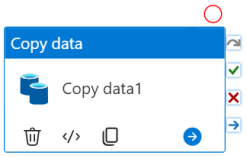
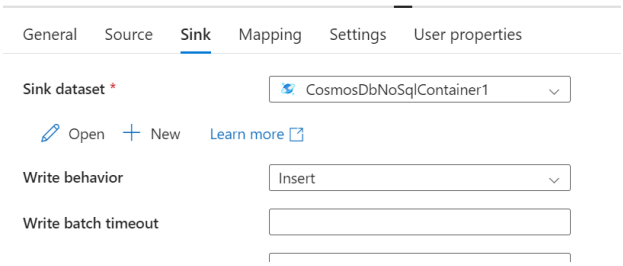
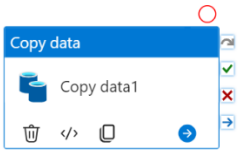
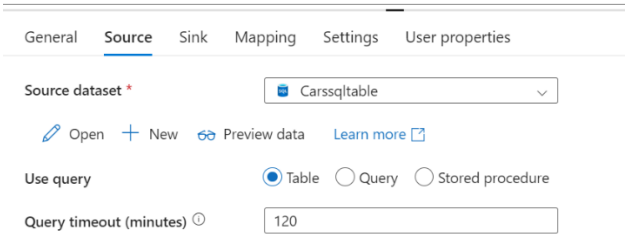
Assignment 10

1. Design an ADF pipeline to copy data from an on-premise Azure SQL database to Azure Cosmos DB, ensuring data consistency and performance optimization. Pick correct options of partitioning for better performance.

