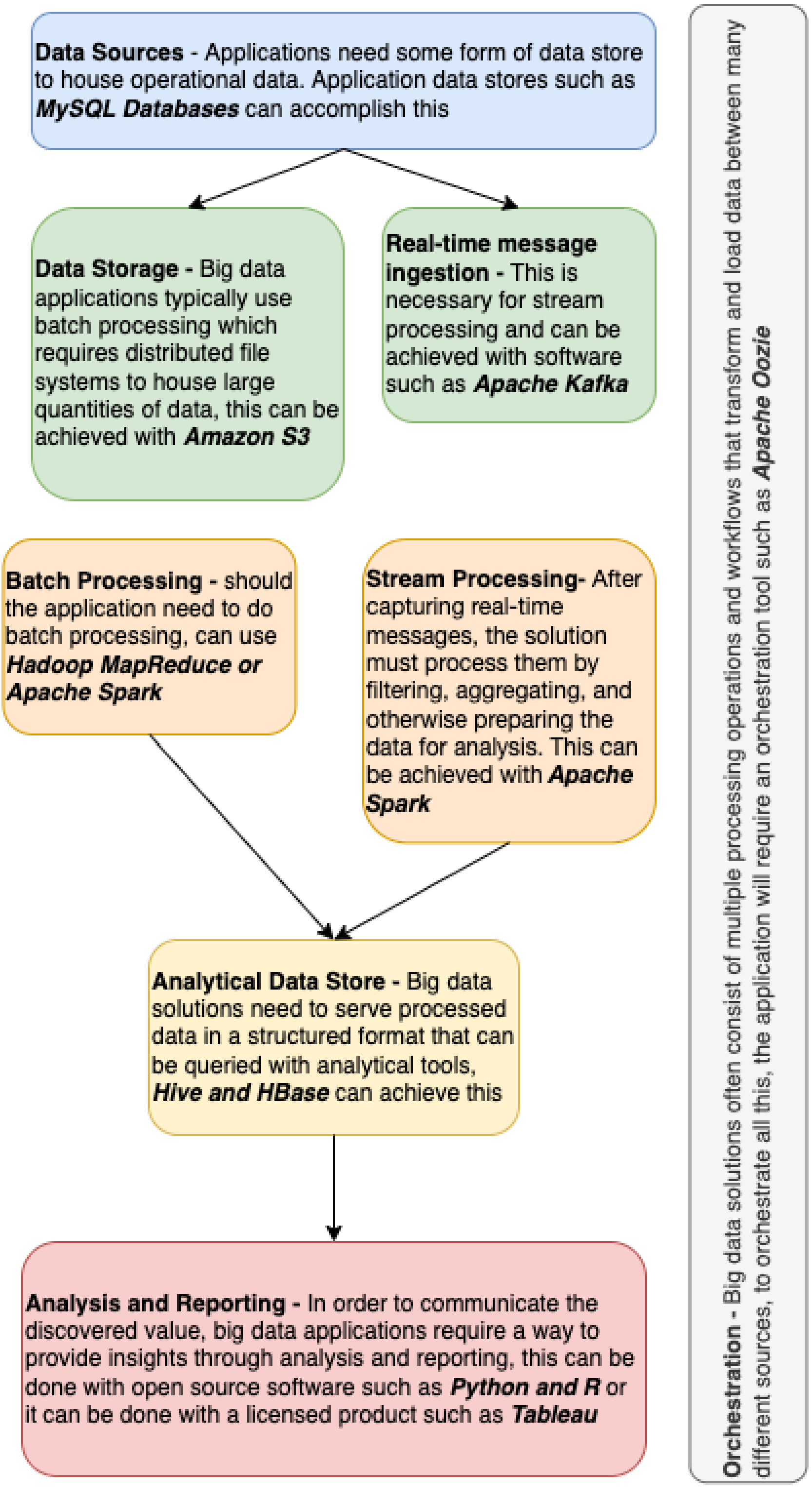
Question 2:



