# DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

## MIT 805: Big Data

MIT805 Semester Exam

Matimba Shingange

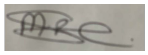
Student Number: u12264017

Date: 23 November 2021

**Git Hub Link**: *https://github.com/u12264017/U12264017---MIT805-Exam*

**Disclaimer:**

"The University of Pretoria commits itself to produce academic work of integrity. I affirm that I am aware of and have read the Rules and Policies of the University, more specifically the Disciplinary Procedure and the Tests and Examinations Rules, which prohibit any unethical, dishonest, or improper conduct during tests, assignments, examinations, and/or any other forms of assessment. I am aware that no student or any other person may assist or attempt to assist another student, or obtain help, or attempt to obtain help from another student or any other person during tests, assessments, assignments, examinations, and/or any other forms of assessment."



M R Shingange

**Signature** **Initials**

## Question 1

## About the data:

- I am using the MUBI movies dataset. This data could be described as the "Netflix for art movies". MUBI was founded in 2007 by Efe Çakarel who wanted to create a social network for cinema lovers.
- MUBI users have a selection of thirty movies on a daily rotating basis. Unlike many SVOD platforms relying on recommendation systems, the MUBI selection is human-curated.
- I have 2 files, the movies_data file, and movie_ratings_data.
- The MUBI movie data was last updated on the 25th of April 2020. This file has 196,628 rows with 10 columns. It is a 51.63 MB size file.
- The MUBI movie rating data was last updated on the 26th of April 2020. This file has 15,519,997 rows with 13 columns. It is a 2.12 GB size file
- User IDs were anonymized. This dataset does not contain personal identifiable data. Data from MUBI users who set their profile in private mode are not in this database.
- This dataset was created with MUBI API.
- This table goes back to 2008 and has 15 million records.
- Use cases examples for this dataset include text analyses on critics, recommendation systems, predicting movie popularity.
- Every row of the data represents a movie that was viewed.

## Big Data V's on MUBI Movies Dataset:

- **Volume**
  The MUBI Movies dataset has data records that go as far back as 2008, with the last update for 2021 made on the 26th of April 2021. The dataset has 15 million movie records. There is also a data file that contains movie metadata. With the two files merged, there are 18 columns. The data takes close to 2.12 GB of storage on storage. From this, we can clearly see that we are working with big data which is represented by the large volumes.
- **Veracity**
  There are two basic sources of this data. The data comes from movie streaming by users and there is also metadata about the movies that are being streamed on the platform. The data can be sourced from movie producers. Also, the data does not represent all the existing movies, it is only movies for which the licensing permission has been acquired by MUBI movies site owners.
- **Variety**
  The data about movie streaming and ratings given to the movies is accessed and stored in real-time. Whereas information about the movies is only loaded when the permission and licenses have been acquired. This shows that the data is available in both real-time and batch processing forms.
- **Value**
  The movies dataset has use cases ranging from text analyses on critics, recommendation systems, predicting movie popularity. The data has got good business value as they can also determine the streaming times with most traffics. This

will ensure that the streaming platform owners increase available resources around the peak streaming times. They can also use this data to identify the popular movie release years. This will ensure that the movies they have access to are able to movies of popularity which will ensure even more traffic and revenue is generated. Also, the producers with the highest views will also share information about which movies to the source to ensure more viewership.

## Methodology:

In this section of the exam, I am required to download and use a dataset with a file size of at least 1GB. The file I downloaded is a 2.12 GB file size, this file contains 15 million records. For simplicity of the application. I am using 7 million records of this file. I also merged the movies data and ratings data files on the movie id. This ensures that the metadata of the movies is also included, and more insights can be obtained.

The method I used to process this data, is to create Python written Map-reduce scripts. These scripts take in a data file which is the merged file created from the movie's ratings and movie data. This file contains 23 columns, however, I had to remove the URL columns since python was not able to use this data correctly.

**Environment:**

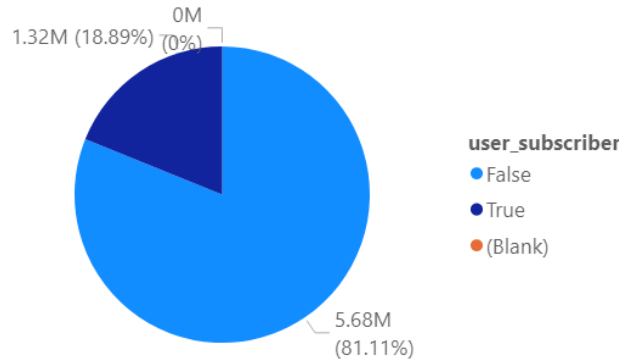
I downloaded a Horthonworks sandbox virtual machine from Cloudera. The Sandbox is a straightforward, pre-configured, learning environment that contains the latest developments from Apache Hadoop, specifically the Hortonworks Data Platform (HDP). From here I downloaded Python, pip, and MRJob. This pre-configured VM is easy to work with and it has a web interface where you can load small-size files and you can run a few hive queries. You are also able to monitor all the nodes running, to start and restart your nodes as well. I am running a mixture of single and multiple map-reduce steps for my data. For all the visuals there's a Map-reduce code for each. Which is a total of 10 Map-reduce algorithms. All the Map-reduce codes read data from a CSV input file and return the output to another CSV output file. Which is named with the map-reduce algorithm name. The data file was too big to load from the sandbox web interface, for this reason, I downloaded a PSCP agent to copy and transfer files between my Sandbox VM and local computer.

