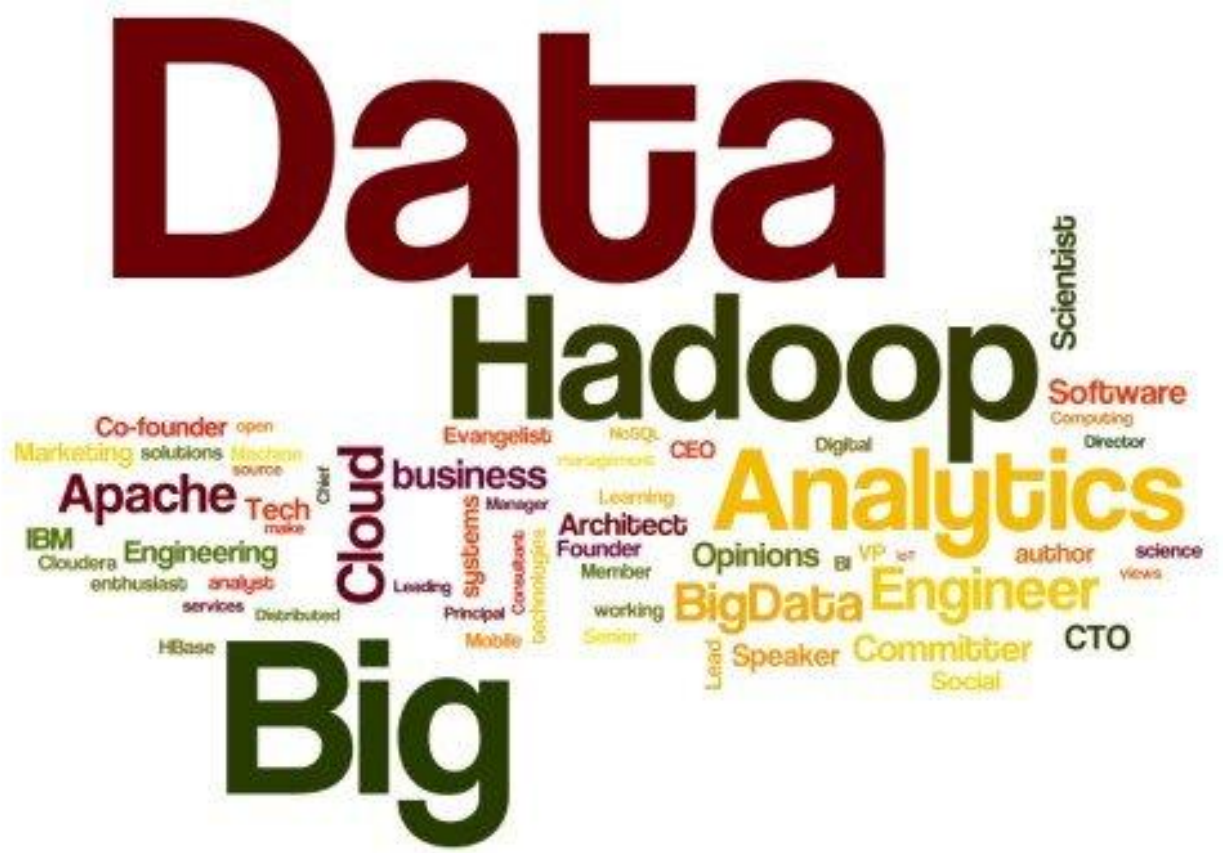

Hadoop and Big Data Integration

[Thiago Petcov]

[Big Data Integration]

[David McQuaid]

[30/04/2021]

Word Count:2574

Contents

List of Figures

Glossary of Acronyms

HDFS Hadoop Distributed File System

SQL Structure Query Language

Introduction

This consultation report was commissioned by Delta Social Media. They are in the process of acquiring a Cloud Based Distributed Big Data Storage solution. For the past few years, as technology has continued to grow and change, start-up companies such as Facebook, Google and Amazon have used cloud computing to build huge scalable solution systems. Delta Social Media are intending to follow a similar path. The benefits for the company in terms both financial and technical are immense. They will be able to make a profit managing data as a public or private host provider for any kind of data and any kind of company or individual. This report was commissioned to research the risks and benefits between Hadoop and Spark in cloud computing to assist the company in their final choice.

