Map Reduce processing challenges in cloud environment

Suyash Mishra[.1],Suryakant yadav [2]
[1] Suyash Mishra ,Noida International University ,
[2] Suryakant Yadav ,NIU.

Abstract. MapReduce is a programming model, proposed by Google initially to process large datasets. It consist of two components Map and Reduce motivated from functional languages e.g. Lisp and Haskell. To achieve desirable result programs task is to define and implement Map and Reduce functions. Map and Reduce should describe the logic encapsulated in both components to achieve desirable result. There are various big data based operations e.g. culling, highlighting, indexing, searching, faceting. It impossible for a single node machine to manage and process such a huge data. Map Reduce regarded as an important function to tackle continuous increasing demands of computing resources required for processing big data sets. MapReduce highly scalability nature permits parallel and distributed processing on multiple computing nodes. This paper talks about MapReduce usages in cloud computing and motivate its use for big data processing into cloud heterogeneous environment

Keywords: Big data, Iot, Cloud computing, Grid, MapReduce.

1 Introduction

Traditional data storage and processing capabilities were limited and was function of infrastructure, storage availability and processing capability, which deemed to be very different from today. Thus, those approaches and databases are facing technical challenges to accommodate Big Data storage and processing demands.

Current rapid advancement and ongoing progress in the social media, robotics, web, healthcare and mobile devices are producing data, which is growing exponentially, and processing such data becomes a huge challenge. Social sites are generating more than 1000 terabytes of new data daily due to millions of user base. Traditional data stores processing capabilities are limited and reliant on available infrastructure, storage and processing capability, which is not as per today. Thus, those traditional approaches are facing tremendous challenge while ingesting Big Data and executing to draw meaningful insight from data. The term “Big Data” can be defined as large and unstructured data sets consists of variety of structured or unstructured data, which is very large and growing very rapidly and unmanaged by traditional databases and approaches.

In an organization huge effort made to manage data, draw meaningful business insights from data, which can be shared further to business for better decision making, stay ahead into business competition, and improve sales and revenues by minimizing processing time and better decision making. Thus, Industry is exploring and investing much too effective harnessing Big Data and analyze to identify beneficial business insights to enable better business decisions and value addition to their business. MapReduce allows on demand nodes addition, along with capability to process big data in parallel fashion on a huge number of computing nodes. Google [2] advocated MapReduce programming paradigm .This was used in many open source projects, but most prominently adopted by Apache Hadoop [3]. MapReduce combination of two components map and reduce which application developers can design as per task they want to achieve. The popularity of MapReduce is due to its simplistic approach to process big data efficiently and offering features like scalability to include multiple high processing nodes to process rapidly growing data, fault-tolerant fashion when any of node went down and simple nature. It is also free from specific programming language syntax or specific data storage system. In the Big Data processing, MapReduce is a core component of the Apache Hadoop software framework. However, this paper aims to gauge main challenges of MapReduce in three different platforms: Individual computing node, multi connected desktop grids and cloud computing.

2 Related work

2.1 MapReduce: MapReduce is a programming paradigm for harnessing distributed computations on big data sets on clusters of machines. Other way MapReduce represents to a programming design, which executes on a computational cluster to mine the business insight or results from a large datasets, which cannot processed with traditional approaches and takes many hours to get result. MapReduce is a combination of two primitive functions map and reduce functions. The Map applies task and returns a list of results while Reduce function shuffles, sort and combines intermediate results from Map in parallel manner. MapReduce Model splits the input dataset into independent small datasets called as chunks, which are further processed by map and reduce .Submitted task execution using map and reduce is performed on Data Nodes and interim result is stored on same node locally. In the MapReduce Model, programs are executed on the large cluster of commodity hardware automatically in parallel manner. Hadoop itself take care of split task, processing log and nodes where sub tasks .Hadoop take care of node failures, and manages inter-machine communication making programmer or user not worried about node failure by providing efficient fault tolerant mechanism and scalability.

