Cataschevastica

Data Warehouse, Analytics and Visualization

Requirements specification document

May 2024

Training Partner:

Contents

Abstract

The Online Transaction Processing (OLTP) Cataschevastica Database DB, which has been designed in the previous assignment, is used for handling transactional workloads in the real-time or near real-time environment of the company. In this project, a Data Warehouse (DW) will be designed for analytical purposes. The trainees will analyse the needs, define the fact and dimension tables, design its structure, implement the Extract Load Transformation (ETL) procedure, perform analytical calculations, provide visualisations and produce managerial dashboards. For the analysis of data from various factories the data will be concentrated in a delta lake managed by Databricks. Finally, the visual reports and dashboards will be designed using the Power BI Desktop tool.

1. Project general description

The Cataschevastica company already has a web application and an OLTP Database. The analysis of this database has been already given in the previous document. While the OLTP database excels in transactional processing, it is not suitable in facilitating complex analytical queries. By separating the concerns of transactional processing and analytical querying, a enables businesses to derive actionable insights from their data. This project deals with the design and implementation of the DW as a solution.

The working team will identify the fact and dimension tables. The fact tables include the measures, i.e. quantitative data or metrics about a business process or activity in the company. Each row in a fact table typically has a surrogate key, which is a unique identifier for that row. This surrogate key is used for joining and referencing dimensional data within the data warehouse. The dimension tables provide the context or descriptive information surrounding the quantitative data stored in the fact table. Dimension tables typically contain attributes or characteristics that describe the dimensions of the business, such as time, products, customers, locations, etc.

The DW is expected to have the star architectural schema, if there is one fact table. If more than one fact tables will be identified, then a constellation (or galaxy) architectural schema will be designed. The normalisation of dimension tables is used if it provides support for drill-down and roll-up operations.

The dimension tables are smaller in size compared to fact tables, as they contain descriptive attributes rather than numerical measures. Also, they will include type-2 supporting retention of historical values. The calendar (date) dimension table will include the Greek public holidays.

It is expected to create a complete BI pipeline, consisting of a DW, ETL packages to automate the creation and loading of the DW, as well as, dashboards which will load data directly from the tabular model based on the DW.

2. Project development roles

For the purposes of this exercise, as a team you will have to share the roles of the Business analyst, Data engineer and Project manager. The instructor has the role of the Product Owner.

3. Functional requirements

- Integration of data from multiple sources such as databases, spreadsheets, flat files, etc.
- Ability to create and maintain tabular models (star, snowflake, constellation schema) to facilitate efficient querying and analysis.
- Extract, Transform, Load (ETL) processes to cleanse, transform, and load data into the data warehouse.
- Hierarchies, measures, and dimensions to represent business entities and their relationships.
- Support for historical data storage to enable time-based analysis.
- Already designed and ad-hoc querying and complex analytical queries.
- Integration to create dashboards, reports, and interactive visualisations.

- Data access controls to ensure that only authorised users can access specific data.
- Collaboration features to facilitate teamwork and knowledge sharing among users.

4. Non-functional Requirements

The following requirements must be met for creating the project:

- Use MSSQL Server for a database.
- Use proper coding and architectural standards and conventions.
- Azure Cloud deployment
- Use of Power BI desktop

Furthermore, the proposed solution will exhibit

- Efficient storage mechanisms to store large volumes of data in a structured format optimised for querying and analysis.
- Fast query processing capabilities to provide timely access to data for reporting and analysis.
- Performance optimization techniques such as indexing, partitioning, and caching to improve query performance.

5. Milestones

The milestones are given as a logical separation of the various tasks and will help us to allocate the various stages of implementation in time. It is not obligatory to observe them as ordered, but it is suggested that they be followed so that there is an indication of the progress of the deliverables.

- Data Analysis and Source Identification
- Design the dimensional model (e.g., star schema, snowflake schema) based on business requirements and data analysis.
- Define data transformations and ETL processes to extract, transform, and load data into the data warehouse.
- Develop ETL workflows and scripts to extract data from source systems, apply transformations, and load data into the data warehouse.
- Develop and deploy data warehouse components such as tables, indexes, partitions, and views.
- Creation of the data lake and the spark scripts.
- Create the power BI file, reports and visualisations
- preparation of presentation with analysis and reflections

6. Deliverables

Your deliverables should include all of the following components:

- SQL scripts to create the DW and load it from the staging area

- SQL scripts to implement the ETL processes that support SCD Type 2 for the dimension tables and incremental loading (delta loading) of new rows for the fact table.
- A .bak file containing a backup of the loaded DW.
- Python Spark script to create a Datalake and perform analytical and visualisation operations.
- A Power BI .pbix file to analyse the fact tables, which will draw data through a live connection with the tabular model you created in the previous step. Visualisation reports will be included in the file.
- A PowerPoint presentation summarising your work, for use during the final presentation of your project.

A Github repository for keeping and sharing your work is recommended but not mandatory.

The deadline for submitting your files is the end of the day before the presentation day. The deadline is an absolute business requirement. The timestamp of the submission files will be used for confirmation.