The internet is constantly growing and changing. A large number of industries are use Hadoop to analyse their data, whether it be a college, media company, art gallery , music label, and so on. As global internet traffic grows so does the amount of data. This became a problem for many industries and Hadoop was one of the first companies to offer a solution. But what is data? Data is a set of facts. People want to save these facts. It could be a simple picture or complex financial data. Companies always want to know more about their users. By analyzing what the user is doing on their network or website, companies can find immense value in gathering data. This data can tell companies a lot about specific regions can describe and predict person'sbehavior or can be used to dictate where the industry can improve.

Technology Discussion

Expanding data storage needs

Hadoop is very scalable and affordable.It can grow without requiring programmers to re-architect their algorithms applications. A user can add as much as they need in the cluster. Whenmore nodes are added to the system, it automatically grows.Scalability can keep all data alive forever. Hadoop uses disk storage.Users who have a lot of hardware failurehave to consider this extra costwhen preparing the company's budget. They also have to budget for more ram. Hadoop uses copy files.There will not be only one copy of the file in one cluster. If there is hardware failure or the hard drive goes down there is always way of getting the data back. The data will not disappear.

Hadoop/ MapReduce Advantages

Positiveparallel processing

Hadoop spreads information during the cluster in the hard drive and in the slave nodes and master nodes. Should one of the clusters go down, the data has been spread in blocks so the data can be rebuilt. That means companies will always have a copy available to use.

Hadoop/MapReduce limitations

Does not do real- time processing

Hadoop is based on batch processing of data. The data is checked periodically and because of this it can take a long time to process jobs based on the amount of data to be read and the number of nodes in the cluster. We cannot recommend for data live analysis.

Not suitable for filtering and other small-scale operations

Like operations on Structure Query Language (SQL), select options or joint tables to search for specific data you may have to rewrite the operation. To accomplish this in MapReduce requires extra development by the user.

Network issues

Processing large amounts of data takes time. The more data to be processed, the more time it will take to copy and can cause network issues by slowing it down. Hadoop works far better can on a local node.

General Performance

Hadoop can be slow because it runs operations on the disk and cannot deliver data analyses in real time. That means no real time analytics which can be an important point for some clients.

Spark Advantages

General Performance

Generally speaking, performance on spark can be much better. It can run up to 100 times faster in memory (https://spark.apache.org/). See figure 1 below.

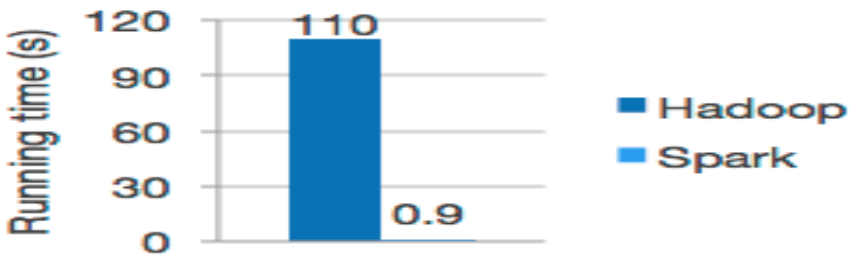

Figure 1 - Running times comparison

Real-time Processing

Spark runs live. It uses ram to store and run data. It delivers real time analytics. A good example of a use for this is for bank transactions such as credit card processing ormachine learning security analyses.

Memory Usage

Spark uses a different way to manage data because everything is processed within the ram. The original data stays available on the disk. This protects the data.

Spark Disadvantages

Cost

As mentioned above, Spark requires large amounts of RAM to run effectively. While Spark is free and open source, the overall cost of setting up a Spark cluster is more expensive than Hadoop (Kalron, 2020).

Methodology

To understand and properly review these technologies, it was decided to download them and run tests on the different parts.

Step 1

Run a practical exercise on Hadoop with their distributed file system. Hadoop was set up on our network on Ubuntu using a super user account. Folders were created to process the MapReduce data. A directory was also set up to store data information. Some java files were created within the framework. While running this practical exercise, the command line

“exploded” and created problems. This was noted as a potential issue for clients such as Delta Social Media when attempting to set up their Hadoop account. The layout of the Hadoop architecture is seen in Figure 2 below.

