Title: Cloud computing platforms for Big Data Analytics

Author: Usman Basharat

Date: 29th November 2019

Course: COMP-1680 Clouds, Grids and Virtualisation

Word Count: 3,233 (excluding references and table of content)

School of Computing and Mathematical Sciences

Table of Contents

Introduction

This paper will discuss different cloud computing platforms that are used for big data analytics. In general terms, big data refers to large sets of data and cloud computing refers to the platforms of processing large sets of data. I will be discussing the concept of cloud computing and what different types such as SaaS, PaaS and Iaas. A discussion on the types of platforms, such as Amazon Web Services (AWS), Google Cloud and Microsoft Azure. All these platforms will discuss how big data analytics will be used on these platforms. Adding onto this, I will be discussing the different types of big data analytics techniques that are used. I will also be comparing the different platforms too. Furthermore, a conclusion will be discussed of personal thoughts and summary of the opinions on cloud computing with big data analytics. Lastly, future directions will be discussed on what will happen with cloud computing and big data analysis.

What is Cloud Computing?

Cloud Computing is a remote network of servers that is hosts, manages, and processes data. This has services including servers, storage, database, networking, software, analytics and intelligence. Many different companies including Microsoft Azure, Google Cloud and AWS offer many different services. Drobox is an example that offer users storage use to upload files such as photos, documents, and videos to be uploaded to the "cloud". A basic free account can let users store up to 2GB. This can be accessed remotely on different servers. However, one issue that concerns all users is the security of this. Incomplete control over this data makes it sensitive and security is a issue that needs top priority.

Many companies have adopted the style of using Cloud Computing as this has variety of benefits of using this service. The big shift from traditional use is mainly aimed at the agility of accessible data. With cloud, accessing data makes it easy to do and can be accessed anywhere. In addition, moving to cloud computing reduces the costs of having IT systems by consistently updating current hardware and software. Hence, this saves time and money by also improving efficiency by switching to Cloud. This service is pay as you go basis, so companies can only pay for the service that they only need. However, different companies have different regimes of payment. This will be discussed further in the report.

Figure 1 explains the structure of how cloud computing is structured. It is essentially in a "cloud" that includes the applications, infrastructure and platforms that run on it. This is how a general example of how one is structured. Many different companies have a detailed structure of this that is looked going to be looked at in Amazon, Google and Microsoft.

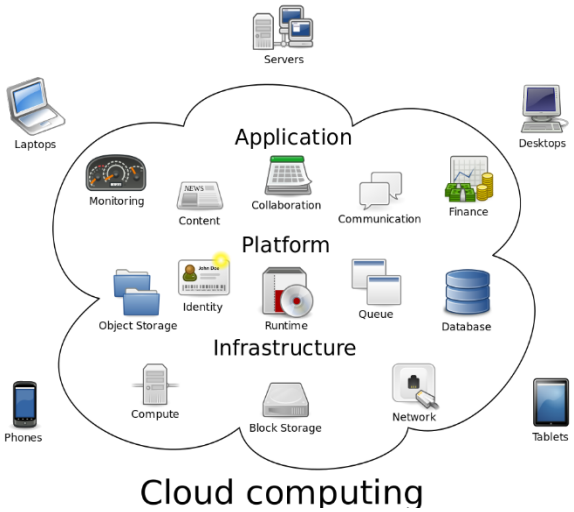

Figure 1 shows the structure of cloud computing.

Cloud Computing have three different variations of deployment such as:

- **Public Cloud**

Individual organisations share, own and operate by third-party cloud services.

- **Private Cloud**

Individual organisations that are on the infrastructure of cloud and not shared by anyone else.

- **Hybrid Cloud**

Individual organisations cloud deployment is split between private and public deployment. Infrastructure of cloud service is provided, but the security of sensitive data is protected by the organisation. Different organisations have different ways of showing the ways of deployment in their service.

Public Cloud	Hybrid Cloud	Private Cloud
+ Services are owned by a third-party provider	+ Combination of public and private cloud	+ Only be used by one organisation
+ Pay-as-you-go method as settings and cost is less than	+ Greater flexibility of deployment and infrastructure.	+ Much more secure as resources are not shared
- Shared responsibility of security	- Can be extremely expensive and harder to implement a cross-platform cloud service	+ Greater flexibility to improve and control the infrastructure
- Resources are shared on one platform infrastructure		- More control creates the implementation to be harder to do.

