IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Special Issue-4, June 2016 E-ISSN: 2347-2693

Cloud Based Big Data Processing Approaches

Rashmi. G[1*], Dr. S. Sathish Kumar [2]

[1*]Research Scholar, Visvesvaraya Technological University, Belagavi, India
[2]Associate Professor, CSE, RNS Institute of Technology, Bengaluru, India

Available online at: www.ijcseonline.org

Received: May/29/2016 Revised: Jun/07/2016 Accepted: Jun/17/2016 Published: Jun/30/2016

***Abstract*- Nowadays Big data processing in cloud has become a challenging task. This paper describes the basics of cloud computing and current big-data processing approaches in cloud such as batch-based, stream-based, graph-based, DAG-based, interactive-based, visual-based and summarizes the strengths and weaknesses of these approaches in order to help the future big data research scholars to select the appropriate processing technique.**

Keywords—Big data, Cloud, Data processing in cloud, Big data processing, Big Data processing approaches

I. INTRODUCTION

Day by day, data is being produced with increasing speed by various sources. Mobile devices, global positioning system, computer logs, social media, sensors, and monitoring systems are all generating huge volume of data that is beyond the processing capability of traditional systems called Big data. Managing and mining such data to extract useful information is a significant challenge [1]. Big data is huge and complex structured or unstructured data that is difficult to manage using traditional technologies such as database management system (DBMS). Call logs, financial transaction logs, social media analytics, intelligent transport services, location-based services, earth observation, medical imaging, and high-energy physics are all sources of big data.

Cloud computing means that instead of all the computer hardware and software you're using sitting on your desktop, or somewhere inside your company's network, it is provided for you as a service by another company and accessed over the Internet. Exactly where the hardware and software is located and how it all works doesn't matter to you.

Most of us use cloud computing all day long without realizing it. When you sit at your personal computer and type a query into Google, the computer on your desk isn't playing much part in finding the answers you need: it's no more than a messenger. The words you type are swiftly shuttled over the Net to one of Google's hundreds of thousands of clustered PCs, which dig out your results and send them promptly back to you. When you do a Google search, the real work in finding your answers might be done by a computer sitting in California, Dublin, Tokyo, or Beijing; you don't know and most likely you don't care.

The same applies to Web-based email. Once upon a time, email was something you could only send and receive using a program running on your Personal Computer sometimes called a mail client. But then Web-based services such as Hotmail came along and carried email off into the cloud. Now, we are all used to the idea that emails can be stored and processed through a server in some remote part of the world, easily accessible from a Web browser, wherever we happen to be. Pushing email off into the cloud makes it supremely convenient for busy people, constantly on the move.

Cloud computing uses computing infrastructures such as data centers and computing farms and software frameworks such as Hadoop, MapReduce, HDFS, and storage systems to optimize and manage big data[1]. Because of the importance and usability of cloud computing in daily life, the number of cloud resource providers has increased. Cloud resource providers offer a variety of services, including computation and storage, to customers at low cost and on a pay-per-use basis.

Cloud computing is one of the best potential solutions to dealing with big-data. Many big data generators have been adapted to cloud computing. According to a survey by GigaSpaces[3], only 20% of IT professionals said their company had no plans to move their big data to the cloud, which indicates that most companies dealing with big data have turned to the cloud[2]. Several applications such as cloud-agent-based urban transportation systems, MapReduce for traffic flow forecasting, and cloud-enabled intensive FCD computation framework [4,5], have been significant in bringing forward the cloud computing paradigm.

II. CATEGORIES OF BIG DATA PROCESSING APPROACHES

Figure. 1 shows various Big data processing approaches in cloud computing environment.

