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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

A Survey of Qubole Data Service on Big Data Analytics and Cloud Computing

SaravanaKumar P [1], Tamil Selvan T [2]

Assistant Professor, Information Technology, SRM University, Sikkim, India [1,2]

Abstract: The rise of cloud computing and cloud data stores have been a precursor and facilitator to the emergence of big data. Cloud computing is the commodification of computing time and data storage by means of standardized technologies. Analytics over the huge volume of data is now possible with big data. Data keep on accumulated on every minute from multitude data sources such as social media, mobile devices, and sensors. In order to extract insights from diverse information feeds from multiple, often unrelated sources, data need to be correlated or harmonized to a common level of granularity. Loading Unstructured Data into Data warehouse getting complex. A strategy for fetching the unstructured data into Hadoop Distributed File System is discussed. Data cleansing and profiling of extracted data is important to overcome data quality concerns. Big data can be analysed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis.

Keywords: Qubole Data Service (QDS), Big Data Analytics, Cloud Computing, Business Intelligence, Hadoop.

I. INTRODUCTION

Big Data Projects focus on scaling or adopting hadoop for data processing. Mapreduce has become a de facto standard for large scale data processing. Tools like hive and pig have emerged on top of hadoop which make it feasible to process huge data sets easily. Hive for example transforms sql like queries to mapreduce jobs. It unlocks data set of all sizes for data and business analysts for reporting and Greenfield analytics projects. Data can be either transferred to or collected in a cloud data sink like Amazon's S3, e.g. to collect log files or export text formatted data. Alternatively database adapters can be utilized to access data from databases directly with Hadoop, Hive, and Pig. Qubole is a leading provider of cloud based services in this space. They provide unique database adapters that can unlock data instantly, which otherwise would be inaccessible or require significant development resource. One great example is their mongoDB adapter. It gives Hive table like access to mongoDB collections.[1] Qubole scales Hadoop jobs to extract data as quickly as possible without overpowering the mongoDB instance.

Ideally a cloud service provider offers Hadoop clusters that scale automatically with the demand of the customer. This provides maximum performance for large jobs and optimal savings when little and no processing is going on. Amazon Web Services Elastic MapReduce, for example, allows scaling of Hadoop clusters. However, the scaling is not automatically with the demand and requires user actions. The scaling itself is not optimal since it does not utilize HDFS well and squanders Hadoop's strong point, data locality. This means that an Elastic MapReduce cluster wastes resources when scaling and has diminishing return with more instance. Furthermore, Amazon's Elastic MapReduce requires a customer to explicitly request a cluster every time when it is needed and remove it when it is not required anymore. There is also no user friendly interface for interaction with or exploration of the data.

This results in operational burden and excludes all but the most proficient users.

II. WHY QDS?

Qubole Data Service (QDS) is an auto-scaling cluster, improved I/O optimization, faster queries and support for hybrid pricing realize significant cost savings while accomplishing tasks faster.[2] Hadoop cluster is ready within minutes post signup, letting you focus on building sophisticated data pipelines, running queries, scheduling jobs and monetizing your big data.

A. *Handoop as a Service*

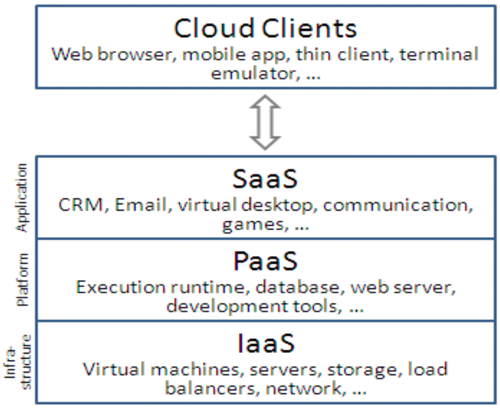

Cloud Clients
Web browser, mobile app, thin client, terminal emulator, ...
SaaS
CRM, Email, virtual desktop, communication, games, ...
Application
PaaS
Execution runtime, database, web server, development tools, ...
Platform
IaaS
Virtual machines, servers, storage, load balancers, network, ...
Infra-structure

Fig 1. Handoop Architecture

Fig. 2 shows an example of Hadoop as a Service, as offered by Qubole Data Service (QDS) is a cloud computing solution that makes medium and large-scale data processing accessible, easy, fast and inexpensive. This is achieved by eliminating the operational challenges of running Hadoop,[6] so you can focus on business growth with its unlimited scale and on-demand access to compute and storage capacity, cloud computing is the perfect match for big data processing. Qubole's Hadoop as a Service offering has several advantages over on-premise solutions.