Table 1 shows the advantages and disadvantages of using public, private and hybrid cloud

Cloud Computing Models

Software as a Service (SaaS)

Under the SaaS service, most companies use this service for their products. This service does not require companies to install on individual devices. Most cases, the software that is provided would be free. However, they would be limitations of access dependant on the company's decision to extend these services. A typical application of SaaS would be Dropbox, JIRA and many other related software (Yandong and Yongsheng, 2012). The biggest advantage of this use is that this is easy to set-up. SaaS company manage the service including network, data, storage. However, the disadvantage of using this service is that the there no control to the infrastructure. This would lead to a massive impact if the company makes a drastic change in their procedure.

Platform as a Service (PaaS)

Under the Platform service, this allows companies to provide a platform for individuals to build applications. This facilitates deployment of the application without having to deal with the rest of the infrastructure. Please refer to Figure 2 to look at this in more detail (Yandong and Yongsheng, 2012). The biggest advantage of using PaaS is the fact of the control over the software. However, the disadvantage is if anything is wrong with the operating system; it would affect the software drastically too i.e. crash. An example of using PaaS would be Google App Engine.

Infrastructure as a Service (IaaS)

Under the Infrastructure service, this is a delivery of hardware and software. In other words, this is a service of hosting. Unlike PaaS, this service will provide for the storage and server side of the database. However, the users would be responsible for the application, security and database of the system (Yandong and Yongsheng, 2012). Please refer yourself to Figure 2 for clarification. The benefits of using IaaS would be the control over the software and hardware of the application. An example of this would be Microsoft Azure.

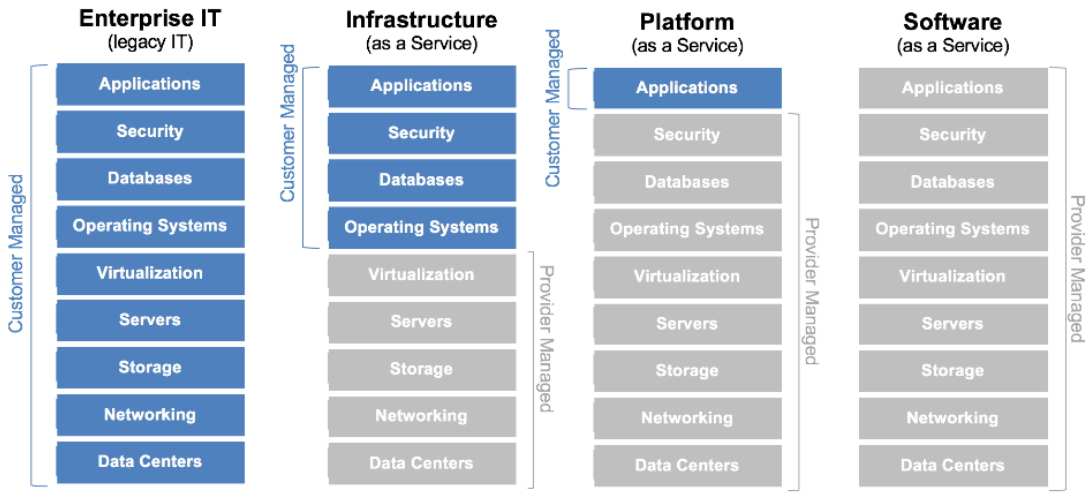

Figure 2 shows the difference infrastructure of services in Cloud Computing (The Enterprise Cloud Blog, 2019)

What is Big Data?

Big data refers to examination of large sets of data to uncover information using various techniques such as hidden patterns, market trends and many others (Rouse, M). Current industries use these techniques use these techniques to uncover information. The concept of big data refers to the four Vs factors that we can understand better of the concept of big data (Chan, 2018):

- Volume – the amount of data accumulated by public, private and any other organisations on a daily basis is extremely large.
- Velocity – considering the large amount of data gathered, it's given that this data will pile up. This matters on the process of examination and the outcome of its useful information.
- Variety – the types of data gathered such as databases, tweets, emails, images and others.
- Veracity – refers to the certainty of the data. Using analysis tools to separate bad data to the good one.

Gathering these large amounts of data offers variety of benefits to companies using variety of tools for benefaction. These large sets of data are organised into two analytics such as structured and unstructured data. Unstructured data involves text, videos, files, emails and social media data. Structured data is "data you're probably used to dealing with" such as computer-generated, or human-generated (Kaufman and Halper, 2019).

All this information is gathered for tools to be used by Google Cloud. This is where Big Data Analytics and Google Cloud complement each other. As Priya summarises the difference perfectly with "big data is information while cloud computing is the means of getting information" (Anand, 2018).

What are Analytics?