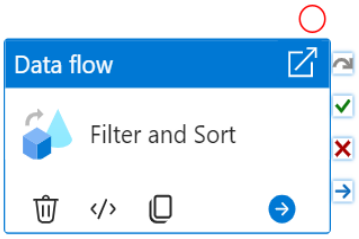


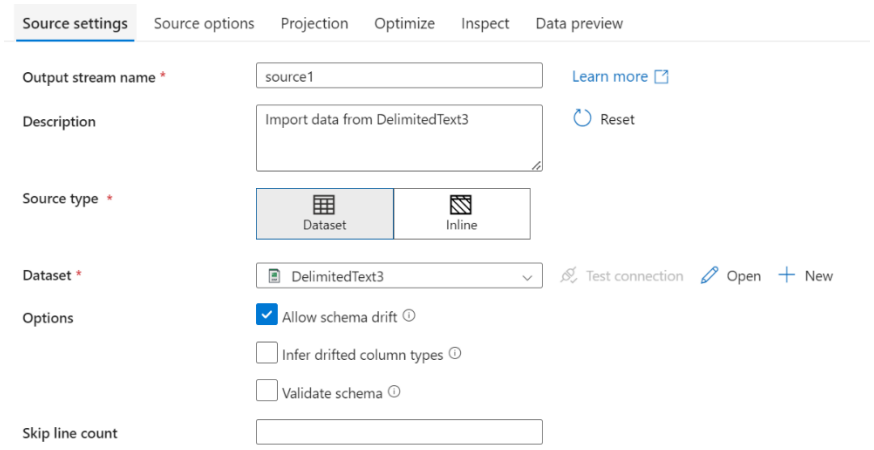


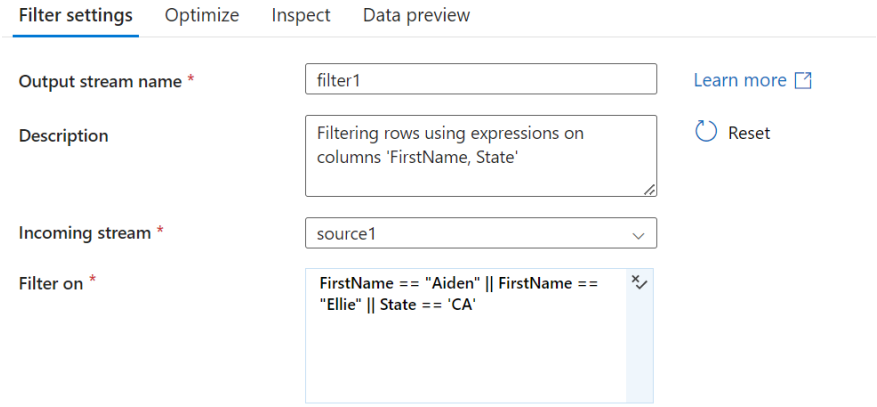


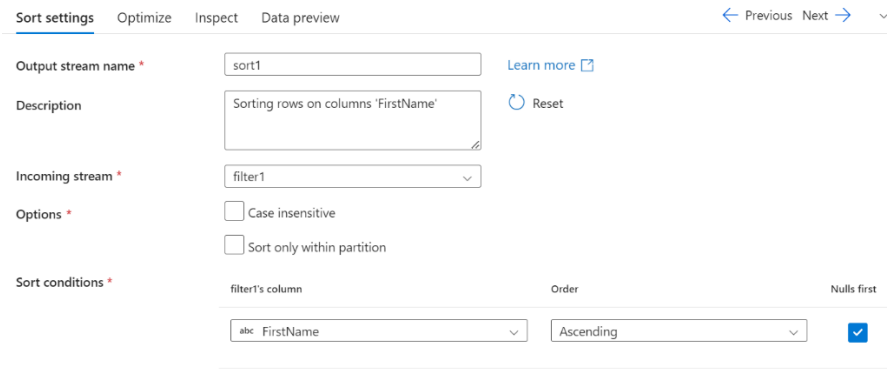


2. Create Pipeline using Azure Data Flow in Azure Data Factory to apply Filter and Sort transformations on datasets.

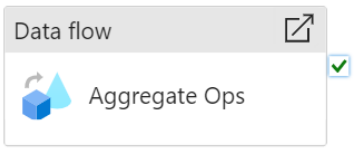


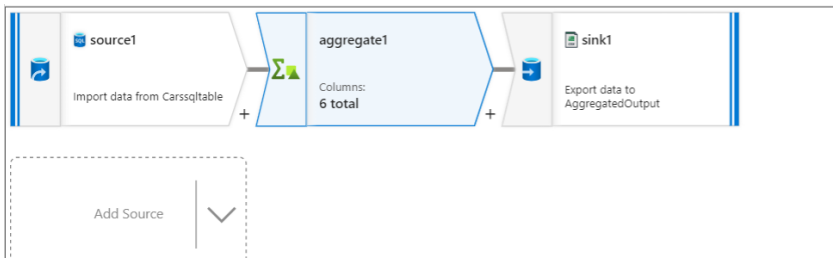






3. Design an ADF pipeline to implement aggregate operations, such as sum, average, max, min and count, within an Azure Data Flow.





Source settings | Source options | Projection | Optimize | Inspect | Data preview

Output stream name * source1 Learn more

Description: Import data from Carssqltable — Reset

Source type * Dataset | Inline

Dataset * Carssqltable — Test connection | Open | New

Options: Allow schema drift; Infer drifted column types; Validate schema

Sampling * Enable | Disable

Aggregate settings | Optimize | Inspect | Data preview — Previous

Output stream name * aggregate1 Learn more

Description: Aggregating data by 'MPG_City, Make' producing columns 'Sum of MSRP, Average of MSRP, Stddev of MSRP, — Reset

Incoming stream * source1

Group by | Aggregates

Columns	Name as
MPG_City	MPG_City
Make	Make

Sink | Settings | Errors | Mapping | Optimize | Inspect | Data preview

This sink currently has Single partition set in Optimize. This will make your data flow execution longer. The recommended setting partitioning.

