Ollscoil na Gaillimhe
University of Galway

Scoil Ghnó agus Eacnamaíochta J.E. Cairnes
J.E. Cairnes School of Business and Economics

EXAMINATION SCRIPT COVER PAGE

Module Name and Code: Data Science and Big Data Analytics and MS5016

Student ID: 22225743

In submitting this script, I am aware that it is my responsibility to adhere to the examination guidelines. Please tick (Yes/No) for the following:

	Yes	No
I have read the module owners' guidelines, description, and expectations for this exam and the submission process.	☒	☐
I have read the examination guidelines.	☒	☐
I am aware of the University Academic Integrity Policy https://www.universityofgalway.ie/media/registrar/docs/QA220-Academic-Integrity-Policy-Final.pdf and confirm the declaration below.	☒	☐
I have saved the files for submission (e.g., .docx, .xlsx) **following strictly the format and naming required** (e.g., 12345678_MS5106_ExamAssignment_PartA.docx).	☒	☐

Declaration for this Exam Submission:

I hereby declare that the work submitted is entirely my own work. It has not been taken from the work of others, except to the extent that such work has been cited and acknowledged within the text of my work. This works is not done in whole or in part by a machine or through Artificial Intelligence, such as ChatGPT. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own.

Overview:

A very successful online retailer, EU-Trade & Co., bases its pricing strategy on evaluating the costs of goods among other online retailers. The business deals with enormous amounts of data and needs to process it effectively.

Eu-Trade & Co. wants to deploy big data technologies to address these problems, and they are aware of tools like Hadoop, Spark, and Clusters. However, they are unsure about the implementation process and worried about the complexity and cost of doing so.

Additionally, all of the data is kept in conventional databases and is accessed by SQL queries. But the company's lack of technical resources means they are in need of guidance on the simplest big data technology.

Introduction to Big Data:

Big data simply refers to the enormous amount of data that organizations collect and use for analytics, visualization, and the development of prediction models. This data has a wider diversity, arrives in higher volumes, and moves at a faster rate. We generated 2.5 quintillion tons of data every day with 6,500 million devices in 2016 and it is projected to increase to 20,000 by 2025, according to **Bernard Marr and Gartner**, respectively. The three Vs—volume, velocity, and variety—are mostly used to describe the type of data. Eu-Trade & Co. is having the same issues because of the volume (size of the data), velocity (speed of data generation), and variety (types of data) of the data sets, which make it impossible for traditional warehouses and data processing software to store and handle them.

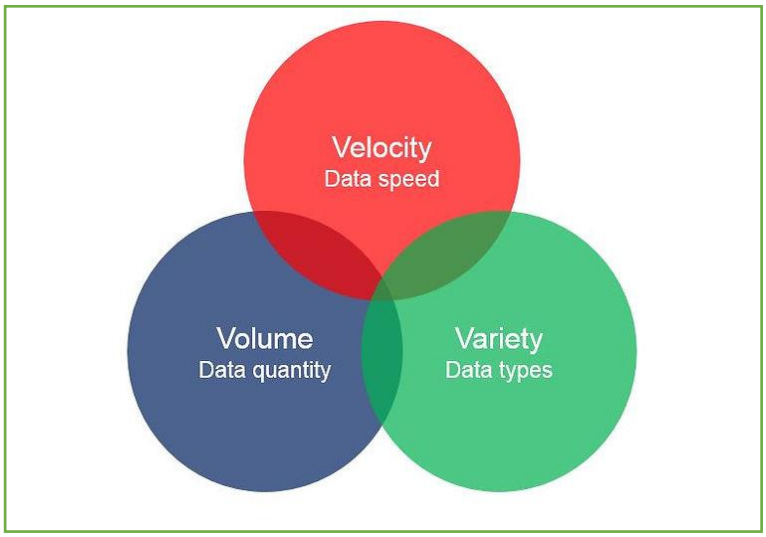

Figure 1: The 3 Vs of Big Data

1. Building and Implementing a Big Data Architecture

It might be difficult to create and implement a big data architecture, but it is crucial for businesses that handle massive amounts of data. Big data frameworks are available, each with its own benefits and drawbacks, and they can all be used to develop a big data platform. These critical functions include meticulous design, execution, and maintenance.

Let's look at the main steps required for the implementation,

1.1. Data Ingestion: Data from several sources are gathered during the intake process. Whether the data is structured or unstructured, it is only worthwhile if it holds important information. All data gathered from online transactions is structured data, claims EU-Trade & Co.

