ETL On AZURE

The use of Azure Data Factory, an ETL tool in Azure to assist us in performing Extract, Transform, and Load (ETL) processes for data

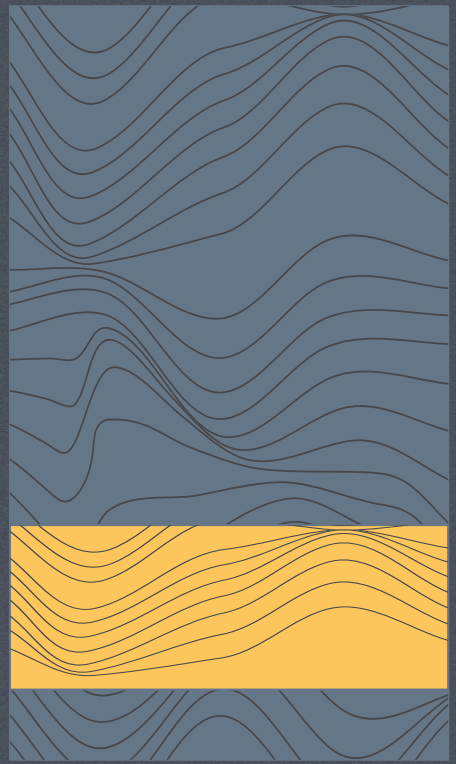
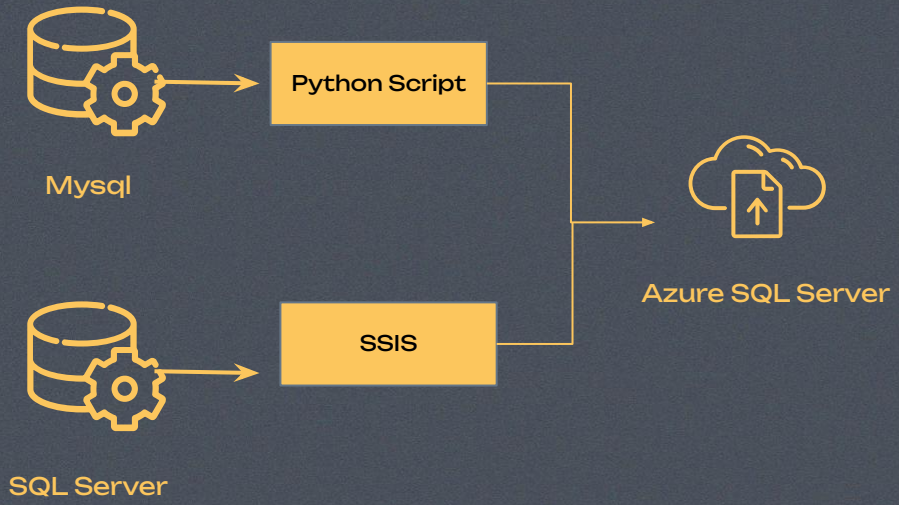


TABLE OF CONTENTS

Scenario 1



01

Scenario 1

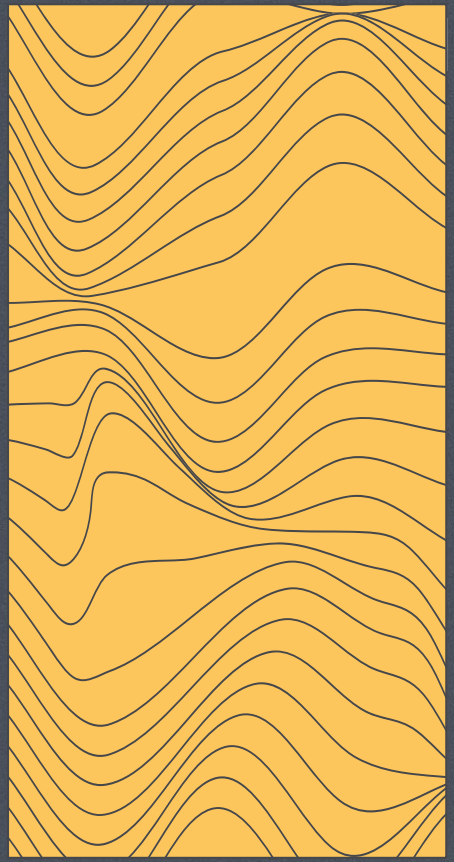
Python Script

On a python script

- Data from MySQL will be ingested into SQL Server Azure
- The Target Table is an empty table, with the addition of Source and Synch Date columns (Last time data was synch)
- Every time data is ingested into SQL Server Azure, the Last_synch column on MYSQL will be updated based on when the data was updated

Note: Weaknesses using Python Scripts takes long time to ingest data

Click to go to Script



Scenario 1

Data Workflow Mysql to Sql Server (Local)