**Programming Language:**

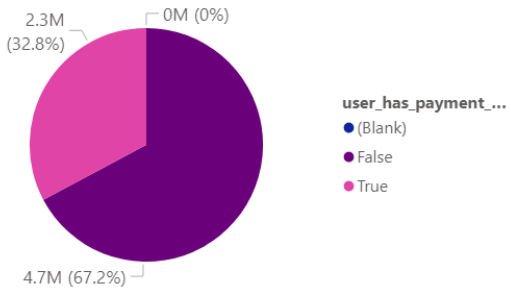
I used Python to write my Map-reduce algorithms. This was entirely because I am familiar with python, and I have used it a number of times this year. Also, python has got some good support online. I also used python matplotlib and seaborn libraries for the visualization of my Map-reduce outputs. The output from map-reduce is of smaller size and I was then able to process them through python as opposed to the original data file.

## Data Visualization

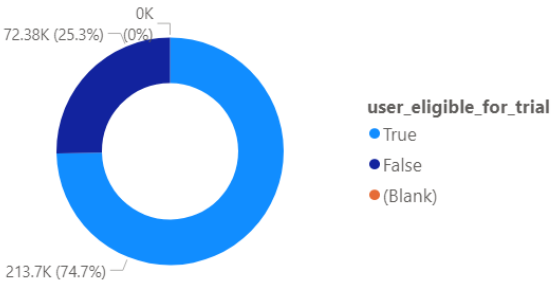
### Total Movies by User Subscription



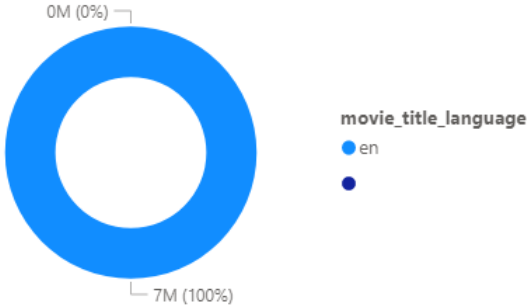

### Total Movies By User Payment



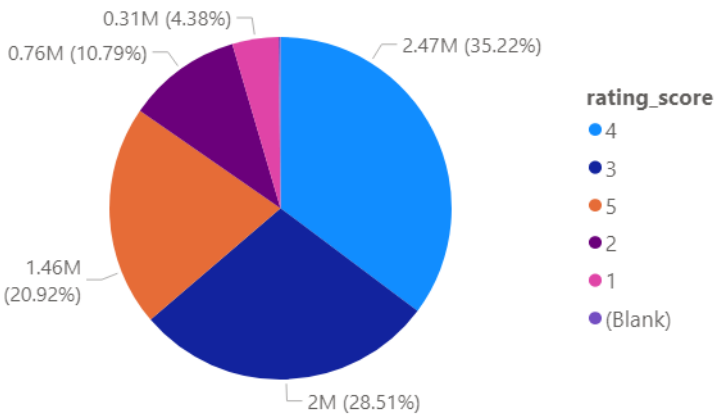

