RAMAIAH UNIVERSITY
OF APPLIED SCIENCES

Assignment

Course Code	CSC409A
Course Name	Data Analytics
Programme	B.Tech
Department	CSE
Faculty	FET

Name of the Student	Satyajit Ghana
Reg. No.	17ETCS002159
Semester/Year	VIII/2021
Course Leader(s)	E. Ami Rai

Declaration Sheet

Student Name	Satyajit Ghana		
Reg. No	17ETCS002159		
Programme	B.Tech	Semester/Year	08/2021
Course Code	CSC409A		
Course Title	Data Analytics		
Course Date		to	
Course Leader	E. Ami. Rai		

Declaration

The assignment submitted herewith is a result of my own investigations and that I have conformed to the guidelines against plagiarism as laid out in the Student Handbook. All sections of the text and results, which have been obtained from other sources, are fully referenced. I understand that cheating and plagiarism constitute a breach of University regulations and will be dealt with accordingly.

Signature of the Student		Date	

Submission date stamp (by Examination & Assessment Section)	

Signature of the Course Leader and date	Signature of the Reviewer and date

Contents

List of Figures

1 Question 1

Solution to Question No. 1 Part A

1.1 Introduction to Data Analytics and its applications

The analysis of data to extract knowledge is the subject of a vibrant area known as data analytics, or simply “analytics”. The definition adopted here is:

> **Analytics** The science that analyze crude data to extract useful knowledge (patterns) from them

In ‘Competing on analytics’, Thomas Davenport defines analytics as “the extensive use of data, statistical and quantitative analysis, exploratory, predictive models, and fact-based management to drive decisions and actions.”

This process can also include data collection, organization, pre-processing, transformation, modeling and interpretation. Analytics as a knowledge area involves input from many different areas. The idea of generalizing knowledge from a data sample comes from a branch of statistics known as inductive learning, an area of research with a long history. (Moreira, J, 2019)

Taxonomy of Data Analytics

- Descriptive analytics: summarize or condense data to extract patterns
- Predictive analytics: extract models from data to be used for future predictions.

For example, a sales report of a company, say Pepsi. This report will tell you how many units of Pepsi were sold, where they were sold, what price and a lot of other things. All of this is information coming from the data. All you are doing is slicing and dicing the data in different ways, looking at it from different angles, along different dimensions etc. There is very little statistics involved in descriptive analytics and so you don’t really need to be a statistical wiz to be able to do effective descriptive analytics.

While descriptive analytics is a very powerful tool, it is still giving us information about the past. Whereas, a business owner’s primary concern is the future. If I run a hotel, I want to be able to predict how many of my rooms will be occupied next week. If I am a drug company, I want to know which of my under-test drugs is most likely to succeed. This is where predictive analytics comes in.

Coming to Data Deluge, *WE’RE DROWNING IN DATA.* Supermarkets, credit cards, Amazon and Facebook. Electronic medical records, digital television, cell phones. The universe has gone wild with the chirps, clicks, whirs and hums of feral information. And it truly is feral: According to a 2008 white paper from the market research firm International Data Corp., the amount of data generated surpassed our ability to store it back in 2008. The cat is out of the bag. In 2010, the amount of digital information — from highdefinition television signals to Internet browsing information to credit card purchases and more — created and shared exceeded ***1 zettabyte*** for the first time. In 2011, it approached 2. The amount has grown by a factor of nine in five years, according to IDC, which pointed out in its 2011 report that there are "nearly as many bits of information in the digital universe as stars in our physical universe." (Stanford Medicine, Summer 2012)

1.2 Illustration with real world examples

We'll look into various real-world examples of how Data Deluge happens and affects starves information retrieval.

1. "Impressions are down. What's going on?"

This power company was concerned about impressions, believing that the quantity of people who viewed their ads was a crucial metric. To maximize impressions, they had cast a wide net with their keywords, hoping to reach as many people as possible. So, they were dismayed to see that their ads weren't getting as wide of an audience as they had hoped. That said, the company was quite happy with the click-through rate (CTR) of some of their keywords. While their average CTR was only .09%, the CTR for many of their low-impression keywords averaged .60% - almost 7 times the aggregate CTR. After examination, many of their low-impression/high-CTR keywords had nothing to do with what the company offered:

- Keywords related to gas prices had high CTRs, but were used almost exclusively by people seeking gasoline for their cars, not gas heating for their homes
- "Transformers" as a keyword attracted fans of the movies and toys, not people interested in electrical transformers
- More mystical keyword choices included "Power Rangers," "Power Ball numbers," "Monster Energy Drink," and "juicers"

Yes, these keywords were delivering clicks and impressions, but they weren't the impressions and clicks that would turn into customers and revenue. Worse, the company was spending almost 25% of their paid search budget on these non-relevant keywords.

