Azure Project (ADF+DBX+CI/CD) Execution Plan:

****

Project Use Case:

[Client Name] is a FMCG manufacturer company in India. It is currently operational in a few cities in India. They want to expand business to other cities as well.

[Client Name] is currently facing a problem where a few key customers did not extend their annual contracts due to service issues. It is speculated that some of the essential products were either not delivered on time or not delivered in full over a continued period, which could have resulted in bad customer service. Management wants to fix this issue before expanding to other cities and requested their supply chain analytics team to track the 'On time' and 'In Full' delivery service level for all the customers daily basis.

The Supply Chain team decided to use a standard approach to measure the service level in which they will measure 'On-time delivery (OT) %', 'In-full delivery (IF) %', and OnTime in full (OTIF) %' of the customer orders daily basis against the target service level set for each customer.

Project Introduction

==========

Prerequisites for project execution

(YouTube video links for each activity given below for reference)

1.Creation of azure account with valid subscription (Free trial/Pay as you go)
(2) How To Register Microsoft Azure Free Trial Account - YouTube

2.MySQL/SQL Server Setup in local machine or laptop
https://youtu.be/GwHpIl0vqY4?si=yyys7PAZYdZLiKK-

3.SFTP Server Setup in local machine or laptop
Azure SHIR and SFTP Server setup | Azure Real Time Data Engineering Project | ADF | Databricks

4.Azure fundamentals (ADF, ADLS, ADB), PySpark, SQL
https://www.youtube.com/@WafaStudies

==========

Intro about ADO Activities (Devops)

Step-1: Creating azure devops organization (CleverStudiesIT) and project (fmcgnov23)
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)

User Story - 1: Creating resource groups or env's in Azure

Resource Group creation:
https://learn.microsoft.com/en-us/azure/azure-resource-manager/management/manage-resource-groups-portal

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
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 Data Factory 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/
Key Vault 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

User Story - 3: Creating required containers
(Landing, bronze, silver, gold, metadata, log) inside storage account

==========

==========

Sprint-2: Establishing the connectivity (Infra Setup)

Intro on What is 1. Integration Runtime, 2,Link Services,3.Dataset

User Story - 1: Creating self-hosted IR for on prem SQL server and SFTP
User Story - 2: Creating linked services for
Task-1: KEY VAULT
Task-2: SFTP
Task-3: MYSQL
Task-4: ADLS GEN2
Task-5: ADB
User Story - 3: Create secrets for MySQL and SFTP Credentials
User Story - 4: Creating SPN for connectivity between adb to adls gen2
User Story - 5: Setting up Spark Configuration for cluster

==========

Sprint-3: Source data Ingestion

User Story - 1: Setting up input data source
Task-1: Create and load source tables in MySQL database

YouTube link: https://youtu.be/ftlJoXEBmis?si=UiYjpFk0PvYRFYRb

User Story - 2: Creating datasets in azure data factory
Task-1: [ds_input_mysql] MySQL input dataset(source)
Task-2: [ds_output_parquet] PARQUET output dataset(target)
Task-3: [ds_metadata_adb_deltalake] ADB delta lake input dataset for metadata lookup

User Story - 3: Creating folder structure in databricks, setting up input metadata, log tables in databricks hive_metastore by running the databricks notebooks

User Story - 4: Creation of ADF Pipeline (MySQL to Landing) with logging, parameterization

1a_pl_source_mysql_ingestion

Short discussion on adf git versioning
Sprint review discussion

==========

==========

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

User Story - 1: Setting up input data source
Task-1: Placing source files(.csv) in SFTP directory

User Story - 2: Creating datasets in azure data factory
Task-3: [ds_input_sftp] SFTP input dataset(source)
Task-4: [ds_output_csv] CSV output dataset(target)

User Story - 3: Creation of ADF Pipeline (SFTP to Landing) with logging, parameterization
1b_pl_source_sftp_ingestion

User Story - 4: Creating required directories inside the container, creating tables in bronze schema
User Story - 5: Creating ADF PL (Landing to Bronze) with logging, parameterization (Full Load)-MySQL Source
2a_pl_mysql_landing_to_bronze

User Story - 6: Creating ADF PL (Landing to Bronze) with logging, parameterization (Full Load)-SFTP Source
2b_pl_sftp_landing_to_bronze

User Story - 7: Removing headers from the source file, if it comes as a column values, adding derived columns (timestamp, load_id) and taking inserted data count.
Creation of budget

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

Sprint review discussion
User Story - 1: Archiving the source files (Moving Landing to Archive folder) after bronze layer table ingestion.
- Task 1. Creation of silver, gold layer tables.
- Task 2. Import silver, gold ingestion scripts to databricks environment.

User Story - 2: Creating ADF PL (Bronze to Silver) with logging, parameterization
3_pl_bronze_to_silver

User Story - 3: [WALKTHROUGH] Data normalization (by using the joins), data validation (duplicate check), data types updating etc.
User Story - 4: Creating ADF PL (Silver to Gold) with logging, parameterization
4_pl_silver_to_gold

==========

==========

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

Sprint review discussion
User Story - 1: Configuration of workflow in logic app for email alerts
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

User Story - 3: Creating Triggers, running adf pipeline for loop parallelly (via scheduling)
User Story - 4: Unit testing and end to end ADF PL's run through Triggers

==========

Sprint-7: Prod deployment

User Story - 1: Creating prod environment [resource groups, other azure resources]
User Story - 2: Creating SHIR's for prod
User Story - 3: Creating and executing Devops ADO Pipeline to deploy Databricks deliverables
User Story - 4: Creating and executing Devops ADO Pipeline to deploy ADF deliverables

==========

Project Q&A Session - Completed