

# Pós-graduação em Data Science & Business Analytics

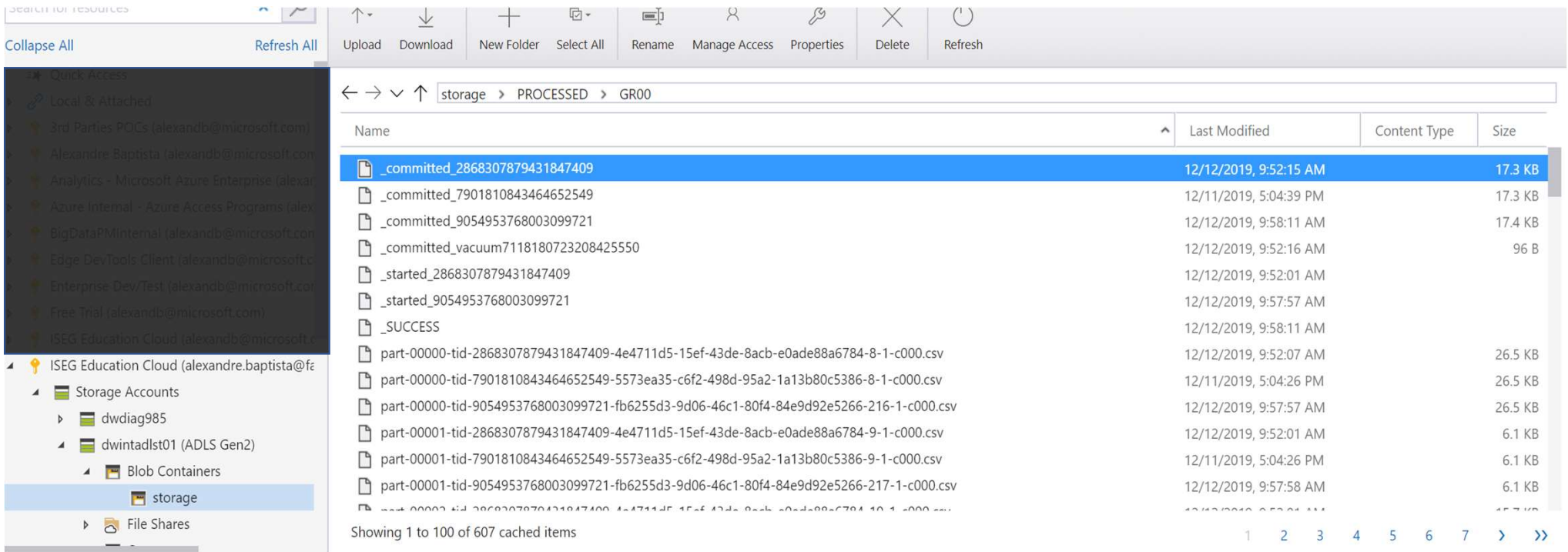
# Index

- Azure Data Factory Concepts
- Azure Data Factory Practical Experience
  - ADLS to DW process
- Final Project Q&A

# Tópico

## Azure Data Factory Concepts

### Data Flow Generated Data for Customers



Search for resources

Collapse All Refresh All

Upload Download New Folder Select All Rename Manage Access Properties Delete Refresh