2. "We don't have keyword-level data."

After working with members of this company's digital marketing team to build out keyword-level data and connect it to revenue, so it was astonishing to hear this. However, the two people we had worked with—the company's "stewards of data"—had left the company. When they left, not only did they take all their data-related knowledge with them, they left a void in the company for data-driven conversations. None of the remaining members of the digital marketing team knew about:

- The systems the two stewards had been using

- How to access those systems and harvest that data
- How to drill down in the data to get keyword-level information and tie it back to revenue
- The importance of regular conversations about data

(Elizabeth et al, 2018)

1.3 Discussion on the barriers for adoption

We are living in an age of information. Staggering amounts of information are collected, stored, and widely disseminated. Yet, we may be less informed and less knowledgeable than ever. This paradox of increasing information, yet decreasing knowledge and insight, has many possible causes, some of which are subtle and difficult to identify, and even more difficult to remedy. The fundamental issue is quantity crowding out quality, leading to an abundance of poor-quality information which may not be a good substitute for scarce but high-quality information. Information is not unique in exhibiting this paradox.

There are three fundamental reasons why quantity may crowd out quality. The most obvious is the production cost problem where the emphasis on quantity shifts the emphasis and resources away from quality. It is costly to produce quality information, and it is difficult to do both quality and quantity. When quality does not pay in proportion to its high cost, quantity wins over. This is also the most common explanation for non-information examples, but explanation for information products involves two other reasons. The second reason is the obsolescence problem. Information is not neutral with respect to the physical world, but it is an agent of change. Information is useful precisely because it is used to change the environment and subjugate nature and society to our purposes. But as information is used to change the environment to take advantage of new opportunities, our existing information about the environment becomes obsolete, leading to a loss of information. The net effect may be positive or negative, but it is increasingly negative as we will show, in a fast-changing information-intensive society. The third reason is the competition problem when information is used as a competitive weapon against others, to mislead and confuse others, leading to a loss of knowledge on their part. Information is power, because it can be used to control others and exploit them, by controlling their information sources, and consequently their behavior. (Orman, 2015)

Real-time expert systems and ANN would give erratic predictions for inputs that were dependent on each other or outside of the range used to develop the system. Furthermore, one could not drill down to determine the major contributors to the strange result. These systems fell into disuse as soon as the developer left the scene. There are so many failures of expert systems, it's difficult to keep track of all of them.

Top 10 Failures of Expert Systems

10. Failure to say you should have bought control valves instead of those cheap on-off valves

9. Failure to say you should have bought Coriolis meters instead of those cheap rotameters
8. Failure to explain why expert systems failed
7. Failure to explain what engineers will do when all the manufacturing is offshore
6. Failure to predict the next layoff
5. Failure to predict the last and next economic crises
4. Failure to explain what is really said in congressional bills
3. Failure to predict your drug costs under the Medicare prescription plan
2. Failure to predict what the cost of medical care will be under the new healthcare plan
1. Failure to figure out where the governor of South Carolina was last June.

(Greg McMillan, 2010)

1.4 Discussion on the future of analytics

The reliance on data driven decision making will continue to grow. Just like the widespread usage of metrics and reports today, companies will start expecting to see some predictive analytics insights as part of regular dashboards.

As analytics becomes more and more prevalent in the corporate consciousness, a basic awareness and understanding of analytical techniques will become a required skill for career growth at the middle to senior management tiers, irrespective of industry and function. There will also be an increased demand for some super specialized roles. These will require intensive expertise with programming and technology to support the actual analytics implementation.

In the next decade we will witness technological advances that will play an increasingly important role in the ability of companies to mine data for real time insights and actions in the context of the rapid pace of data produced and the variety of data that is being captured.

The future of data analytics will see data discovery and preparation change, in a practice known as augmented data preparation and discovery. Machine learning automation augments and streamlines data profiling, modelling, enrichment, data cataloguing and metadata development, making the data preparation process more flexible. Traditional methods often involve rule-based approaches to transform data. However, augmented data preparation makes the process more flexible because it automatically adapts fresh data, especially outlier variables.

