ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 5, Special Issue 12, April 2018

ISSUES, CHALLENGE AND SOLUTIONS: HADOOP & MAPREDUCE

1. Mr. S.Thirumurugan 2. Dr. P.Thambidurai 3. Dr. K. Elangovan 4. Mr. A. Antony Prakash

1. *Research Scholar , Department of Computer Science ,Bharathiar University, Coimbaitore. vstmurugan@gmail.com*
2. *Professor of CSE and Principal, PKIET, Karaikal* ptdurai@pec.edu
3. Asst. Professor in Computer Science and Engineering, Bharathidasan University, Trichirappalli. elangovan.k@csbdu.in
4. Asst. Professor in Information Technology, St Joseph's College, Trichy. aantonyprakash@gmail.com

Abstract: In Big Data Hadoop and MapReduce is a software framework which is process vast amount of data. MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. The reason for this is the high scalability of the MapReduce paradigm which allows for massively parallel and distributed execution over a large number of computing nodes. This paper gives the ideas about MapReduce issues and challenges handling in Big Data. The identified challenges are grouped by Storage of data, Data Analytics, Online Processing and privacy and security.

Keywords: Big data, Hadoop, MapReduce, Security

I. INTRODUCTION

Enormous Data turns into a noteworthy subject in the territory of ICT. It is clear that Big Data implies business openings, yet real research challenges as indicated by "McKinsey and Co1" Big Data is "the following limit for development, rivalry and productivity". The activity of Big Data gives not just an immense feasible for rivalry and augmentation for singular organizations, however the correct utilization of Big Data additionally can enhance efficiency, development, and intensity. Volume of information is a major firm, yet speed and assortment of information is additionally contains difficult issue. Investigation of examination: The New Path to Value is half of the best-accomplishing associations; it turned out, utilized computerized information for their long haul techniques, as opposed to just a fifth of the underperformers.

To have the capacity to extricate the advantages of Big Data, it is pivotal to know how to guarantee canny utilize, administration and re-utilization of Data Sources, including open government information, in and crosswise over Europe to fabricate helpful applications and administrations. The Digital Agenda stresses the significance of augmenting the advantages of open information, and particularly the requirement for opening up open information assets for re-utilize (Action 3 of the Digital Single Market Pillar). Open Sector Information (PSI) is the single biggest wellspring of data in Europe. Its assessed showcase esteem is €32 billion. Re-utilized, this open information could produce new organizations and employments and give buyers more decision and more incentive for cash, as the DAE brings up. The European Commission has just pushed for a few activities at specialized level (basically identified with information arranges so as to advance interoperability and re-utilize) yet additionally at administrative and strategy levels, by cultivating straightforwardness and accessibility of information and empowering access. The EU Open Data procedure, which is a revision of the Public Sector Information Directive, supports more transparency and reuse of open segment data2. Despite the fact that the procedure is still in a beginning period there is as yet a need to investigate the two sides of the issue as open information can be a danger for security.

In Horizon 2020, Big Data discovers its place both in the Industrial Leadership, for instance in the action line "Content advancements and data administration", and the Societal Challenges, identifying with the requirement for organizing information in all areas of the economy (wellbeing, atmosphere, transport, vitality, and so forth.).

Business reasons

1. Conceivable of new imaginative plans of action

2. Focal points of Competitive gives new bits of knowledge emerge

Innovation reasons

3. Capacity of information keeps on developing exponentially.
4. Discover the information in different structures.

5.With respect to don't meet the conventional arrangements

Financial reasons

6. The expenses of information frameworks keep on rising as a rate of the it spending plan

7. New standard equipment and open-source programming Offer money saving advantages

II LITERATURE REVIEW

The Hadoop and mapreduce [1] demonstrate was proposed by

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 5, Special Issue 12, April 2018

"Senior member and Ghemawat at Google". Mapreduce is Programming model is intended for vast volume of information. Do the trick it to state here that a large number of these sorting out information administrations are MapReduce motors particularly intended to enhance the association of huge information streams[2]. Arranging information administrations are, in all actuality, a biological system of instruments and advances that can be utilized to accumulate and collect information in planning for additionally preparing. In that capacity, the apparatuses need to give incorporation, interpretation, standardization, and scale.

MapReduce and Hadoop are dispersed figuring conditions where everything is preoccupied. The detail is dreamy out so the designer or examiner does not should be worried about where the information components are really found. [3]

Huge information accept dissemination. Practically speaking, any sort of MapReduce will work better in a virtualized situation. You require the capacity to move workloads around in light of prerequisites for figure power and capacity. [4], [5] MapReduce motor is parallelized and arranged to keep running in a virtual situation, you can decrease administration overhead and consider developments and compressions in the undertaking workloads. MapReduce itself is characteristically parallel and circulated. By epitomizing the MapReduce motor in a virtual compartment, you can run what you require at whatever point you require it. With virtualization, you increment your use of the advantages you have officially paid for by transforming them into non specific pools of assets. [6] and delineate [7]. "MapReduce uses the Google File System(GFS) as a basic stockpiling layer to peruse information and store output"[8].