storage > PROCESSED > GR00

Name	Last Modified	Content Type	Size
_committed_2868307879431847409	12/12/2019, 9:52:15 AM		17.3 KB
_committed_7901810843464652549	12/11/2019, 5:04:39 PM		17.3 KB
_committed_9054953768003099721	12/12/2019, 9:58:11 AM		17.4 KB
_committed_vacuum7118180723208425550	12/12/2019, 9:52:16 AM		96 B
_started_2868307879431847409	12/12/2019, 9:52:01 AM		
_started_9054953768003099721	12/12/2019, 9:57:57 AM		
_SUCCESS	12/12/2019, 9:58:11 AM		
part-00000-tid-2868307879431847409-4e4711d5-15ef-43de-8acb-e0ade88a6784-8-1-c000.csv	12/12/2019, 9:52:07 AM		26.5 KB
part-00000-tid-7901810843464652549-5573ea35-c6f2-498d-95a2-1a13b80c5386-8-1-c000.csv	12/11/2019, 5:04:26 PM		26.5 KB
part-00000-tid-9054953768003099721-fb6255d3-9d06-46c1-80f4-84e9d92e5266-216-1-c000.csv	12/12/2019, 9:57:57 AM		26.5 KB
part-00001-tid-2868307879431847409-4e4711d5-15ef-43de-8acb-e0ade88a6784-9-1-c000.csv	12/12/2019, 9:52:01 AM		6.1 KB
part-00001-tid-7901810843464652549-5573ea35-c6f2-498d-95a2-1a13b80c5386-9-1-c000.csv	12/11/2019, 5:04:26 PM		6.1 KB
part-00001-tid-9054953768003099721-fb6255d3-9d06-46c1-80f4-84e9d92e5266-217-1-c000.csv	12/12/2019, 9:57:58 AM		6.1 KB