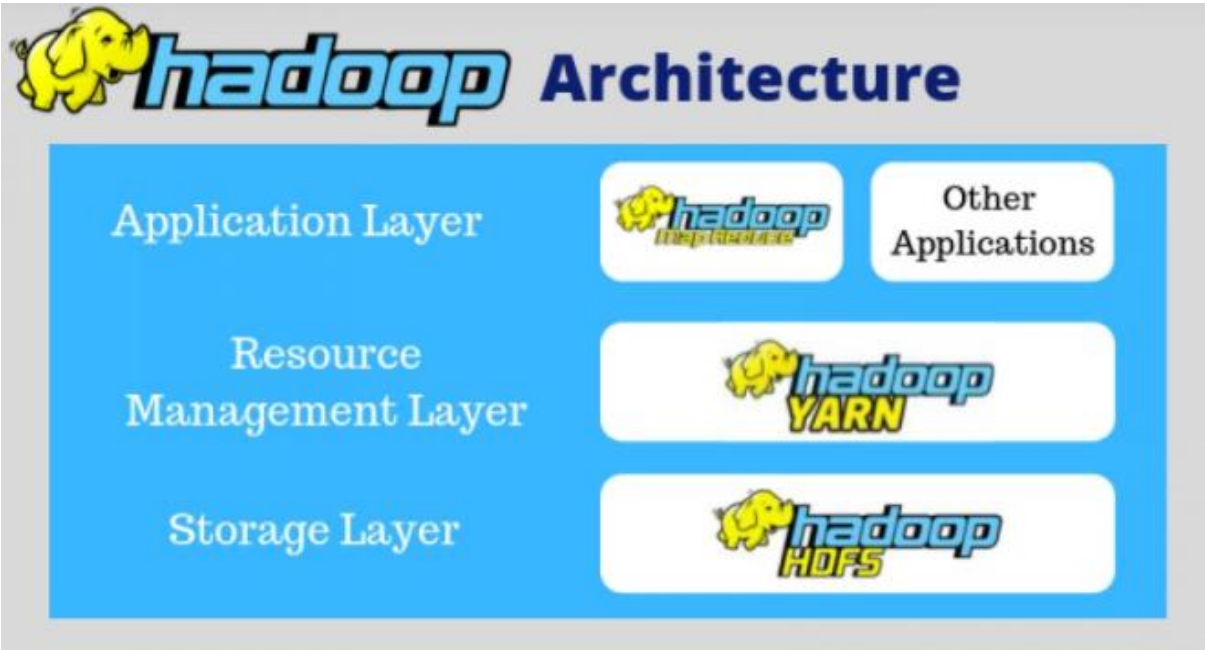

Figure 2 - Hadoop Architecture

Step 2

Further research was conducted on the Hadoop set up during the practical exercise. A number of layers were found. On the first frame is the Hadoop Distributed File System (HDFS). This is the file structure used by Hadoop. It defines how the file is read, how it should be managed and controls storage while working with a wider group of nodes.It also allows access to data anywhere in the cluster to copy, delete, read and write.

On the second layer, Hadoop saves multiple copies of the same files in different areas in the cluster. It processes specific data sets on individual nodes in order to speed things up.

On the HDFS, files are saved in folders. HDFS multiples the computers and file systems. It allows access to all ofthose computers then a user can easily get to the file. There is no need to connect to them directly and therefore less software work involved. In case of failure, HDFS has multiples copies on the cluster that means the data will be saved.

Step 3

MapReduce

MapReduce was made to manage large amounts of data. Users can often have large amounts of data distributed on many different machines. With MapReduce, all of this data can be processed at the one time.

MapReduce has two main tasks:

Mappers that process the data on a given node.

Reducers that take large amounts of data and reduce them.They essentially select the information that really matters, generating meaningful data with the objective to answer the users queries or questions. See figures 3 and 4 below for illustrations of this.

