



Pós-graduação em Data Science & Business Analytics

Index






- Azure Data Factory Concepts
- Azure Data Factory Practical Experience
 - Organize our Data Factory Artifacts
 - Create a Data Flow
 - Derived Column Activity
 - Select Activity
 - Sort Activity
 - Sync to ADLS Processed stage
- Projecto Final

ADF – Data Factory Concepts








Multiple input/Output

- **Join** - Use the join transformation to **combine data from two sources or streams in a mapping data flow**. The output stream will include all columns from both sources matched based on a join condition
- **Conditional split** - The conditional **split transformation routes data rows to different streams** based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.
- **Exists** - The exists transformation is a **row filtering transformation that checks whether your data exists in another source or stream**. The output stream includes all rows in the left stream that either exist or don't exist in the right stream. The exists transformation is similar to SQL WHERE EXISTS and SQL WHERE NOT EXISTS.
- **Union** - Union will **combine multiple data streams into one**, with the SQL Union of those streams as the new output from the Union transformation. All of the schema from each input stream will be combined inside of your data flow, without needing to have a join key.
- **Lookup** - Use Lookup to **add reference data from another source to your Data Flow**. The Lookup transform requires a defined source that points to your reference table and matches on key fields.




Multiple inputs/outputs

-  Join
-  Conditional Split
-  Exists
-  Union
-  Lookup

Schema modifier

-  Derived Column
-  Select
-  Aggregate
-  Surrogate Key
-  Pivot
-  Unpivot
-  Window

Row modifier

-  Filter
-  Sort
-  Alter Row

Destination






-  Sink

ADF – Data Factory Concepts








Schema modifier

- **Derived Column** - Use the derived column transformation to generate new columns in your data flow or to modify existing fields.
- **Select** - Use this transformation for column selectivity (reducing number of columns), alias columns and stream names, and reorder columns.
- **Aggregate** - The Aggregate transformation defines aggregations of columns in your data streams. Using the Expression Builder, you can define different types of aggregations such as SUM, MIN, MAX, and COUNT grouped by existing or computed columns.
- **Surrogate Key** - Use the Surrogate Key Transformation to add an incrementing non-business arbitrary key value to your data flow rowset. This is useful when designing dimension tables in a star schema analytical data model where each member in your dimension tables needs to have a unique key that is a non-business key, part of the Kimball DW methodology.
- **Pivot** - Use Pivot in ADF Data Flow as an aggregation where one or more grouping columns has its distinct row values transformed into individual columns. Essentially, you can Pivot row values into new columns (turn data into metadata).
- **Unpivot** - Use Unpivot in ADF mapping data flow as a way to turn an unnormalized dataset into a more normalized version by expanding values from multiple columns in a single record into multiple records with the same values in a single column.
- **Window** - The Window transformation is where you will define window-based aggregations of columns in your data streams. In the Expression Builder, you can define different types of aggregations that are based on data or time windows (SQL OVER clause) such as LEAD, LAG, NTILE, CUMEDIST, RANK, etc.). A new field will be generated in your output that includes these aggregations. You can also include optional group-by fields.




Multiple inputs/outputs

-  Join
-  Conditional Split
-  Exists
-  Union
-  Lookup

Schema modifier

-  Derived Column
-  Select
-  Aggregate
-  Surrogate Key
-  Pivot
-  Unpivot
-  Window

Row modifier

-  Filter
-  Sort
-  Alter Row

Destination

-  Sink

ADF – Data Factory Concepts






Row modifier

- **Filter** - The Filter transformation allows row filtering based upon a condition. The output stream includes all rows that matching the filtering condition. The filter transformation is similar to a WHERE clause in SQL.
- **Sort** - The Sort transformation allows you to sort the incoming rows on the current data stream. The outgoing rows from the Sort Transformation will subsequently follow the ordering rules that you set. You can choose individual columns and sort them ASC or DEC, using the arrow indicator next to each field. If you need to modify the column before applying the sort, click on "Computed Columns" to launch the expression editor. This will provide with an opportunity to build an expression for the sort operation instead of simply applying a column for the sort.
- **Alter Row** - Use the Alter Row transformation to set insert, delete, update, and upsert policies on rows. You can add one-to-many conditions as expressions. These conditions should be specified in order of priority, as each row will be marked with the policy corresponding to the first-matching expression. Each of those conditions can result in a row (or rows) being inserted, updated, deleted, or upserted. Alter Row can produce both DDL & DML actions against your database.