2.2 Big data and cloud computing: Big data can be thought as fast growing and having verity of data, which cannot be processed efficiently The concept of big data is high in demand and major buzz now a days due to advancement of technologies and IoT devices both into academics and industries. MapReduce is used to deal with big data processing and derive business insights; data can be structured or unstructured in. Ability and need to harness the Big data and extract the meaningful information has led organization to invest into Big data and mine information which can be used for predicting user behaviors, for financial organization to identify fraudulent transition early warning sign or potential default behavior for dollar loss saving or increasing revenue. Big data provide opportunities to store historical clinically huge scientific datasets and perform analysis on vast historical data rather than doing data analytics on couple of years for more accuracy.
As Big, data is started adopted rapidly across the business organizations and it is proving its efficacy will make organizations move towards the big data for managing and using value out of data, which was previously considered as waste or unprocessed. People thinks of big data mostly for data storage of huge data but it is equally great to process and extract information from it The five different properties define the charerestic of big data (famous as five V) as Volume, Variety, Velocity, Value and Veracity .

a. Size of dataset is known as Volume, which can be processed, and accommodated big data system.

b. Variety represents type of data, which is produced from various sources e.g. audio, logs, video chat data.

c. Velocity means the pace of data generation or pace of execution should be very fast.

d. Value means the true value of data (i.e., the appropriate information and value data can provide). Storing huge amounts of data is wastage of time and capability and have no meaning, as valued insight cannot extracted out of it.

e. Data integrity and reliability of data termed as Veracity refers .This data confidentiality, integrity, and availability. Users expects data and results are trustworthiness.

Cloud computing is an advanced programming approach which gives view to users that they have virtually infinite on-demand resources available to its end users with setting up roles as per their requirement. Cloud's power to virtualize resources makes efficient and cost effective hardware procurement, need minimum interaction with cloud service providers and allows users to access petabytes of data storage and high processing capability, and high availability in a pay-as-you-go basis (González-Martínez et al., 2015). This approach makes users not to worry about resource management and delegates cost and resource management overhead from the user to the cloud server provider. This gives an opportunity to business organization to use what they want and pay what and how much they user and power in hand to increase number of services. Also there us great cost saving and boon to small organization who wish to utilize minimum services but have a spent much now adopting cloud offered services they will pay as per use and don't have to bother about infrastructure and resource management. This throws a great opportunity because initial IT infrastructure and procurement is reduced which includes hardware procurement and maintenance, software licenses, IT resources. Cloud computing provides an easy option to secure resources pay-as-you-go basis. This allows users or organizations to easily procure resources with the cloud service vendor. Broadly Cloud offered services are categorized into three hierarchies based on the kind of services they Iaas (Infrastructure as a Service, PaaS (Platform as a Service) and SaaS (Software as a Service).

Using these services facilitate various services to user and reduces pain areas and cost of incurred .Other than above mention three main services cloud offers few more services e.g. Big Data as a Service (BDaaS), Database as a Service (DBaaS) and Analytics as a Service (AAAS). Cloud is considered as most viable and efficient framework for big data processing as this allows and manages on demand resource and makes powerful framework for big data.