### Unique Users By Trial Eligibility



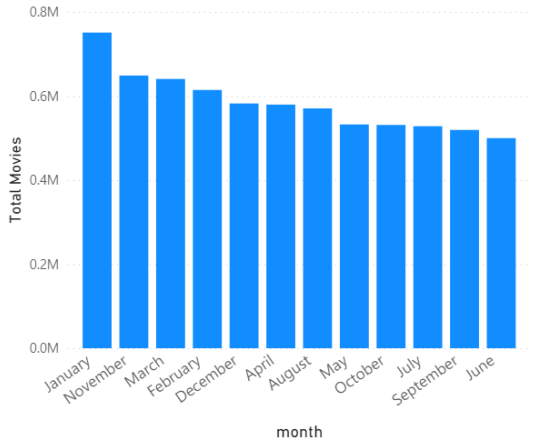

### Total Movies By Language

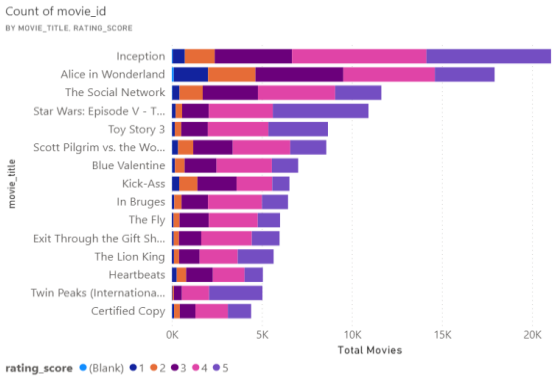



### Total Movies By Rating Score

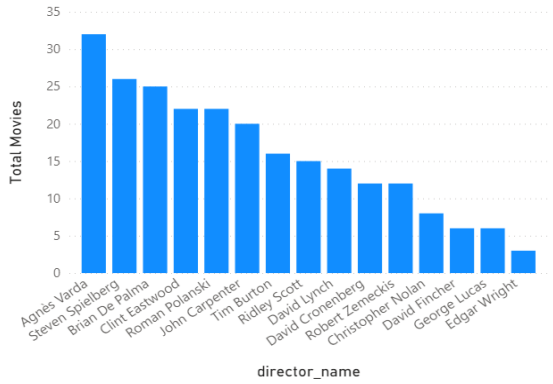


Total Movies by Month

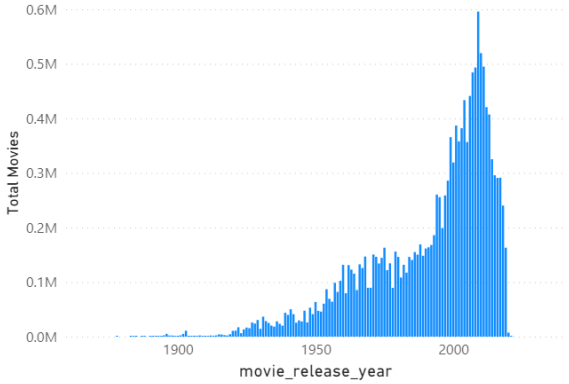


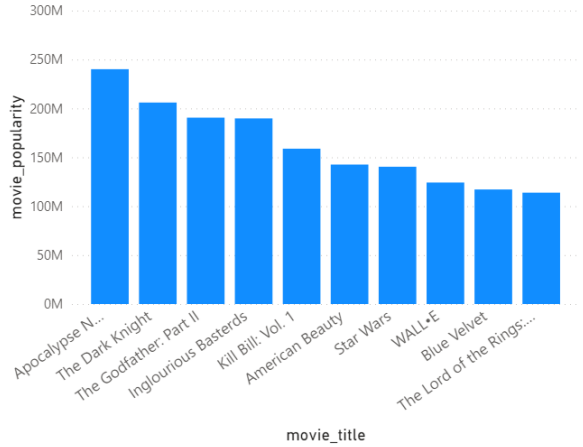

Top 15 Movie Directors



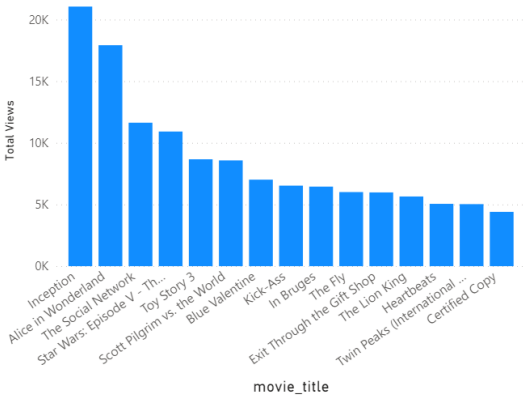

Total Movies By release Year



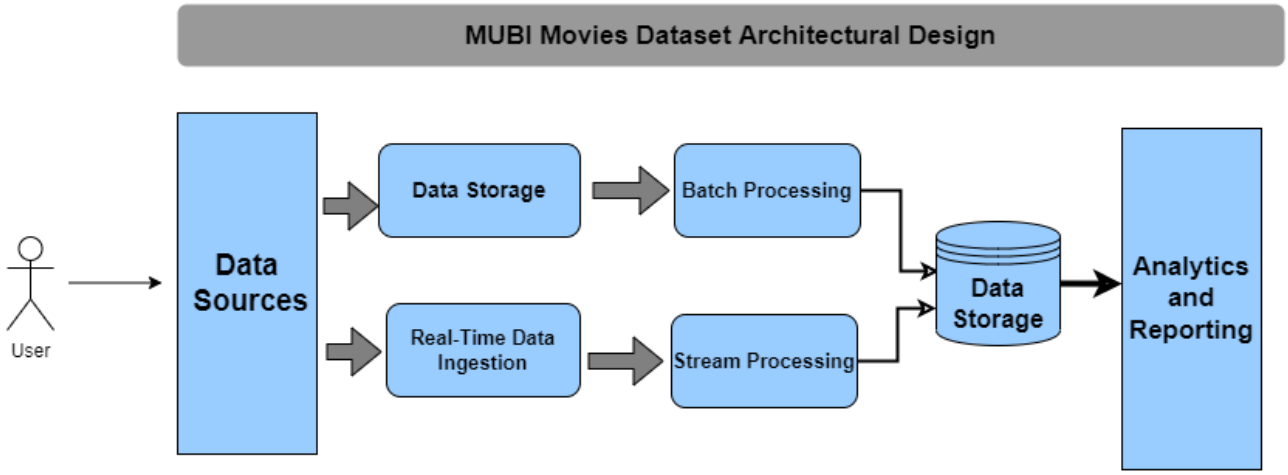

movie_popularity
BY MOVIE_TITLE




Total Movies Views



## Question 2




Big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems. For this question, I have created an architecture for the MUBI Movies data set. The architecture aims to show the flow of data from the movie's dataset.

