Skills Evaluation Exercise

Dear Candidate

Thank you for applying for the position of Data Scientist. For the purposes of evaluation and in the interests of fairness to determine your level of skill, you are required to complete a written assessment. This assessment contains three (3) sections, which must be completed within 24 hours.

Section 1 - General

1. What is Business Intelligence (BI)?

   BI is refers to technologies, applications and practices for the collection, integration, analysis, and presentation of business information. In other words, BI systems combine data gathering/collection and cleaning, data storage, and knowledge management together with data analysis process to evaluate and transform complex data into meaningful, actionable information, which then can be used to support operational insights and decision-making. To perform BI activities, this requires BI software tools and environments that is consist of a variety of technologies, applications, processes, strategies, products, and technical architectures that be can be used to enable the collection, analysis, presentation, and dissemination of internal and external business information.

   Reference: https://www.omnisci.com/technical-glossary/business-intelligence

2. What is self-service business intelligence (SSBI)?

   This involves the business systems and data analytics that allow and give business end-users (i.e., user of the system) access to an organization's information without direct organizational IT involvement. Hence, SSBI gives end-users the opportunity and ability to do more with their data without necessarily having technical know-how or skills. With SSBI, it allows end-users find it easier and more flexible to analyze their data, make decisions, plan and forecast on their own.

3. What are the different stages and benefits of Business Intelligence?
4.

   Since BI is about collecting, extracting and transforming data into valuable insights while utilizing a wide variety of applications and technologies for collection, storage, and analysis of such data. The different stages of BI include: Data sourcing, Data analysis, Situation awareness, Risk assessment, and Decision support. See below for more explanation;

   Data Sourcing: BI process requires collecting and gathering data, such as texts, images, tables, web pages, and more, from multiple sources like scanners, computer file access, web searches, etc.

   Data analysis: The then collected data goes through the process of data mining or knowledge discovery wherein it is analysed to gain useful insights. Hence, this stage helps to decode current

trends, integrate and then summarize disparate information, validate models of understanding, and predict future trends.

Situation awareness: In this stage, BI helps in filtering out irrelevant or unnecessary information, and ensure that the relationship with the remaining information is established with the context of the organization and its internal processes. Therefore, the key steps here is the information that's based on the relevance to current tasks and the summaries of all the relevant data are taken to understand and make informed decisions.

Risk assessment: In this stage, the BI tools help to decide on plausible actions and decisions that need to be made at different times. That is, this stage help business owner to understand the current analysis and future risk, cost, or benefit of taking any action, assisting them to make the best decisions where possible.

Decision support: This stage of the BI assist business owner to improve on marketing and sales while enhancing customer experience. Furthermore, this stage of the BI helps in making the best use of available information from analysing point of view to making better business decisions.