3 MapReduce: Processing of Big data

MapReduce is a programming framework having great ability to execute and large volumes of data in parallel manner by breaking the job into a set of independent small processing units. At one side, MapReduce offers developers to customize and develop Map and Reduce task s per their requirement along with concealing many system level information and complexities from users to let them focus on their task .Data processing is conducted in two phase first phase is Map and second phase is Reduce. Map phase takes split data file as its input, which can be located in any of data nodes of distributed file system and contains the key/value data. The split data file can be present in same location where Map function or not in same location. If data and the Map function are not present in the same node, then the system will try to move the data split file to the Map function on the node where it exists. Hence MapReduce adheres concept of "Moving data closer to compute" to reduce processing time. Further Reduce function is executed to all values that assigned same intermediate key and gives output key/value pairs as the result. The MapReduce framework is based on master and slave architecture. One node acting as single master node, which keeps a JobTracker along with several slave nodes to runs assigned task and Tasktrackers maintain assigned task's progress status, which executes one task in each node of cluster. The JobTracker acts as interface between users who submits the job and the MapReduce framework. Users submit map/reduce jobs to the JobTracker and jobs are accepted and resources allocation is done based on several scheduling techniques e.g. first come/first-served basis. The JobTracker records the assignment of map and reduce tasks to the slave nodes processing tasktrackers. The Tasktrackers accepts and run tasks based on instruction from the JobTracker, manages data transfer between the maps, and reduce phases.
Whole processing can be elaborated in detail as per below processing sequence it has been explained in figure 1.

a. Map – Input data is first accepted by master node, which divides data file into smaller datasets, and move them to slave nodes. A slave node may reiterate the process leading to a multilevel tree structure. Map accept one pair of data with a type in one data and gives a list of pairs in a different domain:.
b. Map logic execution – Map logic will be executed once for each key value, generates output grouped by key values.
c. Reduce phase execution – the MapReduce system identify a node to run Reduce function, assigns the key value to each processor, and provides that node with all the Map-generated data allocated by same key value.
d. Execute Reduce code – Reduce is executed only once for each key value produced by the Map stage.
e. Produce the result – Final output is generated by consolidating all the Reduce output, and sorts them by key.

Below diagram elaborates above-mentioned steps using word count example following big data MapReduce processing..

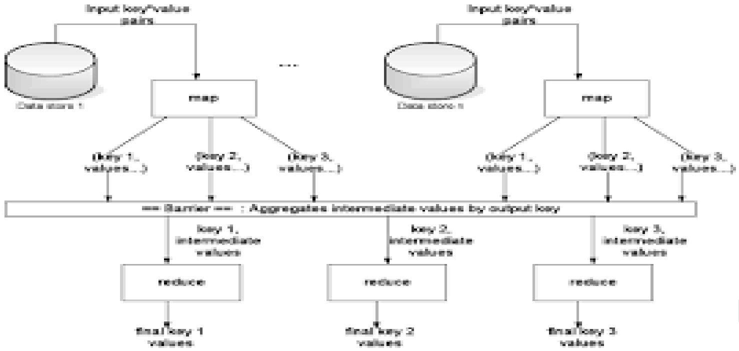

Figure1.

In a homogeneous environment where all nodes are considered as possessing similar computing power and storage capacity. In case of node failure, MapReduce attempts to reruns failed tasks on a different idle/less-occupied node. There may be case when a node is processing a task slowly is termed as straggler; Map Reduce then kick off a speculative copy of straggler task ("backup task") on another node, which can finish task quickly. In absence of speculative execution, a job will be very slow and run for longer downgrading the performance. Stragglers can arise due to numerous reasons due to faulty hardware and in appropriate configuration. It is proved by various researchers that speculative execution can increase job response times by 44%.In Heterogeneous environment this problem is deteriorate the execution of speculative execution due to difference in nodes storage and processing capability.

4 MapReduce: Processing of desktop grid Computing

Grid computing assumed to be the combination of computer resources spanned across multiple places with a single objective. The grid is a distributed system consists of non-interactive nodes involving a number of input data files. Grid computing is different from traditional high executing computing nodes in such a way that grid computers each node executes a different task or application. Grid nodes are diverse hence; form more heterogenenity.A single grid can be assigned to an application having similar processing requirement.

Foster and Kesselman [9] has defined that desktop Grids are systems that combine the computation capability of idle nodes, taking benefits of the fact that these machines are available during most part of the day (85% during day time, in addition, as much as 95% during the night [9]. This environment also comes with verity of concepts.