Machine learning augments data discovery because the algorithms allow data analysts to visualise and narrate relevant findings easily. Machine learning also paves the way for several functions like clusters, links, exceptions, correlations predictions and data exceptions without having to rely on end-users to generate all these results. Augmented data preparation and discovery will play a huge role in the future of data analytics because it streamlines data preparation and discovery, giving analysts large sets of clean data. (Michael Dixon, 2019)

1.5 Stance taken and justification

Successfully managing the “data deluge” will allow scientists to compare the genomes of similar types of cancers to identify how critical regulatory pathways go awry, to ferret out previously unknown and unsuspected drug interactions and side effects, to precisely track the genetic changes that have allowed evolving humans to populate the globe, and even to determine how our genes and environment interact to cause obesity, osteoporosis and other chronic diseases.

I do believe there are many factors that can make the data so overpowering that information retrieval becomes difficult or even impossible in some cases, but there are a few measures that can be taken to mitigate so,

1. Countering data deluge by using the right data

The primary challenge for any company is to select the right information that serves its customers. Data is growing rapidly and it’s difficult for marketing analysts to collect and analyse data on the go. Additionally, it’s important to make sure that the data collected by you really reflects the purchase decisions taken by your customer base. Data deluge is going to increase in the coming time, but it is manageable if companies learn to select the right amount of data as it will help them build better and reliable customer relationships in the long run.

2. Countering data deluge by controlling costs

Traditionally, companies have always been convinced by the theory of throwing money for technology solutions. This, however, can no longer be called a sensible strategy because of the exponential growth in data that calls for sensible analysis and handpicking only “useful” data. Companies can train its data scientists to reject duplicate data and cull useful information that can save a lot of money in the long run.

3. Countering data deluge by updating and auditing policies

According to the 2011 McKinsey Global Institute report, many large U.S companies have more data stored as compared to the U.S Library of Congress and that has become a cause of concern for data managers. The only way to counter this is by having effective data retention and data destruction policies. Companies can have such policies that allow auto-deletion of inessential data after a specific period of time. Also, there must be policies which allow retrieval of data such as important emails, files, and documents in times of litigation. Companies must also diligently follow both internal as well as government compliances in the process. Moreover, while data should be transparent and available to all employees, they should be given the right to access it only for a limited time. (Naveen Joshi, 2017)

2 Question 2

Solution to Question 1 Part B

2.1 Introduction

The Data analytic lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.

2.2 Discuss data preparation phase tools

Phase 2: Data Preparation

- Steps to explore, preprocess, and condition data prior to modeling and analysis.
- It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
- Data preparation tasks are likely to be performed multiple times and not in predefined order.
- Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, Storm, Spark etc.

1. Apache Spark

Spark is a framework for parallel processing of Big Data. Spark is designed to use the basis of Hadoop MapReduce with some modifications that enables it to perform more efficiently than Hadoop MapReduce. Spark has its own streaming API and independent processes for continuous batch processing across varying short time intervals.

Spark runs up to 100 times faster than Hadoop in certain circumstances, however it still uses Hadoop distributed file system. This is the reason why most of the Big Data projects install Spark on Hadoop so that the advanced Big Data applications can be run on Spark by using the data stored in Hadoop distributed file system. So, we can consider Spark as an extension of Hadoop, which has some features for real-time analytics like being fast, simple, and supportive of applications such as machine learning, stream processing, and graph computation. Xu, Wu, Xu, Zhu, and Bass implement Spark into their idea for real-time data analytics as a service. It is able to support both stream and batch processing while Hadoop is made mostly for batch processing.

Spark provides many real-time processing and evaluation options that Hadoop alone cannot. Therefore, to manage the data for their architecture, they utilize Spark specifically. Though Bilal et al. are making use of a graph database, Neo4J, to store datasets, Spark is the graph processing system being used. Their use of Spark will allow them to process the waste data and analyze it efficiently. The research on distributed computing engines shows that Spark has consistent scalability for large datasets. Yan, Huang, and Yi show Spark is scalable to process seismic data with its in-memory computation and data locality features. (Yadranjiaghdam, 2016)

2. Apache Storm

Storm is another real-time computation system. It is a task parallel distributed computing system which can reliably process unbounded streams of importing data. Storm uses an independent workflow, Directed Acyclic Graphs, in its platform. Storm utilizes Zookeeper, a minion worker to manage its processes, instead of running on Hadoop clusters. Many of the explored resources make use of Storm with their new contributions to real-time data analytics. Storm, unlike Hadoop alone, can continue to analyze data as it arrives. As Storm is a complex event processing system that has the ability to detect important event occurrences, it is the processing system that Jones utilizes to detect crucial events through the processing of Twitter feeds. (Yadranjiaghdam, 2016).

