

P4- Implementation
Streaming Service Analysis
Team 6

Abhishek Tikam Ramchandani – 002743745

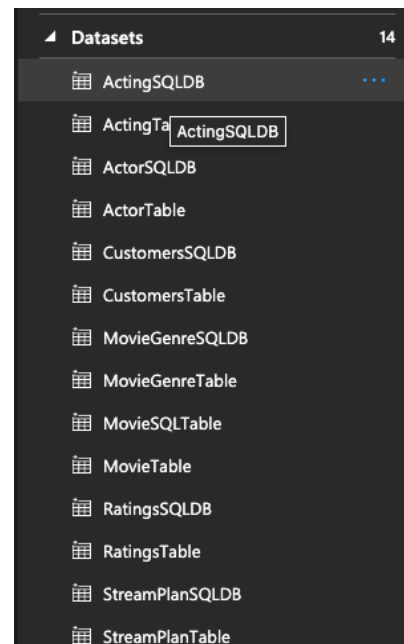
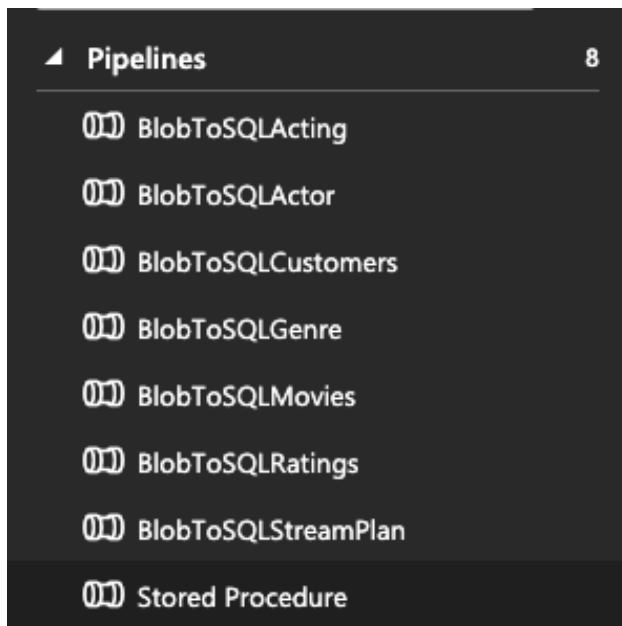
Jurreyah Firdaws Mohammed – 002747514

Atharva Uplenchwar – 002990536

Samhitha Mereddy – 002796140

In this phase of our project implementation, we have utilized linked services in Azure Data Factory to connect and manage data across different storage and database services. As we now have cleaned datasets that we transformed using Alteryx and Python in Phase-3, we established linked services to Azure Blob Storage and Azure SQL Database, which acted as the sources and destinations for our data. Blob Storage was primarily used for handling large volumes of structured + unstructured data, whereas SQL Database served as our structured data's destination storage system.

The core of our Azure Data Factory implementation involved setting up Data Pipelines and Scheduled Triggers. These pipelines were designed to efficiently transfer and transform data between the source and sink database, “xxxTable” & “xxxSQLDB”, respectively.



Showing 1 - 11 items

<input type="checkbox"/> Pipeline name	Run start ↑	Run end	Duration	Status
<input type="checkbox"/> Stored Procedure	12/7/2023, 9:48:47 PM	12/7/2023, 9:48:51 PM	5s	✓ Succeeded
<input type="checkbox"/> BlobToSQLActing	12/7/2023, 9:41:27 PM	12/7/2023, 9:41:44 PM	18s	✓ Succeeded
<input type="checkbox"/> BlobToSQLRatings	12/7/2023, 7:19:52 PM	12/7/2023, 7:20:08 PM	17s	✓ Succeeded
<input type="checkbox"/> BlobToSQLRatings	12/7/2023, 7:07:25 PM	12/7/2023, 7:07:42 PM	17s	✓ Succeeded
<input type="checkbox"/> BlobToSQLCustomers	12/7/2023, 4:20:07 PM	12/7/2023, 4:20:24 PM	18s	✓ Succeeded
<input type="checkbox"/> pipeline1	12/7/2023, 3:49:27 PM	12/7/2023, 3:49:44 PM	18s	✓ Succeeded
<input type="checkbox"/> BlobToSQLGenre	12/7/2023, 3:45:12 PM	12/7/2023, 3:45:29 PM	18s	✓ Succeeded
<input type="checkbox"/> BlobToSQLActing	12/7/2023, 3:38:33 PM	12/7/2023, 3:38:51 PM	18s	✓ Succeeded
<input type="checkbox"/> BlobToSQLMovies	12/7/2023, 3:34:01 PM	12/7/2023, 3:34:18 PM	18s	✓ Succeeded
<input type="checkbox"/> BlobToSQL	12/7/2023, 12:33:51 AM	12/7/2023, 12:34:11 AM	21s	✓ Succeeded
<input type="checkbox"/> BlobToSQL	12/7/2023, 12:03:31 AM	12/7/2023, 12:04:19 AM	48s	✓ Succeeded

