**Capstone Project**

**Business Background:**

The pinnacles are a huge brand in chocolates & confectionary domain who sell their products on various retailer websites across multiple countries. They collect the search rank data from these retailers to understand the search performance across different retailers, manufacturers, brands, categories, and search terms in terms of different metrics for a given country. The first step is to extract and clean the data using PySpark in Databricks. We need to create a pre-validation script to check the schema of the datasets, to check if the latest data has been uploaded. The search rank data that is collected from the retailers, contains only the products which appear in first page of search. To present the insights on this data at dashboard, we need to fill in the data points for the dates on which product did not appear in the search and set the rank as 100 for these products.

**Data Extraction & Quality:**

Pinnacles has data stored in different data sources such as CSV, Parquet etc. The company wants to extract the data from all these sources and store it in a centralized data lake for future analysis. After the data has been extracted, the next step is to ensure the quality of the data. The pinnacles want to ensure that the data is accurate, complete, and consistent. Create the script to perform the below steps for a given country.

**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 Glod layer in ADLS storage.
3. Create a pipeline in ADF,
   1. Add copytool activity to inject the data from github, place the raw data in Bronze layer.
   2. Create a notebook activity, the notebook activity will read the data from Bronze / Silver L1 layer.
   3. Below mentioned transformations needs to be performed on this data read and they need to be written to Gold layer.
   4. Data transformation logic has been mentioned below. Once the processing in ADB notebooks is complete, use web activity and Logic apps to trigger an email to yourself with a success message.
   5. Once this transformed data is written onto Gold layer, create views using serverless synapse and Read the data from synapse view.
   6. Do the necessary visualization on the transformed data using SQL pool / spark pool.

**Transformation Logic:**

As stated above, you will be reading the data from Bronze / Silver layer, Transforming the data and writing it to Gold layer.

1. Extract data from different data sources such as CSV, Parquet etc, by define the schema for the dataset.
2. Using the above schema read the data and the data frame in the delta format.
3. Verify the schema.
4. Check the datatypes.
5. Cache the dataframe.
6. Verify the first few records.
7. Clean the data by removing duplicates, null values, and invalid data.
8. Check for data accuracy by validating the data against the business rules.
9. Ensure the data is complete by checking for missing values and filling them with appropriate values.
10. Ensure the data is consistent by checking for data type consistency and format.
11. Verify the total number of rows and columns.
12. Verify the summary statistics.
13. Fill in the data for the dates on which product did not appear in the search and set the rank for those dates as 100.   
    E.g. Product with retailer\_product\_id ASIN01 appeared first in search on 12-05-2021 and it appeared in the search till 20-05-2021. On 15-05-2021 product appeared at second page, so we do not have the search-rank data available for this date, so we need to generate the data for this with rank set to 100 specifying products did not appear in the search.
14. Find the maximum and minimum values in each column.
15. Find if there are any duplicate values in the columns.
16. create a table/view on the spark dataframe to run SQL queries.
17. Complete the preprocessing steps by bringing any required analysis and give the insights you have identified.

**Technical Solution Approach:**

1. Extract the data from different sources using PySpark's built-in API.

2. Store the data in a centralized data lake such as Delta Lake, which provides ACID transactions and scalable storage.

3. Create the pyspark script for checking pre-validation rules.

4. Fill in the missing dates data using pyspark.

5.Implement the incremental load approach for fact data

6.Create a partitioned based delta table (Country specific)

Note: Code should be scalable to handle new markets.

**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!!!