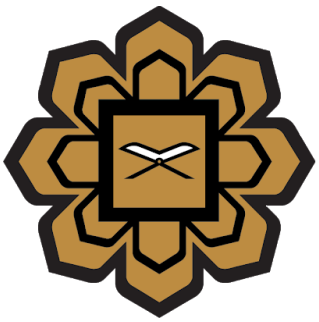

الجامعة الإسلامية العالمية ماليزيا
INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
يونيورسيتي إسلام انتارابغسا ملايسيا

Big Data Analytics

CSC 3303

Sem II 2018/2019

Group Assignment 2

Group members :

Abdirahman Abdullahi	1432401
Asem Hamood Al-abdali	1513599
MHD Khaled Maen	1523591
Alrefaei Mohammad	1617111

1. Clarify

In this stage, we implement extraction to our data to explore with work, we may need to subset our data so we can have smaller dataset that can score or explore with our developed model.

If we not subset our dataset into smaller subset, our data might not working with big data problems because, big data by natural hard to work with entire dataset due to its large size.

classes of extraction:

Class1 : extract a sample or summary from the whole dataset and use it for as permanent sample.

Class 2: compute on part, which means repeating the computation for many subgroup and fit it into model per individual and after that we can combine the results of all parts .

Class 3: compute on the whole, in some problems require to use the whole dataset at once, this kind of problems are irretrievably big, and this consider most difficult problem because they must run at scale within the data warehouse .

2. Develop

In Order to develop a good model we need powerful tools which being provided by Amazon Web Services (AWS) such as Data Lakes and Analytics.

AWS-powered data lakes can handle the scale, agility, and flexibility required to combine different types of data and analytics approaches to gain deeper insights, data lakes can be build by applying some other AWS services as shown in Figure 1:

- **Data Movement:** Upload the data to the server in real-time.
- **Data Lake:** Store limitless amount of data**.**
- **Analytics:** Interactive analysis, big data processing using Apache Spark and Hadoop.
- **Machine learning:** Predict the outcome of the data

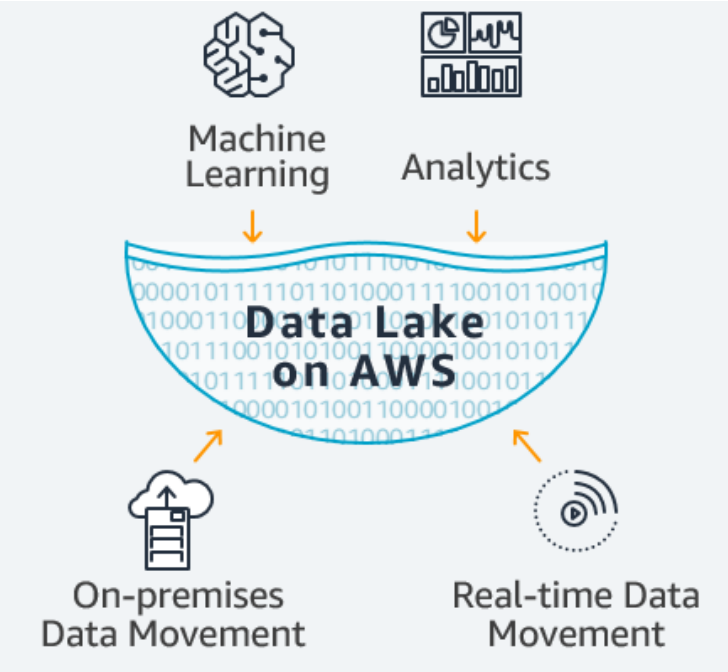

Figure 1 Data Lake on AWS

In that by applying the previous steps on our dataset we build our first version of our machine learning model using Big data.

3.Productize

The transformation from developing customer-specific software to product software is growing in scientific industries. After developing the model of this project, we come to the stage where we need to test our model and let other to try it and productizing services is a very powerful way of building efficiency. The projects are especially built to meet the customers' specific requirements and therefore they are the main stakeholders within this phase. After the software is developed our model will be ready to use and implemented by the customer.

Recent studies on productization showed the important of transforming customer-specific software to a business product software. According to (Artz, van De Weerd, & Brinkkemper, 2010), the process of productization added more valuable contribution in the research of software business.

Moreover, since we are running a service-based business the project's objective is to satisfy the customers' needs on the best succeeded restaurants, we can provide value to our model if the product software used widely and online cloud services is a powerful tool to meet product development challenges. Microsoft azure particularly provides smart tools for building manufacturing industry solutions.

4. Publish

It is known also as the closeout phase or the completion phase and it is the final phase of the lifecycle of an analysis project , it is about releasing the final achievements to the client, and releasing the project's resources and documentation to the business. In our project, we provided the new constructed plan to the development team. After creating the model, we will analyze its performance and determine if it is met the project's goals. In this phase we also may analyze the performance of the team along with timeline and quality.

After the work is done, if our model is working and is a predictive model, it will be used by the client.

References

Data Lakes and Analytics | AWS. (n.d.). Retrieved from https://aws.amazon.com/big-data/datalakes-and-analytics/

Artz, P., van De Weerd, I., & Brinkkemper, S. (2010). Productization: The process of transforming from customer-specific software development to product software development. *Software Business First International Conference ICSOB2010 Jyvaskyla Finland June 2010 Proceedings*, *51*, 90–102.

Working with Big Data in R. (2017, July 25). Retrieved from https://www.rstudio.com/resources/webinars/working-with-big-data-in-r/