IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Survey on Data Lake System for Handling Exponential Growth of Multi-Structure Data

Abhyuday Patil

Computer Engineering, B.V.D.U.C.O.E., Pune, India

Abstract: Data Lake is project in which aim to store the big data in system. Big Data, it needs the perfect technology in place to acquire the data, store it, combine it and enrich huge volumes of unstructured data in raw format. Big data analysis can said to be as the analysis of a special sort of data which comprise of structured, semi structured and unstructured data. Data Lake is one of the empowering data capture and processing capability for Big Data analysis. Data Lake is an easily accessible, flexible enough and scalable large data repository.

Keywords: Big Data, Big Data analytics, Data Warehouse, Data Lake.

I. INTRODUCTION

Big Data is the data that exceeds the processing capability of the conventional database systems. The burgeoning field of data science is fusing with the new business requirement to store and analyse big data. Big data is most often produced due to the sensors, log files, audio messages, video messages, social media websites, network packets and web. Big data is analysis of a special sort of data which comprise of structured, semi structured and unstructured data.
Big data analysis is a continuous, and not an isolated set of activities. Thus you need a perfect set of solutions for big data analysis, from acquiring the data from the source and finally discovering new insights to make decisions and for ongoing analysis. Data Lake is a place to store unlimited amounts of data of any type, schema and format that is relatively inexpensive and massively scalable. For Data Lake we can use Hadoop technology. Hadoop implements a scalable and parallel processing framework that will process exceedingly large amounts of data in a smooth way, and makes it almost impossible to lose any kind of data, as it is replicated across the cluster. Data is coming in at such an overwhelming rate that organizations with traditional approaches cannot hope to capture the data and process the data efficiently.

II. LITERATURE SURVEY

- Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts Data being generated in the academic domain and educational perspective are increasing in an exponential rate. There exists many data some are relevant and some are irrelevant. The knowledge extraction from these data will yield wanted and unwanted information. The challenging task is to extract wanted and relevant information and knowledge from these huge set of data. Over the time of research in the field of data mining, rapid growth is seen in Rough set theory and its applications. This paper discusses the basic approach and concepts of the rough set theory in the field of academic domain for the performance prediction of students in course works

- The Anatomy of a Small-Scale Document Search Engine Tool: Incorporating a new Ranking Algorithm
A search engine is an information retrieval system to help find out the information contained in documents stored on a computer system. The results provided by this kind of a system are usually in form of a list. Search engines basically work on the concept called 'Text-Mining'. Text mining is a variation on a field called data mining and refers to the process of deriving high-quality information from unstructured text. In this paper we are going to depict an intelligent agent based search engine tool which takes the input from user in form of keyword and based on the keyword, find out the matching documents and show it to user (in the form of links). This tool uses a new 'Ranking Algorithm' to rank the documents.

- Performance Impact Analysis of Application Implemented on Active Storage Framework
There is a tremendous increase in usage of digital devices and services provided. As the number of users increased so as the proliferation of data and need of information and knowledge. There has been increasing trend in need for storing and processing these growing data. With the development trend in technology, the storage capacity and processing power has kept well improving. But since the computing power is faster than the spinning device, there is a delay of providing data to processing component from the traditional spinning disks. This delay evolves with a Processing – I/O performance gap. This CPU-I/O gap worsens for application which are data intensive. Active storage framework are proposed to minimize this Processing – I/O performance gap. This research paper provides a test bed which analyses the performance impact on the application

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

III. EXISTING SYSTEM

Data storage is a large storage for data collected from a wide range of data sources. They store currently and past data and are used for creating analytical reports throughout the enterprise or organizations as per requirement.