Big Data needs extremely large techniques to process the volumes of data. Chen and Zhang discuss in their journal an example of Wal-Mart uses techniques to explore patterns from their large transaction data using machine learning and statistical techniques (Philip Chen and Zhang, 2014). Referring to Figure 3, this shows the some of the different techniques within Analytics that are used. All of these will be discussed later in this section.

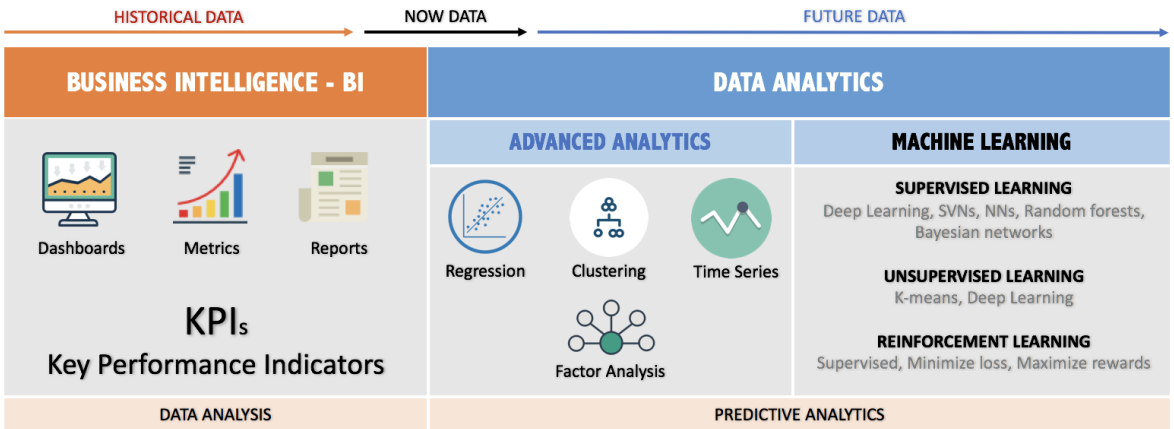

Figure 3 shows the difference between the two (The Asian Entrepreneur Authors & Contributions, 2019)

Business Intelligence (BI)

BI refers to the process of large amounts of unstructured data. The process of this information usually formed in dashboards, metrics, reports gained from historical data for business decisions (Rapid Miner, 2019). Business Intelligence usually asks questions such as:

- What happened?
- When did it happen?
- Who did it?
- How many?

Advanced Analytics

Advanced Analytics discovers the deeper insight into the traditional methods of prediction. This also gives an insight for organisations to make accurate decisions in business strategy (Rapid Miner, 2019). Some of the methods that are used data mining, text mining, statistical/quantitative analysis. Some of the questions are raised when making such analysis:

- What will happen?
- When will it happen?
- Who is going to do it?
- How many are involved?

Please refer yourself over to Figure 4 that shows the difference between Advanced Analytics and Business Intelligence.

	Business Intelligence	Advanced Analytics
Orientation	Rearview	Future
Types of questions	What happened When, who, how many	What will happen? What will happen if we change this one thing? What's next?
Methods	Reporting (KPIs, metrics) Automated Monitoring/Alerting (thresholds) Dashboards Scorecards OLAP (Cubes, Slice & Dice, Drilling) Ad hoc query	Predictive Modeling Data Mining Text Mining Multimedia Mining Descriptive Modeling Statistical / Quantitative Analysis Simulation & Optimization
Big Data	Yes	Yes
Data types	Structured, some unstructured	Structured and Unstructured
Knowledge Generation	Manual	Automatic
Users	Business Users	Data scientists, Business analysts, IT, Business Users
Business Initiatives	Reactive	Proactive

Figure 4 shows a table between the difference of the two (Rapid Miner, 2019).

Machine Learning

Machine Learning is a "method of data analysis that automates analytical model building". As mentioned briefly in Figure 3, Machine Learning has a variety of learning styles that they use. Supervised learning algorithms are trained on the comparison of inputs and outputs. Unsupervised learning goes over data that has no historical data. This must be explored. Reinforcement learning is often found in graphics, gaming and navigation. This discovers the trial and error with the actions that has the greater reward (Insight, 2019).

Cloud Computing Platforms

In this section, I will be discussing different platforms such as Amazon Web Service (AWS), Google Cloud and Microsoft Azure of how they use big data analytics. Various of these companies have different principles and this would be discussed as a comparison at the end of this section. Each platform is going to be discussed of the introduction and what services they use. Also, a section on detail of how each platform uses big data analytics techniques within their organisation. They are many different techniques, and this would only highlight what services the organisation does in this. Lastly, a comparison of all three are going to highlight the advantages, disadvantages, costs, infrastructures, companies that use their services, and general summary of this at the end.