Data on Mysql

	OrderID	OrderDate	PropertyID	ProductID	Quantity
1	2999	2023-05-18 20:00:00	13	35	3
2	3000	2023-05-18 20:00:00	2	29	3
3	3001	2023-05-18 20:00:00	20	68	1
4	3002	2023-05-18 20:00:00	5	1	1
5	3003	2023-05-18 20:00:00	18	1	1

Data Contains Order Details from Property Data

data is loaded into the data warehouse

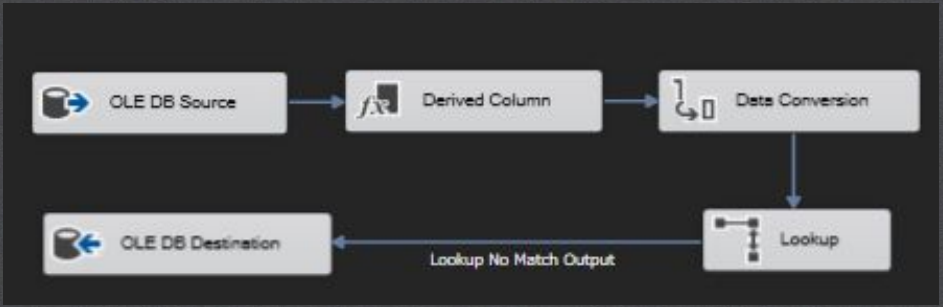
Data on SQL Server

Results | Messages

	OrderID	OrderDate	PropertyID	ProductID	Quantity	source	synch_date
1	2999	2023-05-18 20:00:00.000	13	35	3	MYSQL01	2023-05-24 02:34:04.000
2	3000	2023-05-18 20:00:00.000	2	29	3	MYSQL01	2023-05-24 02:34:04.000
3	3001	2023-05-18 20:00:00.000	20	68	1	MYSQL01	2023-05-24 02:34:04.000
4	3002	2023-05-18 20:00:00.000	5	1	1	MYSQL01	2023-05-24 02:34:04.000
5	3003	2023-05-18 20:00:00.000	18	1	1	MYSQL01	2023-05-24 02:34:04.000

Scenario 1

Data Flow on OrderDetails Dataset



1. OLE DB Source : Extraction Process
2. Derived Column: Perform column transformation by adding the "Source" column (Data Source) And "Synch_date" for the time the data synch
3. Data Conversion: to perform data type conversions in data flow. The Data Conversion component allows you to change the column data type to a different data type according to your needs.
4. Lookup (Optional): to check the data if there is no duplicate data
5. OLE DB Destination : Load Data Location

Scenario 1

Data Workflow SQL Server (Local) to Sql Server (Cloud)

Data on SQL Server

	OrderID	OrderDate	PropertyID	ProductID	Quantity	source	synch_date
1	1	2023-05-18 19:50:00.0000000	17	41	1	NULL	NULL
2	2	2023-05-18 19:50:00.0000000	15	54	2	NULL	NULL
3	3	2023-05-18 19:50:00.0000000	5	71	2	NULL	NULL
4	4	2023-05-18 19:50:00.0000000	19	3	2	NULL	NULL
5	5	2023-05-18 19:50:00.0000000	12	45	3	NULL	NULL

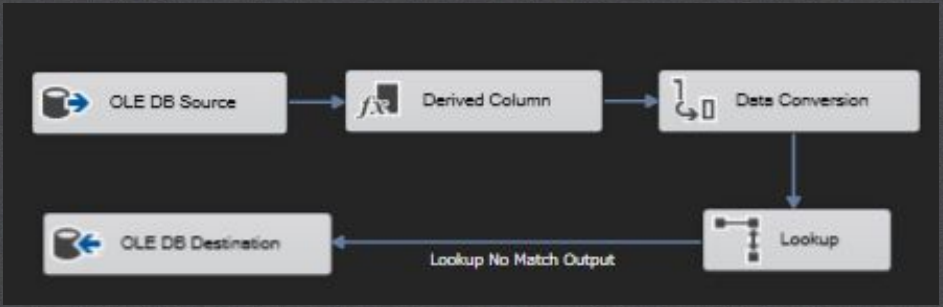
data is loaded into the data warehouse

Data on SQL Server (Cloud)

	OrderID	OrderDate	PropertyID	ProductID	Quantity	source	synch_date
1	1	2023-05-18 19:50:00.000	17	41	1	MSSQL01	2023-05-24 09:48:06.157
2	2	2023-05-18 19:50:00.000	15	54	2	MSSQL01	2023-05-24 09:48:06.157
3	3	2023-05-18 19:50:00.000	5	71	2	MSSQL01	2023-05-24 09:48:06.157
4	4	2023-05-18 19:50:00.000	19	3	2	MSSQL01	2023-05-24 09:48:06.157
5	5	2023-05-18 19:50:00.000	12	45	3	MSSQL01	2023-05-24 09:48:06.157

