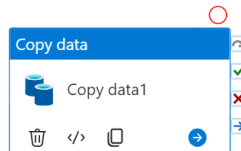


Assignment 10

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.



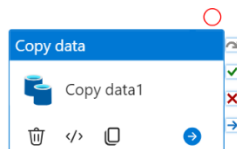
General Source **Sink** Mapping Settings User properties

Sink dataset * CosmosDbNoSqlContainer1

[Open](#) [+ New](#) [Learn more](#)

Write behavior Insert

Write batch timeout



General **Source** Sink Mapping Settings User properties

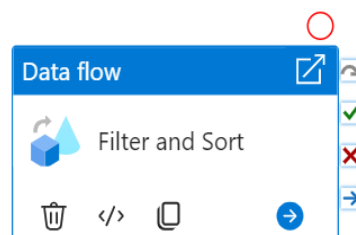
Source dataset * Carssqltable

[Open](#) [+ New](#) [Preview data](#) [Learn more](#)

Use query ☒ Table ☐ Query ☐ Stored procedure

Query timeout (minutes) 120

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



Source settings Source options Projection Optimize Inspect Data preview

Output stream name * source1 [Learn more](#)

Description Import data from DelimitedText3 [Reset](#)

Source type * Dataset Inline

Dataset * DelimitedText3 [Test connection](#) [Open](#) [+ New](#)

Options ☒ Allow schema drift [?](#)
☐ Infer drifted column types [?](#)
☐ Validate schema [?](#)

Skip line count

Filter settings Optimize Inspect Data preview

Output stream name * filter1 [Learn more](#)

Description Filtering rows using expressions on columns 'FirstName, State' [Reset](#)

Incoming stream * source1

Filter on * `FirstName == "Aiden" || FirstName == "Ellie" || State == 'CA'` [✕](#)

Sort settings Optimize Inspect Data preview [← Previous](#) [Next →](#)

Output stream name * sort1 [Learn more](#)

Description Sorting rows on columns 'FirstName' [Reset](#)

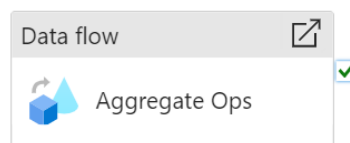
Incoming stream * filter1

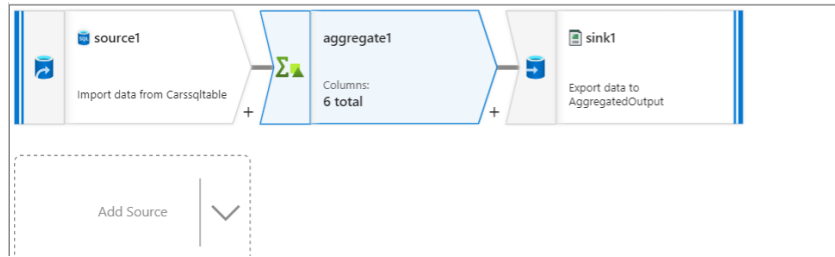
Options * ☐ Case insensitive
☐ Sort only within partition

Sort conditions *

filter1's column	Order	Nulls first
abc FirstName	Ascending	<input checked="" type="checkbox"/>

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 Carssqtable

Reset

Source type *

Dataset

Inline

Dataset *

Carssqtable

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

123 MPG_City	MPG_City	+	🗑
abc 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

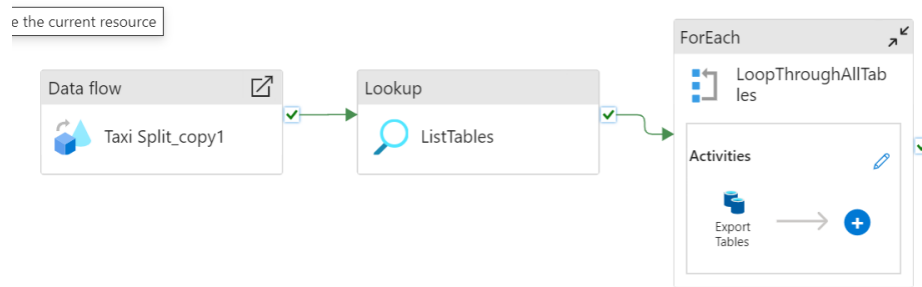
New

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.



General **Settings** User properties

Source dataset * CopySource

[Open](#) [+ New](#) [Preview data](#) [Learn more](#)

First row only ☐

Use query ☐ Table ☒ Query ☐ Stored procedure

Query *

SELECT * FROM adfsqldb.INFORMATION_SCHEMA.TABLES

[Edit](#)

Query timeout (minutes) ⓘ

Isolation level ⓘ Select...

Partition option ⓘ ☒ None ☐ Physical partitions of table ⓘ ☐ Dynamic range ⓘ

i Please preview data to validate the partition settings.

General **Settings** Activities (1) User properties

Sequential ☐

Batch count ⓘ

Items @activity('ListTables').output.value

General **Source** Sink Mapping Settings User properties

Source dataset * CopyTablestoCSV

[Open](#) [+ New](#) [Preview data](#) [Learn more](#)

Dataset properties ⓘ

Name	Value	Type
TableName	@item().table_name	string
SchemaName	@item().table_schema	string

Use query ☒ Table ☐ Query ☐ Stored procedure

Query timeout (minutes) ⓘ

Isolation level ⓘ Select...