Amazon Web Service (AWS)

AWS launched in 2006 by Amazon to provide a "secure cloud service offering functionalities to help businesses grow" (Amazon Web Service, 2019). AWS include variety of services involving different infrastructure as service such as IaaS, PaaS and SaaS. More than over 100 services are included in this and they are separated in groups. Some of these are Storage Database, Networking, Security and many more. Please refer yourself to Figure 5 to find out more in detail of how many category and services Amazon Cloud provide. Figure 5 only shows some of the services provided. As stated, they are many services. To view all of these, AWS have a huge list on their website. However, we are going to be focusing on in particular is Analytics. This will be discussed in detail later on in this section.

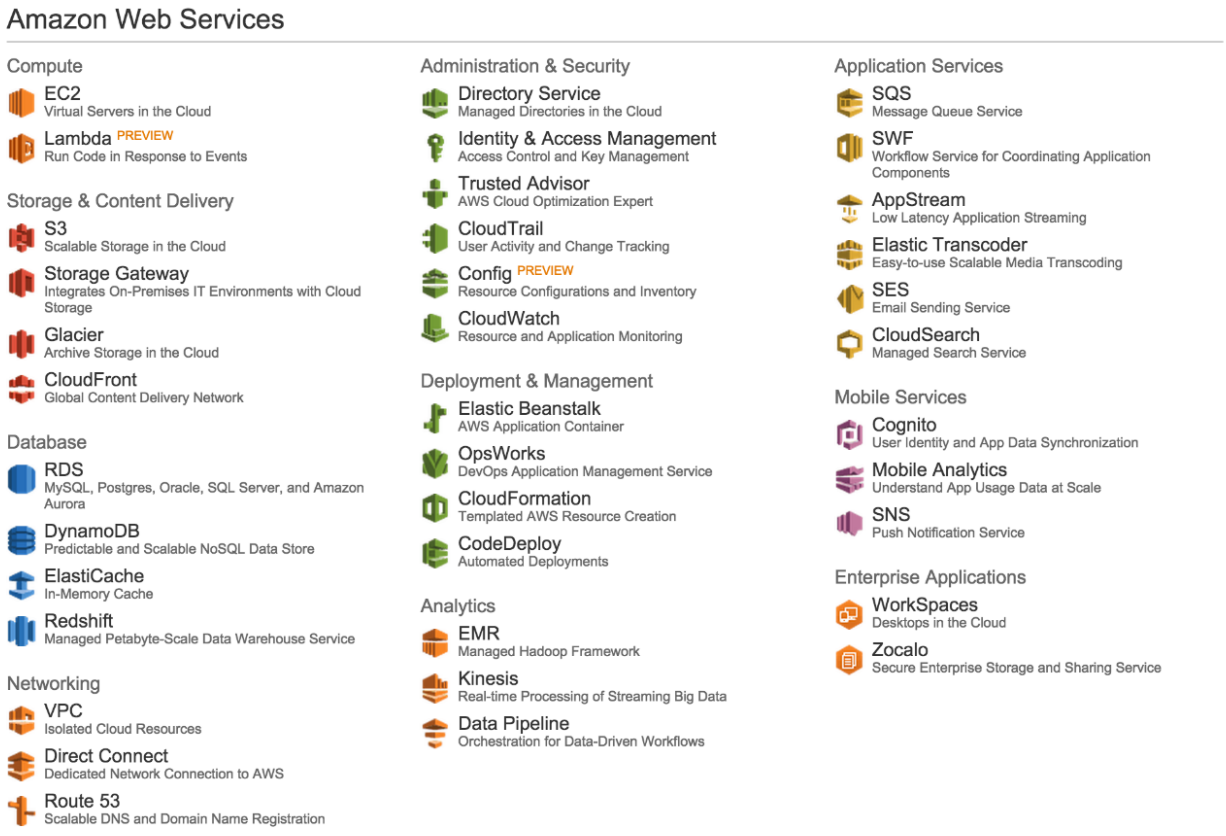

Figure 5 shows some of the services provided by Amazon Web Service

AWS and Big Data Analytics

Amazon Athena is one of the analytics that Amazon run for services. As Amazon states Athena is a query service that allows organisations to run and analyse data using SQL queries in Amazon S3. There is no infrastructure to manage. However, organisations pay for the services dependant on how many queries they run. Therefore, this makes this service serverless. This makes it easy to setup for organisations. There is no worry of the managing services or data warehousing. This is straight to the point of using Amazon S3. The results of these queries that are ran are quick. As Amazon states on their official website, it takes less than "10 minutes to set-up" (Amazon Athena, 2019). This is one of the services that Amazon provide. As Figure 5 shows in the Analytics section, they are also EMR, Kinesis and Data Pipeline that also have not been discussed.

