**Capstone Project**

**Problem Statement**

Assist Superstore in designing and implementing an ETL process to populate a data mart focused on order analytics

**Business Requirements:**

Superstore aims to create a data mart utilizing its orders data to support analytical insights

**Data Understanding**

Data dictionary of the transactional order data:



**Prerequisites**:

All the services discussed in the sessions needs to be configured in your Free Azure subscription account.

**Technical Requirements:**

1. Create a GitHub repo(source), add all the documents related to the capstone project
2. Use the medallion architecture to place the data in the Bronze, Silver and Gold layer in ADLS storage
3. Build Data Processing pipeline using Synapse Analytics service:
   1. Copy data from GitHub repo into Bronze layer
   2. Perform cleaning operations (specified in transformations section) on the input data and load into Silver layer. Use Spark Notebooks.
   3. Gold module will host the dimensional model. Transformations to populate the datasets specified in the ‘Transformation Logic’ section. Use Spark Notebooks.
   4. Create views over all datasets using serverless synapse

**Transformation Logic:**

Silver transformations

While loading data from Bronze to Silver module, perform following cleaning operations:

1. Discard records having blank/empty order\_ID
2. Discard records having blank/empty product\_ID
3. Discard duplicates on order\_ID

Gold transformations

There are three tables in the dimensional model:

1. DIM\_CUSTOMER (dimension table)
2. DIM\_PRODUCT (dimension table)
3. FACT\_ORDERS (Fact table)

Build ETL logic to populate these tables.

In addition, there are some KPIs to be evaluated which are part of FACT\_ORDERS table. Details here:



**General Instructions:**

1. The project must be done by an individual.
2. Queries regarding the project need to be discuss with allotted mentor / SME / Trainer.
3. Design the project as per the problem statement given below.
4. The project evaluation is for 100 marks.

**Submission:**

* Detailed presentation(Fractal PPT template) needs to prepare by taking a screen shot of all the steps mentioned above with your name/id that is present in ADB and ADF on top left corner in a document.
* Also attach the notebook code files (Download the DBC archive file from ADB) in the respective folders you will be creating for submission.
* ipynb, ppt, screenshot, dataset, any relevant document
* System will accept only ZIP file submissions i.e., in .zip format (Max size- 100 MB).
* Review the .zip file before uploading it.
* Please ensure that your submission is complete in all aspects.
* Multiple submission is not accepted.
* We strongly recommend you submit at least 60 minutes before your deadline.
* There will be no extension so please make sure to submit before the deadline.
* Result of capstone project will be declared after the presentation.

|  |  |
| --- | --- |
| Project Start Date | 04-10-2023 |
| Project End Date | 16-10-2023 |
| Project Submission Date | 16-10-2023 |
| Naming Convention for the file | <empid\_firstname\_Capstone\_Project>  F11035\_Kiran\_Capstone\_Project.zip |

Capstone Presentation: -

* Individual viva
* Duration of viva: 30 min
* Mode of Viva : Online
* Date: 17.10.23 to 20.10.23
* Meeting Link & detailed schedule will be shared later.

Assessment Criteria

Participants will be graded on Approach, Solution and Presentation (25%,50%,25%)

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Criteria | | Marks |
| 1 | Approach (25) | Design of the solution | 12.5 |
| 2 | Domain, Azure Services and Technical Understanding | 12.5 |
| 3 | Solution (50) | Best programming practices, Completeness & Readability | 15 |
| 4 | Data Ingestion & Pipeline | 15 |
| 5 | Data exploration | 20 |
| 6 | Presentation (25) | Domain Business understanding | 7.5 |
| 7 | Completeness of presentation | 7.5 |
| 8 | Visualization Approach | 5 |
| 9 | Future Work | 5 |

All the Best!!!