**Data Sources:**

- The main source of the data is users who watch movies on the MUBI site, the MUBI site has a similar design and works to the famous Netflix movies streaming platforms.
- The other source of the data is from the movie's metadata. This information contains details such as the movie name, language, director name, and the director's URL.

**Real-time and Batch Processing:**

- This data is processed through the streaming and batch processing methods.
- Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats.
- Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis
- The Batch processing that is done on the movie's dataset is aggregating and matching users viewing patterns as well as linking the movie's properties to the Metadata that is already stored.
- The other kind of data process that is conducted in batch processing includes the mapping of users' profiles. This includes merging and checking which users are paying through what platforms. Also, flagging those that are using the platform on a trial account.
- The real-time ingestion of the data happens when users are streaming and rating movies. This information is stored as the activities take place.
- After capturing real-time movie streaming, the architectural solution must process them by filtering, aggregating, and otherwise preparing the data for analysis.

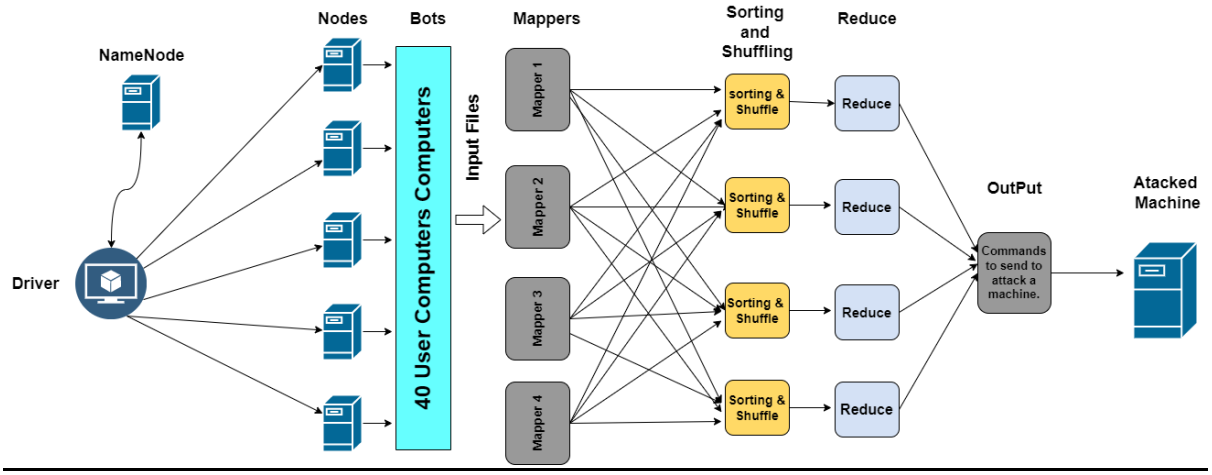
**Data Storage:**

- This is a centralized platform where all the data is stored and saved for easy access.
- The information is stored on separate tables. Some of the tables on the data are used as reference tables, so any form of joining and referencing can be carried out on this part of the architecture.
- Also, in this stage the data is prepared for analysis. The processed data is in a structured format that can be queried using analytical tools.
- The data can be queried through Hive database that provides a metadata abstraction over data files in the distributed data store. Or any form of database tool is used here.

**Analytics Reporting:**

- The goal of this big data solution is to provide insights into the data through analysis and reporting.
- From the reporting end, Microsoft Power BI or Microsoft Excel can be used to visualize and consume the data.
- Python and R can also be used as well.
- This is the part of the architecture that is used by most people in an organization to consume the data and get enabled to make business decisions.

## Question 3




A botnet is a network of compromised computers that are supervised by a command and control (C&C) channel. The compromised computers, or bots, launch attacks designed to crash a target's network, inject malware, harvest credentials, or execute CPU-intensive tasks. A botnet is comprised of 3 main components i.e., The bots, Command and Control Server, and the botnet operator.

A botnet can be used to conduct many types of attacks, such as Phishing, Distributes Denial-of-service (DDoS) attach, crypto-jacking, online fraud, Wide-scale spam attacks, and Spreading malware. Some of the common sources of Botnet attacks are social media, software download, file sharing, and applications. With high-risk devices being those that do not have internet security software and anti-malware software. Some of the effects of botnet attacks include slow computers, high internet bills, and in some instances even legal implications. Some botnet owners use their networks of bots to make extra income by selling to interested parties. For anyone who is part of a botnet, it is sometimes even impossible to notice as everything in the device will be acting as normal.