We focused on creating a seamless flow for specific datasets, particularly our document model (Customers, Stream-Plan & Movie-Ratings) as it. This document represents each Customer's info, the stream plan they are on and the Movies they have rated on IMDB. To do the same, we wrote a Stored Procedure (code below) and scheduled an **Ongoing Data Refresh** that executes this SP on a 24-hour basis. This was also done for our Customers table.

```

CREATE PROCEDURE GetCustomerDetails
AS
BEGIN
    SELECT
        c.CustID as id,
        c.Cust_FName as firstName,
        c.Cust_LName as lastName,
        c.Email as email,
        sp.Description as streamPlanDescription,
        sp.Monthly_Cost as streamPlanMonthlyCost,
        (SELECT
            r.MovieID as movieID,
            r.IMDB_Rating as rating
        FROM Ratings r
        WHERE r.CustID = c.CustID
        FOR JSON PATH) as ratedMovies
    FROM Customers c
    INNER JOIN Stream_Plan sp ON sp.Stream_Plan_ID = c.Stream_Plan_ID
    FOR JSON PATH;
END;

```

Generated JSON Output (Customers, Stream-Plan & Movie-Ratings):

```
netflix-analysis.database.window...
{
  "id": 1,
  "firstName": "Goddard",
  "lastName": "Glassup",
  "email": "gglassup@un.org",
  "streamPlanDescription": "Standard: Unlimited ad-free movies and TV shows, watch on 2 devices at a time, Full HD, download on 2 devices, add u
  "streamPlanMonthlyCost": 15.49,
  "ratedMovies": [
    {
      "movieID": "ts20358",
      "rating": 6.2
    },
    {
      "movieID": "ts273318",
      "rating": 5.9
    }
  ]
},
{
  "id": 2,
  "firstName": "Vito",
  "lastName": "Newlands",
  "email": "vnewlands@e-recht24.de",
  "streamPlanDescription": "Premium: Unlimited ad-free movies and TV shows, watch on 4 devices at a time, Ultra HD, download on 6 devices, add u
  "streamPlanMonthlyCost": 22.99
},
{
  "id": 3,
  "firstName": "Bealle",
  "lastName": "Dobbinson",
  "email": "bdobbinson2@slashdot.org",
  "streamPlanDescription": "Standard with ads: Ad-supported, watch on 2 devices at a time, Full HD, download on 2 devices",
  "streamPlanMonthlyCost": 6.99
},
],
```

To manage and automate the entire data flow, we set up triggers in ADF. These triggers were configured to initiate pipeline runs based on our requirements, ensuring that data processing was timely and consistent. This setup not only streamlined our data management tasks but also provided a scalable and flexible solution for handling our data processing needs in ADF.

Ratings Table will need updating regularly as many people are more likely to rate many movies, so this trigger updates the same every 24 hours.

