

[1] Qutaibah Althebyan, Omar ALQudah, Yaser Jararweh, Qussai Yaseen, “Multi-Threading Based Map Reduce Tasks Scheduling”, 5th International Conference on Information and Communication Systems (ICICS), 2014.
[2] YongLiang Xu, Wentong Cai, “Hadoop Job Scheduling with Dynamic Task Splitting”, International Conference on Cloud Computing Research and Innovation, 2015.
[3] Chen He, Ying Lu, David Swanson, “Real-Time Scheduling in Map Reduce Clusters”, IEEE International Conference on High Performance Computing and Communications & IEEE International Conference on Embedded and Ubiquitous Computing, 2013.
[4] Divya M., Annappa B., “Workload Characteristics and Resource Aware Hadoop Scheduler”, IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), 2015.
[5] Kamal Kc, Kemafor Anyanwu, “Scheduling Hadoop Jobs to Meet Deadlines”, 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010.
[6] Li Liu, Yuan Zhou, Ming Liu, Guandong Xu, Xiwei Chen, Dangping Fan, Qianru Wang, “Preemptive Hadoop Jobs Scheduling under a Deadline”, Eighth International Conference on Semantics, Knowledge and Grids, 2012.
[7] Hong Mao, Shengqiu Hu, Zhenzhong Zhang, Limin Xiao, Li Ruan, “A Load-Driven Task Scheduler with Adaptive DSC for MapReduce”, IEEE/ACM ICGCC, 2011.
[8] Qinghua Lu, Shanshan Li, Weishan Zhang, “Genetic Algorithms based Job Scheduling for Big Data Analytics”, ICIKSIOT, 2015.
[9] Ke Wang, Ning Liu, Iman Sadooghi, Xi Yang, Xiaobing Zhou, Tonglin Li, Michael Lang, Xian-He Sun, Ioan Raicu, “Overcoming Hadoop Scaling Limitations through Distributed Task Execution”, IEEE International Conference on Cluster Computing, 2015.
[10] Xiangming Dai, Brahim Bensaou, “Scheduling for response time in Hadoop MapReduce”, IEEE ICC 2016 SAC Cloud Communications and Networking, 2016.
[11] Rajatish Mukherjee, Partha Sarathi Dutta, Gerdur Jonsdottir & Sandip Sen "Movies2go-An Online Voting Based Movie Recommendation System".
[12] Christoph Beckmann, Tom Gross “Towards a Group Recommender Process Model for ad-hoc Groups and On-Demand Recommendations”.
[13] Yan Tang, Mingzheng Li, Wangsong Wang, Pengcheng Xuan, Kun Geng “Quality-aware movie recommendation system on big data”.

➢ Mapping scheme
➢ Performance optimizations
➢ Optimized data shuffling
➢ Automation and configuration:
[14] Goral K. Godhani, Maulik Dhamecha “Simulation of Genre Based movie recommendation system using hadoop map reduce technique” 2007 IEEE, ICECDS-2017.

[15] Lakshmi Tharun Ponnam, Sreenivasa Deepak Punyasamudram, Siva Nagaraju Nallagulla, Srikanta Yellamati “Movie Recommendation System Using Item Based Collaborative Filtering Techniques” 2016 IEEE.

[16] Daniel Pomerantz, Gregory Dudek “Context Dependent Movie Recommendations Using a Hierarchical Bayesian Model” Springer- Verlag Berlin Heidelberg 2009.