**Synapse & ADF Integration**

**1. Create Synapse Workspace with Dedicated SQL Pool**

**Prerequisites:**

* Azure subscription
* Resource group
* ADLS Gen2 storage account

**Steps:**

1. Sign in to Azure Portal.
2. Navigate to **Create a resource → Analytics → Synapse Analytics**.
3. Provide:
   * Workspace name: synapse-lab-ws
   * Region: Choose closest
   * Storage: Select existing ADLS Gen2
   * File system: synapsefs
4. Enable **Managed Virtual Network** and **Data Exfiltration Protection**.
5. After workspace is created, go to **SQL pools** → **+ New**.
6. Select **Dedicated SQL pool**:
   * Name: dedicatedpool
   * Performance level: DW100c (for lab)
7. Click **Create** and wait for provisioning.

**2. Data Load: Import CSV and Parquet from ADLS**

**Steps:**

1. In **Synapse Studio**, go to **Data → Linked → + Linked Service**.
   * Select **Azure Data Lake Storage Gen2**.
   * Choose **Account Key** or **Managed Identity** authentication.
   * Name: adls-linked.
2. Upload sample datasets to ADLS container rawdata/:
   * sales.csv
   * nyc\_taxi.parquet
3. In Synapse SQL script (dedicated pool):
4. CREATE TABLE Sales\_CSV (
5. SaleId INT,
6. Region NVARCHAR(50),
7. Amount FLOAT,
8. SaleDate DATE
9. )
10. WITH (DISTRIBUTION = ROUND\_ROBIN, CLUSTERED COLUMNSTORE INDEX);
11. COPY INTO Sales\_CSV
12. FROM 'https://<storage>.dfs.core.windows.net/rawdata/sales.csv'
13. WITH (
14. FILE\_TYPE = 'CSV',
15. FIELDTERMINATOR = ',',
16. ROWTERMINATOR = '0x0A',
17. FIRSTROW = 2
18. );

For Parquet:

CREATE TABLE Taxi\_Trips

WITH (DISTRIBUTION = ROUND\_ROBIN, CLUSTERED COLUMNSTORE INDEX)

AS

SELECT \*

FROM OPENROWSET(

BULK 'https://<storage>.dfs.core.windows.net/rawdata/nyc\_taxi.parquet',

FORMAT = 'PARQUET'

) AS [result];

**3. Partition Table: Create a Distributed Table**

**Steps:**

1. Create a **partitioned table** on SaleDate:
2. CREATE TABLE Sales\_Partitioned (
3. SaleId INT,
4. Region NVARCHAR(50),
5. Amount FLOAT,
6. SaleDate DATE
7. )
8. WITH (
9. DISTRIBUTION = HASH (SaleId),
10. CLUSTERED COLUMNSTORE INDEX,
11. PARTITION (SaleDate RANGE RIGHT FOR VALUES (
12. '2022-01-01', '2022-07-01', '2023-01-01'
13. ))
14. );
15. Insert data into partitioned table:
16. INSERT INTO Sales\_Partitioned
17. SELECT \* FROM Sales\_CSV;
18. Verify partition usage:
19. SELECT
20. $PARTITION AS PartitionNumber,
21. COUNT(\*) AS RecordCount
22. FROM Sales\_Partitioned
23. GROUP BY $PARTITION;

**4. Optimize Query: Execution Plan & Joins**

**Steps:**

1. Compare **cross join vs hash join**:
2. SELECT s.SaleId, t.trip\_id, s.Amount
3. FROM Sales\_Partitioned s
4. JOIN Taxi\_Trips t
5. ON s.SaleId = t.trip\_id;
6. Enable **Query Statistics** in Synapse Studio (Run with "Include Actual Execution Plan").
7. Observe:
   * **Data Movement Operations (DMO)**
   * Shuffle moves due to distribution mismatch
8. Optimize by **using the same distribution key**:
   * Recreate tables with DISTRIBUTION = HASH(SaleId) for both.
   * Re-run join and compare execution plan.

**5. Integrate with ADF: Load Data into Synapse**

**Steps:**

1. Open **Azure Data Factory Studio**.
2. Go to **Manage → Linked services → + New**:
   * Add Linked Service for **ADLS**.
   * Add Linked Service for **Azure Synapse Analytics**.
3. Create a **Dataset** for CSV (ADLS).
4. Create a **Dataset** for Synapse table (Sales\_CSV).
5. Build a **Pipeline**:
   * Drag **Copy Data Activity**.
   * Source: ADLS dataset (sales.csv).
   * Sink: Synapse dataset (Sales\_CSV).
6. Configure **Mapping** (auto-detect or manual).
7. Trigger pipeline and monitor in **Monitor Hub**.
8. Validate data landed in Synapse by running:
9. SELECT TOP 10 \* FROM Sales\_CSV;