Apache Storm is based on the 'fail fast, auto restart' approach that allows it to restart the process once a node fails without disturbing the entire operation. This feature makes Storm a fault-tolerant engine. It guarantees that each tuple will be processed 'at least once or exactly once', even if any of the nodes fail or a message is lost. The standard configuration of Storm makes it fit for production instantly. Once the Storm cluster is deployed, it can be easily operated. Besides, it is a robust and user-friendly technology, making it suitable for both small- and big-sized firms.

2.3 Discuss model building phase tools

Phase 4: Model Building –

- Team develops datasets for testing, training, and production purposes.
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – Rand PL/R, Octave, WEKA, Python
- Commercial tools – MATLAB, STASTICA.

1. Python – Scikit-learn

The Python programming language is establishing itself as one of the most popular languages for scientific computing. Thanks to its high-level interactive nature and its maturing ecosystem of scientific libraries, it is an appealing choice for algorithmic development and exploratory data analysis.

One of the popular library used for building models is Scikit-learn, Scikit-learn harnesses this rich environment to provide state-of-the-art implementations of many well-known machine learning algorithms, while maintaining an easy-to-use interface tightly integrated with the Python language. This answers the growing need for statistical data analysis by non-specialists in the software and web industries, as well as in fields outside of computer-science, such as biology or physics. Since it relies on the scientific Python ecosystem, it can easily be integrated into applications outside the traditional range of statistical data analysis. Importantly, the algorithms, implemented in a high-level language, can be used as building blocks for approaches specific to a use case, for example, in medical imaging (Dubois, 2007; Milmann and Avaizis, 2011)

Strong points

– General purpose, open-source, commercially usable, and popular Python ML tools.
– Funded by INRIA, Telecom Paristech, Google and others.
– Well-updated and comprehensive set of algorithms and implementations.
– It is a part of many ecosystems; it is closely coupled with statistic and scientific Python packages.

Weak points

– API-oriented only.
– The library does not support GPUs.
– Basic tools for NNs.

2. Weka3

Weka collects a general purpose and very popular wide set of ML algorithms implemented in Java and engineered specifically for DM (Weka3 2018; Waikato 2018) . It is a product of the University of Waikato, New Zealand and is released under GNU GPLv3-licensed for non-commercial purposes. Weka has a package system to extend its functionality, with both official and unofficial packages available, which increases the number of implemented DM methods. It offers four options for DM: command-line interface (CLI), Explorer, Experimenter, and Knowledge Flow.

Weka can be used with Hadoop thanks to a set of wrappers produced for the most recent versions of Weka3. At the moment, it supports MapReduce but not yet Apache Spark. Clojure (Hickey 2018) users can also leverage Weka, thanks to the Clj-ml library (Clj-ml 2018).

Related to Weka, Massive Online Analysis is also a popular open-source framework written in Java for data stream mining, while scaling to more demanding larger-scale problems.

Strong points

– General purpose, involving wide set of algorithms with learning schemes, models and algorithms.
– It comes with GUI and is API-oriented.
– Supports standard DM tasks, including feature selection, clustering, classification, regression and visualization.
– Very popular ML tool in the academic community.

Weak points

– Limited to Big Data, text mining, and semi-supervised learning.
– Weak for sequence modelling; e.g., time-series

(Giang, 2019)

2.4 Justify with suitable scenarios

Model Building Tools

1. Use case for Python-Scikit-learn

Machine learning for neuroimaging with Scikit-Learn – Alexandre et. Al (2014)

In this paper they have illustrated with simple examples how machine learning techniques can be applied to fMRI data using the scikit-learn Python toolkit in order to tackle neuroscientific problems. Encoding and decoding can rely on supervised learning to link brain images with stimuli. Unsupervised learning can extract structure such as functional networks or brain regions from resting-state data. The accompanying Python code for the machine learning tasks is straightforward. Difficulties lie in applying proper preprocessing to the data,

