ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 3, Issue 2 (Apr. - Jun. 2015)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2015)

Challenges with Hadoop

Tushti Naryani
B.Tech (Information Technology), ITM, MDU

Abstract
New challenges to businesses have been imposed by epoch of 'Big Data'. Incoming data volumes are blowing up in intricacy, range, pace and size, while legacy tools have not kept the rate of knots. Apache Hadoop has appeared on the scene, in recent years. Businesses need to operate efficiently on data, for that they need to take data on board and also use that data for value. This will allow analysts to iterate business questions fast. One of the purposes of Hadoop is to facilitate certain forms of distributed data processing which are batch oriented which lends itself ready for data assimilation. Since it was built on fundamentals, it has limited ability to act as an analytic database. Since big data has risen, so is the analytic database platform. Few years back, DBMS such as Oracle would be leveraged by company for data warehouse. Oracle was built in age where databases rarely exceeded few gigabytes in size. Today, we have a lot of structured and unstructured data increasing rapidly every day. Along with other legacy databases, it can't perform at the current requisite scale. As far as analytic platform is concerned, it allows analysts to use their obtainable tools and skill set at scales unseen formerly.

Keywords
Apache Hadoop, Big Data, Challenges, YARN, Cloudera, Impala

I. Introduction

There is ever growing demand for storage and compute capacity. There is enormous amount of unstructured data which companies accrue to reveal the customer trends, and which needs to be shown as making sense by companies. Structured data fits fine into relational tables and arrays, new unstructured data does not. Such data includes, GPS outputs, Weblogs, Social media updates, Industrial sensor data, images and other media, Computer logs and so forth. Data quickly grows, and it's easier to notice that. The big data industry is still building infrastructure needed to integrate structured and unstructured data. To handle massive data loads, more capabilities are required like scale out compute and storage capacity. All these large- scale requirements are met by Apache Hadoop because computing nodes can be added as required. Commodity servers can be used to work as these nodes and any increase can be easily habndled by conducting massive parallel computing. As the scale out requirements increase by a factor of 10, companies must have analysis in place to meet this change. Increase in customer side transactions represents another area related to data analytics that organizations would like to exploit. Factually, the best practice infrastructure for big data at present, over and over again consists of a processing infrastructure of systems such as Hadoop to get hold of and archive the data, and an analytic platform to enable the exceedingly iterative analysis process. But because Hadoop is still relatively new, there is a great deal of confusion about its strengths and weaknesses. This paper will discuss the shortcoming of this new big data ecosystem.

II. Big Data in current market

The current market condition with big data related to different organizations is as follows:

The number of worldwide email accounts continues to grow from over 4.1 billion accounts in 2014 to over 5.2 billion accounts by the end of 2018. The total number of worldwide email users, including both business and consumer users, is also increasing from over 2.5 billion in 2014 to over 2.8 billion in 2018. [Email statistics report, Radicati group].

1. As of the fourth quarter of 2014, the micro blogging service averaged at **288 million** monthly active users. At the beginning of the 2014, Twitter hat surpassed **255** MAU per quarter. [Twitter statistics]
2. As of the third quarter of 2014, Facebook had **1.35 billion** monthly active users. In the third quarter of 2012, the number of active Facebook users had surpassed **1 billion**. Active users are those which have logged in to Facebook during the last 30 days.
3. Video views on youtube perday 4 billion
4. 323 days worth of youtube videos watched on facebook every minute
5. 2.7 Zetabytes of data exist in the digital universe today. – IBM Infographic
6. The amount of monthly active WhatsApp users worldwide as of January 2015. As of that month, the mobile messaging app announced more than 700 million monthly active users, up from over 400 million in December 2013. The service is one of the most popular mobile apps worldwide. –Source: Statista.
7. There are now more than 2 trillion (2×10^{12}) objects stored in Amazon S3 and that the service is regularly peaking at over 1.1 million requests per second.

IDC expects the Big Data technology and services market to grow at a 26.24% compound annual growth rate through 2018 to reach $41.52 billion. The big data and analytics market will reach $125 billion worldwide in 2015, according to IDC. 2013 was an important year in the evolution of Big Data technology.

III . Hadoop

The concept of Hadoop-based Big Data analytics and applications moving beyond MapReduce-style batch analytics existed before 2013, but this was the year that the structural foundation to such a transition was laid in the form of YARN.YARN, or Yet Another Resource Negotiator, has been in the works for more than three years and made its official debut in October 2013 as part of Hadoop 2.0. While the technical architecture of YARN is outside the purview of this report, the important point is that YARN enables Hadoop to function as a true multi-application framework. Developers now have the structural underpinnings to build real-time and streaming data applications, interactive SQL-style query applications, graph analytic apps, and more. YARN is critical to the future of Hadoop. It ensures that Hadoop

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

will not be relegated to backroom data science projects but will take a prominent (and potentially starring) role in the modern data architecture. YARN was an indirect growth driver for the Big Data market in 2013. As stated above, in 2013 vendors began to crystalize their visions for Big Data in the enterprise. The pending arrival of YARN, among other technology advances, enabled vendors to credibly position Hadoop at the center of their Big Data plans. Complimenting YARN were a number of moves by Hadoop and non-Hadoop vendors to better integrate the open source Big Data framework with existing data management infrastructure and legacy databases. These included:

- Cloudera's Impala, Search and its Enterprise Data Hub;
- Hortonworks' technical partnerships and reseller agreements with Teradata, Microsoft and SAP;
- HP's HAVEn reference architecture and Vertica's new FlexZone feature;
- Pivotal's HAWQ and Data Dispatch offerings for Hadoop;
- IBM's BigSQL feature and BLU Acceleration release;
- Microsoft's PolyBase data-processing framework.