Data Sources - Applications need some form of data store to house operational data. Application data stores such as ***MySQL Databases*** can accomplish this

Data Storage - Big data applications typically use batch processing which requires distributed file systems to house large quantities of data, this can be achieved with ***Amazon S3***

Real-time message ingestion - This is necessary for stream processing and can be achieved with software such as ***Apache Kafka***

Batch Processing - should the application need to do batch processing, can use ***Hadoop MapReduce or Apache Spark***

Stream Processing- After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis. This can be achieved with ***Apache Spark***

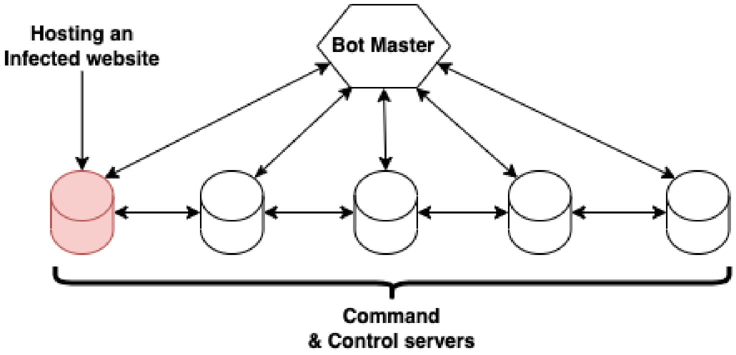
Analytical Data Store - Big data solutions need to serve processed data in a structured format that can be queried with analytical tools, ***Hive and HBase*** can achieve this

Analysis and Reporting - In order to communicate the discovered value, big data applications require a way to provide insights through analysis and reporting, this can be done with open source software such as ***Python and R*** or it can be done with a licensed product such as ***Tableau***

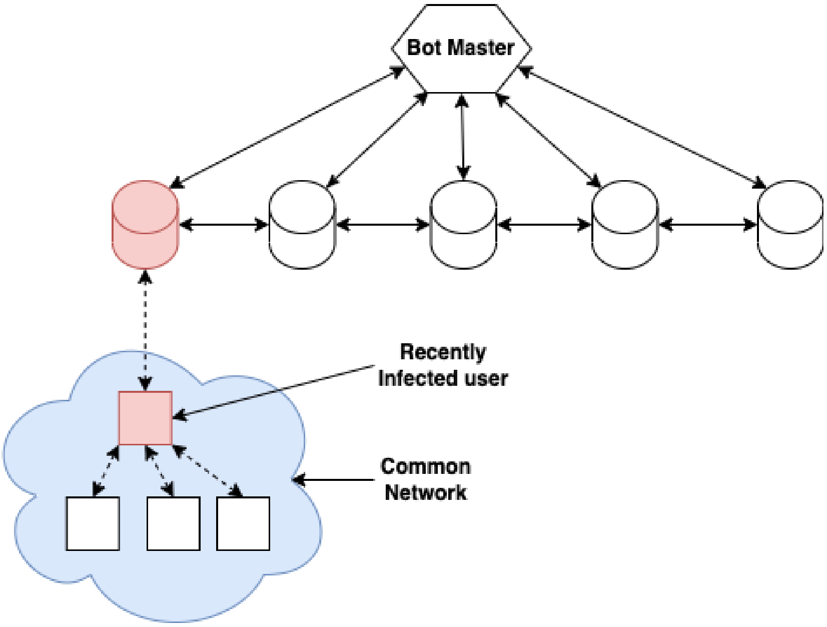
Orchestration - Big data solutions often consist of multiple processing operations and workflows that transform and load data between many different sources, to orchestrate all this, the application will require an orchestration tool such as ***Apache Oozie***

Question 3:

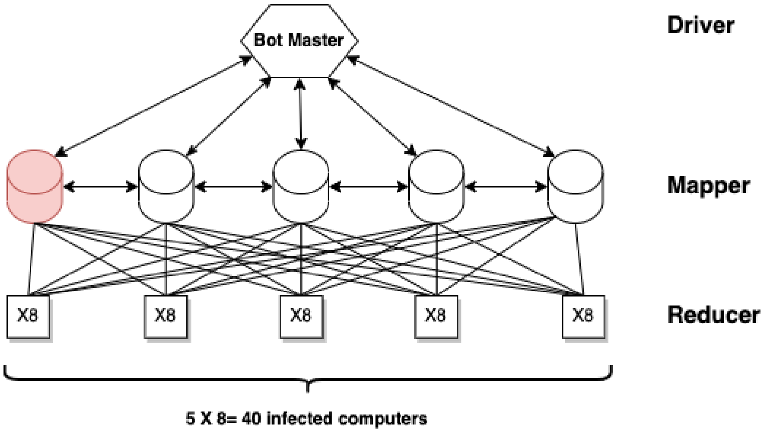
For this question I will simulate the formation of a bot network which executes a DDOS attack via MapReduce. Suppose the Bot Master has 5 command and control servers, these servers can all communicate with each other and each contains a copy of the static IP address of all C&C servers. One of the command and control servers hosts an infected website.



When an unsuspecting user visits the infected site, the botnet stores the users IP address and shares it with the remaining C&C servers, these servers then connect with the infected user and collect the IP addresses of all the other users on the same network as the originally infected user. This process is subsequently repeated until the network has infected 40 hosts.

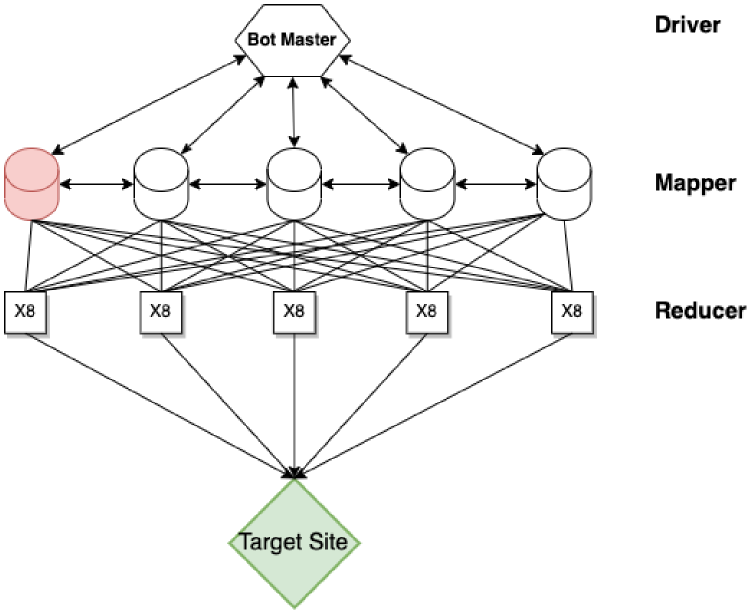


This forms a cluster network of computers on which a hadoop framework can be applied to utilise the cluster for distributed computing, where the botmaster acts as the driver, the C&C servers act as mappers and the infected computers act as reducers.



The DDOS Attack would proceed as follows:

1. The Bot Master initiates the driver class on each of the C&C Servers and sends the link of target website to each of the C&C servers
2. The C&C servers act as Mappers and map this link to each computer as a key value pair structured as follows (key, value) -> (infected users IP address, link to target website)
3. The infected computers act as reducers and execute a command that makes a request to the target website's server
4. If one of the C&C servers fails the others are able to take on its workload as they all have the IP addresses of all 40 computers.



Question 4:

Introduction:

The global production of data is at an all time high, some experts have forecasted that we will produce 74 zettabytes of data in 2021, which is 25% higher than 2020 (CloverDX Blog, 2021). This can largely be attributed to the rise of the internet, smartphones, fitness trackers and IoT devices. Data of this scale often contains vast amounts of hidden insights which can be turned into business value, however, it is incredibly hard to manage, and is often assessed on the "4 V's of Big Data", which relate to the Volume, Velocity, Variety and Veracity of the data. These aspects all demand specialised hardware and technical skills to be able to extract the potential value hidden within the data, and this has led rise to a number of data-focused professions, including, but not limited to: data engineering, data architecture, and data science.