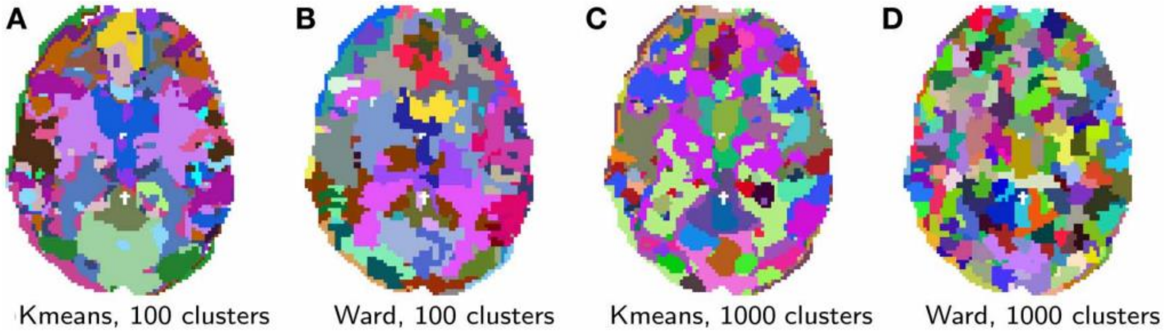

Figure 1 Brain parcellations extracted by clustering using scikit-learn

choosing the right model for the problem, and interpreting the results. Tackling these difficulties while providing the scientists with simple and readable code requires building a domain-specific library, dedicated to applying scikit-learn to neuroimaging data. This effort is underway in a nascent project, nilearn, that aims to facilitate the use of scikit-learn on neuroimaging data.

2. Use case for Weka3

Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification – Ignacio et al. (2017)

State-of-the-art light and electron microscopes are capable of acquiring large image datasets, but quantitatively evaluating the data often involves manually annotating structures of interest. This process is time-consuming and often a major bottleneck in the evaluation pipeline. To overcome this problem, we have introduced the Trainable Weka Segmentation (TWS), a machine learning tool that leverages a limited number of manual annotations in order to train a classifier and segment the remaining data automatically.

To segment the input image data (2D/3D grayscale or color), TWS transforms the segmentation problem into a pixel classification problem in which each pixel can be classified as belonging to a specific segment or class. A set of input pixels that has been labeled is represented in the feature space and then used as the training set for a selected classifier. Once the classifier is trained, it can be used to classify either the rest of the input pixels or completely new image data. All methods available in WEKA can be used.

Data Preparation Tools

1. Use case for Apache Spark

Bioinformatics applications on Apache Spark - Guo, R., Zhao, Y., Zou, Q., Fang, X. and Peng, S., 2018

Among the state-of–the-art parallel computing platforms, Apache Spark is a fast, general-purpose, in-memory, iterative computing framework for large-scale data processing that ensures high fault tolerance and high scalability by introducing the resilient distributed dataset abstraction. They surveyed Spark-based applications used in next-generation sequencing and other biological domains, such as epigenetics, phylogeny, and drug discovery.

Phylogeny reconstruction is important in molecular evolutionary studies but faces significant computational challenges. Before Spark-based tools were created, while several tools had been put forward for phylogeny reconstruction, they did not scale well, and there was a significant increase in the number of datasets. Therefore, in 2016, Xu et al. proposed CloudPhylo, a fast and scalable phylogeny reconstruction tool that made use of Spark. It evenly distributed the entire computational workload between working nodes.

An experiment was conducted using 5,220 bacteria whole-genome DNA sequences. The results showed that CloudPhylo took 24,508 seconds with one worker node, and it was able to scale well with increasing numbers of worker nodes. Moreover, CloudPhylo performed better than several existing tools when using more worker nodes. In addition, CloudPhylo achieved faster speeds on a larger dataset of about 100 Gb generated by simulation.

2. Use case for Apache Storm

Apache Storm Based on Topology for Real-Time Processing of Streaming Data from Social Networks – Batyuk, A. and Voityshyn, V. (2016)

In this paper we represented architectural concept of the Apache Storm based real-time data processing topology.

Experiments with the system allowed concluding the following:

1. The chosen toolset (mostly based on Apache Storm) was convenient in usage and shown its effectiveness on the implementation and testing stages.

2. The implemented topology demonstrated enough flexibility for the sample task. Therefore, it can be evolved in order to be applied for resolving more complex and valuable problems.