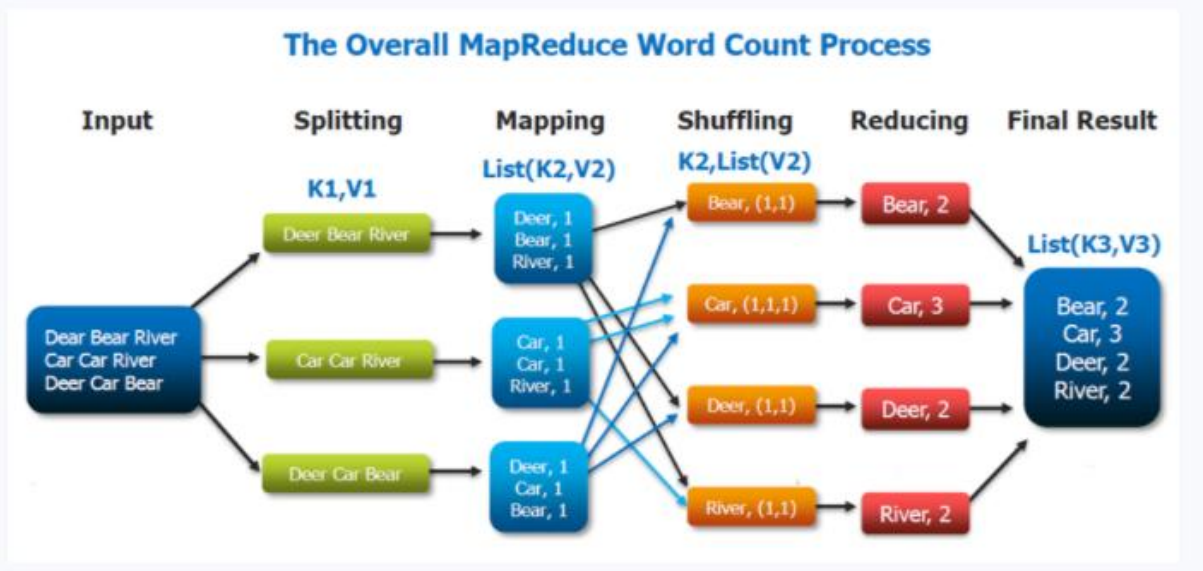

Figure 3 - MapReduce Process 1

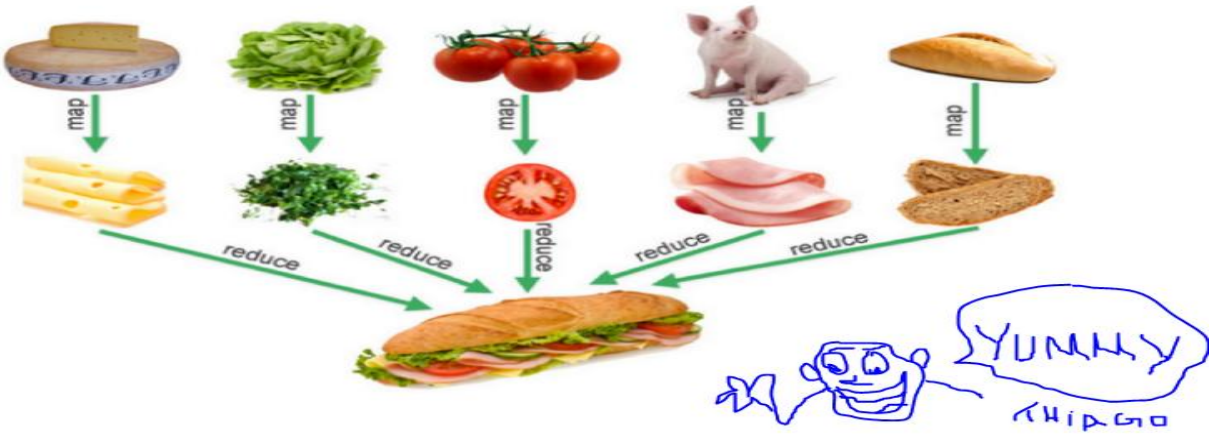

Figure 4 - MapReduce Process 2

Findings and Recommendations

It is difficult to compare these technologies because they are very similar. Both are open-source projects. This means that technically there is no cost for the software, and both should be equal in terms of cost. However, the hidden costs associated are to implement the infrastructure (Krivaa, 2019).Both projects are built and designed in a way that they can run with low total cost of ownership.