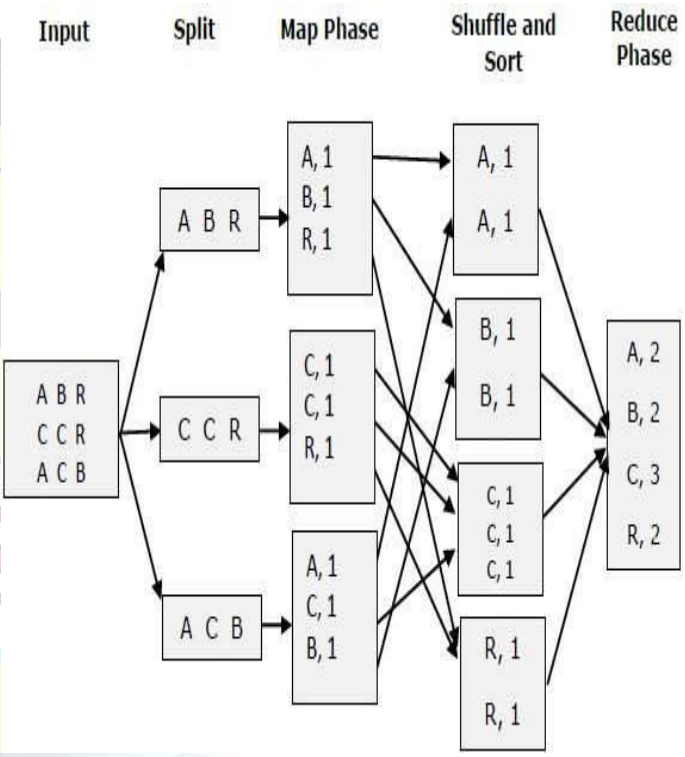
The Hadoop engineering handle expansive informational indexes, versatile calculation. Hadoop enabled huge issues to be separated into littler components with the goal that investigation should be possible rapidly and cost-effectively.By breaking the enormous information issue into little pieces that could [9].

HDFS isn't the last goal for records. Or maybe, it is an information benefit that offers a one of a kind arrangement of capacities required when information volumes and speed are high. Since the information is composed once and after that read commonly from there on, as opposed to the steady read-composes of other record frameworks, HDFS is a fantastic decision for supporting enormous information investigation. The administration incorporates a "NameNode" and different "information hubs" running on an item equipment group and gives the most elevated amounts of execution when the whole bunch is in the same physicalrack in the server farm. [10].

III. Architecture for MapReduce

MapReduce (the algorithm) and a usage of MapReduce. Hadoop MapReduce is a usage of the algorithmdeveloped and kept up by the Apache Hadoop venture. It is useful to consider this usage a MapReduce motor, since that is precisely how it functions. You give input (fuel), the motor changes over the contribution to yield rapidly and effectively, and you find the solutions you require. You are utilizing Hadoop to take care of business issues, so it is fundamental for you to see how and why it functions..

Hadoop MapReduce incorporates a few phases, each with an essential arrangement of tasks getting to your objective of finding the solutions you require from huge information. The procedure begins with a client demand to run a MapReduce program and proceeds until the point that the outcomes are composed back to the HDFS



The MapReduce algorithm has two vital errands that is Map and Reduce.

- Map has two extra capacities to address the inquiries. Since outline diminish need to cooperate to process your information, the program needs to gather the yield from the autonomous mappers and pass it to the reducers. This undertaking is performed by an OutputCollector. A Reporter work too gives data accumulated from delineate so that you know when or if the guide undertakings are finished.

- The Reduce errand decrease accumulates its yield while every one of the assignments are handling. Decrease can't start until the point that all the mapping is done, and it isn't done until the point when all cases are finished. The yield of lessen is likewise a key and an esteem.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 5, Special Issue 12, April 2018

While this is essential for decrease to do its work, it may not be the best yield organize for your application.

Map Stage : The guide or mapper's activity is to process the info information. By and large the info information is as document or catalog and is put away in the Hadoop record framework (HDFS). The info record is passed to the mapper work line by line. The mapper forms the information and makes a few little pieces of information.

Reduce Stage : This stage is the mix of the Shuffle arrange and the Reduce organize. The Reducer's activity is to process the information that originates from the mapper. In the wake of preparing, it delivers another arrangement of yield, which will be put away in the HDFS.