General Source **Sink** Mapping Settings User properties

Sink dataset * CSVFile [Open](#) [+ New](#) [Learn more](#)

Dataset properties [ⓘ]

Name	Value	Type
FirstName	@concat(item().table_schema,'_',item...	string

Copy behavior [ⓘ] Select...

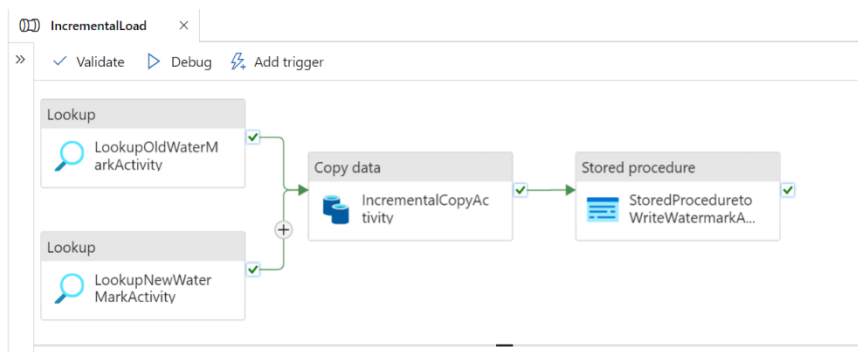
Max concurrent connections [ⓘ]

Block size (MB) [ⓘ]

Metadata [ⓘ] [+ New](#)

Quote all text ☐

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?



General **Settings** User properties

Source dataset * WatermarkDataset [Open](#) [+ New](#) [Preview data](#) [Learn more](#)

First row only ☒

Use query ☒ Table ☐ Query ☐ Stored procedure

Query timeout (minutes) [ⓘ] 120

Isolation level [ⓘ] Select...

Partition option [ⓘ] ☒ None ☐ Physical partitions of table [ⓘ] ☐ Dynamic range [ⓘ]

i Please preview data to validate the partition settings.

General **Settings** User properties

Source dataset * SourceDataset [Open](#) [+ New](#) [Preview data](#) [Learn more](#)

First row only ☒

Use query ☐ Table ☒ Query ☐ Stored procedure

Query * select MAX>LastModifytime) as NewWatermarkvalue from data_source_table [Edit](#)

Query timeout (minutes) [ⓘ] 120

Isolation level [ⓘ] Select...

Partition option [ⓘ] ☒ None ☐ Physical partitions of table [ⓘ] ☐ Dynamic range [ⓘ]

i Please preview data to validate the partition settings.

General **Source** Sink Mapping Settings User properties

Source dataset * SourceDataset [Open](#) [+ New](#) [Preview data](#) [Learn more](#)

Use query ☐ Table ☒ Query ☐ Stored procedure

Query select * from data_source_table wher...

Query timeout (minutes) 120

Isolation level Select...

Partition option ☒ None ☐ Physical partitions of table ☐ Dynamic range

Please preview data to validate the partition settings.

Additional columns [+ New](#)

General **Sink** Mapping Settings User properties

Sink dataset * SinkDataset [Open](#) [+ New](#) [Learn more](#)

Copy behavior Select...

Max concurrent connections

Block size (MB)

Metadata [+ New](#)

Quote all text ☒

File extension .txt

Max rows per file

General **Settings** User properties

Linked service * AzureSqlOutputDB [Test connection](#) [Edit](#) [+ New](#)

Stored procedure name * [dbo].[usp_write_watermark]

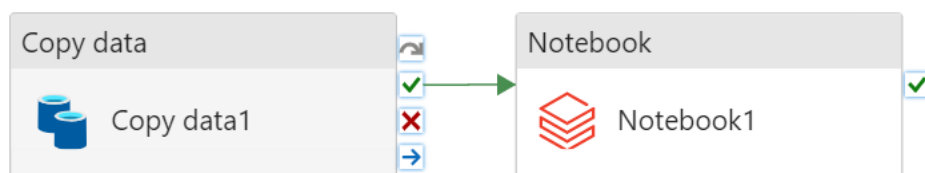
☒ Enter manually

Stored procedure parameters

[← Import](#) [+ New](#) [Delete](#)

<input type="checkbox"/>	Name	Type	Value
<input type="checkbox"/>	LastModifiedtime	DateTime	@{activity('LookupNewWaterMarkAc...
<input type="checkbox"/>	TableName	String	@{activity('LookupOldWaterMarkActi...

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



GeneralSourceSinkMappingSettingsUser 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

Skipped line count

GeneralSourceSinkMappingSettingsUser properties

Sink dataset *

CosmosDbNoSqlContainer2

Open

New

Learn more

Write behavior

Upsert

Write batch timeout

Write batch size

Max concurrent connections

Disable performance metrics analytics

GeneralAzure DatabricksSettingsUser properties

Databricks linked service *

AzureDatabricks1

Test connection

Edit

New

GeneralAzure DatabricksSettingsUser properties

Notebook path *

/Users/sanjose.n@outlook.com/Cosmos ...

Browse

Open

> Base parameters

> Append libraries