I will be using a DDoS cyber-attack example to explain how botnets are used with MapReduce to attack and use one's computer without the owner being aware.

A distributed denial-of-service (DDoS) attack is a malicious attempt to impede normal traffic flow through a server, service, or network by flooding the target or its surrounding infrastructure with an excessive amount of Internet traffic (Anon., n.d.). Multiplying the compromised computer systems used as attack sources makes DDoS attacks effective. DDoS attacks are akin to unexpected traffic jams snarling the highway, making it impossible for regular traffic to reach its destination. Exploited machines include computers, as well as other networked resources such as IoT devices.

A DDoS attack is carried out by networks of computers connected to the Internet. A malware-infected network contains computers and other devices (such as IoT devices) that can be remotely controlled by an attacker thanks to their vulnerability to malware. Bots (or zombies) are individual devices, and a botnet is the collection of bots (Anon., n.d.).

From the above diagram, I am depicting a cyber-attack simulation using MapReduce. The simulation uses 5 NameNodes and a single data node. We have also have a Driver Machine.

- **Driver –** This is the machine owned by the attacker. It is equivalent to the Bot that has created the network of computers for the purpose of attacking a hist server.
- **NameNode –** Master nodes manage blocks and control access to files. These nodes are part of the Apache Hadoop HDFS architecture (Bakshi, 2021). They maintain and manage blocks and files processed by the Data Nodes.
- **Data Nodes –** This is the block server that stores data in the local files.
- Bots – The bots are the computers that have been illegally accessed by the Driver server. These bots' resources are being used without the owners being aware.
- **Input Files –** The input file in this example of a DDoS attack, the files will contain the commands that the driver node would want to send to the host server. The commands will also be accompanied by the IP addresses of the machines being attached.
- **Mappers –** The map-reduce mapper is used to process all input records from a file and generate the output that will be used as input to the sort step.
- **Shuffling Sort** – The map-reduce sort takes as input data from the mapper and sorts the files according to the sorting criteria. The necessary shuffling is also performed. It then sends its output to the reduce step. This step is part of the mapper step. The sorting and shuffling happen concurrently in the mapper step of a map-reduce algorithm.
- **Reduce –** This step comes after the data is mapped, sorted, and shuffles. It takes input from the Mapper step. In the example of a DDoS, the data will be reduced and the maximum commands that need to be sent by one bot computer will the output. This will be to ensure that not a lot of commands that could potentially cause visible slowness of the machine are not sent.
- **Output –** The output data is a list of commands that each bot will be sent with the corresponding IP addresses of the machines being attached.
- **Attacked Machine –** In our example of a DDoS, the machine being attacked is the host in which a lot of commands are being sent such that it services are disrupted with the aim to either cause the host server to crash as an example.

**How the BotNet Architecture works:**

From the above architecture, the driver is the main host who is taking advantage of the network of bots to create a DDoS attack. There are 5 Nodes and 1 NameNode which is used to assign and allocate a data note for use in the attack. All the 5 nodes are used, these nodes are connected to 40 Computers in which case they are our bots. Each computer will have an input file that contains the commands and the I of the machine which is being attacked. Each Computer will perform a MapReduce algorithm. This algorithm consists of the mappers, shuffle & sorting, and the reducer step. The output of the map-reduce is the maximum number of commands that a single machine is supposed to send to the attached machine. In this case, the maximum commands are taken as you would not want to cause the bot machine to run slower. The map-reduce portion is done for all the 40 computers. In this diagram, I am only depicting the view that each computer will perform.

## Question 4

Data science is one of the application fields using machine learning algorithms for problem solving. Learning algorithms are at the heart of data science's ability to perform predictive analysis and extract even more insights from data than traditional methods.

Data science is the practice of extracting and communicating valuable and actionable insights from raw data using a set of analytical techniques and methodologies (Pierson, 2017). The field of data science is highly dependent on statistics, mathematics, and computer science. Data science's goal is to optimize processes and support better data-informed decision-making, resulting in an increase in the organization's value (Pierson, 2017). "Data science generates data insights, which are actionable, data-driven conclusions or predictions that can be used to better understand and improve your business, investments, health, and even your lifestyle and social life" (Pierson, 2017).

The demand for data insights is increasing at an exponential rate, every industry is being compelled to embrace data science. Some of the real-life application examples of data science include but are not limited to, Insurance fraud detection, prediction, and prevention of local criminal activities, spam email detection, targeted marking campaigns, recommender systems, and votes analysis, etc. The adoption of data science insights in businesses is akin to seeing in the dark.