Output stream name * sink1 Learn more

Description: Export data to AggregatedOutput — Reset

Incoming stream * aggregate1

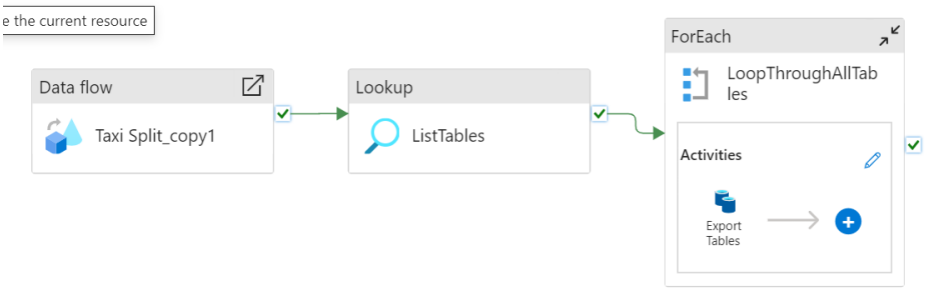
Sink type * Dataset | Inline | Cache

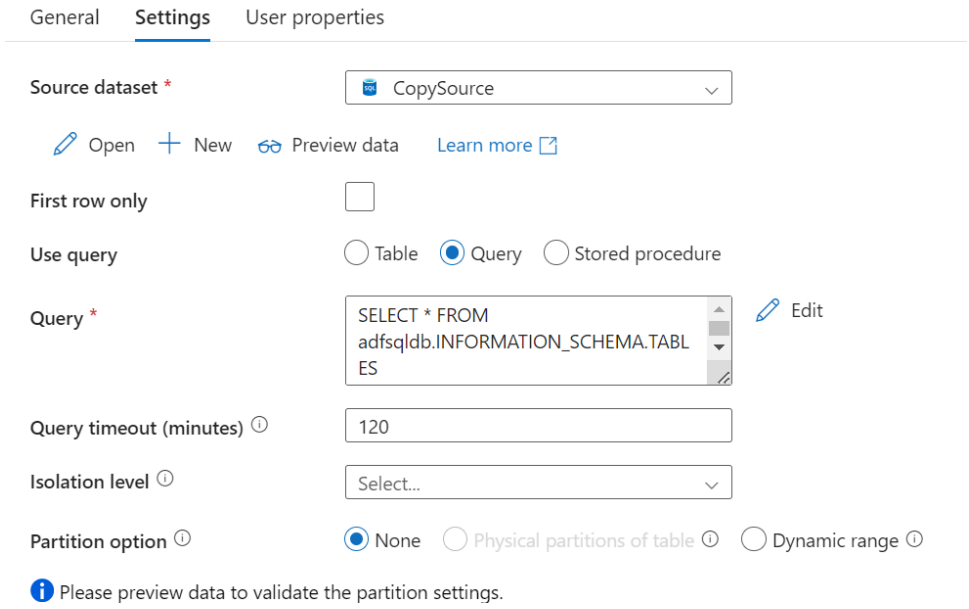
Dataset * AggregatedOutput — Test connection | Open | New

