A Review of MapReduce for Big Data Analytics in Cloud Computing Environment

Mr. Sunil Gupta
Associate Professor
CSE Department, IET ALWAR (RAJSTHAN)
sunilgupta8764@gmail.com

ABSTRACT

Big data stream is a very popular and increasing the day by day for the various real world applications such as social media, marketing, transportation, banking, fraud detection and many more area. It is compromise high velocity and high volume data also a verified data. In this paper we review the various tools and techniques for the big data stream which is used in various real life application, Map reduce is a very play a vital role for big data analytics.

Keywords:- Big Data, MapReduce, Cloud Computing, Hadoop, Data Analytics.

INTRODUCTION

Big data processing underpins many efforts of modern computing usage. Given the size of data and the amount of resources that go into processing them, there is a decided motivation to optimize the processing as much as possible [10]. To serve the increasing demands of various data analytics applications, major cloud providers like Amazon, Microsoft and Google each deploy from tens to hundreds of geo-distributed datacenters; AT&T has thousands of datacenters at their PoP locations [7].

MapReduce jobs are usually executed on clusters of commodity PCs, which require large investment in hard-ware and management. Since a cluster must be provisioned for peak usage to avoid overload, it is underutilized on average. Thus, cloud becomes a promising plat-form for MapReduce jobs because of its flexibility and pay-as-you-go business model.

The rest of this paper is organized as follows in section II we discuss about the rich literature survey for the big data analysis for the various number of application and real life application services which used in a cloud computing environment, we also discuss here hadoop and map reduce techniques for the cloud computing services for the various types of cloud computing models.. here we also describe the about comparative study in a table no. 1. And finally in section III we define the overall summary with conclusion of this study and review paper.

II COMPARATIVE STUDY

Sr. No.	Ref No.	Author Name	Publication	Year	Title	Objective of paper
1.	[1]	Lin Gu, Deze Zeng, Song Guo, Yong	IEEE	2016	A General Communication Cost Optimization	In this paper, they first propose a general modeling framework that describes all representative inter-task relationship

		Xiang and Jiankun Hu			Framework for Big Data Stream Processing in Geo-Distributed Data Centers	semantics in BDSP. Based on our novel framework, they then formulate the communication cost minimization problem for BDSP into a mixed-integer linear programming (MILP) problem and prove it to be NP-hard. We then propose a computation-efficient solution based on MILP.
2.	[4]	Peng Li, Member, Song Guo, Shui Yu and Weihua Zhuang	Journal	2010	Cross-cloud MapReduce for Big Data	In this article they discuss about geo-distributed cloud architecture environment that provides MapReduce services and tasks based on the big data analysis collected from various sources all over the world.
3.	[7]	Chien-Chun Hung, Leana Golubchik and Minlan Yu	ACM	2015	Scheduling Jobs Across Geo-distributed Datacenters	In this paper, they propose novel job scheduling algorithms that co-ordinate job scheduling across datacenters with low overhead, while achieving near-optimal performance.
4.	[6]	Qifan Pu1, Ganesh Ananthanarayanan, Peter Bodik, Srikanth Kandula, Aditya Akella, Paramvir Bahl and Ion Stoica	ACM	2015	Low Latency Geo-distributed Data Analytics	They develop Iridium, a system that focuses on minimizing response times of geo-distributed analytics queries. Their techniques focus on data transfers in these queries that happen across the WAN.
5.	[2]	Xinyu Wang, Zhou Zhao and Wilfred Ng	IEEE Transaction for BIG DATA	2016	USTF: A Unified System of Team Formation	In this paper, they first compare and contrast all the state-of-the-art team formation algorithms. Next, we propose a benchmark that enables fair comparison amongst these algorithms. They then implement these algorithms using a common platform called the Unified System for Team Formation (USTF) and evaluate their performance using several real datasets. They also present a case study that shows the performance of different algorithms in a range of real world cases.
6.	[5]	Hong Xu and Baochun Li	IEEE	2013.	Joint Request Mapping and Response Routing for Geo-distributed Cloud Services	In this paper they discussed a parallel implementation of the algorithm that is well suited in a cloud environment with abundant server resources. Trace-driven simulations are conducted to evaluate the algorithm's performance. As future work, they plan to

