Data Analyst Assignment by Kotak Sakti

Introduction

As part of the application process for the Data Analyst role at Kotak Sakti, this document addresses a series of questions and tasks designed to demonstrate my skills and understanding in key areas relevant to the position.

The assignment involves a detailed exploration of various aspects of data analysis:

- including understanding the role of a Data Engineer
- executing SQL queries
- mplementing data ingestion and extraction processes using Python and
- orchestrating a data pipeline for a company.

Question 1

Describe a Data Engineer role in an organisation and its main responsibilities A data engineer will often have to handle with multiple types of data to perform many operations using many scripting or coding languages according to the preferences of the organization.

Data engineer will deals with three primary types of data:

- Structured: usually a table-based source system i.e relational database, comma separated file.
- Semi-structured: data that requires flattening prior loading i.e JSON
- Unstructured: data stored as key-value pair or non-relational data i.e noSQL, images, PDFs

Among the key task performed by data engineer includes the following:-

- data integration: establishing link between business operation and data sources

- data transformation: data is transformed into a suitable format through *Extract, Transform and Load* (ETL) process, or through the *extract, load and transform*(ELT) where data is ingested to a data lake and processed using big data techniques. Both ETL and ELT are to support downstream analytical needs.
- data consolidation: is a process to combined extracted data from multiple data sources into one consistent structure to support analytics and reporting - usually done using analytical stores using data lake or data warehouses

Question 2

From the three tables in Appendix A

- a. Show schema generation query
- b. Show the SQL query for number of customers purchasing more than 5 books
- c. Show the SQL query for a list of customers who never purchased anything
- d. Show the SQL query for list of book purchased with the users.

Generated relational database named `bookstore` with 3 tables named as `customers`, `invoices` and `invoiceslines` as showed as the following:-

Question 3

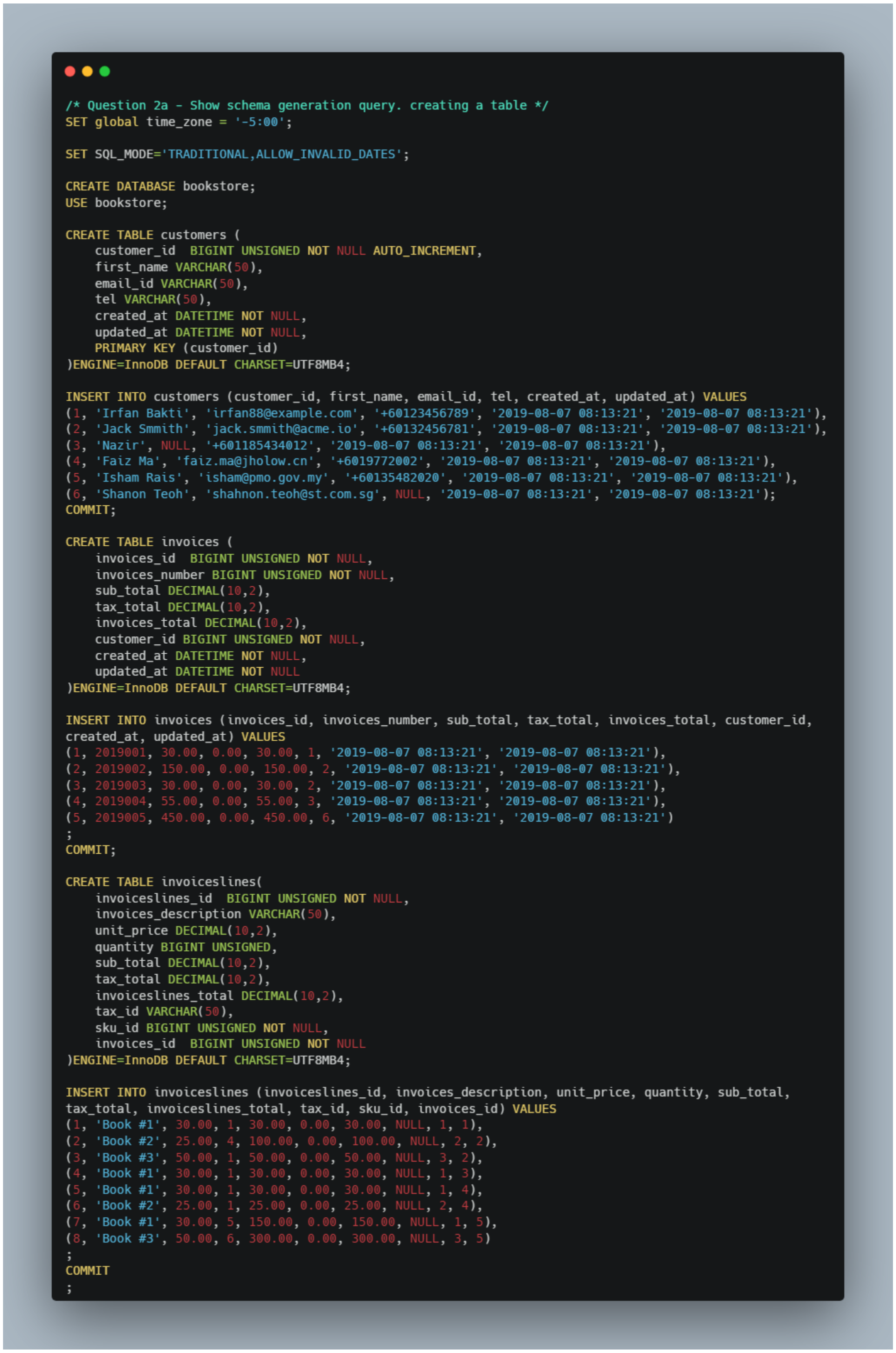
Based on Question 2, implement all queries and the ingestion/extraction process of Appendix A in Python.You can attempt this question in your own development workspace and share GitHub repository or gist URL.