While each of these releases and features is still relatively immature, they served to bolster confidence in Hadoop and related Big Data technologies as a core part of the modern data architecture. This confidence translated into significant investment by Fortune 1000 enterprises in 2013, though the fruits of these investments won't be enjoyed until 2014 and beyond. Large Internet companies like Google and LinkedIn have built their entire businesses off selling data, but now every company has the potential to gather data from its operations–and to make the data and analyses available to customers for a fee. Lucker said traditional companies will increasingly begin monetizing their data. "Companies are beginning to see [their data] as a revenue source in a way they have never seen it before," he explained.GE, for example, began placing sensors on gas turbines, jet engines and MRIs, and provides service to those products on the basis of data analysis. But even retailers such as grocery stores have begun selling their data–and the trend is only in its infancy."These companies used to provide the data as a partnering arrangement to enhance the efficiency of a process, such as their supply chain," Lucker said. "But now they're seeing that all that purchasing and customer behavior data is amazingly valuable to everyone upstream in the process. They realize they're the only ones who can provide it since it's their stores customers are walking through."

IV. Hadoop – Problem Solved

To deal with volume, variety, velocity of big data, standard relational database management systems have proved ineffective. They are effective in dealing with structured data only. Apache Hadoop is open source model which offers capabilities which are aligned precisely with types of systems that store vast amounts of unstructured data, including event, social, web, spatial, sensor data. As a consequence, Hadoop can apply in-depth analytic capability to unstructured data. Hadoop Distributed File System (HDFS) is a connected feature of Hadoop. The file system enabled large amounts of structured and unstructured data t be stored and to be quickly accessed across large server clusters. Unlike RDBS, this does not require complicated transformation and schema changes that traditional databases require. It has capability to store data in its raw form and has minimal, if any, data model restrictions. Easy scalability of Hadoop makes it ideal for analytical workloads which are not like real time transaction processing of a relational database. Hadoop runs on commodity hardware and storage, it's less expensive to employ than conventional RDBMS. In Hadoop, computers offer their own local computation and storage. Framework utilizes a process where data is written on one occasion then read several times in hefty volumes. It has an ability to rapidly process vast amounts of data in parallel and capacity to scale to enormous number of nodes offering built in redundancy, which offsets individual node failure.

V. Gaps In Hadoop

Narrow and limited functionality of database is not the only reason that Hadoop has not taken over the world yet. Since the software is under active development Hadoop Map-reduce and HDFS are quite rough in manner. Being open source, it is notorious for having variable quality. Some of it is even unusable. Reason behind this behavior being, economics of open source development provides absolutely no incentive for software quality and suitability. Open source can be considered as avenue for software engineers for trying their hands on.

Though, there exists small community of experienced software developer which is incented to do so by goodwill. The firms which sell open source solutions are the ones who base their business model on providing implementation services, so they have a few set of incentives to make the software pretty easier to set up. Quality assurance is quite difficult, rather impossible because of distributed nature of open source. Conclusively, some needs are met but with unpredictable quality and usability. Narrowing it to Hadoop, it was conceived to solve very exact problem i.e. enabling distributed MapReduce processing arbitrary sized clusters of low cost hardware. A distributed File system, HDFS and a set of gears to execute distributed MapReduce java programs, were built to enable this.

Management of cluster is an added shortcoming. Operations such as distributing software, collection logs, debugging and so forth are too rigid in a cluster. Since there are single master nodes, it limits scaling and there is one point of failure. Multiple dataset joins are used which increase slowness. They are often tricky as well. Since there are no indices, often entire dataset gets copied in the process which makes the system slow.

An added contemplation is that HDFS was rationale designed to speed the processing of various web documents, and use MapReduce framework to this processing. Being a filesystem, it means that it does not entail a schema. And while it designs for redundancy, it also does not constrain itself. Designing efficient storage was not the point because it was purpose built to implement on arbitrary sizes. From the strengths of HDFS come the weaknesses. Optimizing the data flow needs to be taken in to account by developer since there is no available optimizer. There is absolutely no notion of transaction consistency or recovery checkpoints because this was designed to be a filesystem. This means that the answer you get from a Hadoop cluster may or may not be 100% accurate, depending on the nature of the job. The answer you get from a Hadoop cluster may or may not be 100% accurate, depending on the nature of the job.

While Hadoop is a powerful framework for certain types of distributed problems, it requires specialized expertise to use it effectively. If you're more interested in the end result, you may be better served buying purpose built software, rather than doing it yourself.