Destination

- **Sink** - After you transform your data flow, you can sink the data into a destination dataset. In the sink transformation, choose a dataset definition for the destination output data. You can have as many sink transformations as your data flow requires.




Multiple inputs/outputs

-  Join
-  Conditional Split
-  Exists
-  Union
-  Lookup

Schema modifier

-  Derived Column
-  Select
-  Aggregate
-  Surrogate Key
-  Pivot
-  Unpivot
-  Window

Row modifier

-  Filter
-  Sort
-  Alter Row

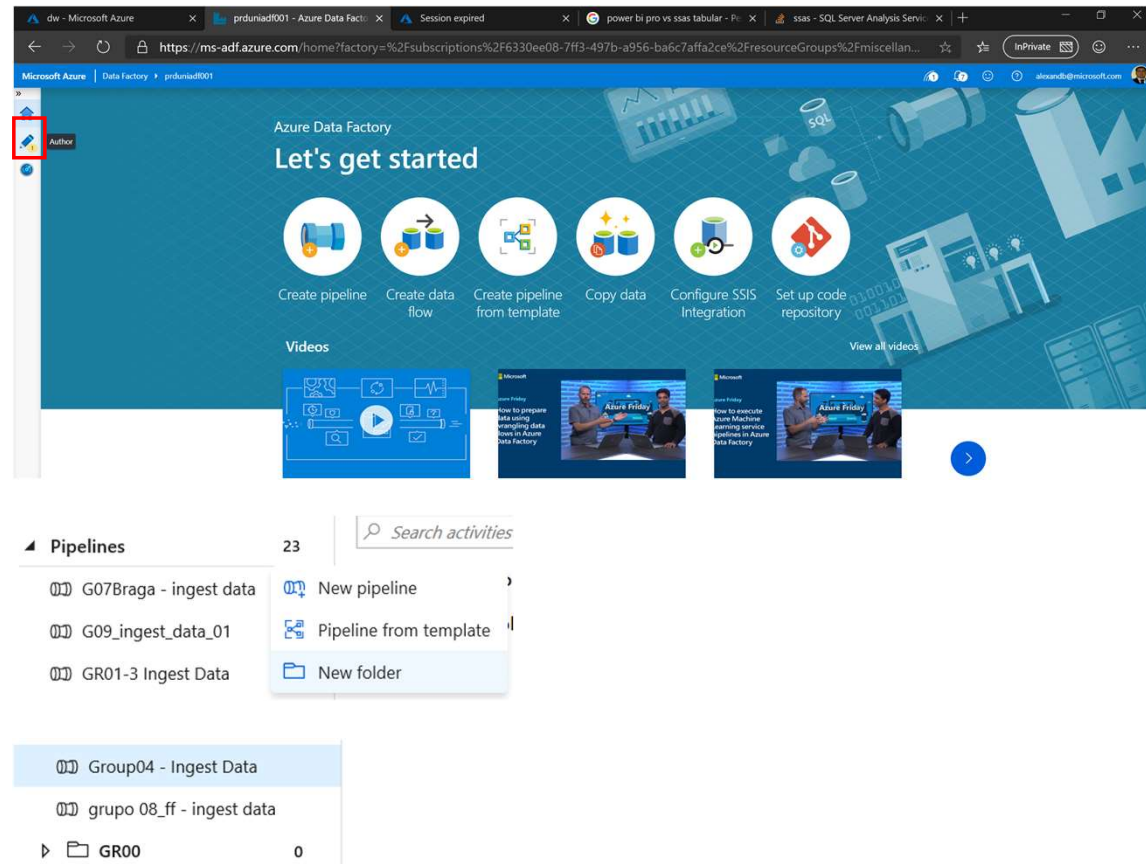
Destination

-  Sink

ADF – Organize Data Factory Artifacts

In this task, you will organize the Data Factory Artifacts

1. **Open** Data Factory
2. Choose the **Author** pen icon on the Right
3. In the Factory Resources pick the **...** option right to the Pipelines section
4. Select **New Folder**
5. Name the new Folder **/iXXXXXX**
6. **Move your pipelines** into the created Folder
7. Perform the same operation in **Datasets sections**
8. **Create the Group Folders** for the Dataflows section too