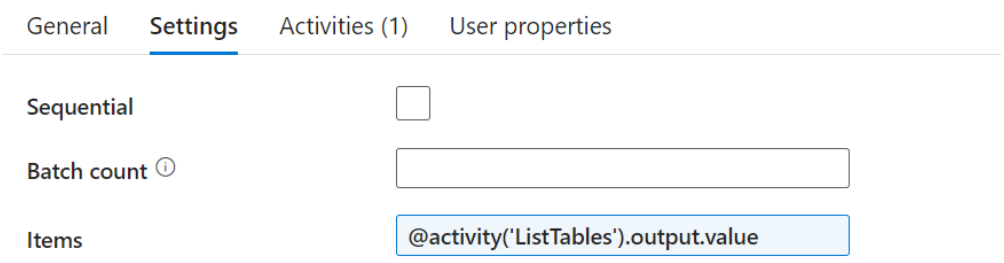
Skip line count

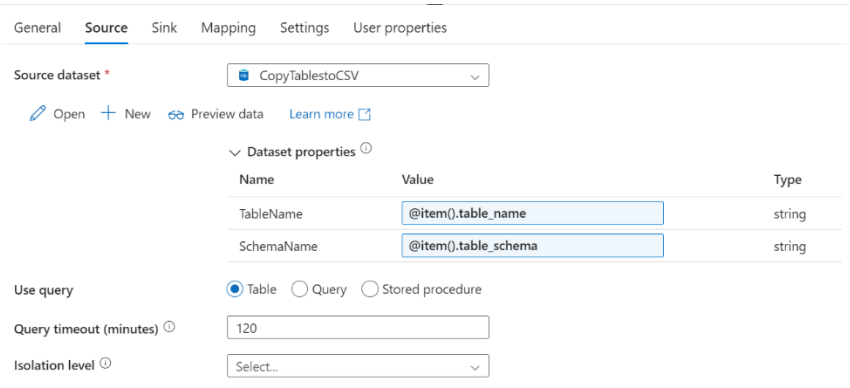
Options: Allow schema drift

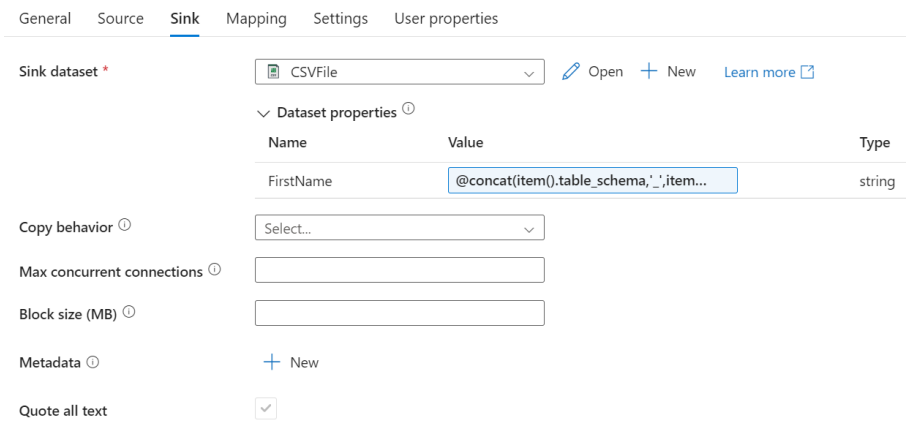
4. Create best approach to bulk copy data from multiple homogenous sources into Azure SQL Database using ADF pipelines. Show usage of Lookup, For Each Loop and Expressions in Azure Data Factory.