Heterogeneity .There is diverse set of hardware and software configurations that affects the processing capability of individual nodes and the performance of the grid as a whole. Therefore, less occupied nodes are assigned tasks to process grid tasks. As additional aspects that can help to characterize this environment, we could cite:

There are below problems running MapReduce deployments on desktop grids in comparison homogeneous environments.

I. More tasks/job failures probability, when users interrupt the execution of nodes.

II. Distributed file system nodes will also become unavailable may cause missing data besides task failure.

III. Excessive launch of speculative tasks along with data distribution due to heterogeneity of nodes.

There were efforts made to resolve above highlighted challenges but not entirely, only one of above were or a specific combination of these issues. Like Zaharia et al. [20] propose a new scheduling algorithm to minimize the negative impact of the speculative execution on heterogeneous environments, but do not consider data distribution, load balance and volatility.

4 MapReduce: Application in cloud for processing big data

Cloud computing is an advance programming suited for meeting parallel and distributed computing. It consists of interconnected and virtualized computers network. With its advancement, new MapReduce implementations developed to deploy and run MapReduce in cloud environment.

I. Cloud services are different from normal services in below manner.

II. Based on "PAY AS YOU GO", means can be sold for hour's days or monthly usages.

III. Elastic in nature, wherein users can dynamically change demand to get more or less services as much as they need.

The key technical challenges identified on cloud systems are as below.

Virtualization: Virtualization is a kind of making a virtual image of a server or services e.g. operating system, storage media or networking resources having an intent to be used across multiple machines simultaneously. Prime objective of virtualization is to take care of workload management by redefine traditional computing by optimizing the processing, efficient and economical Virtualization technology is hardware optimizing cost effective technology that is rapidly advancing the fundamental way of computing. Hence maintaining efficient virtualization is key for Cloud based systems.

Multitenancy: A phenomena in cloud systems, where the code location or data location is not known and copy of same resource is allocated to other users known as tenant and these phenomena known as multitenancy. This makes software managing easy and cost effective by sharing resources, cost .This makes high availability of resources and services, which are hosted on shared resources are must be always available for other users. Data Management is an important aspect for storage clouds, where data distribution across various resources is done it self.

Security: Data security, Compliance and privacy is an extremely important feature in cloud system to protect sensitive data and code. Main challenge for MapReduce in cloud is load balancing .In a traditional Hadoop there are two nodes name (Master node)node and data nodes(slaves) .However in cloud environment this name node is a cloud server and slaves nodes may also be distributed server. Data files are already loaded into cluster nodes. When the MapReduce starts, execution NameNode picks up JobTracker to allocate tasks, TaskTracker monitor which DataNodes and progress will finally process. There are many Map programs, which can execute on each Data Node and the interim results will be send to next process, which combines and produces result.

Load balancing: It plays vital role in distributing the load uniformly across the available /underutilized nodes. When a node became over utilized than what it can process .Though load balancing load is shared to available nodes is not so important in a MapReduce processing, it becomes essential to process large files when hardware resources comes at a cost and usages is critical. Hence, load balancing optimizes hardware optimization in resource-critical scenarios with a significant uplift in performance.AWS EMR allows user to run and submit their jobs via web services rather than taking traditional steps to run over Hadoop makes analyst life easy. This offers immense opportunity to dig more onto unstructured big data generated over cloud EMR merits e.g. reliability, convenient to use, cost effective, scalable, and secure in nature makes this best choice in cloud environment.
There are below areas where Amazon EMR can prove instrumental to benefit analysts and industry.

Clickstream Analysis-Click stream analytics empowers user to collect, process, takeout summarized outlook of website visitor .This helps organizations to know user segment, user click behavior and based on analysis appropriate, and effective advertisements can be sent to user. This helps organization to know their customers better and offer product of their need make customers more satisfied and retain them with business. Amazon EMR is suitable and used for click-stream data analytics.