						more thoroughly study its impact on existing wide-area traffic engineering schemes.
7.	[10]	William Culhane, Kirill Kogan, Chamikara Jayalath and Patrick Eugster	IEEE	2015	Optimal Communication Structures for Big Data Aggregation	They consider two cases of the problem aggregation of (1) single blocks of data, and of (2) streaming input. For each case we determine which metric of "fast" completion is the most relevant and mathematically model resulting systems based on aggregation trees to optimize that metric.
8.	[3]	Xiaomeng Yi, Fangming Liu, Jiangchuan Liu and Hai Jin	IEEE	2014	Building a Network Highway for Big Data: Architecture and Challenges	In this article, they take a close look at the unique challenges in building such a network infrastructure for big data. Their study covers each and every segment in this network highway: the access networks that connect data sources, the Internet backbone that bridges them to remote datacenters, as well as the dedicated network among datacenters and within a datacenter.
9.	[11]	Bjørn Magnus Mathisen, Leendert W. M. Wienhofen and Dumitru Roman	Journal	2016	Empirical Big Data Research: A Systematic Literature Mapping	In this paper they present the current status of empirical research in Big Data. Method: We employed a systematic mapping method with which they mapped the collected research according to the labels Variety, Volume and Velocity.
10.	[16]	Karthik Kambatla, Giorgos Kollias, Vipin Kumar and Ananth Grama	Elsevier	2014	Trends in big data analytics	In this article, they provide an overview of the state-of-the-art and focus on emerging trends to highlight the hardware, software, and application landscape of big-data analytics. In the future, as the data sizes continue to grow and the domains of these applications diverge, these systems will need to adapt to leverage application-specific optimizations.
11.	[9]	Bikas Saha, Hitesh Shah, Siddharth Seth,Gopal Vijayaraghavan, Arun Murthy and Carlo Curino	ACM	2015	Apache Tez: A Unifying Framework for Modeling and Building Data Processing Applications	In this paper, they introduce Apache Tez, an open-source frame-work designed to build data-flow driven processing runtimes. Tez provides a scaffolding and library components that can be used to quickly build scalable and efficient data-flow centric engines. Central to our design is fostering component re-use, without hindering customizability of the performance-critical data plane.

12.	[19]	Anupam Das, Cristian Lumezanu, Yueping Zhang, Vishal Singh, Guofei Jiang and Curtis Yu	Journal	2013	Transparent and Flexible Network Management for Big Data Processing in the Cloud	In this paper they introduce FlowComb, a network management frame-work that helps Big Data processing applications, such as Hadoop, achieve high utilization and low data processing times. FlowComb predicts application network transfers, sometimes before they start, by using software agents in-stalled on application servers and while remaining completely transparent to the application.
13.	[17]	Mohammad Hammoud and Majd F. Sakr	ACM	2011	Locality-Aware Reduce Task Scheduling for MapReduce	In this article they presents about Locality-Aware Reduce Task Scheduler, it is a practical techniques which is used so for improving the MapReduce performance. The LARTS techniques attempts to collocate reduce tasks minimum with the maximum required data from various sources computed after recognizing input data network locations and their sizes. LARTS adopts a cooperative paradigm and reduce the chances of to occurrences in a minimum way.

Table 1: Shows the comparative study of Big Data analytics for various application in cloud computing environment and services.

III CONCLUSION

Today, Hadoop is a booming ecosystem for large-scale data processing, blessed with an ever growing set of application frame-works, providing diverse abstractions to process data. In this paper we present a comparative study of various big data analytics for various number of applications using Haddop and other tools for cloud computing services.