Google Cloud

Google first introduced Cloud in 2008 by launching a "is a set of cloud computing platforms that provide infrastructure tools and services for users to build applications and services on top of" (Forrest, 2019). Unlike Amazon Web Service, Google provides a short category of services that they cover are in Figure 5.

Figure 6 demonstrates the list of various services that are provided by Google. However, the one that is interesting in this is Big Data. All comparison of services between these are shown further in the report.

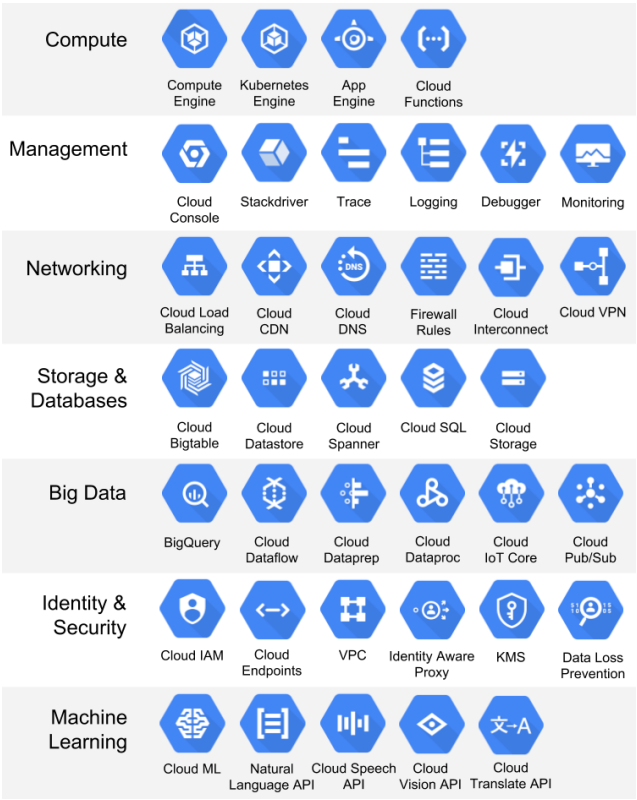

Figure 6 shows the list of services for Google Cloud.

Google Cloud and Big Data Analytics

Cloud Dataflow provides a data processing stream for applications. Various project applications that Google are associated with are Java and Maven, Python and Java and Eclipse. Google also provide their own templates that users can use to execute the current project. This enables high-volume computation tasks to be easily divided by using relevant tools that are provided by Cloud Dataflow. In comparison with Amazon Athena, Google also provide similar tool to work on which is BigQuery. BigQuery also provides data analysis service by create, load, execute SQL commands. However, payment structure for this are different and divided in four sections such as active, long-term, on-demand and flat-rate. Unlike Amazon Athena, customers can choose to have a variety of options when it comes to purchasing this service. Each payment is different. For example, Queries (flat rate) is $8-500 per 500 slots. Lastly, as shown on Figure 6, they are variety of services in Big Data.

Microsoft Azure

Having launched in February 2010, the latest out of the three, Microsoft Azure provides similar service out of the two platforms that have been discussed. This is a set of cloud services to help organisations with their challenges, as stated by Microsoft Azure (Microsoft Cloud Service, 2019). Similar to Google Cloud and AWS, Microsoft Azure offer variety of services that are categorised in Figure 7. These are the services that are available. These are split into different infrastructures as shown in Figure 7. For this report that will be discussed, I am interested in the Analytics and Internet of Things (IoT) category.