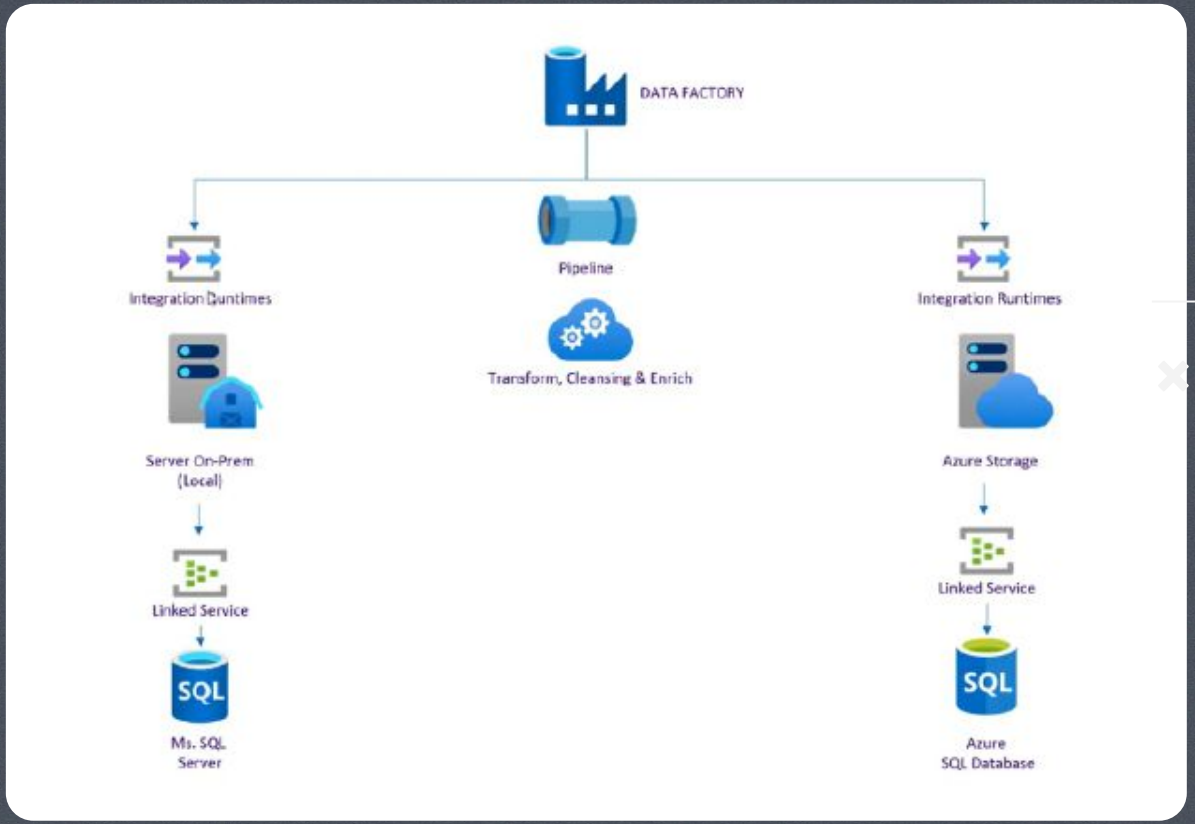
Scenario 1

Data Flow on OrderDetails Dataset



1. OLE DB Source : Extraction Process
2. Derived Column: Perform column transformation by adding the "Source" column (Data Source) And "Synch_date" for the time the data synch
3. Data Conversion: to perform data type conversions in data flow. The Data Conversion component allows you to change the column data type to a different data type according to your needs.
4. Lookup (Optional): to check the data if there is no duplicate data
5. OLE DB Destination : Load Data Location

Scenario 2



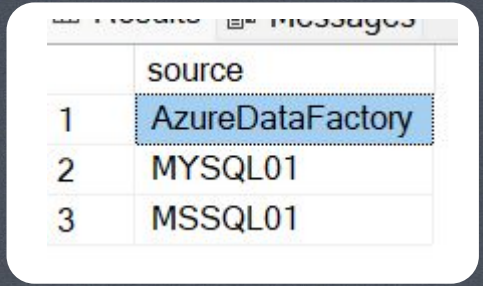
02

Scenario 2

Workflow Explanation

1. Data from SQL Server Local will be extracted
2. Then use Linked Service to connect external data resources to the Azure Data Factory service
3. The Server on-premises setting in Azure refers to using Azure as a platform for hosting and managing servers that are physically located in local or on-premises environments.
4. Integration Runtimes are used to connect and manage access to data resources that reside in different locations and different technologies.
5. Pipelines in Azure are logical constructs used in Azure Data Factory to manage end-to-end workflows or data processes. Pipelines allow you to organize and direct data-related activities into structured series of actions.
6. On the Pipeline we can also set schedules in the form of triggers to debug or copy data

Results



	source
1	AzureDataFactory
2	MYSQL01
3	MSSQL01

There are 3 data sources that have been successfully loaded into the Azure SQL Database

- MYSQL
- Microsoft SQL Server
- Azure Data Factory

Muhammad Khairul

Thanks!

Click!