Academic year 2023/2024
Autumn Term
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. It consists of 2 parts, Part A needs to be completed as **a team work** and Part B is **an individual** coursework. The coursework contributes **60%** to the overall CS7079 module mark.

Coursework Deadlines: The deadline for the Part A is week 6, and Part B is week 12.

Late submission will be penalised according to the University regulation.

Coursework Specification:

CASE STUDY SCENARIO: Blue Sky Online Consumer Electronics Retailer

Blue Sky Online is an online consumer electronics retailer founded in 1990. It conducts online business across UK and Europe via different channels such as Amazon, eBay, Tesco and its own website. Currently, the company offers more than ten thousand products organised in about ten categories and in more than 500 brands.

The dynamic and highly competitive nature of the consumer electronics retail industry means that businesses in this industry are experiencing different decision-making challenges in relation to pricing, targeting high value consumers, consumer satisfaction and product offerings. Retailers are always under huge pressure to invest in new information and communication technologies to gain competitive advantages and optimise business processes.

Blue Sky has a huge customer base across the UK and Europe. Marketing and Customer Relationship managers couldn't utilise the customer base since it is not maintained properly, and some important information are missing. The details of customers and transactions are managed by different applications such as spreadsheet and google docs.

The company is keen to manage and analyse customer interactions to improve its business relationships with customers, facilitate customer retention and drive sale growth. As a data warehouse developer, you are tasked to design dimensional data models and build a data warehouse based on the following reporting and data analysis requirements of the company compiled by a business analyst.

The implemented data warehouse should consolidate customer information and sale transactions into a central corporate repository so that managers can produce reports

and perform analysis that would enable them to provide useful information for their decision-making activities.

Data and analysis requirements:

1. Maintain customer database.

2. Analyse purchase or transaction history based on demographic data such as age group, income group, marital status and postcode/City/Country.

3. Identify customer preferences in terms selling channels/ payment payments/.

4. Identify the best and most profitable customers at different periods (month, quarter and year).

Description of Data Sources:

The data source includes customer details from a relational database and purchased web services, details of selling channels, payment types, text file of generated date and customer transaction. The required data sources will be uploaded to web learn. It is part of your task to understand the structure of the provided data sources using profiling techniques.

Part A: Task 1 below will be completed working as an agile development team and the product will be a design on Conceptual Level; the architecture and the components of the designed architecture needs to be explained clearly and justification needs to be given.

Tasks:

1. Analysis & Design of Dimensional Data Model

1.1 Identify and define the grain (detail level) of a central fact table(tables) to represent the main business activity(activities). Explain the hierarchies, if there are any? **(7 Marks)**

1.2 Identify dimensions of central fact table/s based on the available data source that will meet the reporting and analysis requirements mentioned in the business case scenario. **(7 Marks)**

1.3 Identify attributes of dimensions and measures of fact table/s based on the available data source descriptions that will meet the reporting and analysis requirements mentioned in the business case scenario. . **(8 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 four simple star schemas in one page. **(8 Marks)**

PART B: From the above Conceptual Design, a physical design needs to be implemented as individual work.

2. Create a Relational Database to store dimensions and fact tables using Microsoft SQL Server Management Studio.

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)**

3. Populate the implemented data warehouse based on the provided data sources using SQL Server Management Studio.

3.1 Implement **ETL (extract, transform, and load)** processes to populate dimension tables. (5 **marks)**

3.2 Implement **ETL (extract, transform, and load)** processes to populate a fact table**.** (5 **marks)**

4. Migrate test data from the data warehouse to an Apache Hadoop platform for further analysis of Big Data using Hortonworks Data Platform (HDP)

4.1 Populate the data warehouse database with some test data.
(5 Marks)

4.2 Export the data warehouse database data into an external data file.
(5 Marks)

4.3 Migrate the data file from the file system to Apache HDFS.
(5 Marks)

4.4 Create a suitable data structure for loading the data file into HIVE
(5 Marks)

4.5 Demonstrate the use of Apache Pig for manipulating the loaded data
(5 Marks)

5. Written Report

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

- An introduction section that summarise 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 Pig commands and their results**). (3 Marks)**
- Provide a personal reflective conclusion of what you have learnt from your overall coursework, including the reflection of working in agile environment. **(3 Marks)**