However, the terms "data science" and "big data" are still fuzzy concepts in the minds of many. A Lot of organisations insist that"data science" and "big data" are just buzzwords to throw around in corporate meetings, and that a software engineer more than meets the requirements to act as any of the recent data professions. On the other end of the spectrum, many researchers believe that data science can give a distinct competitive advantage to organisations through various means such as: increasing the chances of making the right decision, improved risk management, opportunity identification and much more (Barham, 2017).

Problem Statement:

Many people do not fully understand what data science entails, how it is connected to big data, and why it is the key to extracting value from big data. In this paper we explore the concept of "Modern Data Science" and "Big Data" and how the two fields interlink to create value within organisations by addressing two research questions, namely:

- What is modern data science and how is it connected to big data?
- Why is data science the key to getting value out of big data?

Objectives:

To adequately address the problems detailed above, four objectives were defined in an attempt to gain a better understanding of the concepts mentioned in the problem statement and how to adequately answer them in a coherent and clear way.

- Define "Modern Data Science"
- Define "Big Data"
- Investigate the challenges faced when attempting to extract value from Big Data.
- Investigate how Data Science can be utilised to extract value from Big Data.

Findings:

To address the first two objective points, the definitions of the terms “Modern Data Science” and “Big Data” were researched. Oracle, a global computer technology corporation and pioneer in the development of modern databases, defines data science as follows: “*Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data*” (Oracle, 2021). Furthermore, the term “Big Data” can be defined as “*data that are too large and/or complex to be effectively and/or efficiently handled by traditional data-related theories, technologies, and tools*” (Cao, 2017). Once these terms were accurately defined, the next two objectives could be investigated.

To understand the challenges faced with extracting value from big data we can circle back to the 4 V’s mentioned in the introduction. Volume is one of the most obvious challenges associated with big data, the size of a data set both in terms of memory usage and number of observations can hinder processing performance if not managed properly. Furthermore, the storage of big data is one of the first things that needs to be addressed when attempting to analyse big data. Velocity refers to the speed at which data is generated, high-velocity data can be generated by the millisecond (such as IoT data) and requires distributed processing techniques to extract value from it. Variety assesses the structural and data variation within big data. Data can come in a wide range of structured and unstructured forms and can create storage and analysis challenges for data scientists. Lastly, veracity refers to data quality and the inherent unreliability of data sources. This creates a number of challenges such as data uncertainty, dirty or noisy data and difficulty in tracing the origins of various data.

The skills associated with the data science profession can adequately address these challenges in the following ways. First, data scientists are able to clean data by filtering out noise and outliers. This is important as uncleaned data cannot be modelled or generalised very accurately. Next, data scientists are able to explore, visualise and mine data to uncover hidden patterns that may be useful in the modelling of the data. Modeling the data follows this step and is where a large portion of the business value is generated, these models are then optimised and the results are validated, for use in business operations. Therefore, the statistical and programming background of a data scientist plays a key role in combining two very different fields in a way that is able to tackle big data and extract true business value from it.

Conclusion:

To summarise, data science is a relatively new and misunderstood profession, however this doesn’t take away from the fact that it can be incredibly powerful in any organisation if done correctly. Data science brings domain knowledge from a wide array of fields and allows organisations to create insights and make sense of huge quantities of operational data that would have otherwise gone to waste. The world is only going to generate more data in the future and the way in which organisations use this data will become critical to their success.

Question 5:

Position 1: Big Data Architect

Responsibilities

- Design and architect end-to-end systems and processes - use analytical skills to meet reliability, scalability, security requirements
- Be forward thinking to identify and build the architecture runway needed to keep our capabilities and infrastructure modern and able to leverage new technologies
- Pair regularly and lead by example, staying current with the technology stack and being integrally involved with the code base
- Apply a test-first mentality and impress upon the team the importance of a healthy suite of tests (unit, integration, smoke tests, etc.)

