Assignment-6

Sai Hemanth Nirujogi
G01065588

Commercial Big Data Solutions

The Big data analytics software/solutions mainly focus on delivering efficient analytics for organizations who deal with huge amounts of data and providing them with analytics to gain deeper insights, by turning the unstructured big data to valuable information. All types of organizations from small to big experiencing a rather sizable flow of data every day. Most of the information is untouched because the traditional applications used for analysis can no longer take the volumes. But an organization cannot keep allowing to ignore the loss of valuable information in the data that is obtained from different sources. The big data cannot be analyzed just using the traditional warehousing and data management techniques. That is the reason why big organizations like Facebook, Twitter, Amazon etc, started to embrace big data solutions. These big data solutions are created to handle the massive amounts of unstructured variety data generated every single day.

A Big Data solution is a platform that delivers features and functionalities of big data application in one single solution for managing, operating, deploying and developing the organization's data obtained from different sources. The solution should be cost-effective, easily operational and scalable. Some of the top big data solutions are AWS, Microsoft Azure, IBM Big Data, Arcadia Data, Google BigQuery, Oracle Big Data Analytics, Vertica, Cassandra, MapReduce Converged Data Platform, Hortonworks Data Platform, Splunk Big Data Analytics, MongoDB, HP Big data, Dell Big Data, Cloudera Enterprise Big data etc. [1]

TPCx-BB (Big Data Benchmark)

Big Bench is used to compare the different Big data engines/solutions. It specification based benchmark with an open source implementation, which measures all the big data characteristics (4Vs). It is an extension of TCP-DS, supports multiple engines, and multiple table formats.

MapReduce

“MapReduce is the heart of Hadoop. It is the programming paradigm that allows for massive scalability across hundreds and thousands of servers in a Hadoop cluster. MapReduce refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes input data and processes it to produce key/value pairs. The reduce job takes key/value pairs and then combines the final results. The second is reduce job which is always performed after the map job. The reducer combines the map results and outputs into a single value, the overall maximum”. [2]

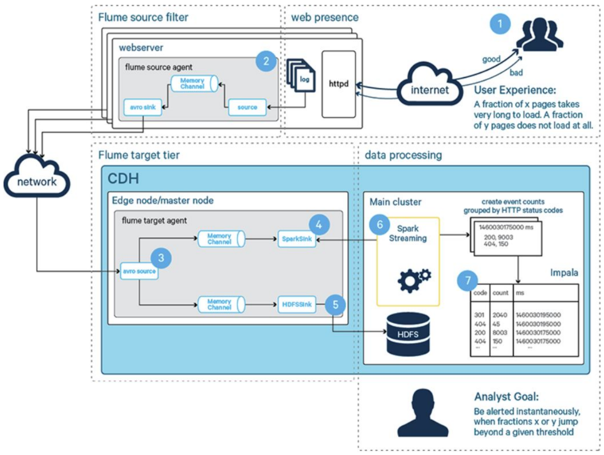
JCPenney is one of the most popular stores out there, who use MapReduce. They are successful in incorporating different departments for e-commerce, marketing, and Information Technology to process the data from different sources. They use MapReduce to diagnose performance bottlenecks. Hadoop components like Hadoop Common, HBase, HDFS and Yarn are used to handle the data in a scalable way with security aspects of data. For in-memory high-speed cluster computing, they use technologies like Apache Spark and Storm. [3]

Limitations of MapReduce

- The central storage cannot be scaled.
- High bandwidths are required to send the data to the processing nodes.
- The cost of servers that needs to be added with the growing data is very costly.
- Performance is affected by the low bandwidths, because of the slow processing speeds.
- It does not support real-time data processing.

Cloudera

Cloudera is the first company to develop and distribute Apache Hadoop-based software with the largest user database. It is an open-source Apache Hadoop distribution, CDH (Cloudera Distribution including Hadoop) targets enterprise-class deployments of that technology. This open-source software powers the data processing engines of world’s largest and popular websites like Facebook, Oracle, Yahoo, and Google.



Solution design to detect and report web-traffic anomalies in near real-time [4]

Cloudera was used to detect and report we traffic anomalies in near real-time which helps in the websites performance and availability efficient and fix the website overburdened backing data issue. Web servers generate up to 20 million user sessions every day, which may result in a lot of HTTP GET requests per second. Using Cloudera solutions, a high scalability, and availability even for large implementations of web-traffic analysis. Cloudera also supports Impala which offers real-time massively parallel processing of Big data to Hadoop. [4]

Cloudera claims to help customers in profiting from all of their data, understand the client needs, provides a vastly scalable database management system which runs on Hadoop clusters. It provides customers with the individual behavior of the different types of unstructured and structured data stored in the cluster. It also provides recommendations based on the trends in the data to the clients who store the data and using the data for their businesses. Cloudera delivers the service and support for integrating Hadoop in the database system with Cloudera Manager, saving a lot of time and manpower in figuring out the implementation process. [5]

References

1. "Top 53 Bigdata Platforms and Bigdata Analytics Software - Compare Reviews, Features, Pricing in 2018." *PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices*, 7 Aug. 2018, accessed on 3 Nov. 2018 www.predictiveanalyticstoday.com/bigdata-platforms-bigdata-analytics-software/.
2. "MapReduce." *MapReduce - Gppd-Wiki*, 3 Nov. 2018 gppd-wiki.inf.ufrgs.br/index.php/MapReduce.
3. *J.C. Penney Corporation, Inc. Careers*, 4 Nov. 2018 jobs.jobvite.com/jcp/job/olgU5fwk.
4. "How-to: Detect and Report Web-Traffic Anomalies in Near Real-Time." *Cloudera Engineering Blog*, 4 Aug. 2016, 4 Nov. 2018 blog.cloudera.com/blog/2016/06/how-to-detect-and-report-web-traffic-anomalies-in-near-real-time/.
5. Hadoop, Apache, et al. "Internet of Things Solutions | IoT." *Cloudera*, 4 Nov. 2018 www.cloudera.com/solutions/improve-products-and-services.html.