The big part of data science success lies with data. If there's no data, there is no analysis or predictive modelling that can be done to improve business performance. W. Edwards Deming even said, "Without data, you are just another person with an opinion.". "Data is generated in every social media interaction we make, every file we save, every picture we take, and every query we submit; it's even generated when we do something as simple as asking a favorite search engine for directions to the closest ice-cream shop" (Pierson, 2017). Big data is defined as a collection of information that is too big for the processing capacity of traditional database systems due to being too large, moving too quickly, or not meeting the structural requirements of traditional database architectures (Pierson, 2017). Nowadays, a large amount of big data is generated by automated processes and instrumentation.

"Although valuable insights can be derived from a single data source, it is often the combination of several relevant sources that provides the contextual information needed to drive better data-informed decisions" (Pierson, 2017). The three Vs of big data are velocity, variety, and volume (McKinsey Global Institute, 2011). These V's are what characterizes big data and are used as the basis of its definition. However, there are additions to the Vs of big data with more research being done. You can only identify the value of your data after it is summarized and analyzed considering the problem being solved. In data science, the quality of the data you have, and its usability are high priorities. Data is a powerful tool of the 21st century. The ability to reveal trends and prevent things from happening simply by looking at data is without a doubt one of the most powerful things in technology today.

Looking back into the world of information action technology, information used to be sourced from limited sources. In the modern world, almost every activity you perform creates data in

one fashion or another. There are various sources of data that yield what is called structured, semi-structured, and unstructured forms of data (Kivenson, n.d.). With the volumes in which this kind of data is generated we now find ourselves seating with what is big data. Which is basically large volumes of data. It is one thing to be able to own and generate data, but it is essential to also be able to use the information generated.

The usability o the large volumes of data lie with data science. As defined above Data science is aimed at curating, using, and ensuring that organizations are able to extract maximum value from their data resources. This field of Data Science relies fully on data for its success. This shows the connectivity between modern data science and big data. Big data can exist without Data Science; however, it is not usable without the methods and application of data science techniques. While on the other hand, Data Science success relies solely on the existence of information.

With the machine and Artificial intelligence techniques. Insights are able to be generated from information such as text, audio, video files, and even pictures. One of the popular applications of data science is face recognition. This is now widely used and adopted on smartphone devices. The use of face detection for security or privacy locking devices is popular. This is able to allow users to use devices seamlessly without the need to capture their passcodes on devices over and over again. The face recognition models were trained on image data, as this is crucial and its ability to work accurately is essential. Lots of image data were needed in order to train and curate the AI techniques. This is an example of how the value from image data is extracted through the use of data science techniques. In this example, we can note that the existence of images was of no value even if this was collected and saved on a database. Through the use of data science, these images were made usable.

From the above examples and detailed definitions of big data and data science, we are able to see the relation between the two fields. Also, the importance and some of the benefits that can be attained from the interlinking of the fields. With the easy-to-understand visualization methods that are presented by the data science fields. Organizations are able to attain value, predict their business gains as well as prepare for any changes that the data insinuates. This is a powerful aspect of how data science is key to getting value out of big data. The application of data science is able to work and extract as much value from any kind of data that could be generated by different sources. Traditionally without the modern data science techniques, this was almost impossible, not every data generated value.

In conclusion, we that modern data science is key to generating and extracting groundbreaking value from any form of data. This has brought about so many changes to the way data is used and treated. We also note how the use and continuous development in the field of data science is now an integral part of how organizations are able to extract value and use the data that they can generate. The future of data science and big data is very promising with so many other developments still being done in the fields, we can expect even greater benefits from these fields in years to come. The way of work is also going to be changed through proper implementations of data science and value extraction on big data.

## Question 5

In this section, I am required to as the head of the Data Division for South African Health Product Regulatory Authority to define the job description of 5 data experts that I would hire. The aim of hiring the experts will be such that they assist in managing the organization of data, protecting users' privacy as well as ensuring the organization extracts maximum value from the data. To do this, I am creating a detailed job description for a Chief Data Officer, Data Engineer, Data Scientist, Data Analyst, and Technical Project Manager.

### 1. Chief Data Officer.

**Job Description:**

As a corporate officer, the chief data officer is responsible for enterprise-wide data management, information analysis, data mining, information trading, and other means by which information is utilized as a strategic asset. CDOs are senior executives who are responsible for the firm's overall data and information strategy, governance, control, policy development, and effective exploitation of data and information. CDOs are responsible for information governance, data quality, data life cycle management, information privacy, and the exploitation of data assets to bring value to the business. Besides developing data procedures and policies, the Chief Data Officer will also collaborate with various departments to gather, organize, protect, and analyze data.

**Job Requirements:**

- **Experience:**
  - 15+ years relevant operational or project-related experience of technology platform management in a data/BI environment.
  - 10+ years of business management of a technical data platform.
  - Experience in the Health Regulatory sector.
  - Good knowledge and understanding of the interaction between data, technology, and business applications for insights and improving decision-making.

- **Academic:**
  - Bachelor's degree in Engineering, Science, or Information Technology.
  - MBA or postgraduate business degree.