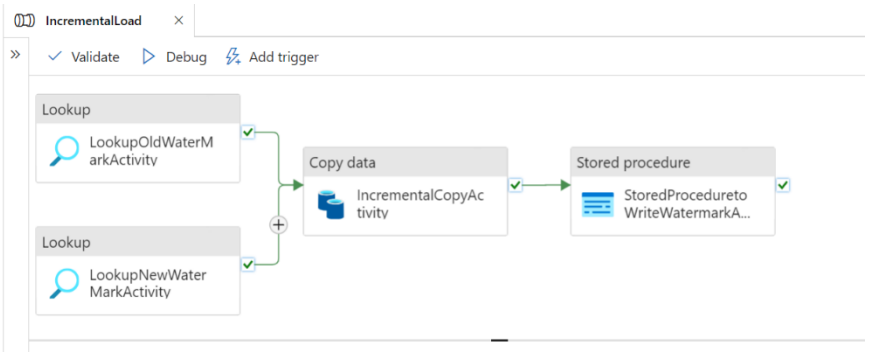


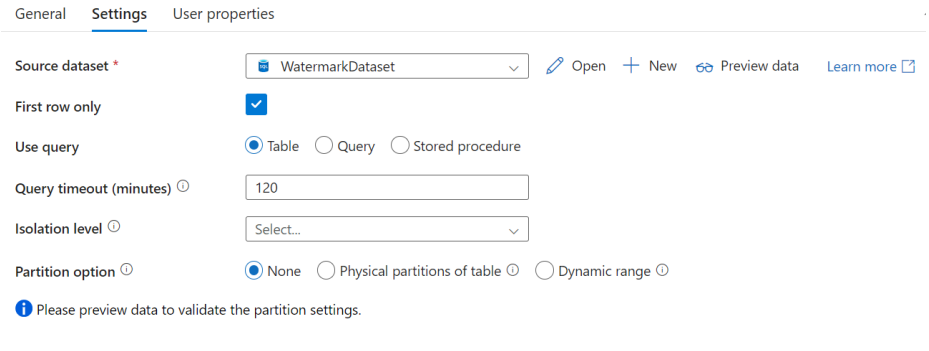


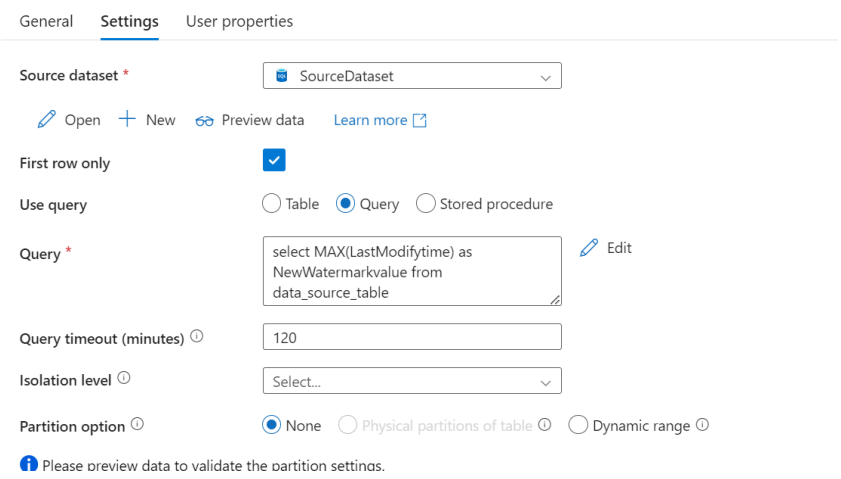


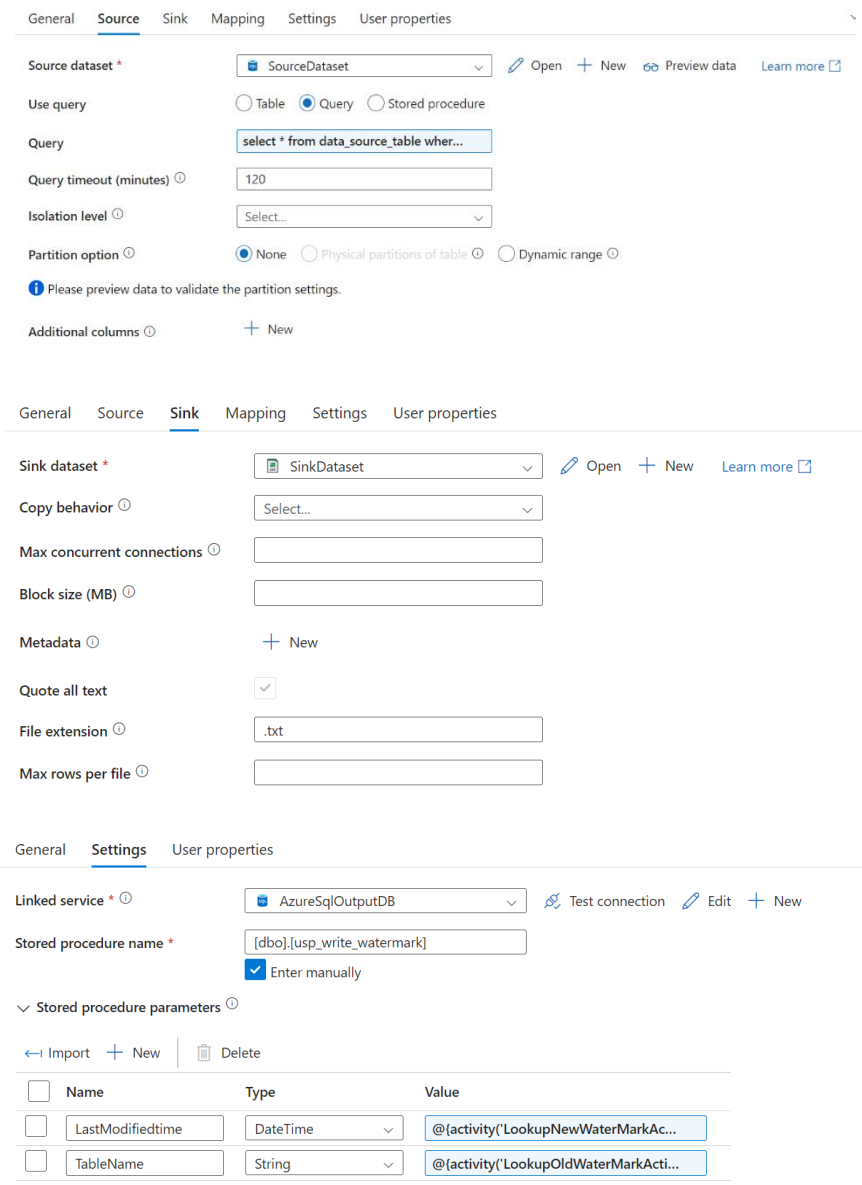


5. Implement incremental load Pipeline in Azure Data Factory for handling datasets, ensuring efficient insert/upsert/updates to the target storage without re-inserting the entire dataset?