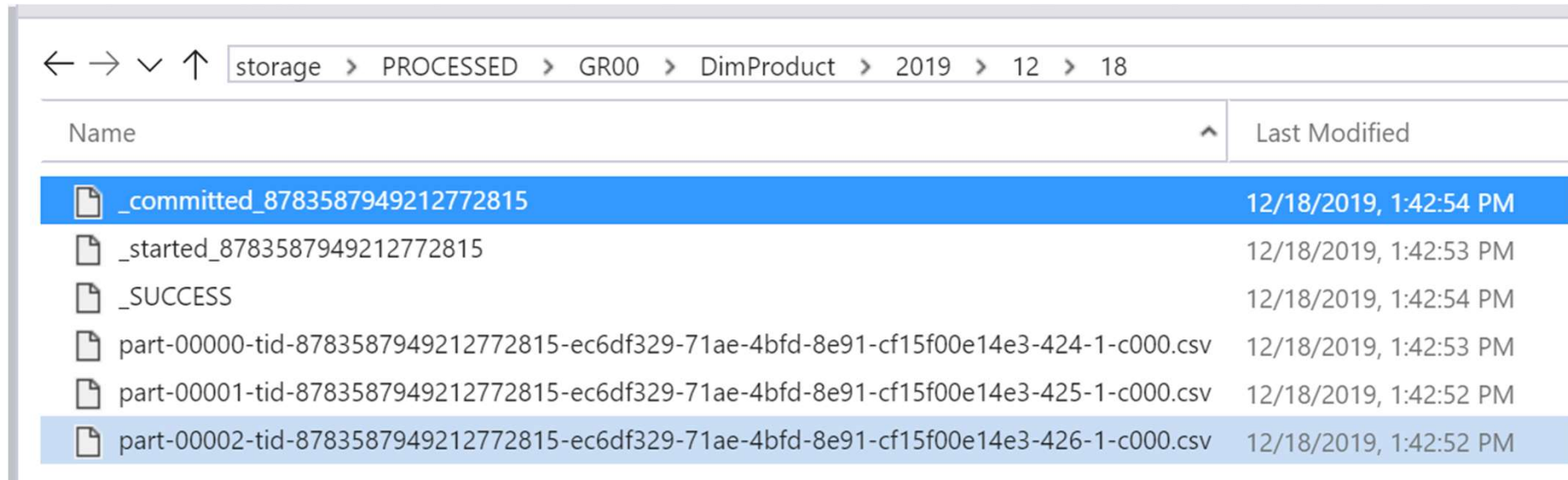
Showing 1 to 100 of 607 cached items

1 2 3 4 5 6 7 > >>







# Tópico

## Azure Data Factory Concepts

### Partition by date exemple



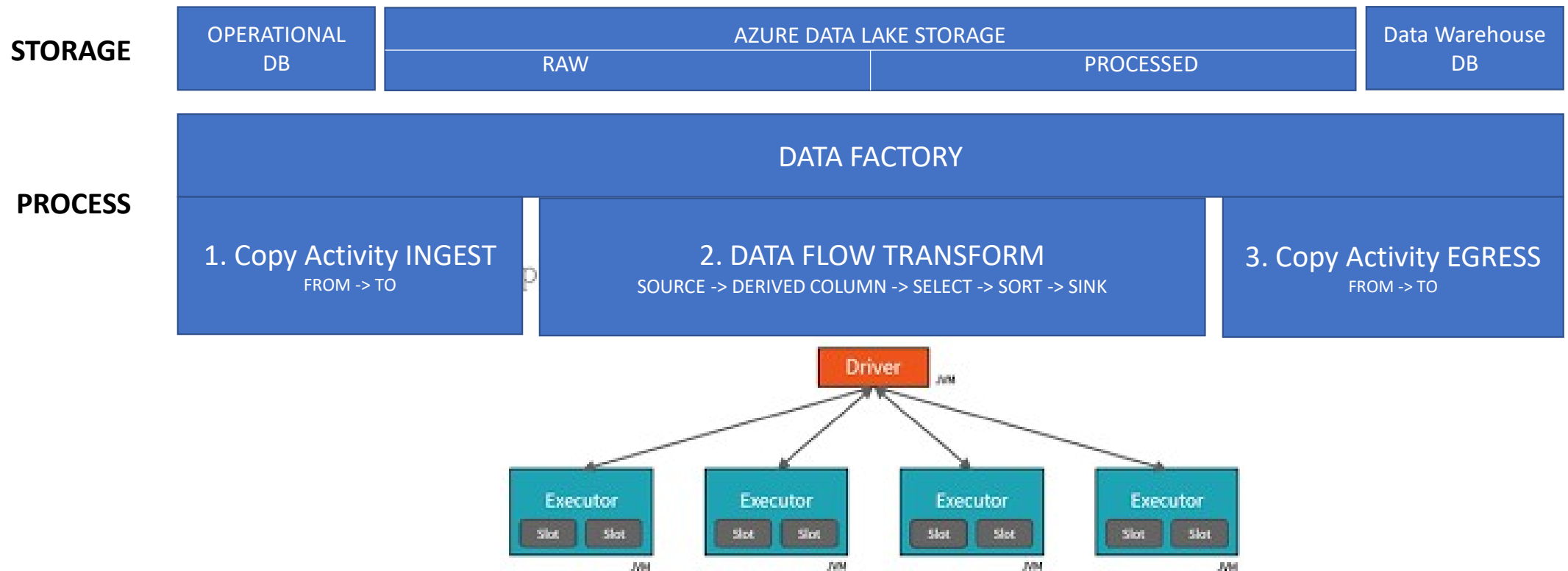
The screenshot shows the Azure Data Factory file explorer interface. The breadcrumb navigation path is: storage > PROCESSED > GR00 > DimProduct > 2019 > 12 > 18. The table below lists the files in this directory, including their names and last modified timestamps.

Name	Last Modified
 _committed_8783587949212772815	12/18/2019, 1:42:54 PM
 _started_8783587949212772815	12/18/2019, 1:42:53 PM
 _SUCCESS	12/18/2019, 1:42:54 PM
 part-00000-tid-8783587949212772815-ec6df329-71ae-4bfd-8e91-cf15f00e14e3-424-1-c000.csv	12/18/2019, 1:42:53 PM
 part-00001-tid-8783587949212772815-ec6df329-71ae-4bfd-8e91-cf15f00e14e3-425-1-c000.csv	12/18/2019, 1:42:52 PM
 part-00002-tid-8783587949212772815-ec6df329-71ae-4bfd-8e91-cf15f00e14e3-426-1-c000.csv	12/18/2019, 1:42:52 PM