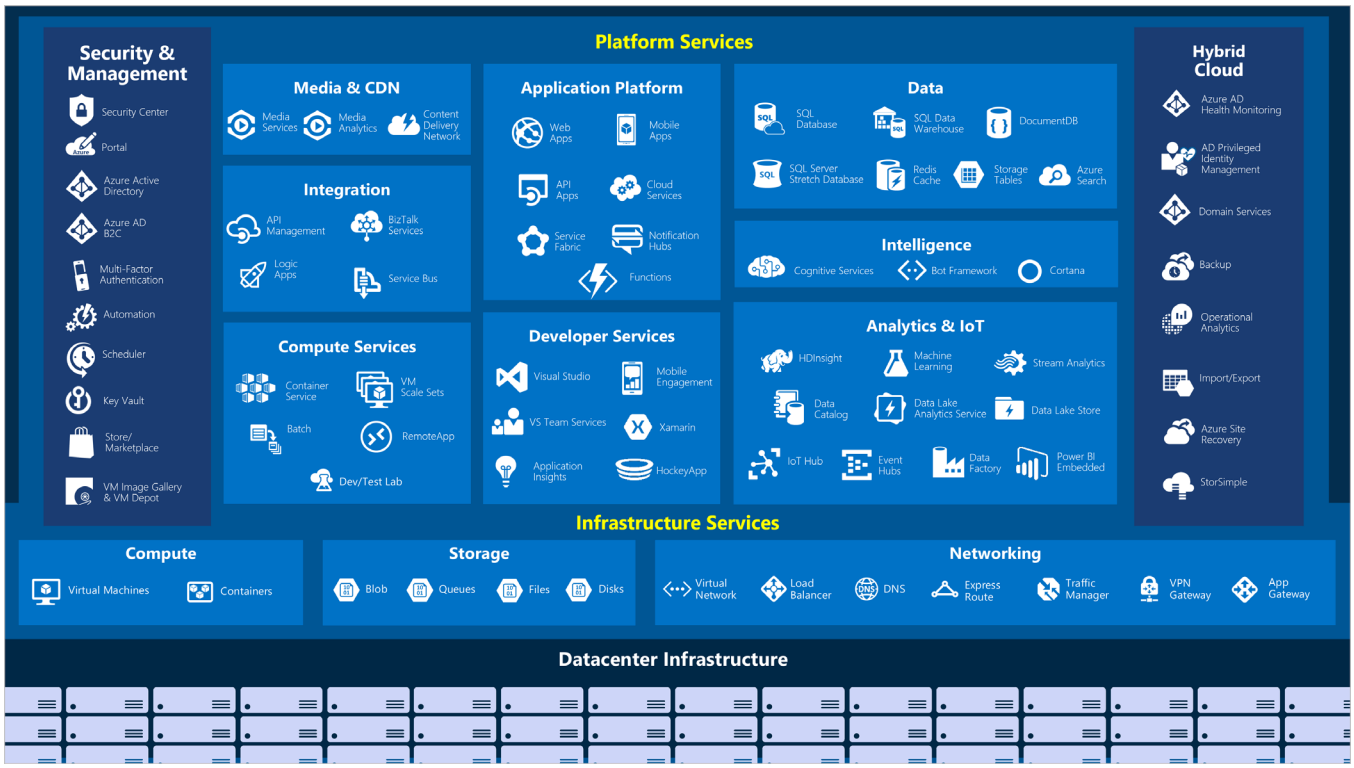

Figure 7 shows the list of services for Microsoft Azure.

Microsoft Azure and Big Data Analytics

Data Factory collect data information from all resource data types that have been discussed to transform and analyse data. This is a cloud-based extract, transform and load (ETL) system that allows you to create workflows for data collected. Different tools can be used to enhance this service by using SQL Database. The ETL service in cloud works by the following diagram:

ingest → control flow → data flow → schedule → monitor

As Figure 7 shows different tools that Big Data uses Machine Learning by providing various tools to enhance capabilities of designing predictive models. Microsoft Azure provides a drag and drop tool with popular models to create a unique model that can be published (Microsoft Documentation, 2019). Unlike Google and Amazon that provide SQL Query separate, Microsoft provide this as a package known as SQL Database.

Comparison of Cloud Computing Platforms

	Amazon Web Service	Google Cloud	Microsoft Azure
Advantage	• Greatest global reach • Record of reliable services shown in Figure 8. • Large range of services shown in in Figure 5. • Ideal for larger companies	• Most cost-effective • Ideal for apps and software-based users • High durability	• Offers hybrid that uses public and private together • Ideal for start-up and developers • Benefits for Windows users
Disadvantage	• Limitation of use • Complex pricing for users to understand	• Storage expensive • Complex pricing for users to understand	• Expertly maintained and managed • Requires expert knowledge
Cost for Big Data Analytics (SQL Query)	$5 per 1TB	$0.02 per GB, per month	$0.745 per database per month
Cost for Storage (monthly)	$0.023 / GB – S3 Standard	$0.026 /GB - Standard Storage	$0.0618 per GB - Standard
PaaS	✓	✓	✓
IaaS	✓	✓	✓
Saas	✓	✓	✓
Organisations	Netflix, Samsung, Kellogs	Spotify, HSBC, Domino's	Linkedin, Starbucks

Table 2 shows the comparison between different cloud companies.