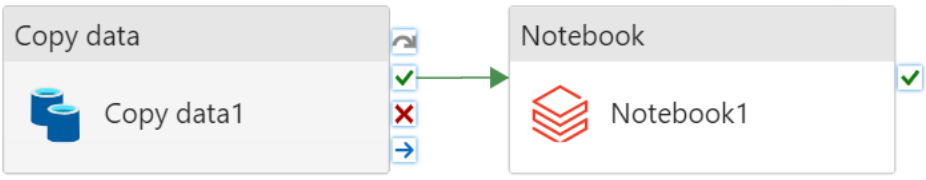


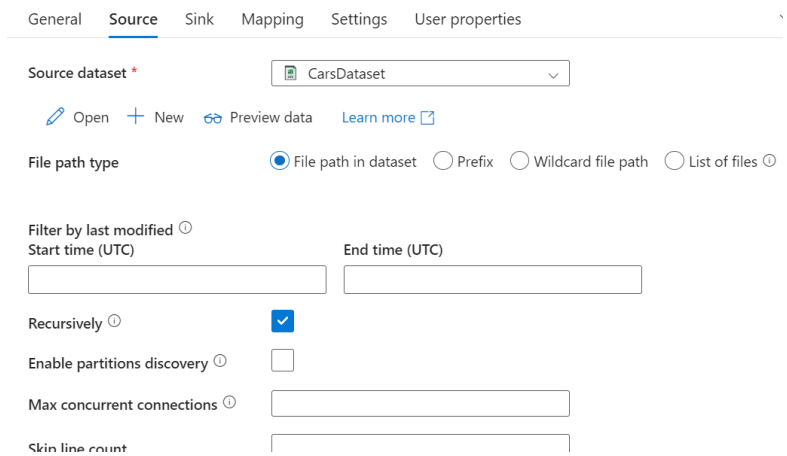


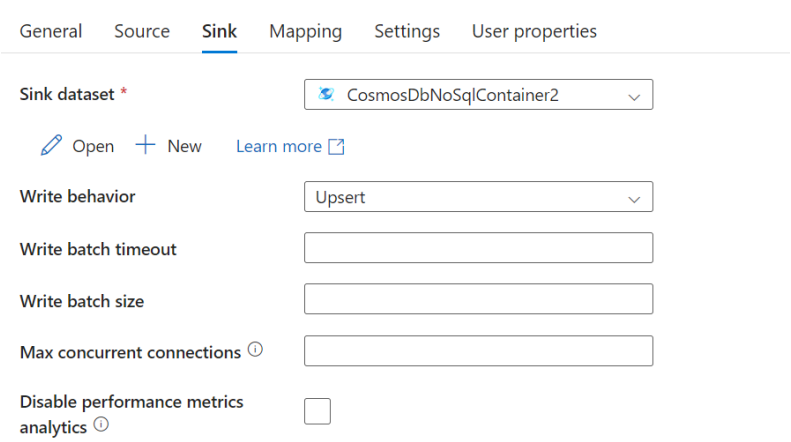


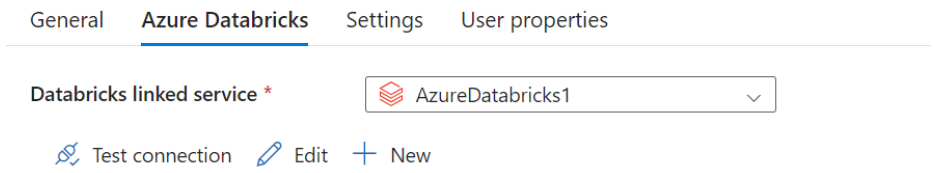


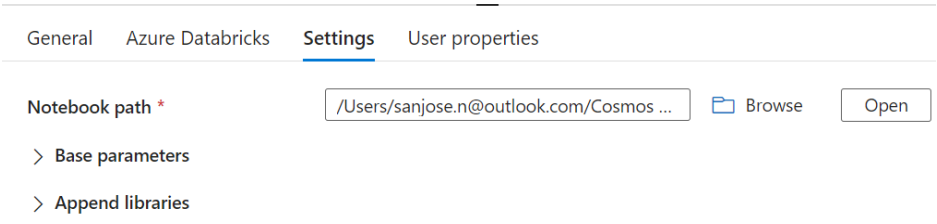
6. What are the key steps to connect Azure Databricks to Cosmos DB for real-time analytics and data transformation using spark and Databricks.




General
Source
Sink
Mapping
Settings
User properties
Source dataset *
CarsDataset
Open
New
Preview data
Learn more
File path type
File path in dataset
Prefix
Wildcard file path
List of files
Filter by last modified
Start time (UTC)
End time (UTC)
Recursively
Enable partitions discovery
Max concurrent connections
Skip line count


General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
CosmosDbNoSqlContainer2
Open
New
Learn more
Write behavior
Upsert
Write batch timeout
Write batch size
Max concurrent connections
Disable performance metrics analytics


General
Azure Databricks
Settings
User properties
Databricks linked service *
AzureDatabricks1
Test connection
Edit
New


General
Azure Databricks
Settings
User properties
Notebook path *
/Users/sanjose.n@outlook.com/Cosmos ...
Browse
Open
Base parameters
Append libraries