# Tópico

## Azure Data Factory Concepts

### Overall ETL Process



# Tópico

## Azure Data Factory Concepts

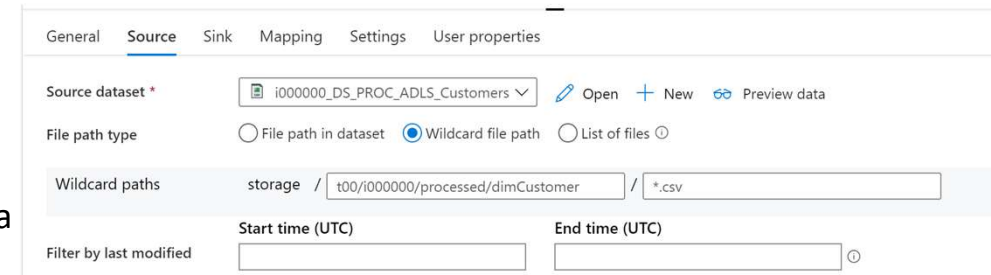
### ConfigParameters

	SourceTable	DestinationTable	active	lastExecution	lastExecutionDW
1	DimCurrency	DimCurrency	1	NULL	NULL
2	DimGeography	DimGeography	1	NULL	NULL
3	DimDate	DimDate	1	NULL	NULL
4	DimProduct	DimProduct	1	NULL	NULL
5	DimProductCategory	DimProductCategory	1	NULL	NULL
6	DimProductSubCate...	DimProductSubCate...	1	NULL	NULL
7	FactInternetSales	FactInternetSales	1	NULL	NULL

# ADF Copy Activity – Configure Source for DW

In this task, you will set the Source settings for Copy Activity to DW

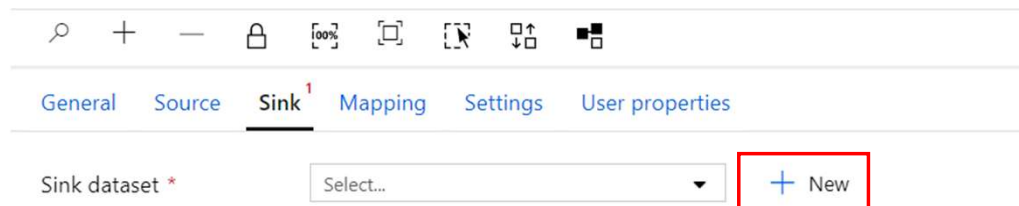
1. In the Data Factory pane pick the ... option right to the **Pipelines** section
2. Select **New pipeline** option
3. Rename the new Data flow **IXXXXXX - Egress To Azure DW**
4. Add **Fetch ADLS processed data and store it in DW** to the description
5. Drag a **Copy Data Activity** available in the Move & Transform section
6. Set the Name for the Copy activity to **Egress Customers To Azure DW**
7. Set the Description for the Copy activity to **Fetch ADLS processed Customers data and store it in DW**
8. Navigate to the **Source TAB**
9. Select the **iXXXXXX\_DS\_PROC\_ADLS\_Customers** for the Dataset name
10. Make sure the **Recursively** option is selected
11. Set the value **i000000/processed/dimCustomer** and **\*.csv** in the Wildcard file name and **"wildcard file path"** option






# ADF Copy Activity – Configure Source for DW

In this task, you will set the Source settings for Copy Activity to DW

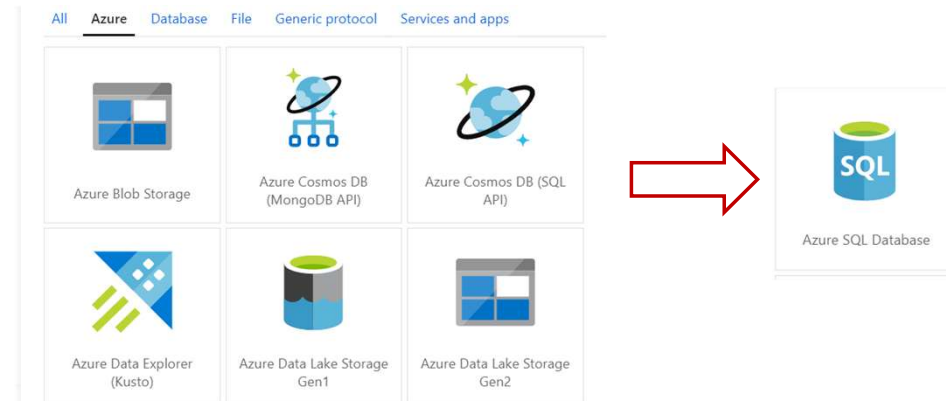
1. Press the **Preview Data** button and ensure Customers Data Is readable and ready to be used
2. Choose the **Sink TAB Window**
3. Press **New** button next to the Sink Dataset Drop Down List
4. In the New Dataset Window select the **Azure Tab**
5. Select the **Azure SQL Database** and Select **Continue**
6. Press the **Continue** button