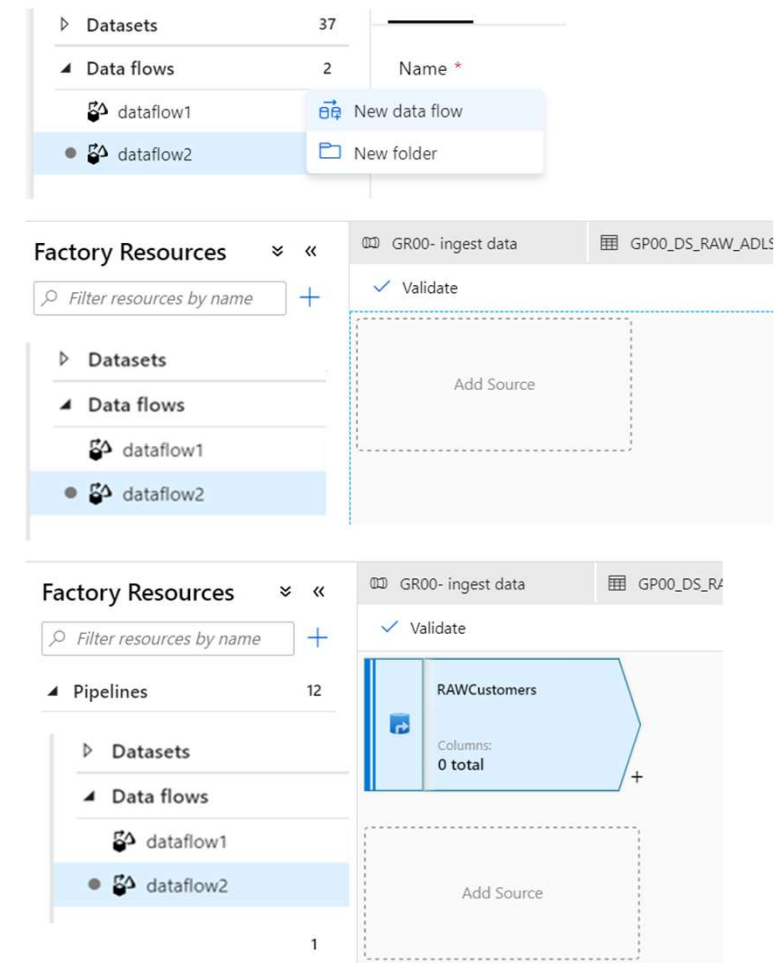


The screenshot shows the Azure Data Factory Author interface. The top section is titled 'Let's get started' and includes buttons for 'Create pipeline', 'Create data flow', 'Create pipeline from template', 'Copy data', 'Configure SSIS Integration', and 'Set up code repository'. Below this is a 'Videos' section. The left sidebar shows 'Factory Resources' with a search bar and a list of resources: Pipelines (23), Datasets (66), Data flows (0), and a folder GR00 (0). The 'Pipelines' section is expanded, showing a list of pipelines: G07Braga - ingest data, G09_ingest_data_01, and GR01-3 Ingest Data. A context menu is open over the 'Pipelines' section, showing options: 'New pipeline', 'Pipeline from template', and 'New folder'.

