**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.
* Fetch data that has been processed in last 24 hrs from source & load into ADLS raw<participant\_name>/Unzippeddata/file.
* 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

Important links:

Datawarehouse:

[OLAP Operations in Data Warehouse | Slice Dice Roll-up Drill-down Pivot | Example of location, time - YouTube](https://www.youtube.com/watch?v=4NKbZMsddBI)

[Choose between slowly changing dimension types - Training | Microsoft Learn](https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types)

Azure:

<https://github.com/deeksharm/DP203>

[Integration runtime - Azure Data Factory & Azure Synapse | Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime)

[Mapping data flow transformation overview - Azure Data Factory & Azure Synapse | Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/data-flow-transformation-overview)

[AZ-900: Microsoft Azure Fundamentals - Testprep Training Tutorials](https://www.testpreptraining.com/tutorial/microsoft-azure-fundamentals-az-900)

[Study guide for Exam AZ-900: Microsoft Azure Fundamentals | Microsoft Learn](https://learn.microsoft.com/en-gb/credentials/certifications/resources/study-guides/az-900)

https://sonraisecurity.com/blog/what-is-azure-governance

Powershell:

[13 Powershell Basics - Command Piping - By Value - YouTube](https://www.youtube.com/watch?v=Z0X04OYzf0c&list=PLFoX_td1iTj8Aqs8oE0W6licmsBV2QUbO&index=15)

Hadoop:

[BigData & Cloud Computing Tutorials | CloudDuggu](https://www.cloudduggu.com/)

Synapse:

[Dedicated SQL pool (formerly SQL DW) architecture - Azure Synapse Analytics | Microsoft Learn](https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/massively-parallel-processing-mpp-architecture)

[Memory and concurrency limits - Azure Synapse Analytics | Microsoft Learn](https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/memory-concurrency-limits)

Databricks:

[Connect to Azure Data Lake Storage Gen2 and Blob Storage - Azure Databricks | Microsoft Learn](https://learn.microsoft.com/en-us/azure/databricks/connect/storage/azure-storage)

Compute:

[Compute | Databricks on AWS](https://docs.databricks.com/en/compute/index.html#:~:text=These%20are%20the%20types%20of%20compute%20available%20in,used%20to%20reduce%20start%20and%20autoscaling%20times.%20)

Compute polices:

[Compute policy reference | Databricks on Google Cloud](https://docs.gcp.databricks.com/en/administration-guide/clusters/policy-definition.html)

[Default policies and policy families - Azure Databricks | Microsoft Learn](https://learn.microsoft.com/en-us/azure/databricks/administration-guide/clusters/policy-families)

Mounting:

[Mounting cloud object storage on Azure Databricks - Azure Databricks | Microsoft Learn](https://learn.microsoft.com/en-us/azure/databricks/dbfs/mounts)

Pyspark:

[API Reference — PySpark master documentation (apache.org)](https://spark.apache.org/docs/latest/api/python/reference/index.html)