The screenshot displays the Azure Data Factory (ADF) console. On the left, the 'Activities' pane shows a list of activities including 'Move and transform', 'Synapse', 'Azure Data Explorer', 'Azure Function', 'Batch Service', 'Databricks', 'Data Lake Analytics', 'General', 'HDInsight', 'Iteration & conditionals', 'Machine Learning', and 'Power Query'. The main workspace shows a pipeline named 'StreamPlanSQLDB' with a 'Copy data' activity named 'RatingsCopyData'. The 'Edit trigger' pane on the right is open, showing the configuration for a 'ScheduleTrigger' named 'RatingsTrigger'. The trigger is set to start on '12/7/2023, 10:05:00 PM' in the 'Eastern Time (US & Canada) (UTC-5)' time zone. The recurrence is set to 'Every 1 Day(s)'. The 'Advanced recurrence options' section shows 'Execute at these times' with 'Hours' and 'Minutes' fields. The 'Schedule execution times' field is set to '22:05'. The 'Status' is set to 'Started'.

Similarly, for Customers table as new customers can join the platform every day:

CustomersTableCustomersSQLDBBlobToSQLCustom...RatingsTable

ValidateDebugTrigger (1)

Stored procedureDocument-SP

ParametersVariablesSettingsOutput

Pipeline run ID: Saa6d37b-e99f-438e-8f09-bf36cc6f132a

All status

Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration
Document-SP	Succeeded	Stored procedure	12/7/2023, 9:48:48 PM	2s

Edit trigger

Name *CustomersTrigger

Description

Type *ScheduleTrigger

Start date *12/7/2023, 10:12:00 PM

Time zone *Eastern Time (US & Canada) (UTC-5)

This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.

Recurrence *Every 1 Day(s)

Advanced recurrence options

Execute at these times

Hours

Minutes

Schedule execution times22:12

Specify an end date

Annotations

New

Successful runs of our Scheduled and Manual Triggers:

Microsoft AzureData FactoryADMS

Search factory and documentation

»«

Dashboard

Runs

Pipeline runs

Trigger runs

Change Data Capture (previ...

Runtimes & sessions

Integration runtimes

Data flow debug

Notifications

Alerts & metrics

Pipeline runs

TriggeredDebugRerunCancel optionsRefreshEdit columnsListGantt

Filter by run ID or nameLocal time : Last 24 hoursPipeline name : AllStatus : AllRuns : Latest runsTriggered by : AllAdd filter

Showing 1 - 3 items

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Param
Stored Procedure	12/7/2023, 10:12:00 PM	12/7/2023, 10:12:05 PM	5s	CustomersTrigger	Succeeded	Original	
BlobToSQLRatings	12/7/2023, 10:05:00 PM	12/7/2023, 10:05:17 PM	18s	RatingsTrigger	Succeeded	Original	
> BlobToSQLRatings	12/7/2023, 9:58:47 PM	12/7/2023, 9:59:01 PM	15s	Manual trigger	Succeeded	Rerun (Latest)	

Here are some reference screenshots to show that the tables update & populated successfully:

```
18 SELECT * FROM Movies
19 --done pipelining
20 CREATE TABLE Acting(
```

Results		Messages			
	MovieID	Title	Year	Movie_Rating	IMDB_Rating
1	tm1000037	Je suis Karl	2021	R	5.4
2	tm1000147	Zone 414	2021	R	4.9
3	tm100015	Takers	2010	PG-13	6.2
4	tm1000166	Wave of Cinema: Surat Dar...	2020	PG-13	7.5
5	tm1000185	Squared Love	2021	PG-13	5.1
6	tm100027	Alibaba Aur 40 Chor	1979	PG-13	6.2
7	tm1000296	New Gods: Nezha Reborn	2021	PG-13	6.8
8	tm1000551	Namaste Wahala	2020	PG-13	5.1
9	tm1000599	The Last Forest	2021	PG-13	7.3
10	tm1000619	Radhe Shyam	2022	PG-13	5.3
11	tm1000797	Ride or Die	2021	R	5.6
12	tm100106	My Amnesia Girl	2010	PG-13	6.7
13	tm1001095	Roohi	2021	PG-13	4.3
14	tm1001097	Beauty	2022	R	3.9
15	tm1001108	Biggie: I Got a Story to ...	2021	R	6.8

Row Count validation:

```
8
9 SELECT count(*) FROM dbo.Movies;
10
```

Results		Messages	
	(No column name)		
1	5489		

```
5
6 SELECT count(*) FROM dbo.Ratings;
7
```

Results		Messages	
	(No column name)		
1	1000		