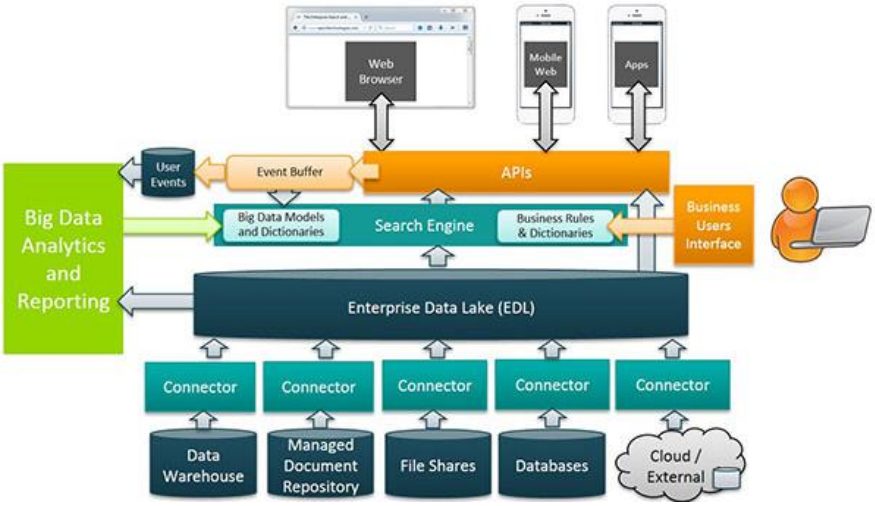
There are some drawback in existing System:-

- The conventional Data storage systems are not designed to integrate measure and handle this exponential growth of multi-structured data. But with the emergence of Big Data, there is a need to combine together data from various sources and to generate a powerful meaning of it.
- With traditional approaches, optimization for analytics is too much time consuming and incurs huge cost. These methods fails when there are new requirements.
- It is difficult to identify what type data is available and to integrate the data to answer any questions. Manual recreation of data is error-prone and consumes lot of time which is a big problem.

IV. PROPOSED SYSTEM

The Data Lake systems are designed to integrate measure and handle this exponential growth of multi-structured data. This method Captures and stores raw data from various source at low cost. It can store various types of data in the same directory. It performs various modifications, and on the data.

Architecture diagram of Data Lake



V. ADVANTAGE

- Scale as much as you can: HDFS based storage in Hadoop gives the flexibility to support large clusters while maintain efficient performance. The Hadoop for underlying storage makes the Data Lake more scalable than Data warehouse by any order of magnitude.
- Plug in disparate data sources: unlike Data Warehouse that can ingest only structured data, Hadoop supported Data Lake has an ability to ingest multi-structured and massive data sets from variant sources. This is one huge benefit and enables quick integration of data sets.
- Store in native format: In Data warehouse the data is pre prepared into some format during ingestion phase. But Data Lake skips this phase, and provides iterative and immediate access to raw data. This provides the analytical insights.
- Do not worry about schema: Traditional Data warehouses support schema for storing the data. But the Data Lake uses Hadoop's simplicity in storing data based on schema less write and schema based read modes.
- Administrative resources: The Data Lake works better than Data warehouse in reducing the resources necessary for pulling, transforming, aggregating and analyzing the data in an efficient way.

VI. CONCLUSION

Big Data is the data that exceeds the processing capability of the conventional database systems. It can be said that the data is too big, and moves too furious that it do not fit the structure of the relational database architecture. Data Lake is a huge repository which is not just limited to one type or source of data but all the data belonging to an organization in variant schemas and formats. Storing all this data at one place will increase availability and reusability of data among different departments and business units. Large and growing volume of data is stored in the Data Lake. It is used as a multi-tenant service and stores sensitive data and it mainly works on schema on read.