**Responsibilities:**

- Reduce costs and redundancies the result from duplication of data functions.
- Leverage opportunities to monetize data and insight.
- Ensure compliance with regulatory and privacy requirements.
- Manage the transition to the modern data platforms and the decommissioning of legacy data platforms to gain the benefits of simplification of the platform environment.
- Actively and continuously evolve the platform technologies, architecture, and standard patterns to enable the business portfolio teams with solutions in a changing business.

## 2. Big Data Engineer

**Job Description:**

Data engineers are responsible for designing, building, and maintaining datasets used for projects involving big data. This requires them to work closely with both data scientists and analysts. With this opportunity comes tremendous technical challenges around ingesting, managing, and understanding high-volume streams of heterogeneous data. The role requires an individual that is experienced in Hadoop and has very excellent coding and troubleshooting experience.

**Job Requirements:**

- **Experience:**
  - 3+ years of industry experience with a proven track record of ownership and delivery
  - Experience developed scalable low-latency, distributed data processing solutions.
  - Excellent verbal and written communication
  - A very strong data analytical skills and database experience.
  - Excellent troubleshooting skills
  - Skills using Presto or HIVE
  - Understanding of Big data concepts and the implementation thereof.
- **Academic:**
  - Bachelor of Science degree in Computer Science (Masters or Ph.D. are advantageous) or related discipline.

**Responsibilities:**

- Solve interesting low-latency, distributed systems challenge to handle our ever-increasing scale.
- Build data processing solutions that are core to our platform.
- Deploy distributed data components and configure jobs to run under those components/services.
- You'll use Hadoop stack to build data pipelines, like SPARK, KAFKA, HIVE, and Presto.

## 3. Data Scientist

**Job Description:**

Data scientists are integral members of the analytics team.For large-scale analyses, these professionals utilize advanced mathematical and programming skills, as well as technologies (such as statistical modeling, machine learning, and artificial intelligence). A data scientist's work is typically focused on informing and guiding data projects. They are also responsible for ensuring that advanced analytics are used and implemented in growing the business. Data mining/data analysis skills, experience using various data tools, and knowledge of building and implementing machine learning models are required of the data scientist.

**Job Requirements:**

- **Experience:**
  - Coding knowledge and experience with several languages: Python, JavaScript, SQL, R, Julia, etc.
  - Knowledge and experience in statistical and data mining techniques: GLM/Regression, Random Forests, Gradient Boosted Trees, NLP, social network analysis, etc.
  - 5-7 years of experience manipulating data sets and building statistical models,
  - Experience working with and creating data architectures.
  - Experience with distributed data/computing tools: Map/Reduce, Hadoop, Hive, Spark etc.
  - Experience in Extraction, Transformation, and Loading (ETL) processes.
  - Experience visualizing/presenting data for stakeholders using: Google Big Query, Google Data Studio, etc.
- **Academic:**
  - Master's or Ph.D. in Engineering, Data Science, Statistics, Mathematics, Computer Science, or another quantitative field.

**Responsibilities:**

- Provisioning of Data Visualizations tools to clearly communicate findings and enable some self-service functionality where suited.
- Predictive and classification models using supervised and unsupervised learning
- Explain and break down complex mathematical/statistical concepts and correlate the effects thereof to real-world scenarios in the business.
- Research and Development of new data analysis technologies and validating their value within our environment.
- Investigate tools for transformation of data into usable formats depending on the use case, paying special attention to real-time vs batch data.
- Leverage business knowledge to create solutions that enable enhanced business performance.

## 4. Data Analyst

**Job Description:**

Data Analysts are the human factor in translating numbers into easy-to-understand outcomes and suggestions. They're required to collect, process, and analyze data for a variety of business concerns ranging from product pricing to employee productivity anything requiring data to make better business decisions. Data analysts use data to perform reporting and direct analysis.

**Job Requirements:**

**Experience:**

- At least 2 years of experience as a data analyst
- Strong experience in either of the following: SQL, R, Python (Pandas, NumPy, Matplotlib), SAS, SPSS
- Experience using and maintaining business intelligence tools.
- Strong statistical skills
- Experience in creating reports that are used for business decision-making.

**Academic:**

- Undergraduate degree in Computer Science, Statistics, Mathematics, or related fields.

**Responsibilities:**

- Designing and maintaining data systems and databases; including fixing coding errors and other data-related problems.
- Mining data from primary and secondary sources, then reorganizing said data in a format that can be easily read by either humans or machines.
- Using statistical tools to interpret data sets, paying particular attention to trends and patterns could be valuable for diagnostic and predictive analytics efforts.
- Demonstrating the significance of their work in the context of local, national, and global trends that impact both their organization and industry.
- Collaborating with programmers, engineers, and organizational leaders to identify opportunities for process improvements, recommend system modifications, and develop policies for data governance.
- The data analyst will drive the value generation that can be acquired from data for business improvement.

## 5. Project Manager:

**Job Description:**

The project manager is responsible for the day-to-day management of the project and must be competent in managing the six aspects of a project, i.e., scope, schedule, finance, risk, quality, and resources. The Project Manager will be responsible for planning, organizing, and directing the completion of specific projects for an organization while ensuring these projects are on time, on budget, and within the scope.

