Test for DWH / BI Analyst

Instructions

Questions are in English. We will appreciate if you answer in English. Please do not use external information sources (books, internet), use only your current knowledge. Questions are organized into sections and ordered by difficulty (roughly); we start with easy questions and proceed to more complex. You can choose to answer questions in any order. If you do not know the answer, please indicate this by writing N/A.

Section A. Describe Yourself

Evaluate yourself on the scale 1 (basic knowledge) to 5 (excellent knowledge). You can provide additional information. Fill other areas into empty lines, for example other languages.

Area	Knowledge
SQL	
- Postgres	4
- MySQL	2
ETL tools	
1. Manual ETL with Python and SQL scripts. Can build DWH from scratch, schedule and run build scripts and queries	5
2. Airflow	1
3. Talend	1
Reporting / Analytical Tools	
- Power BI (provides the best value for money)	4
- Tableau	2
- Looker	1
Programming languages	
- Python	3
- R	2

What database design principles do you know?

The question is not very specific:
There are following database designs are being used:

1. Snowflake Schema;
2. Star Schema;
3. Mixed Schema;

Snowflake schema is a schema where there is no Central table or driving table mechanism.
Star Schema: There is one central table or driving table and other tables are connected to that table. The Design will look like a star.

A database has to comply with Normalization Rules:
- First Normal Form;
- Second Normal Form;
- Third Normal Form;
- BCNF;
- Fourth Normal Form;

A database has to comply with Integrity Rules:
- Entity Integrity Rule;
- Referential Integrity Rule;
- Business logic Integrity;

Database design is represented by three-layer (or three-schema) architecture:
physical database design:

- conceptual schema => logical schema
- internal schema
- external schema

How large was the largest database that you worked with as developer / analyst (on the IT / delivery side), or user (on the business / client side)? Specify the number of records / dimensions.

The question is not very clear:
- The largest database i have worked with as analyst was over 80 Gb, 20 schemas, ca. 200 tables.
- The largest table i have worked with as analyst 30 columns x ca. 87 mio rows, 30 Gb.

What are the most valuable books / courses have you completed in the last 3 years related to data warehouses and business intelligence?

Courses:
- Data Analytics - Mining and Analysis of Big Data, Alison;
- Machine Learning with Python: A Practical Introduction, eDX;
- Python Programming, Stepik.

Books:
- The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Ralph Kimball, Margy Ross;
- Beginning Databases with PostgreSQL From Novice to Professional, Neil Matthew, Richard Stones;
- Python Crash Course, 2nd Edition: A Hands-On, Project-Based Introduction to Programming, Eric Matthes;
- Automate the Boring Stuff with Python: Practical Programming for Total Beginners,Al Sweigart.

Section B. Practical assignment

<u>Please see file Section_B_MB.pdf</u>

You received the following information in several .csv files:
- List of customers with their personal data
- List of accounts that belong to the customers
- List of transactions from/to these accounts

You need to:
- Model the data, so that all information can be stored in a relational database. The choice of data types, indices and relations is upon your decision
- Provide an SQL script to create such database schema
- Populate the schema with the data from the .csv files.
- Provide a query that returns transactions for the users 345 and 1234, aggregated monthly, sorted by month, for the period from 15.02.2020 till 06.06.2020:

Please note, that these .csv files are extracted manually from the company's data warehouse and they might be malformed. Data inconsistencies might occur.

The resulting report MUST look exactly like shown in the table above in terms of column names and order.

Deliverables:
- [x] Schema creation script (SQL)
- [x] Database population script or ETL project (language or tool of your choice), including all data preparation steps
- [x] SQL query to generate the report