1.2. Data Storage: The data collected in the preceding stage must be stored in a data warehouse, data lakes like HDFS, or cloud-based technologies such as AWS, GCP, and Azure.

1.3. Data Processing: The processing stage entails the translation of data into useful information through operations including filtering, joining, cleaning, and aggregation. After transformation for use further down the line, the transformed data can be stored in the database once more.

1.4. Data Analysis and Visualization: Finally, useful data can be examined to produce metrics that add value to the business, as well as machine learning prediction models or visualization charts using Tableau or Power BI to examine trends and insights that are crucial for making crucial business decisions.

Let's examine the major frameworks listed below with the EU-Trade & Co. challenge as our primary focus.

1.5. Apache Hadoop: The most popular open-source framework for managing massive data sets dispersed over various computer clusters that facilitate local computation and storage is Hadoop. The main components of Hadoop are the **HDFS (Hadoop Distributed File System)** file system for data storage, **YARN (Yet Another Resource Manager)** for resource management, and **MapReduce** for data processing. Processing speed is incredibly quick because Hadoop splits large datasets into manageable bits and distributes them among various nodes (computers) in a cluster.

1.6. Apache Spark: Spark is an open-source solution that is more advanced than Hadoop due to its much faster data retrieval. Spark is a framework for in-memory computing that supports batch processing as well as streaming data (reading data, running operations, and writing results). RDD (Resilient Distributed Dataset) and Data frames, which aid in performing operations on cached data, are the building blocks of Spark.

1.7. Apache Storm: Storm was created primarily for real-time data streaming of big datasets. Apache Zookeeper is used by Storm, a highly scalable, dependable, and fault-tolerant program, to maintain distributed culture state. A Master Node and a Worker Node, which are responsible for monitoring, task distribution, and request processing, respectively, make up Storm, which offers manual scaling. Storm offers high throughput, low-latency processing of streaming data.

1.8. Apache Flink: Flink is a powerful and hybrid cluster-ready framework that supports both batch and stream data processing. With less computing time than Hadoop and Spark due to a feature that lets you impose time boundaries on the dataset to reduce memory issues, but it is a more sophisticated framework. Flink offers both high-throughput batch processing and low-latency stream processing.

To sum up, when building a big data infrastructure, it is important to carefully consider each function's complexity and cost. According to the information provided, while Spark, Storm, and Flink are open source, flexible, and provide more functions, less computing time, and faster data processing than Hadoop, they are complex to set up, difficult to manage, and require significant infrastructure and maintenance costs for deploying and scaling the system.

2. Transition to Hadoop Technologies

Using conventional databases and SQL to process more data is not a practical strategy, according to EU-Trade&Co. The simplest and most straightforward method for issue solving is **HADOOP**. It was established in 2006 by Doug Cutting and Mike Cafarella, and the Apache Software Foundation now oversees its operations.

2.1 How it works: Hadoop is a distributed file system that allows for parallel computation and resource management while dealing with huge and complicated datasets.

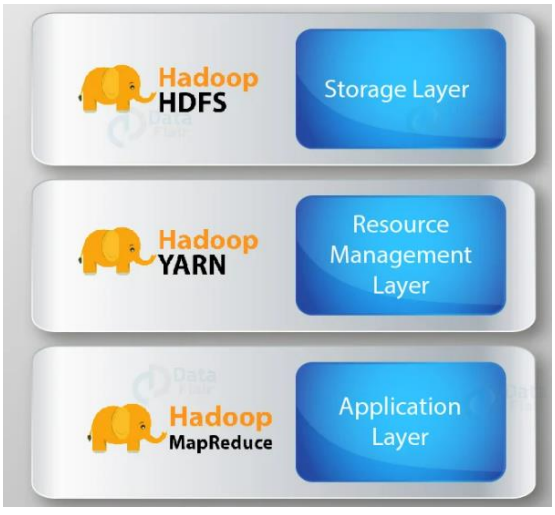

Figure 2: Components of Hadoop

- **HDFS (Hadoop Distributed File System)** is used by Hadoop for storage. The data is divided up into 128MB blocks, and it also has a replication factor set that distributes the same data among several nodes.

- It consists of the Name node, which stores metadata about each data block, and the Data node, which adheres to the master-slave architecture and keeps the underlying data.
- The stored data is processed using the **MapReduce** method, in which the data is read and operations like filtering, sorting, and grouping are carried out in the Mapper phase and sorted and partitioned in the Reduce phase, with the output of both phases being a key-value pair.
- The final and most important component is the **YARN**, which manages resource allocation and task scheduling across nodes.