The following link redirect to Jupyter Notebook file, demonstrating the ingestion process: Jupyter Notebook Kotak Sakti Assignment

Question 4

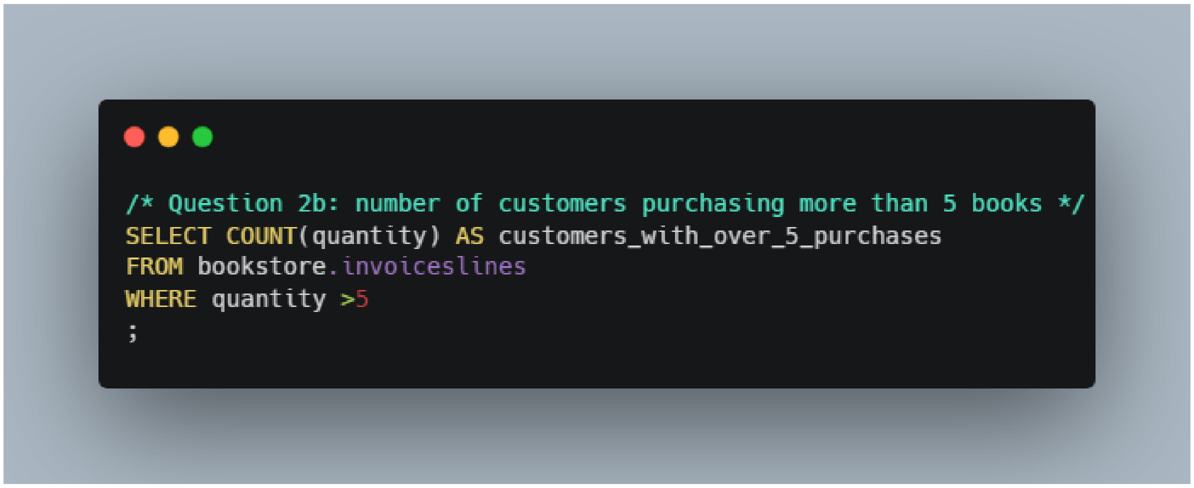
Megah Holdings Berhad is a diversified holding company in a few industries. Each industry has different best practices and different ERPimplementation methods.The revenue optimization team requires a dashboard which displays and analyses daily sales at the end of business day from three business units. Describe a data pipeline with the following input sources:

- a. Retail company ERP - Real time API in XML
- b. Chicken Broiler/Farm ERP - Hourly batch file generation in FTP server
- c. Trading company ERP - Manual Excel files download