ADF Data Flows – Mapping Data Flow

In this task, you will create a Mapping Data Flow

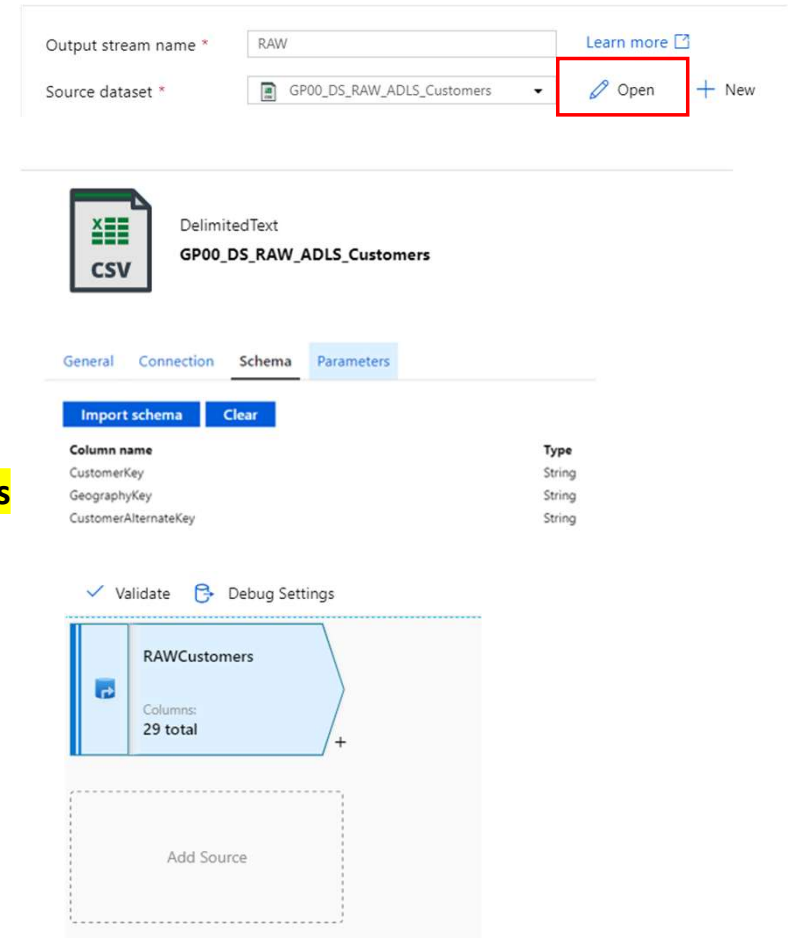
1. In the Factory Resources pick the **...** option right to the **Data Flows** section
2. Select **New data flow** option
3. Select **Mapping Data Flow**
4. Rename the new Data flow **iXXXXXX_TransformSalesData**
5. Add **Dataflow process to prepare the Sales Data** to the description
6. Click on the **Add Source** rectangle inside the main area
7. In the Source Section set the output stream name to **RAWCustomers**
8. Select **iXXXXXX_DS_RAW_ADLS_Customers** in Source Dataset
9. Select **Allow Schema drift** and **Infer drifted column types** to ON
10. Validate that the Columns is set to **0 Total**



ADF Data Flows – Mapping Data Flow

In this task, you will create a Mapping Data Flow

1. For columns to be recognized go to Source settings TAB
2. Press **Open** next to the Source Dataset
3. In the Dataset option select **Schema**
4. Press the **Import Schema** button and ensure fields are populated
5. Navigate back to the Data Flow using the opened TABS
6. Make sure the columns were populated and that you now see **29 Columns** in Total
7. Select the **RAWCustomers** Activity
8. Navigate to **Projections** TAB and check that all columns are set to String
9. This is because data is coming from CSV format with no schema definition



The screenshot shows the 'Source settings' tab for a Mapping Data Flow. At the top, the 'Output stream name' is set to 'RAW'. Below it, the 'Source dataset' is 'GP00_DS_RAW_ADLS_Customers'. A red box highlights the 'Open' button next to the dataset name. Below the dataset selection, there is a 'DelimitedText' icon and the dataset name 'GP00_DS_RAW_ADLS_Customers'. The 'Schema' tab is selected, showing a table with columns: 'CustomerKey' (String), 'GeographyKey' (String), and 'CustomerAlternateKey' (String). Below the table, there are 'Import schema' and 'Clear' buttons. At the bottom, there is a 'Validate' button and a 'Debug Settings' button. Below these buttons, there is a 'RAWCustomers' activity box showing 'Columns: 29 total'. Below the activity box, there is a dashed box labeled 'Add Source'.