Apart from the above-mentioned features, there are some more important components like Hive, HBase, Sqoop, Pig, and Spark which EU-Trade&Co can utilize.

2.2. Hive:

For data management, analysis, and querying, Hive is a SQL-like interface built on top of Hadoop. Commonly known as **HQL (Hive Query Language)**, it is a data warehouse concept that is utilized to work on sizable datasets in a distributed setting. When executing commands, it makes use of an interface called the hive command line or Beeline.

Features: It is highly scalable, employs write-once, and read-many principles, offers quicker and easier data retrieval using a metadata repository to store schema, is fault-tolerant, and supports a wide range of file types.

Limitations: It uses MapReduce to execute queries, which may result in delayed execution depending on cluster usage and prevents it from being suitable for real-time data processing. It also only supports OLAP and not OLTP for deleting or updating operations.

2.3. HBase:

HBase, a **NoSQL database** built on top of Hadoop, is used for real-time applications due to its shorter retrieval time. columnar and scalable vertically. HBase, which internally uses hash tables and stores data in tables with several column families, offers a quicker way to access data stored in HDFS.

Features: HBase is schema-less, which speeds up searching and processing compared to traditional databases. It also offers great stability and scalability because to its column orientation and can store both structured and semi-structured data.

2.4. Pig:

A Hadoop-based tool for analyzing massive datasets is called Pig. It uses **Pig Latin**, a high-level language that works with Pig Engine to convert procedural languages into MapReduce jobs. Pig offers enhanced Java and Python user-defined function creation functionality.

Features: Compared to MapReduce, Pig offers a high level of abstraction with less lines of code. It can perform join, filter, and sorting operations, among many other things. It is extremely flexible, scalable, and Hadoop integration is simple.

Limitations: Pig's lack of clear schema development results in human labor and data propagation. When working with large datasets, the script performs worse than Spark because it is compiled into MapReduce.

2.5. Spark:

Similar to Hadoop MapReduce, Apache Spark is a framework for data processing that runs on top of Hadoop, except it performs in-memory data processing, which is much faster. **RDD (Resilient Distributed Dataset)** is used for this, and Mesos is used for resource allocation rather than YARN.

Features: Spark supports a high level of fault tolerance by doing real-time and batch processing of data in memory (RAM). It is dynamic in nature, supports multiple programming languages, including Spark SQL and PySpark, and permits parallel processing.

Limitations: Due to the fact that all operations are carried out instantly, Spark has a high memory need and may have memory issues. The major drawback is the absence of a file management system, which also contributes to the small-file issue when used with Hadoop and slows down performance.

3. Conclusion

In conclusion, the report covers a number of big data technologies, including Hadoop, Spark, Storm, and Flink. It is desirable for EU-Trade&Co to switch from conventional databases to Big Data Tools as data volumes increase. Our research indicates that using tools like HDFS, HIVE, HBase, Pig, and Spark, Hadoop may be utilized to store, analyze, convert, retrieve, and preserve data. Due to the obvious, straightforward, and low-tech nature of these instruments, EU-Trade&Co can expand their operations both economically and financially.

References:

[1] How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read - Bernard Marr (May 21, 2018) https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=41c6747760ba [Accessed 04/20/2023]

[2] Big Data and Data Analytics in ACCA SBL exam – Tommy L (May 16, 2019) https://www.gotitpass.com/post/big-data-and-data-analytics-in-acca-sbl-exam [Accessed 04/25/2023]

[3] Top 5 Big Data Frameworks in 2021- Javin Paul (August 3rd 2021) https://hackernoon.com/top-5-big-data-frameworks-in-2021 [Accessed 04/25/2023]

[4] How Hadoop Works Internally – Inside Hadoop https://data-flair.training/blogs/how-hadoop-works-internally/ [Accessed 04/20/2023]

[5] HBase Pros & Cons - Ankit Kumar (May 13, 2022) https://www.codingninjas.com/codestudio/library/hbase-pros-cons [Accessed 04/25/2023]

[6] Apache Pig Advantages and Disadvantages – Must know for 2022 https://data-flair.training/blogs/pig-advantages-and-disadvantages/ [Accessed 04/28/2023]

[7] Spark vs Hadoop MapReduce - Donal Tobin (Mar 13, 2023) https://www.integrate.io/blog/apache-spark-vs-hadoop-mapreduce/ [Accessed 04/28/2023]