REFERENCES:-

[1] Lin Gu, Deze Zeng, Song Guo, Yong Xiang and Jiankun Hu “A General Communication Cost Optimization Framework for Big Data Stream Processing in Geo-Distributed Data Centers”, IEEE, 2016, Pp 19-29.

[2] Xinyu Wang, Zhou Zhao and Wilfred Ng “USTF: A Unified System of Team Formation”, IEEE, 2016, Pp 70-84.

[3] Xiaomeng Yi, Fangming Liu, Jiangchuan Liu and Hai Jin “Building a Network Highway for Big Data: Architecture and Challenges”, IEEE, 2014, Pp 1-25.

[4] Peng Li, Member, Song Guo, Shui Yu and Weihua Zhuang “Cross-cloud MapReduce for Big Data”, JOURNAL OF LATEX CLASS FILES, 2010, Pp 1-14.

[5] Hong Xu and Baochun Li “Joint Request Mapping and Response Routing for Geo-distributed Cloud Services”, IEEE, 2013, Pp 1-9.

[6] Qifan Pu1, Ganesh Ananthanarayanan, Peter Bodik, Srikanth Kandula, Aditya Akella, Paramvir Bahl and Ion Stoica “Low Latency Geo-distributed Data Analytics”, ACM, 2015, Pp 421-434.

[7]Chien-Chun Hung, Leana Golubchik and Minlan Yu “Scheduling Jobs Across Geo-distributed Datacenters”, ACM, 2015, Pp 1-14.

[8] Benjamin Heintz, Abhishek Chandra and Ramesh K. Sitaraman “Optimizing Grouped Aggregation in Geo-Distributed Streaming Analytics”, ACM, 2015, Pp 1-12.

[9] Bikas Saha, Hitesh Shah, Siddharth Seth,Gopal Vijayaraghavan, Arun Murthy and Carlo Curino “Apache Tez: A Unifying Framework for Modeling and

Building Data Processing Applications”, ACM, 2015, Pp 1357-1359.

[10] William Culhane, Kirill Kogan, Chamikara Jayalath and Patrick Eugster “Optimal Communication Structures for Big Data Aggregation”, IEEE, 2015, Pp 1-9.

[11] Bjørn Magnus Mathisen, Leendert W. M. Wienhofen and Dumitru Roman “Empirical Big Data Research: A Systematic Literature Mapping”, arXiv, 2016, Pp 1-19.

[12] Priya Gawande and Nuzhaft Shaikh “Improving Network Traffic in MapReduce for Big Data Applications”, ICEEOT, 2016, Pp 2979-2983.

[13] Seref SAGIROGLU and Duygu SINANC “Big Data: A Review”, IEEE, 2013, Pp 42-47.

[14] Alvaro A. Cárdenas, Pratyusa K. Manadhata and Sreeranga P. Rajan “Big Data Analytics for Security”, IEEE, 2013, Pp 74-76.

[15] Paolo Costa, Austin Donnelly, Antony Rowstron and Greg O’Shea “Camdoop: Exploiting In-network Aggregation for Big Data Applications”, USENIX, 2012, Pp 1-14.

[16] Karthik Kambatla, Giorgos Kollias, Vipin Kumar and Ananth Grama “Trends in big data analytics”, Elsevier, 2014, Pp 2561-2573.

[17]Mohammad Hammoud and Majd F. Sakr “Locality-Aware Reduce Task Scheduling for MapReduce”, ACM, 2011, Pp 1-7.

[18] Jun Liu, Feng Liu and Nirwan Ansari “Monitoring and Analyzing Big Traffic Data of a Large-Scale Cellular Network with Hadoop”, IEEE, 2014, Pp 32-39.

[19] Anupam Das, Cristian Lumezanu, Yueping Zhang, Vishal Singh, Guofei Jiang and Curtis Yu “Transparent and Flexible Network Management for Big Data Processing in the Cloud”, USENIX, 2013, Pp 1-6.