To compensate for their lack of control over cluster resources, organizations usually size their clusters based on anticipated peak loads. The goal is to ensure that jobs don't overload the cluster and

lead to massively degraded performance, job failures, or worse. However, because of Hadoop's inefficient, up-front allocation of resources, this strategy is expensive and leaves capacity unused much of the time – and can still fail to prevent undesired outcomes as workloads are often unpredictable.

For monitoring of clusters users are provided with multitude of tools, still administrators are quite a many times left with incomplete view. They are not able to contemplate health of cluster because of the incomplete view. Root cause of problems can't be isolated leading to inefficient behavior because of lack of granular tools. This includes making a guess -restarting to resolve problems and asking users about jobs they submitted. When cluster size grows, businesses start relying on Hadoop, such techniques become unsustainable. Advantages of Hadoop are sufficient and interesting and that's why businesses today are experimenting with it, at least. For instance,

Businesses are putting it to use in ETL and data archival. Good Hadoop functionality and even the staff is easily available for such simple use case.

To extend Hadoop, there are many start ups up-and-coming with goal to make it more enterprise friendly.

Taking into consideration the copies which are kept, several copies of data which is already big is posing as problem because HDFS was built without the notion of efficiency. Generally, three copies of data is made, also because data locality is needed in performance preservation, six copies of data can be seen often and that data is the one which is already "big". In addition to this there is a very limited and restricted support of SQL. To set up Hadoop as a queryable data warehouse, attempts have been made by open source components, but SQL support offered is much limited. Limitation is because of the lack of basic SQL functions such as sub queries, 'group by' analytics and so forth.

Working with Hadoop requires skilled professionals. The Data Mining libraries are intriguing. These are a part of the Mahout, Hadoop project, which are erratically implemented, and requires knowledge of the algorithms in any event. Also, the skills for distributed MapReduce development are needed. Over the years, considerable improvement has been done in Hadoop schedulers. Though improvement is significant but it's still based on pre-allocating resources when job starts. The issue is that a varying mix of different hardware resources is being used by job in its lifetime. Hardware resources aren't limited in standard Hadoop. These factors lead to resource competition that at runtime should be arbitrated, resulting in work not getting completed in time.

In case of data flow and execution, one-input two-phase data flow is quite rigid and pretty hard to adapt. It does not allow for stateful multiple-step processing of records along with inefficient execution. Since there is absolutely no notion for query optimization in HDFS, it cannot pick a cost-based plan which is efficient for execution. Hence we observe that Hadoop clusters are significantly bigger than those which would be requisite for a similar database. The MapReduce framework is disgracefully difficult to leverage for more than simple transformational logic. There are open source components which attempt to simplify this, but they also use proprietary languages. Impala, a new product for use with Hadoop, was announced by Cloudera. It is being positioned as a SQL-like engine which bypasses the Hadoop MapReduce framework and allows business intelligence (BI) tools to execute queries against data in HDFS and HBase. On the face, it looks like an incremental step forward over Hive. Hive relies on MapReduce to execute queries, which degrades query performance significantly. A separate set of processes which bypass MapReduce to read directly from HDFS and HBase data is deployed by Impala. Impala will add a columnar storage engine, cost-based optimizer and other distinctly database-like features which would be covering up for a lot of shortcomings of Hadoop. MapReduce does not make sense as an engine for querying is implied by this move. Also, it's clear that in order to do analytics on big data, it's natural and efficient to use SQL to query a column-oriented MPP database with columnar based storage, a cost-based optimizer and other database-like functions. So reinventing the database is of no use on Hadoop, in particular when platforms already exist that can be an extension of Hadoop for analytics. So, lightweight analytic tools for Hadoop are best used in complement with a fully-baked analytic platform.

VI. Conclusion

For organizations that need to store, process and use data in big volumes, big data is the only solution that helps them. It is under active development hence lacks stability. Though it has various advantages, organizations implementing Hadoop still face a lot of challenges. These challenges are planned to be worked on in future releases. Also, being a filesystem adds to limitations. In spite of challenges, it serves as a very good Big Data solution.As per Markets and Markets research, the Hadoop market was worth $1.5 billion in 2012 and it is expected to grow to about $13.9 billion by 2017, at a CAGR of 54.9%. That is based on the scheme that enterprises are all the time more realizing the weight of Hadoop as it makes analyzing of conventional structured data easier in juxtaposition with unstructured data from fresh sources. A topical Transparency Market Research report predicts wide-reaching Hadoop market would grow to about $20.9 billion in 2018, escalating at a CAGR of 54.7% from 2012 to 2018. Gartner also forecasted that the big data movement will generate 4.4 million new IT jobs globally by 2015, with 1.9 million of those positions being in the U.S. It predicts that over 30 percent of analytics projects will be used to analyze both structured and unstructured data by 2015.

References

[1]. Hadoop: The Definitive Guide by Tom White
[2]. Hadoop in Practice by Alex Holmes
[3]. Pro Hadoop by Jason Venner