ADF Data Flows – Mapping Data Flow

In this task, you will create a Mapping Data Flow

1. In the Projections Window
2. Change **CustomerKey** type to Integer
3. Change **GeographyKey** type to Integer
4. Change **NameStyle** type to Boolean
5. Change **BirthDate** type to Date
6. Change **BirthDate** Format to dd/MM/yyyy
7. Change **YearlyIncome** type to decimal
8. Change **TotalChildren** type to Integer
9. Change **NumberChildrenAtHome** type to Integer
10. Change **DateFirstPurchase** type to Date
11. Change **DateFirstPurchase** Format to dd/MM/yyyy

✓ Validate ⚙ Debug Settings

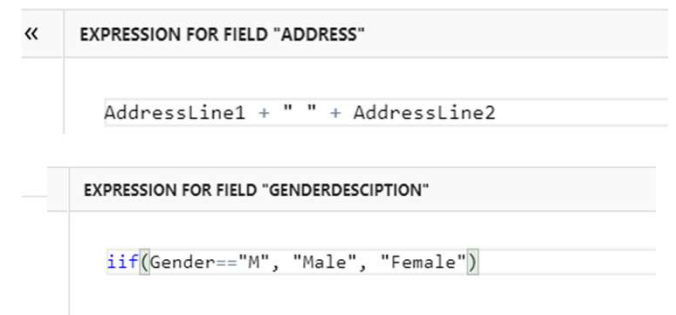
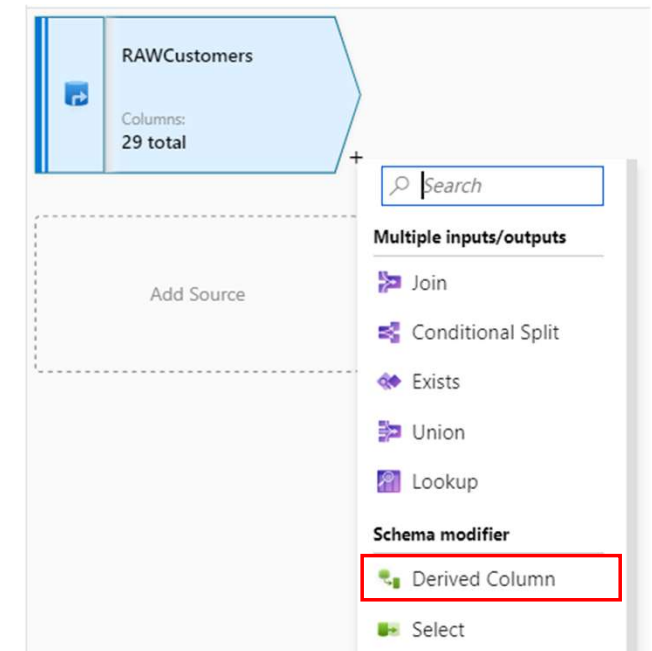
Source settings Source options Projection Optimize Inspect Data preview ●

Column name	Type	Format
CustomerKey	123 integer	Specify format
GeographyKey	123 integer	Specify format
CustomerAlternateKey	abc string	Specify format
Title	abc string	Specify format
FirstName	abc string	Specify format
MiddleName	abc string	Specify format
LastName	abc string	Specify format
NameStyle	☑ boolean	Specify format
BirthDate	📅 date	dd/MM/yyyy
MaritalStatus	abc string	Specify format
Suffix	abc string	Specify format
Gender	abc string	Specify format
EmailAddress	abc string	Specify format
YearlyIncome	e ^x decimal	Specify format
TotalChildren	123 integer	Specify format
NumberChildrenAtHome	123 integer	Specify format
EnglishEducation	abc string	Specify format
SpanishEducation	abc string	Specify format
FrenchEducation	abc string	Specify format
EnglishOccupation	abc string	Specify format
SpanishOccupation	abc string	Specify format
FrenchOccupation	abc string	Specify format
HouseOwnerFlag	abc string	Specify format
NumberCarsOwned	abc string	Specify format
AddressLine1	abc string	Specify format
AddressLine2	abc string	Specify format
Phone	abc string	Specify format
DateFirstPurchase	📅 date	dd/MM/yyyy
CommuteDistance	abc string	Specify format