Hadoop and Spark are not competing against each other. They need each other to complement themselves. Hadoop helps to maintain the datasets and control of data and Spark provides the real-time process analysis, live in memory and high-speed operation analytics Hadoop has multiple integration and hardware support storage. When both work together, they can deliver even better results.

Hadoop

In the short term, Hadoop is the more attractive choice. It is less expensive when the hidden costs are considered. It is open-source and although it requires more memory on disk than Spark, this is a relatively inexpensive commodity.

Spark

Spark is open-source but requires a lot of raw memory to run. This increases the cluster size and also the overall cost. However, for a company as large as Delta Social Media, it is worth investing in disk memory for the long term financial and data processing gains.

Reduction of overall cost and overheads

Hadoop and Spark are helping companies on the market to better understand their customers and their data in a more effective way and to provide better service based on data analysis. Cloud computing is a new solution for big data in the future.The contribution for a small or big company to migrate to Spark or Hadoop infrastructure would be a better financial solution in terms of cut down costs. Either Hadoop or Spark will provide savings down the road for Delta Social Media. In terms of immediate savings, as mentioned above, Hadoop is the more attractive choice.

Review of existing Cloud Based Solutions

As part of the review, case studies were conducted on existing companies that have switched to using cloud computing technologies.

Razor fish case study on Amazon MapReduce implementation

According to Razor fish, the company works with media. Using Amazon Hadoop, there was no extra spending on hardware and no extra staff to work with this new technology. All the testing and development was fast and super efficient. Now the process is all automated. Before Amazon Hadoop, Razor fish was processing their data on a traditional hosting environment that utilized high-cost SAN equipment for storage.

The problem

Razor fish has a cluster with 30 servers and high-end SQL servers. In preparation for a new season, demand for targeted advertising increased. In order to work with this need, theybudgeted 500,000 dollars in extra hardware expenses , plus additionalstaff such as a Senior Operations Database Administrator.They were hoping to increase the processing cycle to be completed in 18 hours (Amazon, 2014).

The solution

To solve this issue of maintaining large datasets and custom target and activities, Razor fish had to change the “old school” infrastructure ofdata storage and analysis. Moving to Amazon cloud, Hadoop Map reduce architecture helped the company to manage the vast amount of data and speed up the application infrastructure levels.

- Benefits: reduce cost and risk of processing delay
- Scalable Amazon infrastructure helped Razorfish, store and process massive (Petabytes) datasets.

According to Mark Taylor, Program director at Razorfish the implementation of Amazon MapReduce, resulted in no upfront investment in hardware, no hardware delays and no extra operations staff were hired. The total cost of the infrastructure was around 13,000 dollars per month. Thanks tothe algorithm and the flexible platform to support it, the first client experienced a 500% increase in return on ad spend.

King gaming industry on Google

King is in the mobile social game world. The objective for this company is to predict whatusers do the most while gaming. With data captured, King can improve the design and engagement of the gamers.

Their infrastructure needs to support and manage hundredsof thousands of live connections per second and also their data warehouse. Jacques Erasmus, CISO, and King saw a potential in Google to handle the workload as Google provide machine learning and artificial intelligence.

With 270 million players a month, as they work with data archives in double petabytes, King conducted research in the hope of finding the best cloud support that suits their needs.

The problem

For years King managed one of the largest Hadoop Clusters in Europe, but the query enginewas suffering from stability issues. Maintaining the rolling infrastructure cost too much in time management. To get rid of something on the network that was not important would be a target to improve in all futures aspects from technical to commercial. If the problem could be fixed, then they could focus on valuable business plans for the company.

The solution

After all the resource costs, the solution for King was to split their data across different platforms to maximize efficiency. The advantage to building a new platform was to unify allits data in one place. After research, King decided to use Google platform. The main reasons were scalability and analytics possibilities. Now King has two great solutions: Data warehousing infrastructure and one single platform for machine learning. King is currently on the process of migration to Big Query.

“Nested Fields in Big Query let us efficiently query our data without reading to join big tables. This has been very useful in allowing our business units to quickly drill down from top levels numbers to very granular data”

- Tom starling, Principal data warehouse engineer at King.

