Project-2:[JANUARY-2024]

GCP,ADF,DATABRICKS,AAD,RBAC,UC,FULL and DELTA LOADS

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Project Introduction [15 min ]**

**Domain:** Telecom

**Project Use Case**

XYZ Corporation, a prominent technology company, recently introduced a new product line in January 2022, alongside other companies in the industry.

However, the company's leadership observed a decrease in both user engagement and revenue growth following the launch in January 2022. The business director at XYZ Corporation has requested the analytics team to generate a comparative analysis report of key performance indicators (KPIs) between the periods before and after the product launch. The management aims to assess the performance during these timeframes and extract insights to facilitate data-driven decision-making for the purpose of improving user engagement and other crucial metrics. Additionally, there is an interest in exploring potential optimizations for the product offerings to enhance user acquisition.

========================================================================

**Prerequisites for project execution [5 min ]**

1.Creation of azure account with valid subscription (Free trial/Pay as you go)

[(2) How To Register Microsoft Azure Free Trial Account - YouTube](https://www.youtube.com/watch?v=Pf99MemtjXw)

2.Azure fundamentals (ADF,ADLS,ADB),PySpark, SQL and more

[Roadmap to become an Azure Data Engineer 2024 (cleverstudies.in)](https://www.cleverstudies.in/courses/Roadmap-to-become-an-Azure-Data-Engineer-2024-658fc2b6e4b084ca8d4a864e)

3.Creating GCP account

[How to create a Google Cloud Platform Account l Create a Free Tier Account on Google Cloud | Edureka (youtube.com)](https://www.youtube.com/watch?v=m5hwU0jD0qc)

========================================================================

**Intro about ADO Activities (Devops) [20 min ]**

Step-1: Creating azure devops organization (CleverStudiesIT) and project (telecom24)  
Step-2: Connect ADO (azure devops organization) to AAD (azure active directory)

Step-3: Create Sprints and User Stories

**========================================================================**

**Sprint-1: Resource creation for Dev (Infra Setup)**

[15 min ] User Story - 1:Creating rsg groups or env's in Azure

Resource Group creation:

<https://learn.microsoft.com/en-us/azure/azure-resource-manager/management/manage-resource-groups-portal>

[45 min ] User Story - 2.Creating Azure resources

Task-1: Creation of Vnet and two Subnets (Private + Public) for adb vnet ingestion.

Task-2: Creation of azure databricks workspace + cluster

Task-3: Creation of azure data factory

Task-4: Configuring Azure GIT version control (Creating Collaboration/dev,feature branch)

Task-5: Creation of Storage account (adls gen2)

Task-6: Creation of Azure Key Vault

Task-7: Creation of Logic App

ADLS gen2 creation:

<https://learn.microsoft.com/en-us/azure/storage/blobs/create-data-lake-storage-account>

Azure DataFactory creation:

<https://learn.microsoft.com/en-us/azure/data-factory/quickstart-create-data-factory>

Azure Databricks workspace creation:

<https://learn.microsoft.com/en-us/azure/databricks/getting-started/>

KeyVault Creation:

<https://learn.microsoft.com/en-us/azure/key-vault/general/quick-create-portal>

Logic App Creation:

<https://learn.microsoft.com/en-us/azure/logic-apps/quickstart-create-example-consumption-workflow>

[15 min ] User Story - 3:Creating required containers

(landing,raw,intermediate,curated,metadata,log,archive) inside storage account

**Sprint-2: Establishing the connectivity (Infra Setup)**

[10 min ] Intro on What is 1.Integration Runtime, 2,Linked Services,3.Dataset

[15 min ] User Story - 1: Creating self hosted SHIR in ADF

[30 min ] User Story - 2: Creating linked services for

Task-1: KEY VAULT

Task-2: ADLS GEN2

Task-3: ADB

[20 min ] User story -3 Create linked services for GCP cloud storage

[10 min ] User Story -4: Create secrets for GCP linked service in Azure Key Vault

[15 min ] User Story - 5:Creating SPN for connectivity between adb to adls gen2

[15 min ] User Story - 6:Setting up Spark Configuration for cluster

[10 min ] UserStory- 7 Creating Secret scope

**Sprint-3 : GCP Source data Ingestion**

[15 Min] Sprint review discussion

[10 min ] User Story - 1:Setting up input data source (Uploading datasets into GCP bucket)

[30 min ] User Story - 2: Creating datasets in azure data factory

Task-1: [ds\_input\_gcp] GCP input dataset(source)

Task-2: [ds\_output\_csv] CSV output dataset(target)