ADF Data Flows – Mapping Data Flow

In this task, you will add derived columns to the Customer Stream

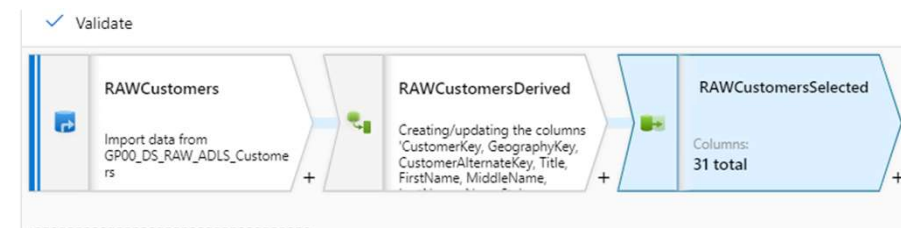
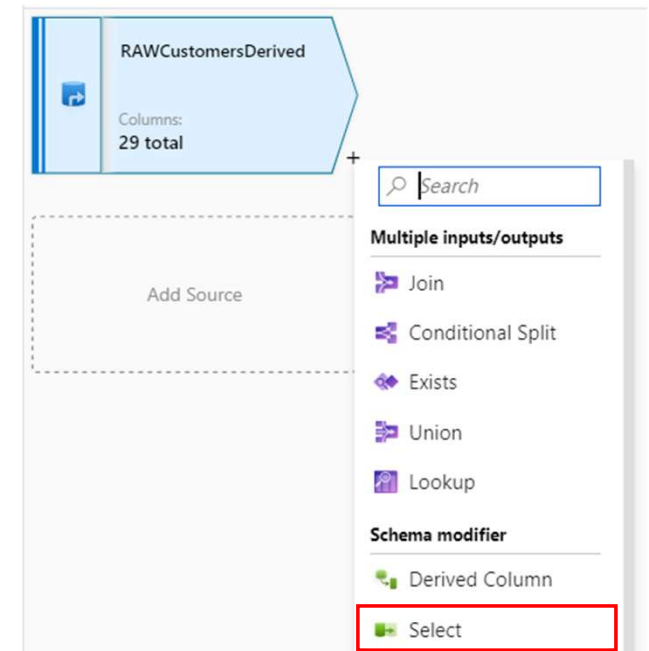
1. In the main pane select the + next to the RAWCustomers Activity
2. Select the **Derived Column** Option
3. Make sure the Derived Column activity is selected
4. Set the Output stream name to **RAWCustomersDerived**
5. In the Derived Column's Settings Name the first column **Address**
6. **Press the Value field** in front of the Column name
7. In the Expression for field "Address"
write **AddressLine1+" "+iif(isNull(AddressLine2), "", AddressLine2)**
1. Press the **Save and Finish** button
2. Press the **+** button in front of the column and choose **Add Column**
3. Name the second column **GenderDescription**
4. **Press the Value field** in front of the Column name
5. In Expression for field "GenderDescription" write **iif(Gender=="M", "Male", "Female")**
6. Press the **Save and Finish** button



ADF Data Flows – Mapping Data Flow

In this task, you will select the Columns from Customers to use

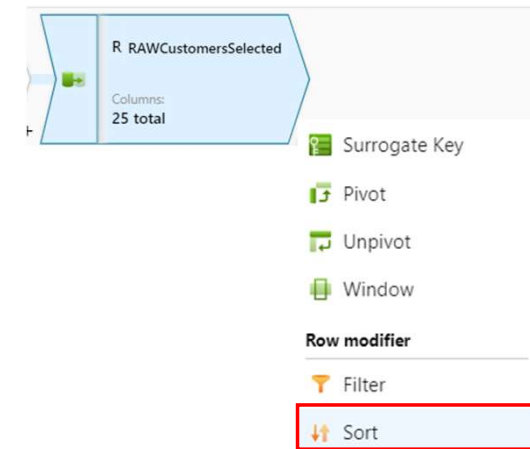
1. In the main pane select the + next to the RAWCustomersDerived Activity
2. Select the **Select** Option
3. Make sure the Select activity is selected
4. Set the Output Stream Name to **RAWCustomersSelected**
5. Remove Column **SpanishEducation** from the Columns to Select
6. Remove Column **FrenchEducation** from the Columns to Select
7. Remove Column **SpanishOccupation** from the Columns to Select
8. Remove Column **FrenchOccupation** from the Columns to Select
9. Remove Column **AddressLine1** from the Columns to Select
10. Remove Column **AddressLine2** from the Columns to Select



