

## **TASK 3**

**Create Incremental Load Pipeline and Automate This on a Daily Basis:** Design an ADF pipeline to perform incremental data loads using techniques like watermarking or change tracking. Schedule this pipeline to run daily to ensure only new or modified records are processed, optimizing performance and reducing load.

### **Step-by-Step Guide:**

#### **Step 1: Prerequisites**

- Source SQL Server has a column like LastModifiedDate or UpdatedAt or shippedDate in each table.

#### **Step 2: Create a Watermark Table (Metadata Table)**

In Azure SQL, create a table to store the last processed timestamp for each table.

```
CREATE TABLE WatermarkControl (
    TableName NVARCHAR(100) PRIMARY KEY,
    LastWatermark DATETIME
);
```

Insert initial values (if needed):

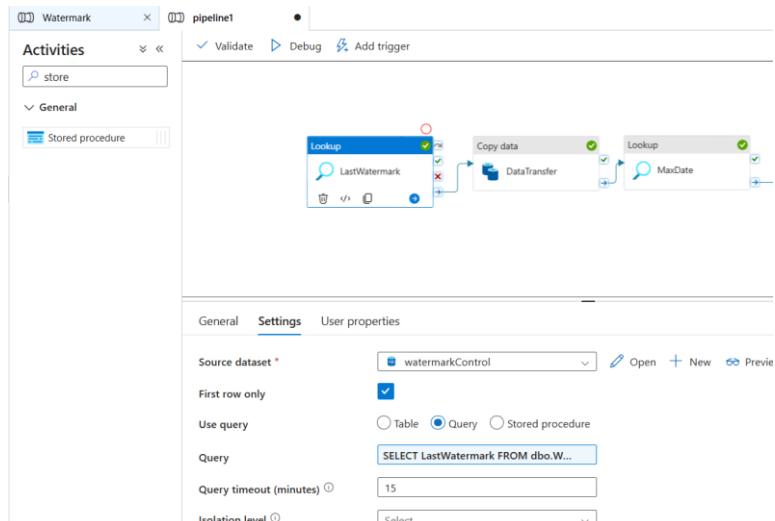
```
INSERT INTO WatermarkControl (TableName, LastWatermark)
VALUES ('sales.orders', '2000-01-01'); -- Initial load date
```

#### **Step 3: Create ADF Pipeline**

##### **a. Lookup Activity: Get Last Watermark**

- Query the control table:

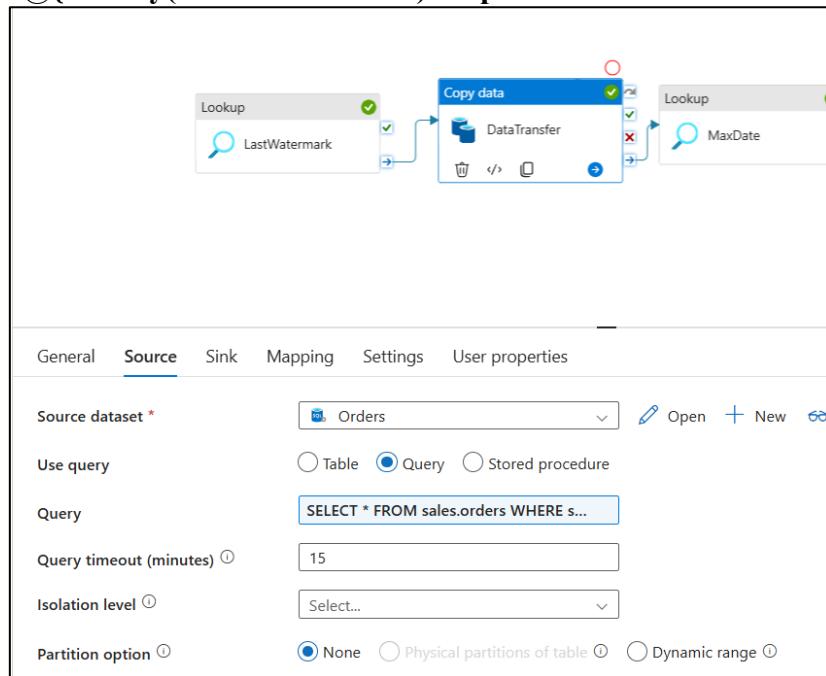
```
SELECT LastWatermark FROM dbo.WatermarkControl;
```