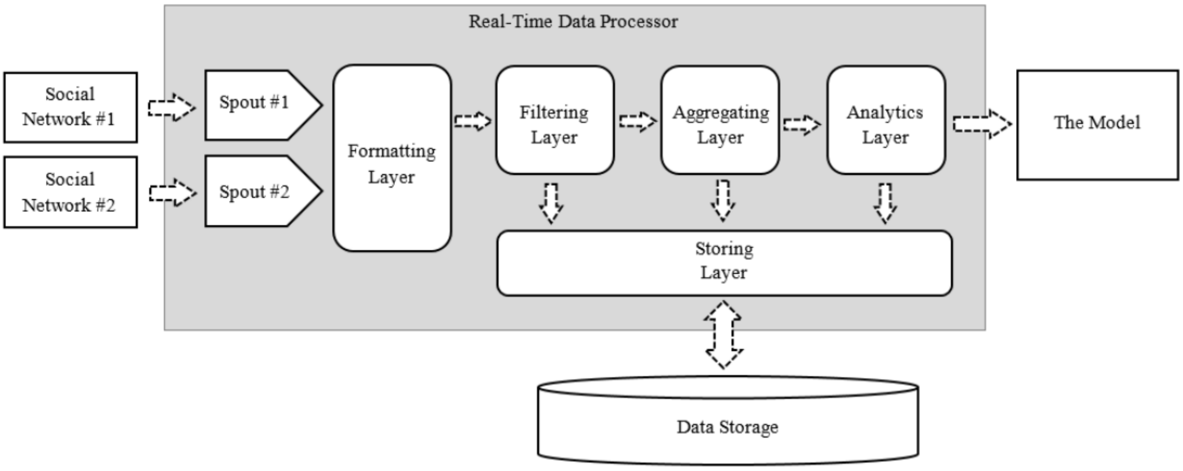

Figure 2 Real-Time Data Processing Topology with Apache Storm

3 Question 3

Solution to Question 2 Part B

3.1 Model different method(s) to address the above issues

3.2 Identify suitable attributes

3.3 Justify your solution by comparison

4 Question 4

Solution to Question 3 Part B

4.1 Recommend a solution

4.2 Discuss issues

4.3 Justification

5 Question 5

Solution to Question 4 Part B

5.1 Introduction to Big Data Platform

5.2 Problem solving approach

5.3 Design and Implementation

5.4 Performance analysis

Bibliography

1. Moreira, J., de Leon Ferreira, A.C.P. and Horváth, T., 2019. A general introduction to data analytics. Wiley.
2. Orman, L.V., 2015. Information paradox: Drowning in information, starving for knowledge. IEEE Technology and Society.
3. Anderson, C., 2008. The end of theory: The data deluge makes the scientific method obsolete. Wired magazine, 16(7), pp.16-07.
4. Controlglobal.com. 2021. [online]

   Available at: <https://www.controlglobal.com/articles/2010/drowninginfo1002/> [Accessed 26 May 2021].
5. Drowning in Data but Thirsty for Insight? Uncommonlogic.com [online]

   Available at: < https://www.uncommonlogic.com/wp-content/uploads/2018/08/Drowning-in-Data-But-Thirsty-for-Insight-ebook-unCommon-Logic-Bulldog-Solutions.pdf> [Accessed 26 May 2021].
6. Data Deluge and How to avoid it, Naveen Joshi, 2017,

   https://www.allerin.com/blog/data-deluge-and-how-to-avoid-it
7. Preparing for the future of data analytics, Michael Dixon, December 6, 2019

   https://seleritysas.com/blog/2019/12/06/preparing-for-the-future-of-data-analytics/
8. https://www.geeksforgeeks.org/life-cycle-phases-of-data-analytics/
9. Yadranjiaghdam, B., Pool, N. and Tabrizi, N., 2016, December. A survey on real-time big data analytics: applications and tools. In 2016 international conference on computational science and computational intelligence (CSCI) (pp. 404-409). IEEE.
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830.
11. Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á.L., Heredia, I., Malík, P. and Hluchý, L., 2019. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review, 52(1), pp.77-124.
12. Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K.W., Schindelin, J., Cardona, A. and Sebastian Seung, H., 2017. Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification. Bioinformatics, 33(15), pp.2424-2426.

13. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B. and Varoquaux, G., 2014. Machine learning for neuroimaging with scikit-learn. Frontiers in neuroinformatics, 8, p.14.

14. Batyuk, A. and Voityshyn, V., 2016, August. Apache storm based on topology for real-time processing of streaming data from social networks. In 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP) (pp. 345-349). IEEE.

15. Guo, Runxin, Yi Zhao, Quan Zou, Xiaodong Fang, and Shaoliang Peng. "Bioinformatics applications on apache spark." GigaScience 7, no. 8 (2018): giy098.