ADF Data Flows – Mapping Data Flow

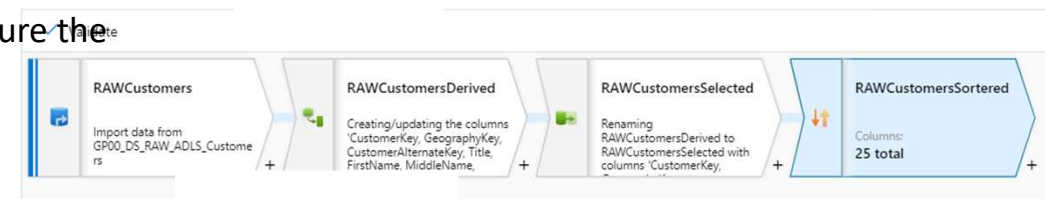
In this task, you will sort the Customers data to be used

1. In the main pane select the + next to the RAWCustomersSelected Activity
2. Select the **Sort** Option
3. Make sure the Sort activity is selected
4. Set the Output Stream Name to **RAWCustomersSorted**
5. In the Sort Conditions select the Column **LastName** and make sure the order is **Ascending**
6. Press the **+** button in front of the column
7. In the Sort Conditions select the Column **MiddleName** and make sure the order is **Ascending**
8. Press the **+** button in front of the column
9. In the Sort Conditions select the Column **FirstName** and make sure the order is **Ascending**



☐ Case insensitive
 ☐ Sort only within partition

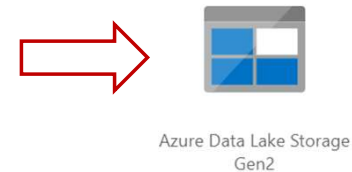
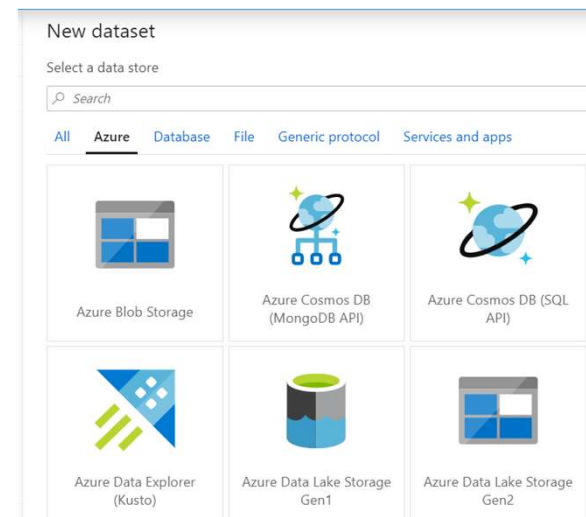
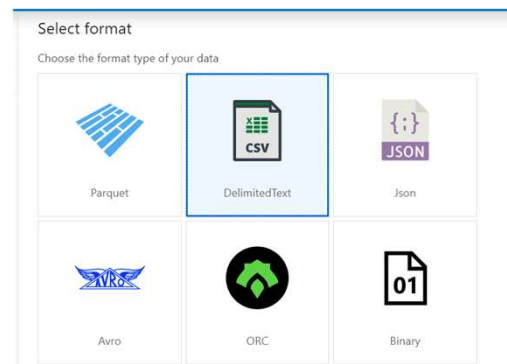
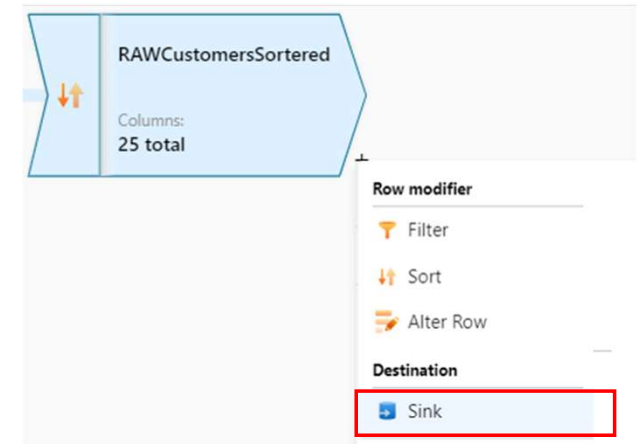
RAWCustomersSelected's column	Order	Nulls first	
abc LastName	Ascending	<input checked="" type="checkbox"/>	+
abc MiddleName	Ascending	<input checked="" type="checkbox"/>	+
abc LastName	Ascending	<input checked="" type="checkbox"/>	+