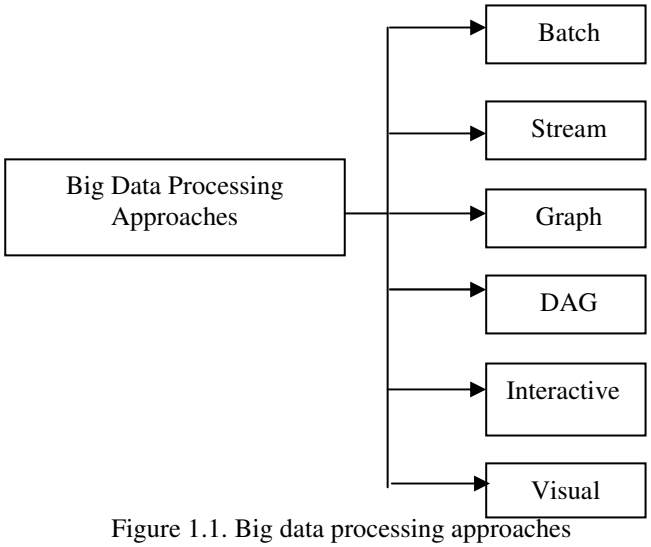

Figure 1.1. Big data processing approaches

A. Batch based approach

This approach follows Map Reduce[21,22] based parallel computing paradigm of cloud computing which is very efficient in processing larger volume of data in batches. Batch jobs are executed without manual intervention. This approach is time consuming as jobs are processed in batches that are sequentially. Depending on the size of the data being processed and the computational power of the system, output can be delayed significantly.

Several tools based on batch processing and run on top of Hadoop are Mahout[9], Pentaho[10], Skytree[11], Karmasphere [12], Datameer[13], Cloudera[14], Apache Hive, and Google Tenzing.

B. Stream based approach

This approach is used to process continuous flow of data streams.

Stream processing tools available are Storm[15], S4[16], SQLStream[17], Splunk, Kafka, SAP Hana, Infochimps, and BigInsights.

C. Graph based approach

This approach works according to the Bulk Synchronous Parallel programming model to run parallel algorithms for processing large-scale graph data.

Hama[6], Pregel[7], and Giraph[18] are common useful graph processing techniques for big-data analytics.

D. Directed Acyclic Graph (DAG) based approach

This approach is used for complex computations which require multiple pair of MapReduce[21] steps of MapReduce model. One pair of map and reduce does one level of aggregation over the data. So a DAG execution model is essentially a generalization of the MapReduce model.

A Directed Acyclic Graph refers to a model for scheduling work in which jobs are represented as vertices in a graph, where the order of execution is specified by the directionality of the edges in the graph. The "acyclic" part just means that there are no loops (cycles) in the graph. In a system which schedules jobs using a DAG, independent nodes in the graph can run in parallel, rather than sequentially. This approach makes it easier for programmers to build more complex multi-step computations, and avoids the scheduling overhead imposed by traditional MapReduce.

Dryad[8] is a scalable parallel and distributed programming model based on dataflow graph processing that can be executed in a distributed way on a cluster of multiprocessor or multicore computing nodes. Dryad computes a job in a directed-graph computation manner, wherein each vertex denotes a computational vertex, and an edge denotes a communication channel. This model can generate and dynamically update the job graph and schedule the processes on the resources. Microsoft Server 2005 Integration Services (SSIS) and Dryad-LINQ are built on Dryad.

E. Interactive based approach

This approach allows users to interact with big data applications for big data processing.
One of the popular tools available is Tableau[19] .

F. Visual based approach

This approach is specially designed for visual big-data analysis.

Talend Open Studio[20] has user's graphical platform that is completely open source software developed in Apache Hadoop. In this platform, programmer can easily build a program for Big Data problem without writing its Java code. Specifically, Talend Open Studio provides facilities of dragging and dropping icons for building up user's task in Big Data problem.

III. SUMMARY OF BIG DATA PROCESSING APPROACHES

Current Big Data Processing Approaches	Strengths	Weaknesses
Batch based	• High volume data is processed in batches. • Batch jobs are configured to run without manual intervention.	consumes more time

Stream based	• It involves continual input data stream . • Data must be processed within small time period or near real time.	It emphasizes on the Velocity of the data
Graph based	• Ease of programming. • Efficiency on graph problems. • Provides distributed and fault-tolerant system • Concurrent computation • Communication • Well-suited to enable automatic memory management for distributed-memory computing	Difficult to process Graphs data
DAG based	• Improves latency. Simpler to implement a fault tolerance. • In the event of a job failure, you can easily backtrack through the graph and re-execute any failed jobs, even at intermediate stages of a computation	Difficult to process DAG data
Interactive based	• Allows users to interact with big data applications for big data processing.	Every time interaction is required between users and applications
Visual based	• Visual big-data analysis. • In Hadoop, programmer can easily build a program for Big Data problem without writing its Java code. • Provides facilities of dragging and dropping icons for building up user's task in Big Data problem	Works only for visual data

Table 3.1. Summary of Big data processing approaches

IV. CONCLUSION

In this paper, we have described the basics of cloud computing and various big-data processing approaches in cloud. We have classified big-data approaches as batch-based, stream-based, graph-based, DAG based, interactive-based and visual-based which help the big-data research scholars to select an appropriate processing technique for future enhancement.