B. *Spark as a Service*

Fig. 2 shows an example of Spark is built on top of Hadoop Distributed File System, but rather than using Hadoop MapReduce, it relies on its own parallel data processing framework which starts by placing data in

DOI 10.17148/IJARCCE.2015.4611

Resilient Distributed Datasets (RDDs), a distributed memory abstraction that performs calculations on large clusters in a fault-tolerant manner.[4] Because data is persisted in-memory, Spark can be significantly faster and more flexible than Hadoop MapReduce jobs for certain applications described below. Spark also adds flexibility to its speed by offering APIs that allow developers to write queries in Java, Python or Scala.

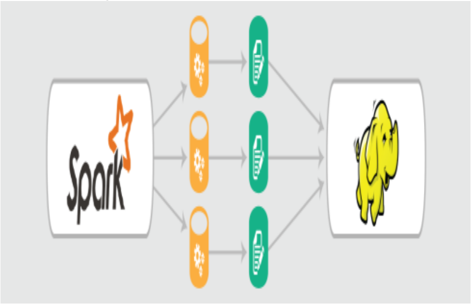


Fig. 2. Spark as a Service

C. *Presto as a Service*

Presto, developed by Facebook, is a real-time SQL engine to query a wide range of data store types including Hadoop. Hadoop as a technology is dominating Big Data projects due to its popularity, attractive ecosystem of tools and reporting formats, and cost effectiveness. SQL on Hadoop has been an important development to democratize Big Data by giving easy data access to business users, analysts, data scientists, and programmers, to previously lock away large data sets. [7]

D. *Pig as a Service*

Pig as a Service makes Hadoop's Big Data processing power scriptable, on demand with a focus on ETL while removing the operational burden and fixed cost of a Hadoop cluster.

E. *Sqoop as a Service*

Fig 3. Shows an example of Qubole Data Service has gained adoption many of our customers asked for import and export facility from their relational data sources into the Cloud (S3). Dimension data from such data sources are an important part of data analysis. Apache Sqoop can import and export data from relational databases over JDBC to HDFS. Sqoop allows importing full table, selected columns – and even allows the flexibility of specifying free-form queries to extract data and write into HDFS. Moreover it can do this using multiple parallel connections where required

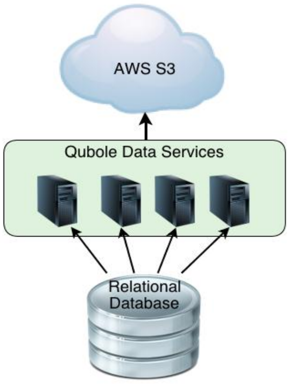


Fig 3. Qubole Data Services

F. *Hive as a Service vs. Traditional RDBMS*

Hive as a Service is different from traditional RDBMS (Relational Database Management System) because it incorporates transparently – independent from user know-how – the benefits of Hadoop and cloud computing, i.e. it auto-scales inexpensively horizontally on demand and offers a pay as you go billing. Data users can concentrate on working with the data and not on technology while enjoying the benefits of combining multiple data sources of any size and a balance between cost and performance. Businesses do not need to hire scarce talents and invest in uncertain technologically complex projects anymore to access Big Data with Hadoop.

G. *QDS Security*

Qubole is a Big Data as a Service (BDaas) offering which runs on all Leading Cloud platforms Like AWS, Microsoft Azure and Google Cloud Platform. Our flagship product, Qubole Data Service (QDS), manages Hadoop infrastructure and allows users to prepare, integrate, and analyse big data in the cloud.