Preview data  

Linked service: GPXX\_LS\_ADLS\_DATA

CustomerKey	GeographyKey	CustomerAlternateKey	Title	FirstName	MiddleName	LastName	NameStyle	Birth
29285	300	AW00029285		Thomas		Adams	false	1954-14
12663	223	AW00012663		Morgan		Adams	false	1952-21
12889	547	AW00012889		James		Adams	false	1977-26
13280	623	AW00013280		Bailey		Adams	false	1981-12

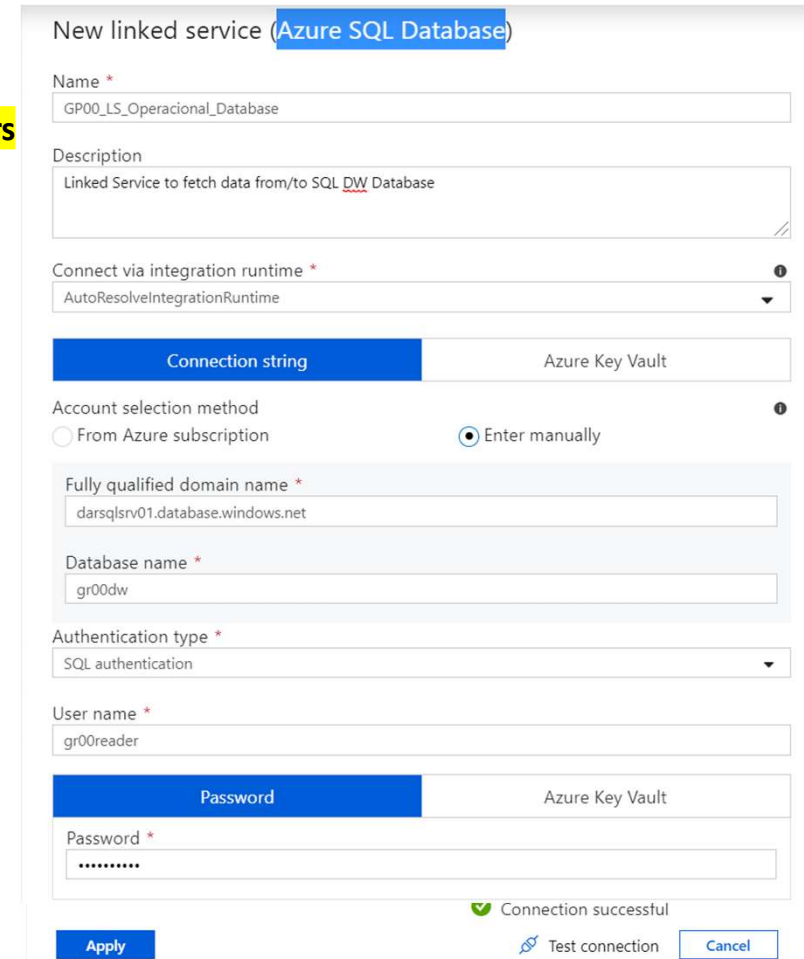




# ADP Copy Activity – Configure Sink

## In this task, you will set the Sink settings for Copy Activity

1. In the set properties window set the name to **IXXXXXX\_DS\_PROC\_ADW\_Customers**
2. In the Linked Service Drop Down List select **New** option
3. In the New linked Service (Azure SQL Database) set Name value to **ixxxxxx\_LS\_DW\_DATA**
4. Set Description value to **Linked Service to manage data inside Azure SQL DW Database**
5. In the Connect Via Integration Runtime select **AutoResolveIntegrationRuntime**
6. In Account Selection Method select **Enter manually**
7. In the Fully Qualified Domain Name insert **--sqlserver dw--**
8. Set the Database Name field value to **--db dw--** and Authentication Type to **SQL Authentication**
9. Set the username to **--user db dw--** and the Password to **--psw db dw--**
10. Press **Test Connection**
11. Press **Create** button