REFERENCES

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

1. Namdeo, Jyoti and NaveenkumarJayakumar. "Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts." International Journal 2.2 (2014).
2. Jayakumar, D.T. and Naveen Kumar, R., 2012. SDjoshi, ". International Journal of Advanced Research in Computer Science and Software Engineering," Int. J, 2(9), pp.62-70.
3. Raval, K.S., Suryawanshi, R.S., Naveen Kumar, J. and Thakore, D.M., 2011. The Anatomy of a Small-Scale Document Search Engine Tool: Incorporating a new Ranking Algorithm. International Journal of Engineering Science and Technology, 3(7).
4. Naveenkumar, J., Makwana, R., Joshi, S.D. and Thakore, D.M., 2015. Performance Impact Analysis of Application Implemented on Active Storage Framework. International Journal, 5(2).
5. Naveenkumar, J., Keyword Extraction through Applying Rules of Association and Threshold Values. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), ISSN, pp.2278-1021.
6. Jayakumar, M.N., Zaeimfar, M.F., Joshi, M.M. and Joshi, S.D., 2014. INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET). Journal Impact Factor, 5(1), pp.46-51.
7. Kakamanshadi, G., Naveenkumar, J. and Patil, S.H., 2011. A Method to Find Shortest Reliable Path by Hardware Testing and Software Implementation. International Journal of Engineering Science and Technology (IJEST), ISSN, pp.0975-5462.
8. Archana, R.C., Naveenkumar, J. and Patil, S.H., 2011. Iris Image Pre-Processing and Minutiae Points Extraction. International Journal of Computer Science and Information Security, 9(6), p.171.
9. Salunkhe, R. and Jaykumar, N., 2016, June. Query Bound Application Offloading: Approach towards Increase Performance of Big Data Computing. In Journal of Emerging Technologies and Innovative Research (Vol. 3, No. 6 (June-2016)). JETIR.
10. Salunkhe, R., Kadam, A.D., Jayakumar, N. and Thakore, D., 2016, March. In search of a scalable file system state-of-the-art file systems review and map view of new Scalable File system. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on (pp. 364-371). IEEE.
11. Naveenkumar, J., Makwana, R., Joshi, S.D. and Thakore, D.M., 2015. Offloading Compression and Decompression Logic Closer to Video Files Using Remote Procedure Call. Journal Impact Factor, 6(3), pp.37-45.
12. Jayakumar, N., Singh, S., Patil, S.H. and Joshi, S.D., 2015. Evaluation Parameters of Infrastructure Resources Required for Integrating Parallel Computing Algorithm and Distributed File System. IJSTE-Int. J. Sci. Technol. Eng., 1(12), pp.251-254.
13. Kumar, N., Angral, S. and Sharma, R., 2014. Integrating Intrusion Detection System with Network Monitoring. International Journal of Scientific and Research Publications, 4, pp.1-4.
14. Jayakumar, N., Bhardwaj, T., Pant, K., Joshi, S.D. and Patil, S.H., 2015. A Holistic Approach for Performance Analysis of Embedded Storage Array. Int. J. Sci. Technol. Eng., 1(12), pp.247-250.
15. Jayakumar, N., 2014. Reducts and Discretization Concepts, tools for Predicting Student's Performance. Int. J. Eng. Sci. Innov. Technol, 3(2), pp.7-15.
16. Salunkhe, R., Kadam, A.D., Jayakumar, N. and Joshi, S., 2016, March. Lustre a scalable architecture file system: A research implementation on active storage array framework with Lustre file system. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on (pp. 1073-1081). IEEE.
17. Naveenkumar, J., SDJ, 2015. Evaluation of Active Storage System Realized Through Hadoop. International Journal of Computer Science and Mobile Computing, 4(12), pp.67-73.
18. Bhore, P.R., Joshi, S.D. and Jayakumar, N., 2016. A Survey on the Anomalies in System Design: A Novel Approach. International Journal of Control Theory and Applications, 9(44), pp.443-455.
19. Bhore, P.R., Joshi, S.D. and Jayakumar, N., 2017. Handling Anomalies in the System Design: A Unique Methodology and Solution. International Journal of Computer Science Trends and Technology, 5(2), pp.409-413.
20. Zaeimfar, S.N.J.F., 2014. Workload Characteristics Impacts on file System Benchmarking. Int. J. Adv, pp.39-44.

DOI 10.17148/IJARCCE.2017.61063