Complexity among prices for users are common for all three of these companies. To get the best understanding, potential organisational users may use third-party resources to get a better understanding of how this could work. However, all three do offer a calculator that users can get estimate for each course. In terms of pricing, Google offer the most cost-effective way by offering users no upfront costs, pay-as-you-go scheme and no termination fees. Also, Google offers various discounts including 30% discount on workloads that run on Compute Engine and Cloud SQL. These various pricing details discounts that makes Google one of the cost-effective companies. In terms of establishments, Amazon was launched since March 2006 and they have been in the sector longer than the rest comparing to Google in 2008 and Microsoft in February 2010. Hence, the reason behind their revenue success in 2019 shown in Figure 8. Security can be a big concern when dealing with issues of cloud computing. When saving data online, organisations face serious risk like losing data and phishing. These great threats need to be at priority of each organisation that deals with data online.

Cloud Computing Trends

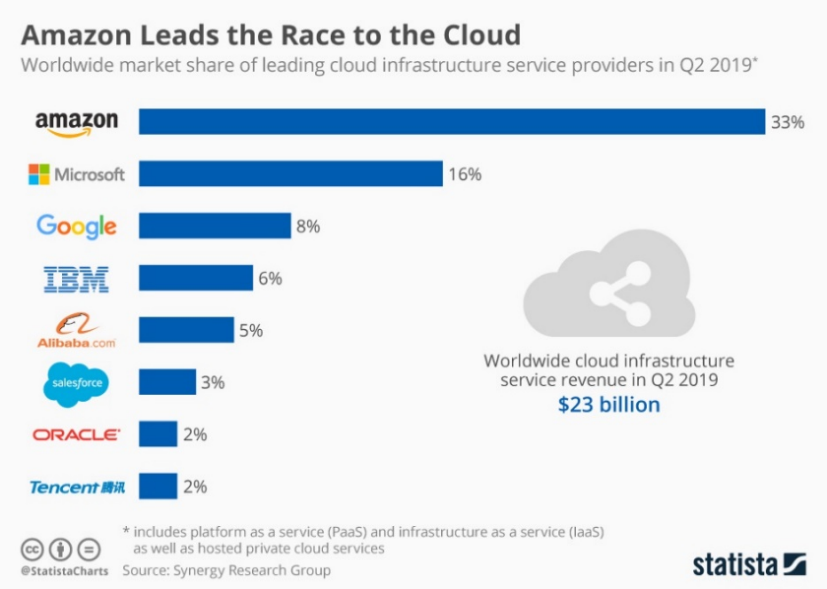

Figure 8 shows the trends in infrastructure service (Market Share Trends, 2018)

Figure 8 shows the race of cloud within 2019 by Amazon leading the way with 33% of revenue in Cloud infrastructure. This shows that this company is leading the way as 33% of organisations and services that are used are trusted and satisfied with the service that has been provided. All of the rest of the companies are looking up to Amazon as a competitor to see what they can change.

Future Directions

Quantum Computing

Organisations are shaping the future for promising research. Quantum computers performs precise calculations by using objects to perform these powerful calculations that classical computers cannot perform. Using these calculations, they can process large data that would benefit large companies such as AWS, Google Cloud and Microsoft Azure. As mentioned, as big data increases, quantum computing has the potential to use this data and process it quickly by getting better results. This would benefit the companies' cloud-tools as mentioned above.

Research Framework

Figure 9 shows us a research framework in how the template needs to be followed. Each organisation needs to able to follow this structure to keep maintenance and standard of each principle and improve aspects as stated of each field. I feel as though this structure needs to be followed for each organisation as the principles are necessary for the relevant field.

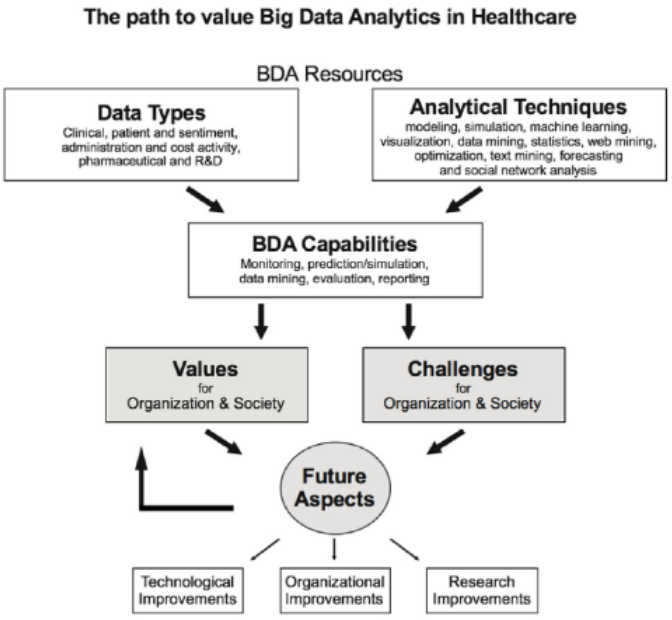

Figure 9 shows the structure of data analytics in Healthcare (Galetsi, Katsaliaki and Kumar, 2019)