New linked service (Azure SQL Database)

Name \*  
GP00\_LS\_Operacional\_Database

Description  
Linked Service to fetch data from/to SQL DW Database

Connect via integration runtime \*  
AutoResolveIntegrationRuntime

Connection string | Azure Key Vault

Account selection method  
☐ From Azure subscription ☒ Enter manually

Fully qualified domain name \*  
darsqlsrv01.database.windows.net

Database name \*  
gr00dw

Authentication type \*  
SQL authentication

User name \*  
gr00reader

Password | Azure Key Vault

Password \*  
\*\*\*\*\*

✔ Connection successful

Apply Test connection Cancel

# ADF Copy Activity – Configure Sink

In this task, you will set the Sink settings for Copy Activity

1. Press **OK living the Table Name empty**
2. In the Sink Pane in the Store Procedure name field select **None**
3. In the Table Options choose **Auto create table**
4. Open Sink dataset and **check edit chechbox** and in the table textbox enter **iXXXXXX** and **dimCustomer**
5. Get back to Sink Tab
6. In the Pre-Copy Script write **IF OBJECT\_ID('iXXXXXX.dimCustomer') IS NOT NULL TRUNCATE TABLE iXXXXXX.dimCustomer**  
(don't forget to change XXXXXX by real number)

## Set properties

Name  
i000000\_DS\_PROC\_ADW\_Customers

Linked service \*  
i000000\_LS\_DW\_DATA

Table name  
Edit

Import schema  
☐ From connection/store ☒ None

Advanced

General Source Sink Mapping Settings User properties

Sink dataset \*  
i000000\_DS\_PROC\_ADW\_Customers Open + New

Stored procedure name  
Select... Refresh  
Edit

Table option  
☐ None ☒ Auto create table

Pre-copy script  
IF OBJECT\_ID('i000000.dimCustomer') IS NOT NULL TRUNCATE TABLE i000000.[Customer]

Write batch timeout

Connection Schema Parameters

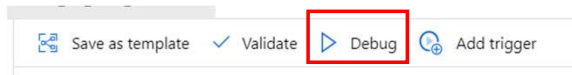
Linked service \*  
i000000\_LS\_DW\_DATA Test connection Edit + New

Table  
i000000 . dimCustomer  
Edit

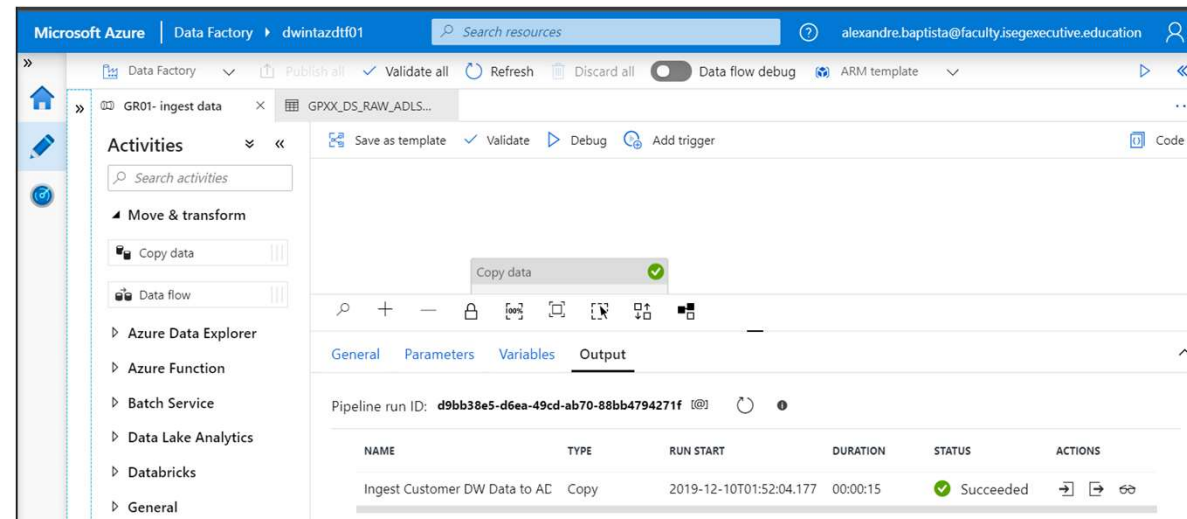
# Execute the Azure Data Factory pipeline