Qualifications

- Degree in Computer Science or Engineering +5 years industry experience
- Experience leading the design, build and support of enterprise-level Internal or external customer-facing applications
- Experience with enterprise-level ‘Big Data’ or Data Warehousing applications
- Experienced in identifying, investigating, proving out and deploying new technologies

Position 2: Data Engineer

Responsibilities

- Design, build, and launch data pipelines to move data across systems and build the next generation of data tools that generate business insights for a product
- Work closely with data architects to analyse user needs and software requirements to determine workability and to offer support for end users on data usage
- Interface closely with data infrastructure, product, and engineering teams to build and extend cross platform ETL and reports generation framework
- Identify data infrastructure issues and drive to resolution

Qualifications

- Master's degree in Computer Science or Engineering with 3 years experience in similar fields
- Experience with Data ETL (Extract, Transform, Load) design, implementation, and maintenance on a large scale
- Programming in Python, Perl, Java, SQL, or PHP
- MapReduce, Spark, Hive, PIG or other

Position 3: Data Scientist

Responsibilities

- Expertise using data modeling skills to identify key product trends and new product opportunities
- Ability to design implement, and track core metrics to analyze the performance of our products
- Ability to create visuals, dashboards, and reports to effectively communicate your insights
- Ability to collaborate with engineers, product managers, and other cross-functional teams

Qualifications

- Bachelors/Masters degree in Computer Science, Statistics or related field.
- Proficiency in structured and unstructured data processing & analysis using tools like SQL, Spark, R etc.
- Programming experience in Python, C++, C# or similar language.
- Familiarity with ML tools & frameworks such as Jupyter notebook, TensorFlow/PyTorch etc.

Position 4: Data Privacy Manager

Responsibilities

- Understand the impact of privacy laws, regulations and standards/trends across the organization and develop strategies to enhance the privacy program's maturity
- Lead development of privacy program governance components (e.g., policies, procedures, standards, frameworks, trainings, notices)
- Lead initiatives showing how privacy technologies can serve as an enabler for privacy program operations.
- Coordinate with a diverse team to meet their unique needs in a fast-paced environment.

Qualifications

- 5 + years of experience in assessing, designing, building, and implementing privacy programs
- Strong knowledge and awareness of domestic and global privacy laws, regulations, and standards such as the GDPR, POPIA, CCPA etc.
- Experience with Privacy policies, notices, contracts, and clauses
- Ability to conduct Privacy training and awareness sessions/workshops

Position 5: Cyber Security Engineer

Responsibilities

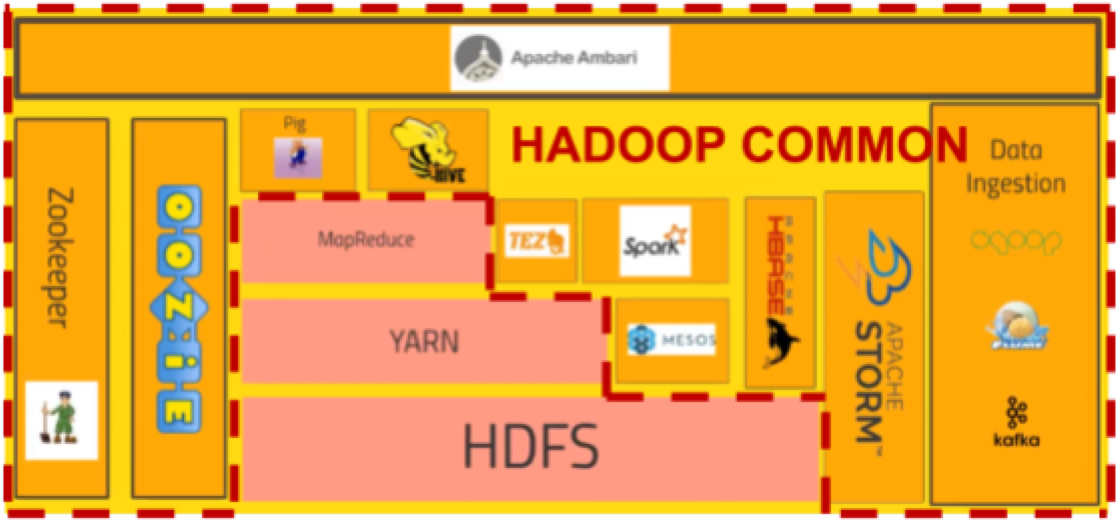
- Monitor server security, perform routine security assessments, and manage software update services
- Daily system monitoring, verifying the integrity and availability of all hardware, server resources, systems and key processes, reviewing system and application logs, and verifying completion of scheduled jobs such as backups
- Build effective monitoring, alerts, and metrics for production security tools and applications
- Ensure system scalability, redundancy, and disaster recovery mechanisms are documented and tested

