

|  |  |
| --- | --- |
| **FIRST SET COURSEWORK QUESTION PAPER:** | **Autumn Semester 2019** |

|  |  |
| --- | --- |
| **Module Code:** | **CS7079NI** |
| **Module Title:** | **Data Warehousing and Big Data** |
| **Module Leader:** | Mr. Mukesh Kumar Sah (Islington College) |

|  |  |
| --- | --- |
| **Coursework Type:** | **Individual** |
| **Coursework Weight:** | This coursework accounts for **60%** of your total module grades. |
| **Submission Date:** |  |
| **When Coursework is given out:** |  |
| **Submission Instructions:** | Submit the following to Islington College RTE department |
| **Warning:** | London Metropolitan University and Islington College takes Plagiarism seriously. Offenders will be dealt with sternly. |

© London Metropolitan University

**CS7079 Data Warehousing and Big Data**

The aim of this coursework is to design, implement, and test a data warehouse based on a given business case scenario, to export data from it and ingest it for further processing on a Big Data platform. It is **individual** coursework. The coursework contributes **60%** to the overall CS7079 module mark.

**Coursework Deadlines:** The deadline for the submission of the coursework is **Week 12.**

**Late submission will be penalized according to the University regulation.**

Coursework Specification:

**CASE STUDY SUPER-MAX**

Super-Max was founded in 2009 to provide advance analytical capabilities to provide consulting services for shopping and business houses. The company use artificial-intelligence and machine learning tools to identify, ingest, and analyse traditionally studied factor including sales rate, inventory management, risk mitigation, predicting high saleable commodities. This analytical solution helps business house to increase their profit and optimize the use of their resource.

## The Challenge

One of the Super-Max’s products uses machine learning models to forecast maximum selling item and help the business houses to properly manage their inventory across various geographical location and stores. To have the best outcome Super-Max need a file storage solution with strong scalable and multi-gigabyte throughput to hold data and interim models for a compute scenario that includes:

* The ingestion and analysis of millions of data points, in hundreds of categories. affecting the delivery process of forecast.
* A cluster of almost 300 Amazon Elastic Compute Cloud (Amazon EC2) instances.
* 150 Kubernetes Pod launches per minute, with a steady state of 1,500–2,000.

## Why Amazon Web Services

Super-Max tested Amazon Simple Storage Service (Amazon S3) for data caching and interim model storage under full-load conditions for a month, and compared its performance to both a leading proprietary parallel network file system (pNFS) solution that is often deployed in high-performance computing (HPC) environments, and a traditional network file system (NFS). These were the results:

* Amazon S3 was 60 percent faster than the traditional NFS and 30 percent faster than the proprietary pNFS.
* Amazon S3 provides a common data source for workloads running on a cluster federated across multiple public cloud providers.

## The Benefits

* High Availability: Super-Max can stand up Amazon S3 for data storage and use those stored data for 24/7.
* Massive scalability: From a cold start, Amazon S3 was able to expand to hold 100 TB.
* Fast throughput: Even connected to 100–200 clients, I/O was clocked at 100–200 MB per second.
* More affordable: Amazon S3 not only outperformed the traditional NFS and the proprietary pNFS it was tested against, it also cost less than either one.
* Hiccup-free: Writes to Amazon S3 were as fast as saving on-premises, with zero timeouts and file corruptions.

**Sales Analysis and Inventory Management:**

Sales analysis is one of the most important business processes of the company. Its main goal is the produce a good analytical solution to efficiently manage the inventory for-casting the items which are more likely be on high demand based of seasons and location of the company outlets.

Managers responsible for managing inventory to meet the following reporting and analysis requirements:

* Monthly/Weekly report of all products sales stats across different outlets.
* Profit analysis and sales trend for different outlets.
* Analysis of stock levels by brand or product for different location.
* A daily stock level of all the products.
* Daily and monthly analysis of top selling products.

**Description of Data Sources:**

For this coursework, data source from Super-Max is considered to build a Data Warehouse for ABC Hyper-mart, which has a many outlet across the country. The data source includes the store information with location , daily sales records for various outlets , product details, warehouse location etc. A sample of data will be provided either as flat files or relational tables. It is part of your task to understand the structure of outlets, location, supplies, products and stock movements.

# Tasks:

# Analysis & Design of Dimensional Data Model

**1.1** Identify and define the grain (detail level) of central fact tables to represent the business activities of the sales management business area. **(10 Marks)**

**1.2** Identify dimensions of central fact table based on the available data source that will meet the reporting and analysis requirements mentioned in the business case scenario.  **(10 Marks)**

**1.3** Identify attributes of dimensions and measures of fact tables based on the available data source descriptions that will meet the reporting and analysis requirements mentioned in the business case scenario.  **(10 Marks)**

**1.4** Use simple star schema to define the structure that shows how the fact tables are related to their dimensions. Create a graphical representation of the simple star schemas in one page. **(10 Marks)**

1. **Create a Relational Database to store dimensions and fact tables using any RDBMS.**

**2.1** Create a database. **(5 Marks)**

**2.2** Create dimension tables (**5 Marks)**

**2.3** Create fact tables. **(5 Marks)**

**2.4** Add appropriate primary keys and foreign keys constraints. **(5 Marks)**

1. **Migrate test data from the data warehouse to an Apache Hadoop platform for further analysis of Big Data using Apache Sqoop.**

**3.1** Populate the data warehouse database with some test data.

**(5 Marks)**

**3.2** Export the data warehouse database data into an external data file.

**(5 Marks)**

**3.3** Migrate the data from the Database to HDFS using Apache Sqoop.

**(5 Marks)**

**3.4** Load the data file into HIVE and demonstrate manipulation on loaded data.

**(5 Marks)**

**3.5** Demonstrate the use of Apache Pig for manipulating the loaded data.

**(5 Marks)**

1. **Written Report**

The report, which you will submit, should be well written, structured and well-presented and it must include:

* An introduction section that summarize the objectives of the course work and business case scenario. **(3 Marks)**
* The Analysis and design of a dimension data model. **(3 Marks)**
* The implementation and testing of the data warehouse (include screen shots to show SQL commands and their results). **(3 Marks)**
* The implementation and testing of the Big Data storage on HDFS (including the Apache Pig and their results**). (3 Marks)**
* Provide a personal reflective conclusion of what you have learnt from your

overall coursework. **(3 Marks)**