In this task, you will run the pipeline and  
**Validate the outcome**

1. Open the Pipeline from the left panel
2. Press the **Debug** button and run the package



3. Make sure the **pipeline runs successfully**
4. Open the Actions buttons and  
explore its content



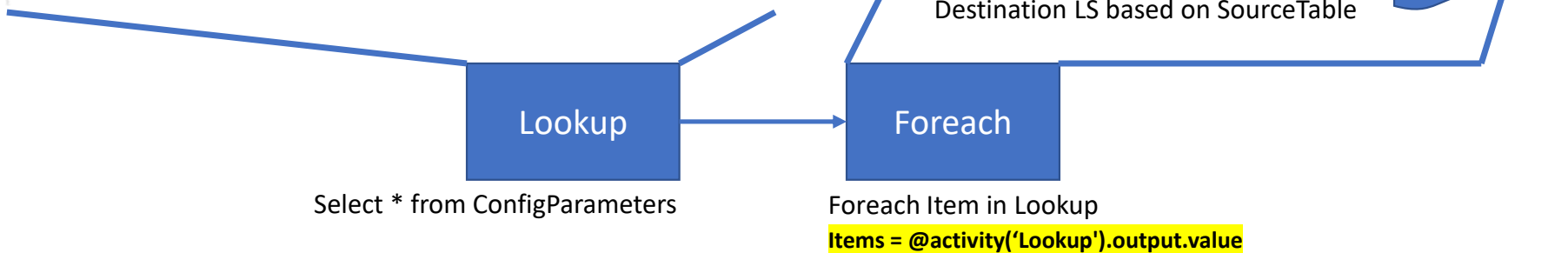
If the pipeline execution fails, use the error action button to understand and correct the problem that occurred. If no error occurred use the pipeline details button mentioned above to understand pipeline's statistics (how long did it took, number of lines that were processed, etc).

# Automating the Egress to DW Pipeline

In this task, you will egress all other tables to DW using yesterdays learnings

- Lookup activity to fetch data from table **ConfigParameters**
- Foreach activity to loop all tables and fetch information from ADLS and Store in DW

	SourceTable	DestinationTable	active	lastExecution	lastExecutionDW
1	DimCurrency	DimCurrency	1	NULL	NULL
2	DimGeography	DimGeography	1	NULL	NULL
3	DimDate	DimDate	1	NULL	NULL
4	DimProduct	DimProduct	1	NULL	NULL
5	DimProductCategory	DimProductCategory	1	NULL	NULL
6	DimProductSubCategory	DimProductSubCategory	1	NULL	NULL
7	FactInternetSales	FactInternetSales	1	NULL	NULL