Conclusion

This report highlights the introduction of how cloud computing works alongside big data analytics. Both of these fit together as *many different* techniques are used in different organisations such as AWS, Google and Amazon. The comparison underlines the differences of each organisation as different organisations have different practices. Future research reveals two ways of the what the future of cloud would include and how to maintain the future. Lastly, the discussion of the techniques was discussed in general. They are *many* techniques that Amazon, Google and Azure cover. However, due to the word limit, this wasn't possible to discuss.

References

Amazon Athena. (2019). *Amazon Athena — Serverless Interactive Query Service - AWS*. [Online] Available at: https://aws.amazon.com/athena/?nc2=h_ql_prod_an_ath [Accessed 10 Nov. 2019].

Amazon *Web Service*. (2019). *Amazon Web Services*. [Online] Available at: https://www.amazon.jobs/en/business_categories/amazon-web-services [Accessed 9 Nov. 2019].

Anand, S. (2018). What's the difference between big data and cloud computing? [Blog] *Quora*. Available at: https://www.quora.com/Whats-the-difference-between-big-data-and-cloud-computing [Accessed 3 Nov. 2019].

Chan, M. (2018). Big Data in the Cloud: Why Cloud Computing is the Answer to Your Big Data Initiatives. *ThornTech*.

Cloud computing terms | Microsoft Azure. [Online] Available at: https://azure.microsoft.com/en-gb/overview/cloud-computing-dictionary/ [Accessed 25 Oct. 2019].

Forrest, C. (2019). *Google Cloud Platform: A cheat sheet*. [Online] TechRepublic. Available at: https://www.techrepublic.com/article/google-cloud-platform-the-smart-persons-guide/ [Accessed 11 Nov. 2019].

Galetsi, P., Katsaliaki, K. and Kumar, S. (2019). Values, challenges and future directions of big data analytics in healthcare: A systematic review. *Social Science & Medicine*, 241, p.112533.

Google Cloud Overview. (2019). *Google Cloud Platform Overview | Overview | Google Cloud*. [Online] Available at: https://cloud.google.com/docs/overview/ [Accessed 11 Nov. 2019].

Insights, S. (2019). *Machine Learning: What it is and why it matters*. [Online] Sas.com. Available at: https://www.sas.com/en_gb/insights/analytics/machine-learning.html [Accessed 4 Nov. 2019].

Kaufman, M. and Halper, F. (2019). *Structured Data in a Big Data Environment - dummies*. [Online] Dummies. Available at: https://www.dummies.com/programming/big-data/engineering/structured-data-in-a-big-data-environment/ [Accessed 3 Nov. 2019].

Market Share Trends. (2018). [Images] Available at: https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/ [Accessed 17 Nov. 2019].

Microsoft Cloud Services. (2019). *What is Azure – Microsoft Cloud Services | Microsoft Azure*. [Online] Available at: https://azure.microsoft.com/en-gb/overview/what-is-azure/ [Accessed 14 Nov. 2019].

Microsoft Documentation. (2019). *Microsoft Azure Documentation*. [Online] Available at: https://docs.microsoft.com/en-us/azure/ [Accessed 17 Nov. 2019].

Philip Chen, C. and Zhang, C. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, pp.314-347.

RapidMiner. (2019). *Advanced Analytics vs Business Intelligence | RapidMiner*. [Images] Available at: https://rapidminer.com/glossary/advanced-analytics-vs-bi/ [Accessed 4 Nov. 2019].

Rouse, M. (2019). Structuring a big data strategy. [Blog] *TechTarget*. Available at: https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics [Accessed 1 Nov. 2019].

The Asian Entrepreneur Authors & Contributors (2019). *Big Data Science in 5 Minutes*. [Images] Available at: https://www.asianentrepreneur.org/big-data-science-in-5-minutes/ [Accessed 4 Nov. 2019].

The Enterprise Cloud Blog (2019). *Who Manages Cloud IaaS, PaaS, and SaaS Services*. [Images] Available at: https://mycloudblog7.wordpress.com/2013/06/19/who-manages-cloud-iaas-paas-and-saas-services/ [Accessed 28 Oct. 2019].

Yandong, Z. and Yongsheng, Z., 2012, August. Cloud computing and cloud security challenges. In *2012 International Symposium on Information Technologies in Medicine and Education* (Vol. 2, pp. 1084-1088). IEEE.