REFERENCES

[1] B. D. Martino, R. Aversa, G. Cretella, et al., "Big data (lost) in the cloud," *International Journal of Big Data Intelligence*, vol. 1, no. 1, pp. 3–17, 2014. doi:10.1504/IJBDI.2014.063840.

[2] A. A. Chandio, F. Zhang, and T.D. Memon, "Study on LBS for characterization and analysis of big data benchmarks," *Mehran University Research Journal of Engineering and Technology*, vol. 33, no. 4, pp. 432–440, Oct. 2014.

[3] GigaSpaces. (2013). *Big Data Survey* [Online]. Available: http://www.gigaspaces.com

[4] Q. Li, T. Zhang, and Y. Yu, "Using cloud computing to process intensive floating car data for urban traffic surveillance," *International Journal of Geographical Information Science*, vol. 25, no. 8, pp. 1303–1322, Aug. 2011. doi: 10.1080/13658816.2011.577746.

[5] Z. Li, C. Chen, and K. Wang, "Cloud computing for agent-based urban transportation systems," *IEEE Intelligent Systems*, vol. 26, no. 1, pp. 73–79, 2011. doi: 10.1109/MIS.2011.10.

[6] S. Seo, E. J. Yoon, J. Kim, *et al.*, "Hama: an efficient matrix computation with the mapreduce framework," in *IEEE Second International Conference on Cloud Computing Technology and Science,* Indianapolis, USA, 2010, pp. 721–726. doi:10.1109/CloudCom.2010.17.

[7] G. Malewicz, M. H. Austern, A. J. C Bik, et al., "Pregel: a system for large-scale graph processing," in *ACM SIGMOD International Conference on Management of Data*, Indianapolis, USA, 2010, pp. 135–146. doi: 10.1145/1807167.1807184.

[8] M. Isard, M. Budiu, Y. Yu, *et al.*, "Dryad: distributed data-parallel programs from sequential building blocks," in *EuroSys'07*, Lisboa, Portugal, 2007.

[9] Apache. (2013). *Apache Mahout* [Online]. Available: http://mahout.apache.org/

[10] Pentaho. (2013). *Pentaho Big Data Analytics* [Online]. Available: http://www.pentaho.com/product/big-data-analytics.

[11] Skytree. (2013). *Skytree The Machine Learning Company* [Online]. Available:http://www.skytree.net/

[12] Karmasphere. (2012). *FICO Big Data Analyzer* [Online] Available:http://www.karmasphere.com/

[13]Datameer. (2013). *Datameer* [Online]. Available: http://www.datameer.com/

[14]Cloudera. (2013). *Cloudera* [Online]. Available: http://www.cloudera.com/

[15] Apache. (2012). *Apache Storm Project* [Online]. Available: http://www.stormproject.net

[16]L. Neumeyer, B. Robbins, A. Nair, *et al.*, "S4: distributed stream computing platform," in *IEEE International Conference on Data Mining Workshops,* Sydney, Australia, 2010, pp. 170–177. doi: 10.1109/ICDMW.2010.172.

[17] SQLstrean. (2012). *SQLstream s-Server* [Online]. Available:http://www.sqlstream.com/blaze/s-server/

[18] Apache. (2011). *Apache Giraph* [Online]. Available: http://giraph.apache.org/

[19] Tableau. (2013). *Tableau* [Online]. Available: http://www.tableausoftware.com/

[20] Talend. (2009). *Talend Open Studio* [Online]. Available: https://www.talend.com/

[21] Aftab A. Chandio, Nikos Tziritas, Cheng-Zhong Xu, "Big-Data Processing Techniques and Their Challenges in Transport Domain", DOI: 10.3969/j.issn.1673-5188.2015.01.007

[22]Poornima Sharma,Varun Garg, Prof. Randeep Kaur, Prof. Satendra Sonare, "Big Data in Cloud Environment", International Journal of Computer Sciences and Engineering, Volume-01, Issue-03, Page No (15-17), Nov -2013, E-ISSN:2347-2693.