III. BEST USE CASES FOR BIG DATA ANALYTICS

A. *Sentiment Analysis*

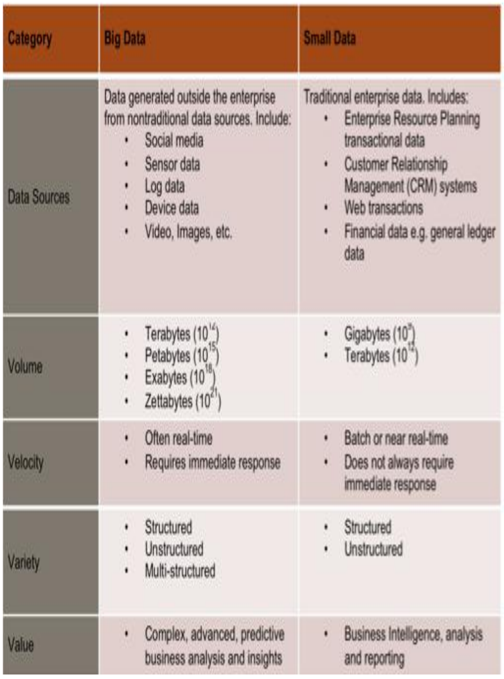
Table 1 shows the Sentiment data, generated through social channels in the form of reviews, chats, shares, likes tweets, etc., often includes comments that can be invaluable for businesses looking to improve products and services, make more informed decisions, and better promote their brands. The key to business success with sentiment data lies in the ability to mine vast stores of unstructured social data for actionable insights. That's a formidable task, requiring sophisticated tools such as Natural Language Processing (NLP) [8] to carry out comprehensive examinations of the sentiments of social media users. Fortunately, big data analytics platforms such as cloud-based Hadoop are up to the challenge.

B. *Ad-hoc Analytics on Big Data sets*

Ad hoc analytics is the discipline of analysing data on an as-needed or requested basis. Historically challenging, ad hoc analytics on big data sets versus relational databases adds a new layer of complexity due to increased data volumes, faster data velocity, greater data variety and more sophisticated data models.

C. Real-Time Analytics

Big data often mention its three vs.: volume, variety and velocity. [5] the most commonly discussed of those three is obviously volume, which isn't surprising given the name big data. However, variety and velocity are just as important in the equation. In fact, velocity is too often overlooked. Companies are so focused on gaining as much data as possible from many different sources, they forget about the analytical element. Our world moves at a remarkable speed, which means what might be relevant today could be stale tomorrow. This is why real-time analytics is such a powerful tool. The faster companies decipher information, the faster they'll find actionable results and outsmart the competition. Real-time analytics don't just help create algorithms that sell dish towels and toy cars. They're also a powerful tool in improving targeted marketing. For online retailers, that means getting the right products in front of the people looking for them, or offering the right promotions to the people most likely to buy. For gamers, that means understanding which types of individuals are playing which game, and crafting an individualized approach to reach them. For example, Microsoft could see who has bought a new game with online capabilities, then offer them an online Xbox subscription right away. Real-time analytics can have a powerful impact on marketing from targeting content to sentiment analysis with the right technology in place. Big data in the cloud puts business intelligence first and technology second. With the power of a Hadoop cluster as a fully managed service, Hadoop as a Service makes using big data for real-time marketing easy. Learn more about Hadoop in the cloud.



Category	Big Data	Small Data
Data Sources	Data generated outside the enterprise from nontraditional data sources. Include: • Social media • Sensor data • Log data • Device data • Video, Images, etc.	Traditional enterprise data. Includes: • Enterprise Resource Planning transactional data • Customer Relationship Management (CRM) systems • Web transactions • Financial data e.g. general ledger data
Volume	• Terabytes (10^{12}) • Petabytes (10^{15}) • Exabytes (10^{18}) • Zettabytes (10^{21})	• Gigabytes (10^{9}) • Terabytes (10^{12})
Velocity	• Often real-time • Requires immediate response	• Batch or near real-time • Does not always require immediate response
Variety	• Structured • Unstructured • Multi-structured	• Structured • Unstructured
Value	• Complex, advanced, predictive business analysis and insights	• Business Intelligence, analysis and reporting