**Job Requirements:**

- **Experience:**
  - At least 5 to 7 years' experience in managing projects in a corporate environment with experience in different types of technology related to banking systems.
  - Experience in implementation of software projects and knowledge of the system development life cycle
  - Experience in operations/ servicing/ contact center capability
  - Knowledge of agile principles
- **Academic:**
  - A technical degree is preferable in Information Technology.
  - A Project Management Diploma is advantageous.
  - Project management certification e.g., PPM

**Responsibilities:**

- Good knowledge of project management theory, and the key areas thereof.
- Ability to grasp concepts of a technical nature quickly, with a great understanding of the underlying business environment.
- Ability to multi-task while managing several projects concurrently.
- Excellent communication skills.
- Ability to manage people, with strong interpersonal and relationship-building skills.

## Question 6

Hadoop is an open-source software platform. It is widely used for storing huge volumes of data and running applications on clusters (Vinod, 2021). In addition to enormous data storage capacity, great computational power, and the ability to handle a variety of jobs and tasks, it also boasts several other benefits. Mainly, it is focused on supporting the rapid growth of big data technologies, which in turn can allow for advanced data analytics, which are crucial in the development of businesses and the value that they can derive from their data.

Platforms or frameworks that help solve the problems associated with big data reside within the Hadoop ecosystem. This system provides components and services that include ingesting, storing, analyzing, and maintaining (Vinod, 2021). Hadoop is primarily intended for processing large volumes of data. In comparison with other systems, its main advantage lies in the fact that it can deal with structured and semi-structured data. The Hadoop ecosystem has several components. For this discussion, I will only at 5 components, and I will then discuss their goals as well as some examples of how they can be applied to a big data project.

The main five core components of Hadoop which I will be discussing include Hadoop distributed file system (HDFS), Yet Another Resource Negotiator (YARN), MapReduce, Apache Hive, and H Base.

1. **HDFS (Hadoop distributed file system)** is the primary storage system of Hadoop. Big data can be stored reliably, scalable, fault-tolerant, and cost-effectively using this Java-based file system (Vinod, 2021). The HDFS NameNode regulates the access to files for clients, while the HDFS DataNode manages the file system's data storage. From these characteristics or goals of HDFS, in a big data project, the ability to regulate clients' access to files is important to ensure the safety and compliance of data protection rules.
2. **Hadoop MapReduce** is a technology that provides data processing. This component provides a framework that enables the data analysis using multiple machines in the cluster.
3. **Hadoop YARN (Yet Another Resource Negotiator)** provides resource management as part of the Hadoop ecosystem. This component is responsible for managing and monitoring workloads. A single platform can handle data stores from multiple data processing engines, for example, real-time streaming and batch processing. The goal of YARN in a big data project would be to ensure that the workload is distributed fairly, and this fair distribution will thus ensure that processes run fast while avoiding processing backlogs. In a real-time processing setting, this will ensure that the processes work efficiently to ensure that data is indeed streamed and processed in real-time.
4. **Apache Hive** is an open-source data warehouse system for querying and analysing large datasets stored in Hadoop files. Its main function is to summarize, query and analyse data. Since big data is a huge amount of data that cannot be processed on normal systems like excel workbooks. In a big data project, Apache Hive is useful for summarising and giving a platform where one can query the big data for a business use case.

5. **Apache HBase** is a component of the Hadoop ecosystem that provides real-time access to read or write data in a Hadoop distributed file system (Vinod, 2021). For a project where data needs to analysed in real-time, this component of the Hadoop ecosystem is very beneficial.

# Bibliography


Anon., n.d. *cloudflare.* [Online]
Available at: https://www.cloudflare.com/zh-tw/learning/ddos/what-is-a-ddos-attack/
[Accessed 23 November 2021].

Bakshi, A., 2021. *Hadoop Distributed File System | Apache Hadoop HDFS Architecture | Edureka.* [Online]
Available at: https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/
[Accessed 23 November 2021].

Eslahi, M., 2012. Bots and Botnets: An Overview of Characteristics,Detection and Challenges. *IEEE International Conference on Control System, Computing and Engineering.*

Kivenson, M., n.d. *Indeed Web Scraping and Analysis.* [Online]
Available at: Rstudio-pubs-static.s3.amazonaws.com
[Accessed 23 November 2021].

Mammunni, S. R., 2020. An overview of botnet and its detection techniques. *IJCRT,* Volume 8.

McKinsey Global Institute, 2011. Big data: the next frontier for innovation, competition and productivity.

Pierson, L., 2017. Data Science for Dummies,. In: *Data Science For Dummies.* New Jersey: John Wiley & Sons, Inc., p. 10.

Vinod, 2021. *Hadoop Ecosystem.* [Online]
Available at: https://mindmajix.com/hadoop-ecosystem
[Accessed 23 November 2021].