Task-3: [ds\_metadata\_adb\_deltalake] ADB delta lake input dataset for metadata lookup

[10 min ] User Story - 3:Creating folder structure in databricks, setting up input metadata,log tables in databricks hive\_metastore by running the databricks notebooks

[30 min ] User Story - 4:Creation of ADF Pipeline (GCP to Landing) with logging, parameterization

**Checks:**

* File availability
* File name (Telecom)
* File format / extension (.csv)

**1a\_pl\_dim\_gcp\_to\_azure\_full\_load**

**Sprint-4 : Data Enrichment activities and data transformations by using databricks -**

[30 min ] User Story - 1:Creation of ADF Pipeline (GCP to Landing) with logging, parameterization

**Checks:**

* File availability
* File name (Telecom)
* File format / extension (.csv)

**2a\_pl\_fact\_gcp\_to\_azure\_delta\_load**

[15 min ] User Story - 2:

Task-1: Importing databricks notebooks which contains create raw table scripts

Task-2: Creating tables in raw schema by running task-1 imported notebook

[20 min ] **Cleaning activities:**

* Bad Records Handling | Permissive
* Adding derived columns like load\_id, last\_inserted\_dttm\_azure, last\_updated\_dttm\_gcp
* Removing record level null values
* Duplicate checks

[30 min ] User Story - 3:Creating ADF PL (Landing to raw) with logging,parameterization

(Full Load)

**1b\_pl\_dim\_sub\_landing\_to\_raw\_full\_load**

[30 min ] User Story - 4:Creating ADF PL (Landing to raw) with logging,parameterization

(Delta Load)

**2b\_pl\_sub\_fact\_landing\_to\_raw\_delta\_load**

**Sprint-5 : Developing business logics by using multiple input datasets, logging, archiving**

[15 Min] Sprint review discussion

[20 min ] User Story - 1:Archiving the source files (Moving Landing to Archive folder) after raw layer table ingestion.

[15 min ] **Cleaning activities:**

* Merge or Upsert
* Incremental load
* Full load

[10 min ] User Story - 2:

Task-1: Importing databricks notebooks which contains create intermediate table scripts

Task-2: Creating tables in intermediate schema by running task-1 imported notebook

[30 min ] User Story - 3:Creating ADF PL (raw to intermediate ) with logging,parameterization

**1c\_pl\_dim\_raw\_to\_intermediate**

[30 min ] User Story - 4: Creating ADF PL (raw to intermediate) with logging,parameterization

**2c\_pl\_fact\_raw\_to\_intermediate**

**Sprint-6 : Creating triggers, email alerts/Unit Testing**

[10 Min] Sprint review discussion

Implementing business logics

* Joins
* Merge
* CTE

[30 min ] User Story - 1: Creating ADF PL (intermediate to curated) with logging,parameterization

**3\_pl\_intermediate\_to\_curated**

[20 min ] User Story - 1:Configuration of workflow in logic app for email alerts

[30 min ] User Story - 2:Email alerting (4 templates)

Task 1:Email template-1: Count mismatch email alert

Task 2:Email template-2: ADF PL In Progress email alert

Task 3:Email template-3: ADF PL error email alert

Task 4:Email template-4: ADF PL Completion email alert

[15 min ] User Story - 3: Creating Triggers, running adf pipeline for loop

[15 min ] User Story - 4: Unit testing and end to end ADF PL’s run through Triggers

**Sprint-7: UNITY CATALOG CREATION -**

[15 Min] Sprint review discussion

[15 min ] Intro on Unity Catalog  
[15 min ] User story -1: Creating Metastore

[45 min ] User story -2: Creating Unity Catalog

Task-1: Create databricks access connector

Task-2: Create external credentials

Task-3: Create external locations

Task-4: Create Schemas

**Sprint-8: UC SYNC & RBAC set up**

[30 Min] User story -1: Create scripts to SYNC tables created in Hive Metastore to Unity Catalog

[30 Min] User store-2: RBAC

Task-1: Create required Groups in Databricks workspace

Group\_Metastore\_Admin

Group\_Workspace\_Admin

Group\_Other

[30 Min] Task-3: Assign groups to respective schemas ,Tables & test users

**Sprint-9 Session: Prod deployment**

[30 Min] User Story - 1: Creating prod environment [resource groups, other azure resources]

[30 Min] User Story - 2: Creating SHIR’s for prod

[30 Min] User Story - 3: Creating and executing Devops ADO Pipeline to deploy Databricks deliverables

[30 Min] User Story - 4: Creating and executing Devops ADO Pipeline to deploy ADF deliverables

**Session-10: Project Q&A** [120 Min]