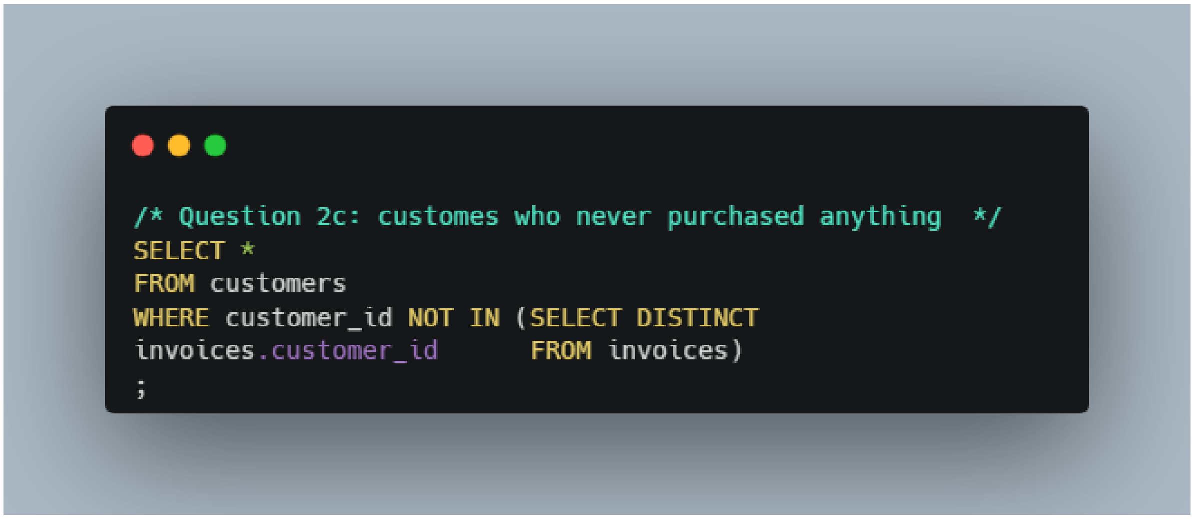
```sql
/* Question 2a - Show schema generation query. creating a table */
SET global time_zone = '-5:00';

SET SQL_MODE='TRADITIONAL,ALLOW_INVALID_DATES';

CREATE DATABASE bookstore;
USE bookstore;

CREATE TABLE customers (
    customer_id  BIGINT UNSIGNED NOT NULL AUTO_INCREMENT,
    first_name VARCHAR(50),
    email_id VARCHAR(50),
    tel VARCHAR(50),
    created_at DATETIME NOT NULL,
    updated_at DATETIME NOT NULL,
    PRIMARY KEY (customer_id)
)ENGINE=InnoDB DEFAULT CHARSET=UTF8MB4;

INSERT INTO customers (customer_id, first_name, email_id, tel, created_at, updated_at) VALUES
(1, 'Irfan Bakti', 'irfan88@example.com', '+60123456789', '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(2, 'Jack Smmith', 'jack.smmith@acme.io', '+60132456781', '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(3, 'Nazir', NULL, '+601185434012', '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(4, 'Faiz Ma', 'faiz.ma@jholow.cn', '+6019772002', '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(5, 'Isham Rais', 'isham@pmo.gov.my', '+60135482020', '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(6, 'Shanon Teoh', 'shahnon.teoh@st.com.sg', NULL, '2019-08-07 08:13:21', '2019-08-07 08:13:21');
COMMIT;

CREATE TABLE invoices (
    invoices_id  BIGINT UNSIGNED NOT NULL,
    invoices_number BIGINT UNSIGNED NOT NULL,
    sub_total DECIMAL(10,2),
    tax_total DECIMAL(10,2),
    invoices_total DECIMAL(10,2),
    customer_id BIGINT UNSIGNED NOT NULL,
    created_at DATETIME NOT NULL,
    updated_at DATETIME NOT NULL
)ENGINE=InnoDB DEFAULT CHARSET=UTF8MB4;

INSERT INTO invoices (invoices_id, invoices_number, sub_total, tax_total, invoices_total, customer_id,
created_at, updated_at) VALUES
(1, 2019001, 30.00, 0.00, 30.00, 1, '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(2, 2019002, 150.00, 0.00, 150.00, 2, '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(3, 2019003, 30.00, 0.00, 30.00, 2, '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(4, 2019004, 55.00, 0.00, 55.00, 3, '2019-08-07 08:13:21', '2019-08-07 08:13:21'),
(5, 2019005, 450.00, 0.00, 450.00, 6, '2019-08-07 08:13:21', '2019-08-07 08:13:21')
;
COMMIT;

CREATE TABLE invoiceslines(
    invoiceslines_id  BIGINT UNSIGNED NOT NULL,
    invoices_description VARCHAR(50),
    unit_price DECIMAL(10,2),
    quantity BIGINT UNSIGNED,
    sub_total DECIMAL(10,2),
    tax_total DECIMAL(10,2),
    invoiceslines_total DECIMAL(10,2),
    tax_id VARCHAR(50),
    sku_id BIGINT UNSIGNED NOT NULL,
    invoices_id  BIGINT UNSIGNED NOT NULL
)ENGINE=InnoDB DEFAULT CHARSET=UTF8MB4;

INSERT INTO invoiceslines (invoiceslines_id, invoices_description, unit_price, quantity, sub_total,
tax_total, invoiceslines_total, tax_id, sku_id, invoices_id) VALUES
(1, 'Book #1', 30.00, 1, 30.00, 0.00, 30.00, NULL, 1, 1),
(2, 'Book #2', 25.00, 4, 100.00, 0.00, 100.00, NULL, 2, 2),
(3, 'Book #3', 50.00, 1, 50.00, 0.00, 50.00, NULL, 3, 2),
(4, 'Book #1', 30.00, 1, 30.00, 0.00, 30.00, NULL, 1, 3),
(5, 'Book #1', 30.00, 1, 30.00, 0.00, 30.00, NULL, 1, 4),
(6, 'Book #2', 25.00, 1, 25.00, 0.00, 25.00, NULL, 2, 4),
(7, 'Book #1', 30.00, 5, 150.00, 0.00, 150.00, NULL, 1, 5),
(8, 'Book #3', 50.00, 6, 300.00, 0.00, 300.00, NULL, 3, 5)
;
COMMIT
;
```

