**Data and Visualization Engineer**

**Requirements:**

**Database: Postgres SQL database, Maria DB**

**Python: v3.9 or above**

**GitHub**

**INSTRUCTION:**

You have **strictly 24 hours** to complete a test, start from the date & time receiving this test, and deliver all your files below by **GitHub**

1. SQL scripts to create database schema
2. Python scripts with batch command file
3. Output files
4. Unit test
5. Instruction to execute your application

## 

## Test Question

The company needs to develop a new **Sales** data warehouse in system and to build a **star schema data** mart for reporting purpose, whereas an existing source system of the company can provide 4 data files: Item referential, Customer referential, Order header and Order line.

Of which the requirements of new Sales data warehouse are specified as per followings;

1. Data ingestion to Data Warehouse area shall be implemented by using **pyspark** or **pandas** libraries
2. All date format **MUST** be m/d/yyyy for all files.
3. The function need to execute 2 times based on 2 sets of source data (day1 and day2)
4. All data files must be archived to the archive folder by **python** after data ingestion completed
5. Data transformation to Data Mart area shall be implemented by using **pyspark** or **pandas** libraries
6. Export all tables in Data mart to parquet format files

**Execution step:**

1. Load **day1 data files (4 files inside)** into a new data warehouse

1. Load **day2 data files (4 files inside)** into an existing data warehouse

## Technical Specifications : Data Warehouse

1. **Item referential** table structure is required per following

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Data Type** | **Key** | **Nullable** |
| item\_id | integer | PK | no |
| item\_description | varchar(50) |  | no |
| item\_status | varchar(1) |  | no |

1. **Customer referential** table structure is required per following

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Data Type** | **Key** | **Nullable** |
| customer\_id | integer | PK | no |
| customer\_number | varchar(50) |  | no |
| customer\_name | varchar(50) |  | no |
| address | varchar(100) |  | no |
| postal\_code | varchar(10) |  | no |
| city | varchar(50) |  | no |
| country | varchar(50) |  | no |
| country\_code | varchar(10) |  | yes |
| telephone | varchar(50) |  | yes |

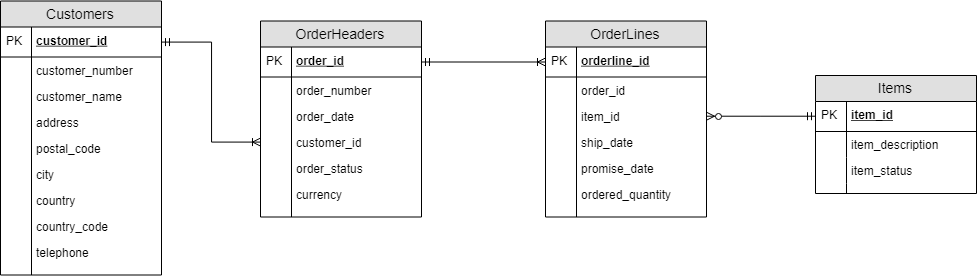
1. **Order header** table structure is required per following

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Data Type** | **Key** | **Nullable** |
| order\_id | integer | PK | no |
| order\_number | varchar(20) |  | no |
| order\_date | date |  | no |
| customer\_id | integer | FK | no |
| order\_status | varchar(10) |  | no |
| currency | varchar(10) |  | no |

1. **Order line** table structure is required per following

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Data Type** | **Key** | **Nullable** |
| orderline\_id | integer | PK | no |
| order\_id | integer | FK | no |
| item\_id | integer | FK | no |
| ship\_date | date |  | no |
| promise\_date | date |  | no |
| ordered\_quantity | integer |  | no |

5. ER diagram of this newdata warehouse schema.



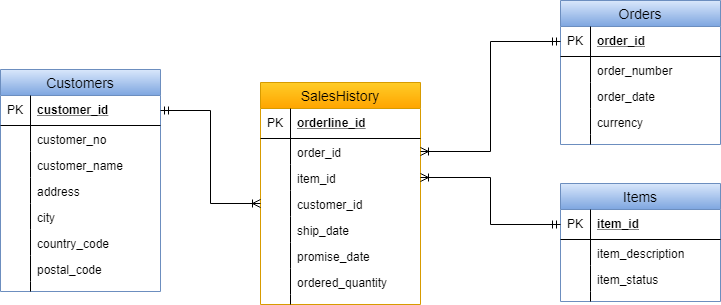
6. No requirement to have data purging or data archiving function at this stage. As the company needs to have tracking information of every new data integration into the table(s) - for auditing purpose (e.g. how many rows are integrated on the time, how long does it take to run.. So forth). Whereas the audit information is expected to be inserted into below **Audit table** ;

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Data Type** | **Key** | **Nullable** |
| id | identity(1,1) integer | PK | no |
| start\_time | datetime |  | no |
| end\_time | datetime |  | no |
| numberrow\_treatment | integer |  | no |
| status | varchar(20) |  | no |

## 

## Technical Specifications : Data Mart

1. Designed sales data mart for reporting tools MUST be followed below diagram



Note:

*The requirement is to keep the Sales data history into a fact table (orange box) and to prepare all related dimensions (blue box). The sourcing data on each table shall be come from the New data warehouse.*

*All dimensions shall be belonged to the* ***Dimensions schema*** *and the fact table shall be belonged to* ***Sales schema.*** *Data type of all fields MUST remain the same as technical specifications of the new data warehouse defined above.*