Incubation Lab – Pepsico Project Simulation

**Business Requirement:**

A company wants to use the Azure cloud platform to store big data in different formats in a cost-effective way. This data will be stored in different formats which need to be consumed by business users for multiple purposes like descriptive analytics. This pipeline should pick data from a source and ingest it into the data lake. The data lake should contain a landing zone, bronze zone (raw data), silver zone (curated data), and gold zone (aggregated data). The business user should be able to effectively & easily use this data using simple & popular tools like Power BI. As a Data Engineer, you need to design and implement an end-to-end solution that can be scalable, cost-effective, and can be maintained easily.

**Getting Ready:**

**You should be ready with the following:**

* Azure Subscription
* **Important:** Please note there would be a fixed budget of USD 150 per month per member.
* Access to Dataset - [Adventure Works sample](https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver15&tabs=ssms)
* Access to [draw.io](https://www.draw.io/)
* Please go through this link- [What is Data Lake?](https://www.talend.com/resources/what-is-data-lake/)
* Understand the term Lakehouse.
* Please check [Batch ETL with Azure Data Factory](https://databricks.com/blog/2020/03/06/connect-90-data-sources-to-your-data-lake-with-azure-databricks-and-azure-data-factory.html)
* Create Azure Key Vault to manage all connections and secrets used in the project

**Milestone 1:**

**Goal:** Understand the **business requirement** and create a high-level **architecture** on Azure

You need to **perform** the following **steps**:

* Please go through the business requirement
* Identify the appropriate Azure services
* Understand the Sales **schema** of the Adventure Works dataset. Refer this [link](https://i0.wp.com/improveandrepeat.com/wp-content/uploads/2018/12/AdvWorksOLTPSchemaVisio.png?ssl=1)
* Create a high-level architectural diagram using [draw.io](https://www.draw.io/)
* The suggested solution should be scalable, cost-effective, and easy to maintain.

**Milestone 2:**

**Goal**: Set up source database & create data dictionary

You need to **perform** the following **steps**:

* **Import** Adventure Works sample database in **Azure SQL**
* **Analyse** this source database (Sales LT Schema)
* Create a data dictionary in Excel (all the tables, columns, data type, profiling statistics).

**Milestone 3:**

**Goal:** Use ADF to copy data from Azure SQL to Landing Zone to bronze zone

You need to **perform** the following **steps**:

* Create required resources in the data lake for landing and bronze zone.
* Using ADF, **import** data from Azure SQL to Landing Zone in Data Lake
* Import only the **Sales** schema tables in Data Lake.
* Then data from the landing zone will be copied to the bronze zone without any change.
* Test the above pipeline.

**Milestone 4:**

**Goal:** Extend the above ADF pipeline to copy data from bronze to silver zone after transformation using **Databricks**

You need to **perform** the following **steps**:

* Extend the pipeline to create one more step
* In this step, the data will be picked from bronze and transformed/processed using data bricks, and then the output will be written to the silver zone.
* Perform basic level transformation on the data.
* Suggest different options to use data bricks in a cost-effective way.
* Test the above pipeline
* Following transformation to be added:
* Ingest ONLY the required columns
* Can we add some NULL handling transformation for all the columns? Below is the common rule for null handling.

STRING -> NA

INT -> -1

DATE/TIMESTAMP -> 1900-01-01

* Column Renaming transformation
* Type Casting with the required data type in silver layer
* Filtering Rules to be applied
* Add audit columns
* Error handling (bad records option OR different Mode option such as PERMISSIVE, DROPMAL FORMED, FAILFAST)

**Milestone 5:**

**Goal:** Make the pipeline **configuration driven**

You need to **perform** the following **steps**:

* The above data bricks job should **read parameters** from a configuration file
* Perform basic validations (Column level) using Databricks
* Register Delta Table
* Merge the incoming file into the target (Insert new records and update existing ones)
* Calculate column level profiling stats and store it in ADLS as well (with every run)

**Milestone 6:**

**Goal:** Createanother ADF pipeline to copy data from silver zone to gold zone after it goes transformation.

You need to **perform** the following **steps**:

* Design & create a simple **star schema** and **aggregate/reporting tables** in Databricks and Azure Synapse Analytics (based on the source data) – Gold zone tables
* Develop and schedule Databricks notebooks to transform the silver zone tables and load data in Gold zone tables (including Synapse tables)
* Test this pipeline

**Milestone 7:**

**Goal:** Integration with Power BI

You need to **perform** the following **steps**:

* Connect PowerBI with Synapse and Databricks to run basic queries and create a very simple dashboard (parameterized).

**Milestone 8:**

**Goal:** Send Email notification alert

You need to **perform** the following **steps**:

* Extend ADF pipeline to send email notification in case of failure or pipeline completion (all tables loaded).

**Milestone 9:**

**Goal**: Design and develop a **SQL Metastore**

You need to **perform** the following **steps**:

* Design and develop a SQL Metastore to capture and manage:
* Job and Job Run details
* Table Details (which needs to be ingested in the lake)
* Name
* Acquisition strategy (incremental or full)
* Watermark to pull incremental data
* PK column
* merge strategy (truncate & load | Append | PK based upsert

**Milestone 10:**

**Goal**: Add **file-based ingestion**

You need to **perform** the following **steps**:

* Create sample data in the given format in the file and ingest
* This data should also go through a similar flow of transformation
* Please refer to file1, file2 for the schema.

**Milestone 11:**

**Goal**: Peer review of code, perform unit testing, and document the results

You need to **perform** the following **steps**:

* Interchange the code and perform peer review
* End to End Black box testing
* Document the results
* Note down the defects

**Reference Document & Architecture:**