Qualifications

- Experience in SaaS based software development working with multiple open source projects and frameworks to deliver intricate & scalable product solutions
- Linux or Windows System Administration experience
- Experiencing managing Kubernetes
- Experience managing and maintaining highly scalable and available network services and APIs, particularly with distributed, evented, or highly-parallel systems

Question 6: Five Significant Goals of the Hadoop Ecosystem

The Hadoop ecosystem is composed of 4 components, 3 of these components are part of Hadoop itself (HDFS, YARN, MapReduce), and the last component is composed of a collection of common utilities and libraries that integrate with Hadoop, this collection is typically known as Hadoop Common (see the diagram below). These components have all been designed with specific goals and uses within applications which will be detailed below.



Hadoop Ecosystem

1. HDFS is the "Hadoop Distributed File System" and was designed to distribute the storage of big data across a cluster of computers or servers. One of the main goals of HDFS is to provide fast recovery from hardware failures. In big data applications this is done by replicating data blocks across multiple servers or computers. In the event of a server crash, a backup of the data can be pulled from a different server to proceed with the task at hand.

2. YARN stands for "Yet Another Resource Negotiator" which is primarily concerned with data processing. YARN manages the resources on the cluster and optimises the use of nodes for various tasks within the cluster. YARN is exceptionally useful in big data applications as it is able to adjust to expanding and contracting workload, YARN handles resource management within the application by separating jobs into daemons, allowing the cluster to be scaled without decreasing the performance of the application.

3. MapReduce is a programming model for processing big data sets with a parallel, distributed algorithm on a cluster. MapReduce uses mappers to distribute data across the cluster and reducers to aggregate this data, some big data applications use MapReduce to create recommender systems to recommend products to customers, this is often seen on online stores such as Amazon or Takealot.

4. One of the most popular libraries within Hadoop Common is Apache Spark, Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark is capable of both real time streaming and machine learning with “Spark Streaming” and “Spark MLlib”, respectively. Many organisations utilise Spark as it is easily able to integrate with Python and R which are popular data science and machine learning languages. A big data application could be set up for fraud detection using Apache Spark to create machine learning models capable of detecting fraud with real-time streaming.

5. Apache Hive is a data warehouse software which provides a SQL-style user interface for interacting with data stored in various databases which are distributed on the file system generated by HDFS. The main goal of Hive allows SQL developers to write Hive Query Language (HQL) statements which are very similar to traditional SQL, this enables them to perform data and query analysis. Many organisations use Hive as it is a simple, and cost effective data warehouse option and allows for quick, ad hoc queries on distributed data sets.

References

Barham, H. (2017). Achieving Competitive Advantage Through Big Data: A Literature Review. *Portland international conference on management of engineering and technology*.

Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 4.

CloverDX Blog. (2021, April 23). *How much data will the world produce in 2021?* CloverDX Blog. Retrieved November 23, 2021, from https://www.cloverdx.com/blog/how-much-data-will-the-world-produce-in-2021

Oracle. (2021). *What is Data Science?* Oracle Cloud. Retrieved November 23, 2021, from https://www.oracle.com/za/data-science/what-is-data-science/