		MapReduce
	Iterative algorithms	Extensions of MapReduce implementation such as **Twister and HaLoop**
Online processing	Latency issues and Performance	communication between phases and jobs
Privacy and Security	Auding , access control, authentication and authorization	**Apache Shiro, Apache Ranger, Sentry** - security analytics as well as privacy policy enforcement with security to prevent information leakage.

IV. Issues and challenges in MapReduce

"MapReduce is a framework using which we can write applications to process huge amounts of data" in parallel, on large clusters of commodity hardware in a reliable manner. MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce.

	Main challenges	Main solution approaches
Storage of data	Schema-free	Map reduce contain cache items, request and reply **NoSQL stores** – MapReduce with various indexing approaches
	Lack of standardized SQL-like language	**Apache Hive** – SQL on top of Hadoop **NoSQL** stores: proprietary SQL-like languages (Cassandra, MongoDB) or Hive (HBase) **Sqoop** – various relational database transfer data
Data Analytics	Scaling complex, machine learning and interactive analytics	Use computationally less expensive, though less accurate, algebra
	Interactive analysis	Map interactive query processing techniques for handling small data, to

The five key challenges of working in Hadoop MapReduce are:

4.1 Lack of information stockpiling and bolster capacities
4. 2 Lack of Deficiency in Tools
4.3 Lack of scientific abilities in database
4.4 Privacy and Security challenges

4.1 Lack of information stockpiling and bolster capacities

Information stockpiling is record and outline free.

Solution

The workaround is to utilize instruments that give in-database MapReduce preparing

Apparatuses you can utilize

MongoBD, Orient, HBase

4. 2 Lack of Deficiency in Tools

Open-source advancement apparatuses are rapidly developing, yet creating arrangements with Hadoop still requires a lot of hand-coding in dialects that couple of information administration proficient know how to utilize, for example, Java, R, Hive, and Pig. This is an abilities issue certainly, however it additionally backs off device improvement.

Solution

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 5, Special Issue 12, April 2018

when you make an application for single capacity or numerous capacities, you make an IPAF (Inter Protocol Acceptability Framework]

Instruments you can utilize

Custom IPAF

4.3 Lack of scientific abilities in database

Doing MapReduce capacities, it is hard proportional complex iterative calculations. It has computational difficulties and measurable.

Solution:

"HaLoop and Twister are the two expansions intended for Hadoop all together for MapReduce usage to better help iterative calculations". The Data pre-handling methodology ought to be taken after utilizing MapReduce.

Instruments you can utilize

R, MatLab, Haloop, Twister

4.4 Privacy and Security challenges

There are issues in inspecting, get to control, confirmation, approval and security when performing mapper and reducer occupations.

Solution

Utilize trusted outsider observing and security examination and in addition protection approach implementation with security to avert data spillage.

V. Conclusion

We have entered a time of Big Data. The paper centers around hadoop and mapreduce preparing issues. These specialized difficulties must be tended to for effective and quick handling of Big Data. The difficulties incorporate the conspicuous issues of scale, as well as heterogeneity, absence of structure, blunder taking care of, protection, auspiciousness, provenance, and representation, at all phases of the investigation pipeline from information obtaining to come about translation. These specialized difficulties are regular over an expansive assortment of utilization spaces, and in this way not financially savvy to address with regards to one area alone. This paper demonstrates the pragmatic issue and capacity issue in mapreduce

VI References

1. Dean J, Ghemawat S: “MapReduce: simplified data processing on large clusters” *Commun ACM* 2008, 51(1):107-113.
2. Lee K-H, Lee Y-J, Choi H, Chung YD, Moon B: “Parallel data processing with MapReduce: a survey” *ACM SIGMOD Record* 2012, 40(4):11-20.
3. Bu Y, Howe B, Balazinska M, Ernst MD: “HaLoop: efficient iterative data processing on large clusters” *Proceedings of the VLDB Endowment* 2010, 3(1–2):285-296.
4. Ekanayake J, Pallickara S, Fox G: “Mapreduce for data intensive scientific analyses” *Proceesings of IEEE Fourth International Conference on eScience* 2008, 277-284.
5. Palit I, Reddy CK: “Scalable and parallel boosting with MapReduce” *EEE Trans Knowl Data Eng* 2012, 24(10):1904-1916.
6. Ekanayake J, Li H, Zhang B, Gunarathne T, Bae S-H, Qiu J, Fox G (2010) “Twister: a runtime for iterative mapreduce” In: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. ACM, pp 810–818.
7. Zhang Y, Gao Q, Gao L, Wang C: “ Imapreduce: a distributed computing framework for iterative computation” *J Grid Comput* 2012, 10(1):47-68.
8. S. Ghemawat et al. “The google file system” ACM SIGOPS Operating Systems Review,37(5):29–43,2003.
9. 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).
10. Bernice Purcell “The emergence of “big data” technology and analytics” Journal of Technology Research 2013.