BRICS
Business Council

TEST PROJECT

Machine Learning and Big Data

CONTENT

1 Introduction

The 21st century is an era of information. The rapid development of computer technology has given birth to big data and machine learning, which have greatly promoted the development of human society.

Machine learning is a kind of artificial intelligence that enables computers to learn from experience and improve their performance without being explicitly programmed. It has been widely used in many fields such as image recognition, natural language processing and so on.

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Big data challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy.

The development of machine learning and big data has greatly promoted the development of human society. Machine learning can help us to process data more effectively and make better decisions. Big data can help us to understand the world better and make more informed decisions.

The development of machine learning and big data will have a great impact on the future of human society.

In this project, the competitor will prepare a big data environment required for the task, extract data from multiple sources, transform data into a format and load data into the data warehouse, analysis and visualize data in the data warehouse, then build a machine learning model with Apache Spark to make predictions.

The project is divided into modules, which shall be performed sequentially. The modules' operation shall be assessed along with the process of the project execution. If a competitor does not comply with the HSE requirements or exposes themselves and/or other competitors to danger, they may be disqualified.

Note that the competitor needs to save his/her answers according to the prompts.

2 Description of Project and Task

Each module must be completed within the pre-determined period of time. This is set in such a way as to provide the competitor an opportunity to solve the problem in a deliberate and concentrated manner. The competitor may complete projects ahead of schedule.

Modules and Completion Timeline:

#	Module	Highest Score	Duration (hours)
1	Preparation for a Big Data Environment	**20**	**3**
2	Build a Data Warehouse	**20**	**3**
3	Data Analysis and Visualization	**30**	**3**
4	Machine Learning	**30**	**3**
	Total	**100**	**12**

3 Instructions to the Competitor

3.1 Module 1: Preparation for a Big Data Environment

A big data environment typically includes a HDFS (Hadoop Distributed File System), Yarn (Yet Another Resource Negotiator), MapReduce, Hive, Sqoop, ZooKeeper, Spark, HBase, Flume, and Kafka.

HDFS is a distributed file system that is designed to run on commodity hardwares. Yarn is a resource manager that is responsible for allocating resources to applications running on a Hadoop cluster. MapReduce is a programming model for processing large data sets. Hive is a data warehouse that enables data summarization, ad-hoc querying, and analysis of data. Sqoop is a tool that is used to transfer data from relational databases to Hadoop. ZooKeeper is a coordination service that helps manage a Hadoop cluster. Spark is a fast and general engine for large-scale data processing. HBase is a columnar database that runs on top of Hadoop. Flume is a tool that is used to collect, aggregate, and transfer large amounts of data. Kafka is a message broker that enables message passing between Hadoop applications.

3.2 Module 2: Build a Data Warehouse.

A data warehouse can be built by importing data with Sqoop, preprocessing data with MapReduce, and analyzing data with Hive.

Sqoop is a tool that is used to transfer data from relational databases to Hadoop. MapReduce is a programming model for processing large data sets. Hive is a data warehouse that enables data summarization, ad-hoc querying, and analysis of data.

3.3 Module 3: Data Analysis and Visualization

Data visualization can be used to display the results of data analysis. Python can be used to create visualizations of data that can be used to understand the results of data analysis.

There are many different visualization tools that can be used, such as Tableau, Qlik, and Power BI. In order to visualize data, it is necessary to use Python to display the results of data analysis.

3.4 Module 4: Machine Learning

Machine learning can be performed by collecting data, preparing and visualizing the data, choosing a Model, training the model, evaluating the model, parameter tuning and making predictions.