**Cohort 3.1:**



Requirements:

Source url: <https://raw.githubusercontent.com/microsoft/sql-server-samples/master/samples/databases/adventure-works/oltp-install-script/SalesOrderHeader.csv>

Staging Raw layer : ADLS/RAW/SalesOrderHeader.zip

Target: ADLS/Refined/<file>(All unzipped file)

Problem statement: Fetch data from source URL & load in staging raw layer as zip file & processed or clean data that has been arrived in last 24 hrs from source & load into ADLS refined container in parquet file as delta table while following below instruction:

Step 1 - Data preparation (use ADF):-

* Fetch data from Source url & load in raw layer.
* Unzip file with help of ADF/ADB under raw<participant\_name>/Unzippeddata/file.
* Fetch data that has been processed in last 24 hrs from source & load into ADLS refined container in parquet file while following below instruction:
* Check if the file is available in the path. If it’s not available, there should be timeout after 1 minute.

Step 2 - Data transformation(use ADB):

* for sourceorderheader file give the schema of column as per below

SalesOrderID,RevisionNumber,OrderDate,DueDate,ShipDate,Status,OnlineOrderFlag,SalesOrderNumber,PurchaseOrderNumber,AccountNumber,CustomerID,ShipToAddressID,BillToAddressID,ShipMethod,CreditCardApprovalCode,SubTotal,TaxAmt,Freight,TotalDue,Comment,rowguid,ModifiedDate

* Drop all entries where color is null.
* Check difference between StandardCost & ListPrice for each sales order.
* Replace Product size s-> 30, M-> 32, L-> 34, XL-> 36, XXL-> 38 to display availble product group by size
* Load final data on delta tables under adls refined container.
* setup email configuration on failure & success of pipeline to get notify, which contains, pipeline name, runid & failure message or success meesage.

Note: Use only two resource : adls, ADF,ADB.

**Deliverables**:

1. Architecture diagram of flow
2. Best Practice in Pipeline
3. Best Practice for Database
4. Best Practice for ETL
5. Target Tables