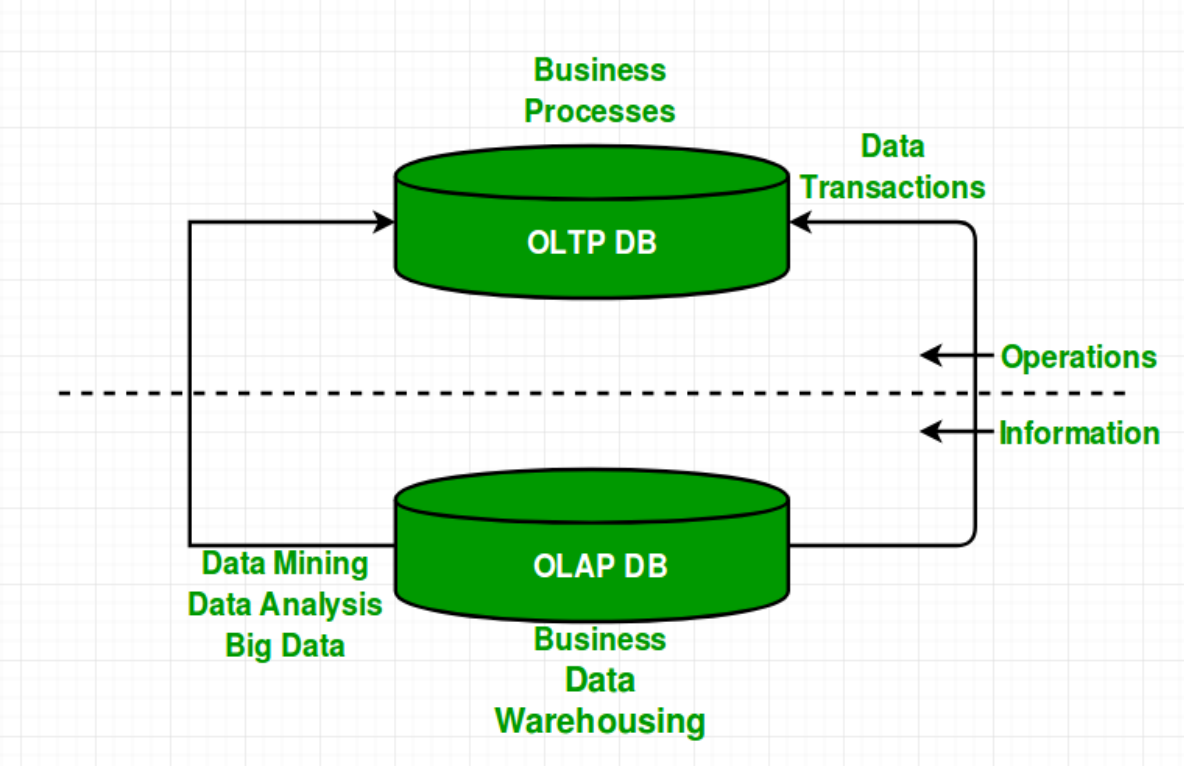
Benefit of Using BI/BI Software tools

I would like to show using a diagram. That is, the diagram show the benefit (advantage) of using BI.

5. List the differences between OLAP and OLTP

To better understand the differences between Online transaction processing (OLTP) and Online analytical processing (OLAP) can be seen in the diagram below. Besides that, OLTP captures,

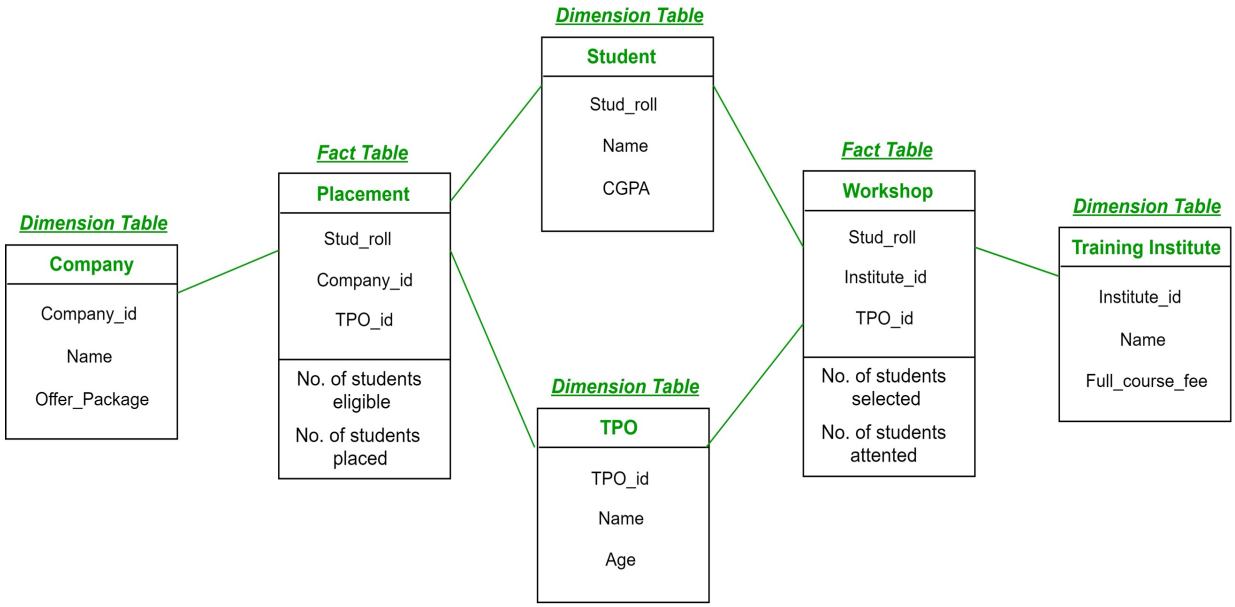
stores, and processes data from transactions in real time while OLAP uses complex queries to analyze aggregated historical data from OLTP systems. In other words, OLTP system is use to capture and maintain transaction data in a database where each transaction involves individual database records that are made up of multiple fields or columns. Examples include Siyenza data activity or retail checkout scanning. Whereas, OLAP applies complex queries to large amounts of historical data, aggregated from OLTP databases and other sources, for data mining, analytics, and business intelligence projects.