The advantage of working with Cloud storage at King was and to have a secure data archive. Cloud data flow can help save down the line for the data warehouse team to ingest data less complex to work within (Carey, 2017).

With the help of Google Cloud Platform King has reliable, scalable data warehousing and it spends less time on management. Now the King team focus on bringing value to the business. According to CISO Jacques Erasmus, there is a massive improvement with their work flow on the new platform. On the old cluster, the analysts used to spend days setting up and building the environment. Now with the help of Google, just with a few clicks it is up and running. Jacques says all the engineers are super happy with the decision of King to move to Google.

Conclusions

Our consultancy will be recommending Spark as the best software to solve the issues that Delta Social Media face.

All the giants of the internet are using some type of software to manage and process their data. This shows how important it is today for large and small companies to gather and store data over the internet. Delta Social Media is another company that with the support of our consultation has the potential to become one of the new giants out there.

By our analysis, Spark is the best software that can keep up with the changing demands of the internet. Spark has excellent software to manage the large amounts of data created by the

company’s 1 million users and has the ability to handle more as the company continues to grow.

At the beginning of our review, Hadoop showed many advantage such as good expanding storage and positive parallel processing but unfortunately showed us some limitations. No real time processing, not suitable for filtering small network issues and on performance it can be slow to process and analyze data. Our engineers found the performance on Spark software to be faster and it has the massive advantage of real time processing. This is an important feature for social media companies such Delta Social Media. Spark has a disadvantage on cost. Setting up a Spark cluster can be more expensive than Hadoop in the short term but in the long term it will return huge savings. This small short term expense is worth the future savings by our analysis.

On reduction of overall cost, both Hadoop and Spark will provide savings down the road. Our researcher analysed the workflow of these two softwares.They both had many benefits in terms of time management and not needing an increase of staffto manage the technology.

Bibliography

Amazon AWS documentation (2014) “Apache Hadoop on Amazon EMR”. Available online. Accessed April 2021. https://aws.amazon.com/solutions/case-studies/razorfish/

Bhardwaj, Y (2020) “Hadoop Architecture”. Available online. Accessed April 2021. https://hackr.io/blog/hadoop-architecture

Bulao, J (2021) “How much data is created every day in 2021?”. Available online. Accessed March 2021. https://techjury.net/blog/how-much-data-is-created-every-day/#gref

Burns, S. (2021) “Majority more conscious of privacy online in last year, survey finds”. Available online. Accessed March 2021. https://www.irishtimes.com/business/technology/majority-more-conscious-of-privacy-online-in-last-year-survey-finds-1.4479563

Carey, S (2017) “12 Hadoop case studies in the enterprise”. Available online. Accessed March 2021. https://www.computerworld.com/article/3412201/12-hadoop-case-studies-in-the-enterprise.html#slide6

Edureka (2018) “Hadoop vs Spark | Which one to choose? | Hadoop Training | Spark Training | Edureka”. Available online. Accessed April 2021. https://www.youtube.com/watch?v=xDpvyu0w0C8

Kalron, A. (2020) “How do Hadoop and Spark stack up” Available online. Accessed April 2021. https://logz.io/blog/hadoop-vs-spark/

Krivaa, K. (2019) “Hadoop vs. Spark: Debunking the Myth” Available online. Accessed April 2021. https://www.gigaspaces.com/blog/hadoop-vs-spark/

Samadi, Y., Zbakh, M. and Tadonki, C. (2017) “Performance Comparison between Hadoop and Spark Frameworks using Hibench benchmarks” *Concurrency and Computation: Practice and Experience* 30(12)

Shaikh, T. (2019) “Batch Processing – MapReduce Paradigm”. Available online. Accessed April 2021. https://blog.k2datascience.com/batch-processing-mapreduce-paradigm-d3c4d08dab6a

Simplilearn (2017) “Spark Tutorial for Beginners | Big data Spark tutorial | Apache Spark Tutorial | Simplilearn”. Available online. Accessed April 2021. https://www.youtube.com/watch?v=QaoJNXW6SQo

Standford University (2012) “Introducing Apache Hadoop: The modern data operating system”. Available online. Accessed March 2021. https://www.youtube.com/watch?v=d2xeNpfzsYI&feature=emb_imp_woyt