Real-time Analytics: Real time data flowing into system require quick storage and process real-time data coming from various sources .This help business to know real time behavior of users and offers services accordingly. There are many tools and vendors e.g. Amazon Kinesis, Apache Kafka, and many more data streams for managing and respond the real time data .For instance Spark Streaming is famous for real time data processing in quick manner and then write processed result back to Amazon S3 or HDFS.

Log Analysis: Amazon EMR helps organization process logs generated by web applications, sensors data and mobile applications. Amazon EMR converts petabytes of un-structured data into semi-structured or structured data so that analytical operations to be performed to derive useful insights about customer's behavior or application functionality.

The AzureMR also provides cloud based MapReduce facility using the HDInsight .A MapReduce job can be submitted using web services, monitor its progress and store output to HDFS for further computation and analysis. The main contribution of AWS and Azure computing is a new architecture, without a server node as below.

6 Conclusion and future work

MapReduce has been regarded as prominent programming paradigm to cope with Big data processing .Though MapReduce offers numerous advantages but there are few trade-offs faced in meeting rapidly growing computing demands of Big Data in heterogeneous environment. The super success and adoption of MapReduce by number of organizations. To utilize MapReduce efficiently in Cloud environment for big data processing there was attempt made to explore and analyses some of the issues of executing MapReduce on two different environments: desktop grid and cloud computing. The challenges in performing MapReduce on these environments share some common tasks, such as dealing with heterogeneity and different location of nodes. Various storage related issues like data locality and data replication etc. Each of Different approaches and scheduling methodology is required in different environments. Future works, aims to identify and explore findings that will allow the adoption of the MapReduce model in a higher and much sophisticated environments e.g. complex clouds or IoT (Internet of things) of platforms.

7 References

[1] Dean, J., Ghemawat, S.: MapReduce: Simplified data processing on large clusters. Year 2004

[2] Hadoop MapReduce, http://hadoop.apache.org/mapreduce/

[3] "Towards Scalable Systems for Big Data Analytics" by H Hu, Y Wen, T Chua, X Li, year 2014

[4] Chang, V. Towards a big data system disaster recovery in a Private cloud. Year 2015

[5] Zhang, L. et al "Moving big data to the cloud" IEEE 2013.

[6] Wu, X. et al. "Data mining with big data" IEEE 2014.

[7] Subashini, S. & Kavitha, V., "A survey on security issues in service delivery models of cloud computing" published in. Journal of Network and Computer Applications 2011.

[8] Hashem, I.A.T. et al." The rise of "big data" on cloud computing" 2014.

[9] I. Foster and C. Kesselman. "The grid: blueprint for a new computing infrastructure." 2004.

[10] https://aws.amazon.com/emr/

[11] https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-use-mapreduce

[12] D. Kondo, G. Fedak, F. Cappello, A. A. Chien, and H. Casanova. "Characterizing resource availability in enterprise desktop grids".

[13] D. Kondo, B. Javadi, P. Malecot, F. Cappello, and D. P. Anderson. "Cost-benefit analysis of cloud computing versus desktopgrids." IEEE 2009.

[14] Case Study Nokia: Using big data to Bridge the Virtual & Physical Worlds. Geller , cloudera, 2012.

[15] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S. H. Bae, J. Qiu, and G. Fox. " Twister: a runtime for iterative mapreduce". ACM 2010.

[16] M. C. Schatz. Cloudburst: highly sensitive read mapping with mapreduce. Bioinformatics, June 2009.

[17]] T. White. Hadoop: The Definitive Guide. O'Reilly Media, original edition, June 2009.

[18] J.Xie,S.Yin,X.Ruan,Z.Ding,Y.Tian,J.Majors,A.Manzanares, and X. Qin. " Improving mapreduce performance through data placement in heterogeneous hadoop clusters" IEEE 2010.

[19] Mahesh, A. et al. Distributed File System For Load Rebalancing In cloud Computing. , 2, pp.15–20. ., 2014

[20] M.Zaharia,A.Konwinski,A.Joseph,R.Katz,andI.Stoica. Improving mapreduce performance in heterogeneous environments.