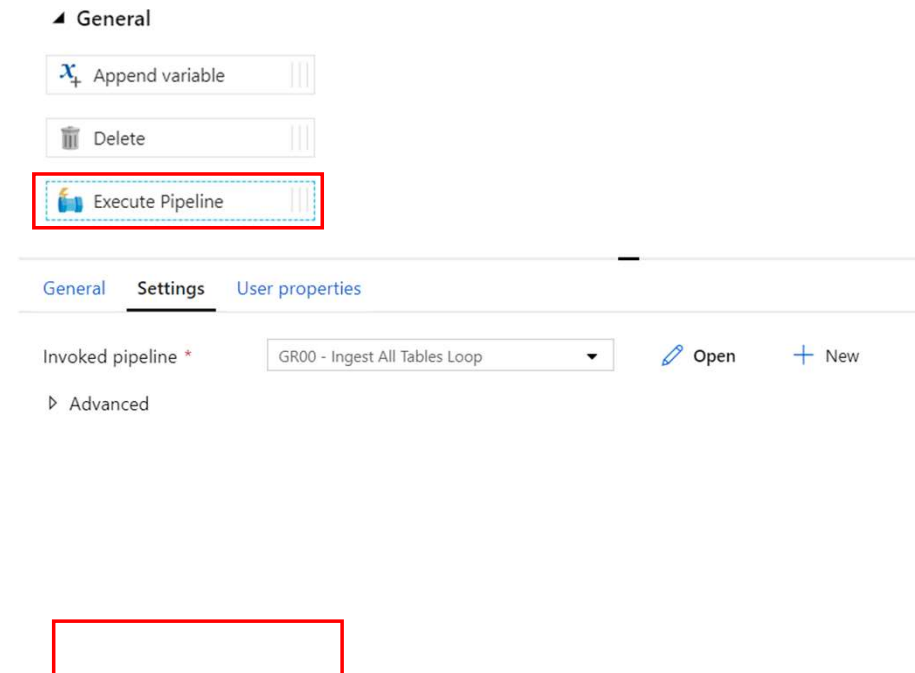


# Tópico

## ADF Data Factory – Main pipeline

### In this task, you will create the main pipeline

1. Create a new Pipeline using the **new pipeline** option
2. Name the pipeline **iXXXXXX - Main Pipeline**
3. Set the Description value to **Pipeline that will invoke all others**
4. Drag the newly created pipeline to the **correct group folder**
5. From the General Section In the Activities Pane Drag **Execute Pipeline** activity
6. Name the new activity **Ingest data**
7. Set the description field to **Ingest Customers**
8. Navigate to the **Settings TAB**
9. Set the Invoked Pipeline to **iXXXXXX – ingest data**

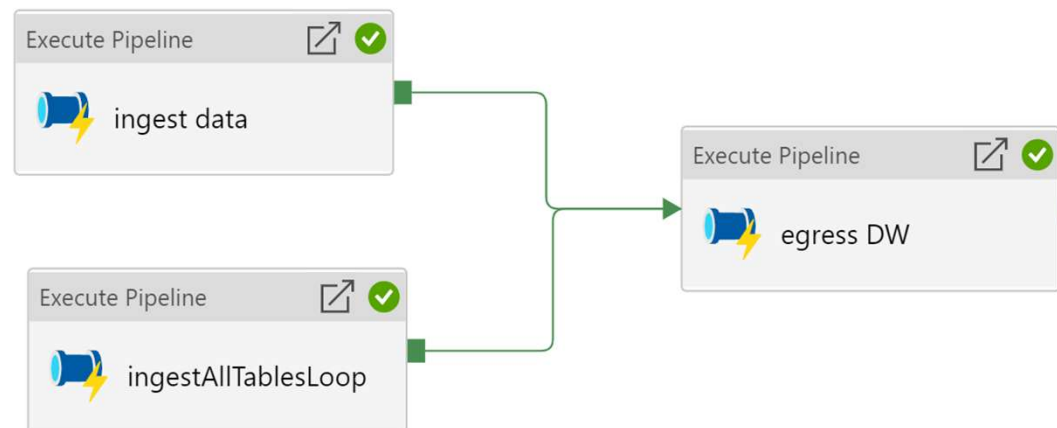


# Tópico

## ADF Data Factory – Main pipeline

### In this task, you will create the main pipeline

1. Do the previous step to add execute pipelines for the created pipelines: '... IngestAllTablesLoop and '... egress to Azure DW'
2. **Connect** the new Activity to the previous ones as described







Executive  
Education

[www.isegexecutive.education](http://www.isegexecutive.education)

Rua do Quelhas, 6  
1200-781 Lisboa

(+351) 213 922 891  
[info@executive.education](mailto:info@executive.education)