Table 1 Comparison of Small and Big Data Analytics

IV. CONCLUSION

Loading Unstructured Data into Data warehouse getting complex. Strategies for fetching the unstructured data into Hadoop Distributed File System is discussed. Data cleansing and profiling of extracted data is important to overcome data quality concerns. Transform phase carried with map reduce frame work. Computation ratio, Network band width and Data locality parameters are monitored with full dump and Incremental load operations. Pig Latin is used to process data from Hadoop Distributed File System and finally load the process data into HDFS file or Data warehouse. Aggregated data from Pig is minimal Subset of Data is Loaded to Data warehouse for Business Analytics and Enterprise Reporting. Based on the Performance related parameters Second Approach Push method with Quality Addressing will be best one for Suggested ETL batch application. Big data solutions may be the most effective means of rooting out the problems. Data is being generated at an astounding rate, so it's time that businesses and analysts use this massive amount of data to detect fraud and stop it in its tracks. Closely monitoring online advertisements while aided by big data will produce impressive results, minimizing the effects of fraud and creating a more secure environment for online businesses. Big data can even be considered a type of advanced fraud detection technology.

REFERENCES

[1] Puneet Agarwal, Gautam Shroff, Pankaj Malhotra, "Approximate Incremental Big-DataHarmonization", IEEE Transaction on Bigdata Congress, Vol 5, Issue No 2, June 2013.

[2] Hongyong Yu, Deshuai Wang Proc., " Mass Log Data Processing and Mining Based on Hadoop and Cloud Computing ", Computer Science & Education (ICCSE 2012)July 14-17, 2012.Melbourne, Australia.

[3] Guozhang Wang, Marcos Vaz Salles, Benjamin Sowell, Xun Wang, Tuan Cao, Alan Demers, Johannes Gehrke, "Behavioral Simulations in Map Reduce", Walker White Procedings of VLDB endorsement Vol 3 No 1 2012.

[4] Aysan Rasooli Oskooei, "Improving Scheduling in Heterogeneous Grid and Hadoop Systems", Open Access Desertions and Theses,May 2013.

[5] Ge Song, Zide Meng, Fabrice Huet, Frederic Magoules, Lei Yu‡ and Xuelian Lin, " A Hadoop Map Reduce Performance Prediction Method", International Journal of Computer Applications & Information Technology, Vol 2, Issue No 2, Mar 2013.

[6] R.Iswarya, P.Saravana Kumar, "NFTaaS on Cloud", International Journal of Latest Trends in Engineering and Technology , Vol 4, Issue 1, May 2014.

[7] K.Suriya Prakash, P.Saravana Kumar., "Ontology Based Search Engine", International Journal of Latest Trends in Engineering and Technology, Vol 4, Issue 1, May 2014.

[8] P.Saravana Kumar, M.Parvathi, M.Kanmani, "Efficient Method for Preventing SQL Injection Attacks on Web Applications Using Encryption and Tokenization", International Journal of Latest Trends in Engineering and Technology, Vol 4, Issue 1, Nov 2014

[9] Debabrata Kar, Suvasini Panigrahi, Prevention of SQL Injection Attack Using Query Transformation and Hashing, IEEE International Advance Computing Conference (IACC),2013.

BIOGRAPHIES

Saravanakumar.P is an Assistant Professor in the Information Technology Department, SRM University, Sikkim, India. He received Master of Technology (M.Tech) CSE degree in 2015 from SRM University, Chennai, India, Master of Philosophy in Computer Science (M.Phil) degree in 2011 from PRIST University, Tanjore, India, and Master of Computer Applications (MCA) degree in 2010 from SRM University, Chennai, India. His research interests are Data Mining, Web Mining, Cloud Computing and Big Data Analytics etc.

Tamilselvan.T is an Assistant Professor in Department of Information Techology, SRM University, Sikkim, India. He received Master of Engineering (M.E) CSE degree in 2010 from Dr.Paul's Engineering College, Affiliated to Anna University Chennai, India, Master of Philosophy in computer science (M.Phil) degree in 2008 from Bharathidasan University, Tiruchirappalli ,India and Master of Computer Applications(M.C.A) degree in 2003 from P.S.G College of Technology, Coimbatore, Affiliated to Bharathiar University,Coimbatore,India. His area of interest is Data Mining, Computer Networks.