Reference: https://www.geeksforgeeks.org/difference-between-olap-and-oltp-in-dbms/

6. Explain Fact and Dimension table with an example.

A fact table is a primary table in a dimensional model. That is, fact table contains business facts or measures, and foreign keys which refer to candidate keys (e.g., normally primary keys) in the dimension tables. In contrary to fact tables, dimension tables contain descriptive attributes or fields that are typically textual fields or discrete numbers that behave like text. Example of dimension tables is when they are joined to fact table via a foreign key. An instance this can be seen in the table below;



7. List the differences between a snowflake schema and star schema.

The following are characteristics of Snowflake Schema:

- The main benefit of the snowflake schema it uses smaller disk space.
- Easier to implement a dimension is added to the Schema
- Due to multiple tables query performance is reduced
- The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

The following are characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are not normalized. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools

Section 2 – Practical exercise - PowerBI

Attached is a spreadsheet with some sample data. This data are daily records of certain indicators from a selection of medical facilities (list of facilities on the second sheet)

From this data please create a PowerBI PBIX with two reports.

1. The first should feature:
    - a graph of values over time
    - values should be limited to [B300 TX_CURR 28 day, B400 Early Missed, B410 Late Missed, and B420 ULTF]
    - the report should have slicers to select one or many of:
        a. months
        b. weeks
        c. districts
        d. facilities
        e. indicators (of the four specified)
2. The second report should feature:
    - A Matrix visual with indicator and facility in rows, And year, week and date in columns and the values of the selected indicators in the body. This needs to drill down on both row and column
    - the report should have slicers to select one or many of:
        a. months
        b. weeks
        c. districts
        d. facilities
        e. indicators

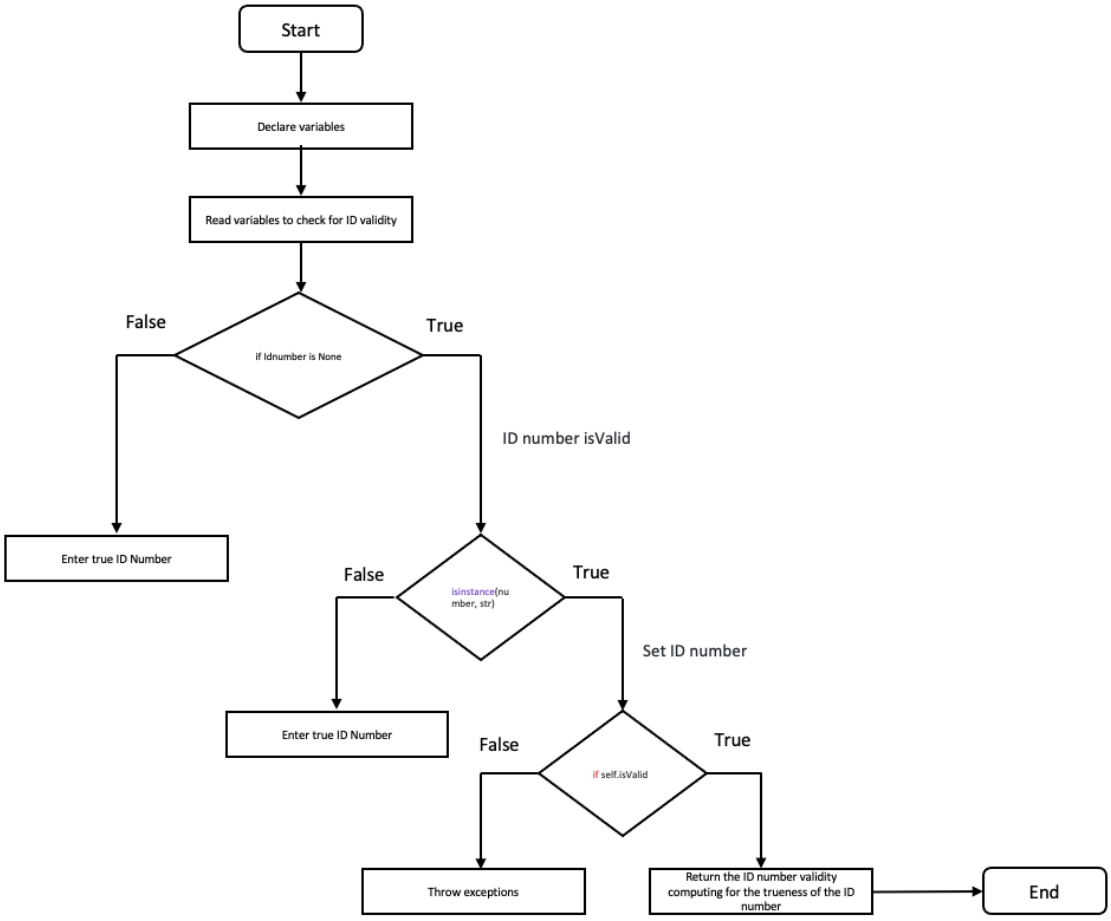
Hints:

- *it will be necessary to pivot the data. This may not be done in the spreadsheet as the source of data is a database link.*
- *it will be necessary to create a calendar table*

SECTION 3 – Database Skills

Problem
Two outputs are expected

1. A logic flowchart, and



2. A SQL procedure.
    a. The procedure will be compiled and tested on a MS SQL database, however you may use whichever flavour of SQL you choose as long as only standard functions are used
    b. The procedure evaluate the id number entered
    c. The DDL of Client demographics table is:

```
CREATE TABLE `clientdataset` (
  `Idnumber` varchar(13) DEFAULT NULL,
  `FirstName` varchar(50) DEFAULT NULL,
  `Surname` varchar(50) DEFAULT NULL,
  `DateOfBirth` date DEFAULT NULL,
  `FolderNumber` varchar(20) DEFAULT NULL,
  `Gender` enum('Male','Female') DEFAULT NULL,
  `RecordNumber` int(11) NOT NULL AUTO_INCREMENT,
  PRIMARY KEY (`RecordNumber`)
) ENGINE=InnoDB AUTO_INCREMENT=16 DEFAULT CHARSET=utf8;
```

d. Validation of ID number is as follows:
    - The number is a numeric field 13 digits long in the form YYMMDDGxxxNRC Where:
        - The first 6 digits shall conform to the date of birth of the subject, in the form YYMMDD where YY = Year in numbers with century removed, MM = Month in numbers, DD day of month.
        - G indicates Gender, where < 5 indicates female and >=5 indicates male
        - xxx is a sequence number in the range of 001 to 999.
        - N is an indication of citizenship where 0 indicates South African, and 1 indicates Foreign.
        - R indicates Race, now no longer used and will generally be 8
        - C is the Mod10 check digit
    - The Mod10 check digit (ISO 2894/ANSI 4.13) is checked in the following manner:
        - The LUHN formula applies some simple arithmetic to a number to calculate a digit that must agree with the check digit, the last digit that appears on the number. Here are the formula's three steps:
        - 1. Beginning with the second digit from the end (on the right), take every other digit and multiply it by two.
        - 2. Proceeding right to left, take each of the digits skipped in step 1 and add them to the result digits from step 1. If the result of doubling a number in step 1 resulted in a two-digit number (such as 7 + 7 = 14), use each of these digits (1 and 4) in adding the digits in step 2.
        - 3. Subtract the result obtained in step 2 from the next higher number that ends in 0. The result must agree with the check digit.

Solution can be find on Github.

https://github.com/boratonAJ/TBHIV_Assessment/upload/main