Figure 1: 2(a) Schema generation query



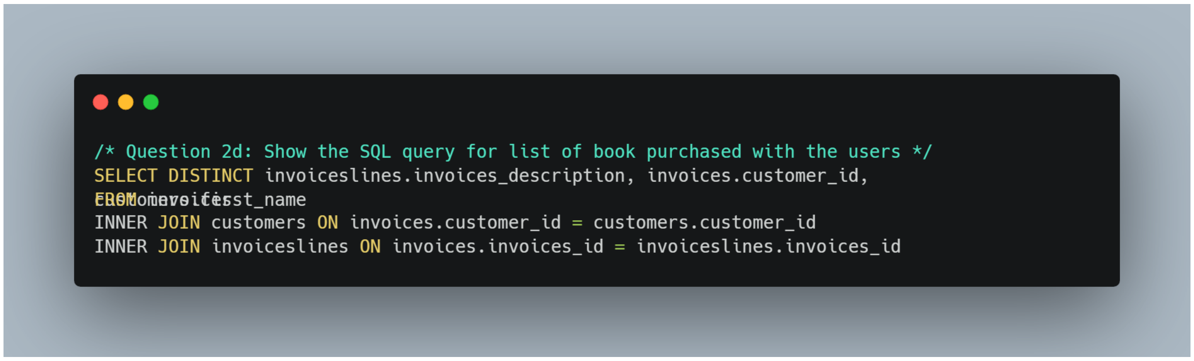
```
/* Question 2b: number of customers purchasing more than 5 books */
SELECT COUNT(quantity) AS customers_with_over_5_purchases
FROM bookstore.invoiceslines
WHERE quantity >5
;
```

Figure 2: 2(b) Number of customers purchasing more than 5



```
/* Question 2c: customes who never purchased anything  */
SELECT *
FROM customers
WHERE customer_id NOT IN (SELECT DISTINCT
invoices.customer_id      FROM invoices)
;
```

Figure 3: 2(c) List of customers who never purchased anything



```sql
/* Question 2d: Show the SQL query for list of book purchased with the users */
SELECT DISTINCT invoiceslines.invoices_description, invoices.customer_id,
customers.first_name
FROM invoices
INNER JOIN customers ON invoices.customer_id = customers.customer_id
INNER JOIN invoiceslines ON invoices.invoices_id = invoiceslines.invoices_id
```

Figure 4: 2(d) List of book purchased with the users.

1. Data ingestion

- For (a): can use Python to consume the XML data from API. As it is semi-structured dataset, Python package `ElementTree` can be used to extract and consume the dataset. The XML format then converted to JSON using `xmltodict` Python The dataset then stored into a dictionary and dataframe, that can be utilised later in the pipeline.
- For (b): The batch file generated can be ingested using Python `ftplib` and can be automated using cron job.
- For (c): for manual file upload, using Python dataframe

2. Data transformation and processing

- For all of the datasets, may follow a *Extract*, *Transform* and *Load* schema. Once extracted according to each dataset, the data is transformed to according schema, quality and business rule. In this Megah Holding, we use csv file schema for later consumption by analyst.
- This transformation can involve cleaning, aggregating, and structuring the data.
- Finally, the transformed data is loaded into the target data warehouse or storage system, often in a format optimized for reporting and analytics.

3. Data storage

- we can export the transformed, structured dataset (i.e csv) into data lake or data warehouses or databases or using cloud solutions i.e Azure Data Factory

4. Data analysis and visualisation

- end-user or data consumers can utilised the analytical storage datasets to create dashboards for the revenue optimisation team.

- Tools like PowerBI or Tableau can connect to data warehouse, data lake or data store that contains the multi-source datasets for visualisation.