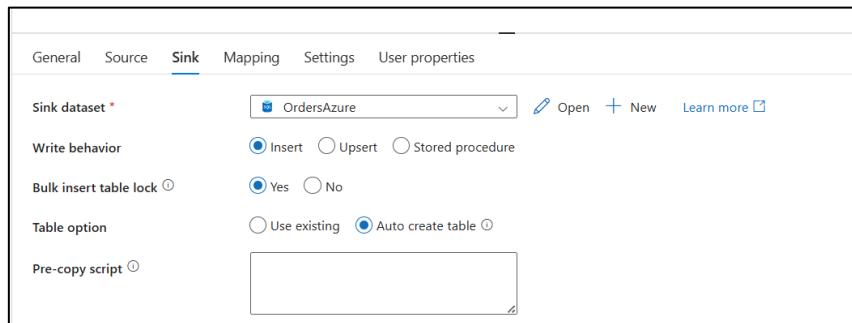
## b. Copy Data Activity: Incremental Data Load

- Source Query:

```
SELECT * FROM sales.orders  
WHERE shipped_date >  
'@{activity('LastWatermark').output.firstRow.LastWatermark}';
```



Sink: Azure SQL.



## c. Lookup Activity: Retrieve Max Date

- Source Dataset: Azure dataset
- Query:

```
SELECT MAX(shipped_date) AS MaxDate  
FROM sales.orders  
WHERE shipped_date >  
'@{activity('LastWatermark').output.firstRow.LastWatermark}'
```

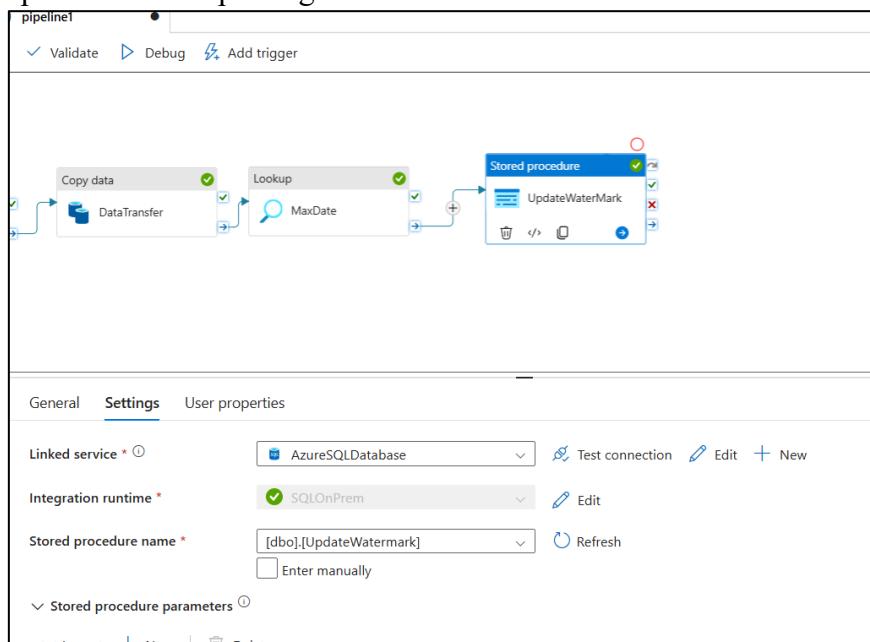
#### d. Stored Procedure / Script Activity: Update Watermark

- Create a procedure in the destination database in Azure SQL:

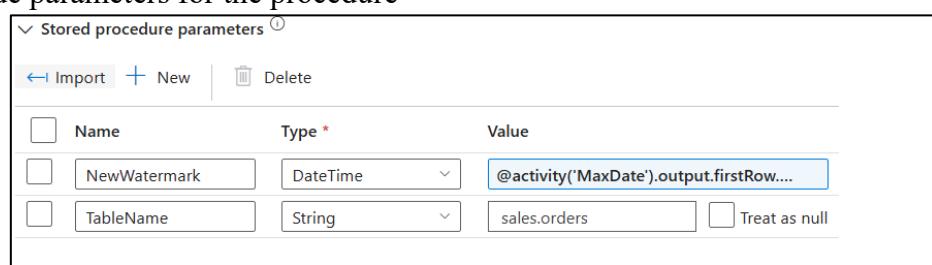
```
CREATE PROCEDURE UpdateWatermark
    @TableName NVARCHAR(100),
    @NewWatermark DATETIME
AS
BEGIN
    SET NOCOUNT ON;

    UPDATE WatermarkControl
    SET LastWatermark = @NewWatermark
    WHERE TableName = @TableName;
END
```

- Select the procedure for updating the watermark.



- Provide parameters for the procedure



NewWatermark: **@activity('MaxDate').output.firstRow.MaxDate**

## Automation: Schedule Daily Trigger

In ADF:

1. Go to your pipeline.
2. Click on **Add Trigger > New/Edit**.
3. Set a **Recurring trigger** with:
  - o Frequency: **Day**
  - o Time: <time>
  - o Start date: Today
4. Link it to your pipeline.

The screenshot shows the ADF Triggers page on the left and the Edit trigger dialog box on the right.

**Triggers Page (Left):**

- Header: date all, Publish all
- Section: Triggers
- Description: To execute a pipeline set the trigger. Triggers represent a unit of processing that determines when a pipeline runs.
- Buttons: + New, Refresh
- Filter: Filter by name, Annotations: Any
- Table: Shows 1 - 2 of 2 items.

Name	Type
EveryLastSaturday	Schedule
UpdateWatermark	Schedule

**Edit trigger Dialog (Right):**

- Name \***: UpdateWatermark
- Description**: (Empty)
- Type \***: ScheduleTrigger
- Start date \***: 7/14/2025, 3:55:00 PM
- Time zone \***: Chennai, Kolkata, Mumbai, New Delhi (UTC+5:30)
- Recurrence \***: Every 1 Day(s)
- Advanced recurrence options**:
  - Execute at these times**:
    - Hours: (Empty)
    - Minutes: (Empty)
  - Schedule execution times**: 15:55
- Buttons: OK, Cancel

## **RESULT**

The screenshot shows the 'Pipeline runs' page in the Azure Data Factory portal. A red box highlights the 'Triggered' tab in the top navigation bar. The main table lists one run for 'pipeline1'. The run details are as follows:

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters
pipeline1	7/14/2025, 3:55:00 PM	7/14/2025, 3:55:59 PM	59s	UpdateWatermark	Succeeded	Original	

At the bottom right of the table, it says 'Last refreshed 0 minutes ago'.

## **ORDERS table updated:**

The screenshot shows the Azure Data Studio interface. The left sidebar shows the database schema for 'bikeStores'. The 'Tables' section has 'sales.orders' selected. The main area displays three queries. Query 3 is active, showing the results of the following SQL statement:

```
1 SELECT TOP (1000) * FROM [sales].[orders]
```

The results table shows the following data:

order_id	customer_id	order_status	order_date	required_date	shipped_date	store_
1	259	4	2016-01-01	2016-01-03	2016-01-03	1
2	1212	4	2016-01-01	2016-01-04	2016-01-03	2
3	523	4	2016-01-02	2016-01-05	2016-01-03	2
4	175	4	2016-01-03	2016-01-04	2016-01-05	1

## **WATERMARK updated for next day:**

The screenshot shows the Azure Data Studio interface, similar to the previous one. The left sidebar shows the database schema for 'bikeStores'. The 'Tables' section has 'dbo.WatermarkControl' selected. The main area displays three queries. Query 2 is active, showing the results of the following SQL statement:

```
1 SELECT TOP (1000) * FROM [dbo].[WatermarkControl]
```

The results table shows the following data:

TableName	LastWatermark
sales.orders	2018-04-02T00:00:00.0000000

A red box highlights the 'LastWatermark' column value '2018-04-02T00:00:00.0000000'.