ADF Data Flows – Mapping Data Flow

In this task, you will save the Customers data to a Sink

1. In the main pane select the **+** next to the RAWCustomersSorted Activity
2. Select the **Sink** Option
3. Make sure the Sink activity is selected
4. Set the Output Stream Name to **RAWCustomersSink**
5. In the Sink Data Set option press the **New** button
6. In the New Dataset Window select the **Azure Tab**
7. Select the **Azure Data Lake Storage Gen 2** and Select **Continue**
8. In the select format Window select **delimited**
9. Press the **Continue** button



Azure Data Lake Storage
Gen2

ADP Copy Activity – Configure Sink

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

1. In the set properties window set the name to **iXXXXXX_DS_PROC_ADLS_Customers**
2. In the New linked Service (Azure Data Lake Storage Gen2) select the Name value to **iXXXXXX_LS_ADLS_DATA**
3. In the set properties press the **Browse** button
4. Select the **Storage/tXX/iXXXXXX/PROCESSED/dimCustomer/** Path
5. Let the **File field** empty
6. Select the **First row as header** option
7. In the Import Schema select **From Connection/store**
8. In the Settings tab check “Clear the Folder” check box
9. Press **OK**

Set properties

Name
GP00_DS_PROC_ADLS_Customers

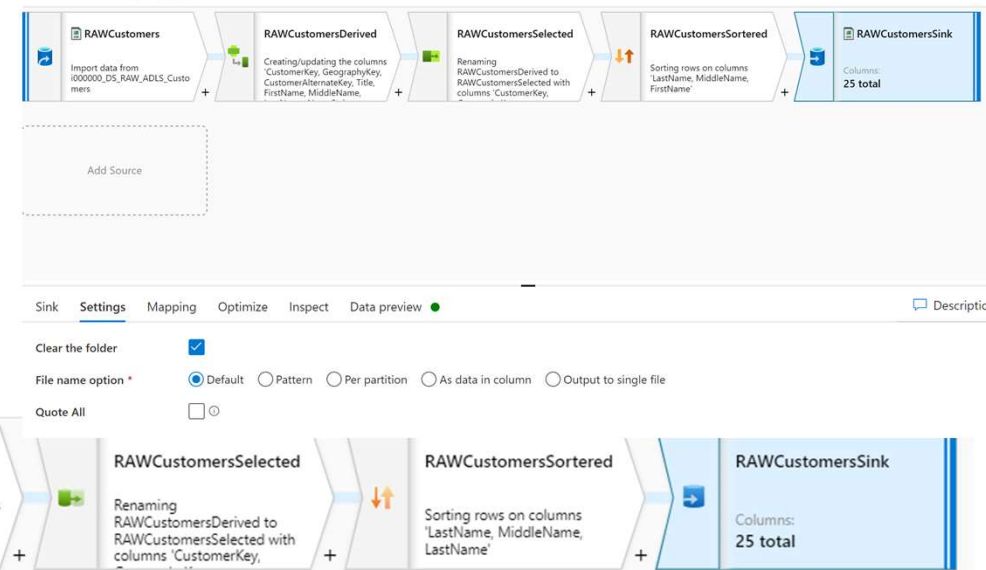
Linked service *
GP00_LS_ADLS_DATA

Edit connection

File path
storage / PROCESSED/GR00 / File Browse

First row as header ☒

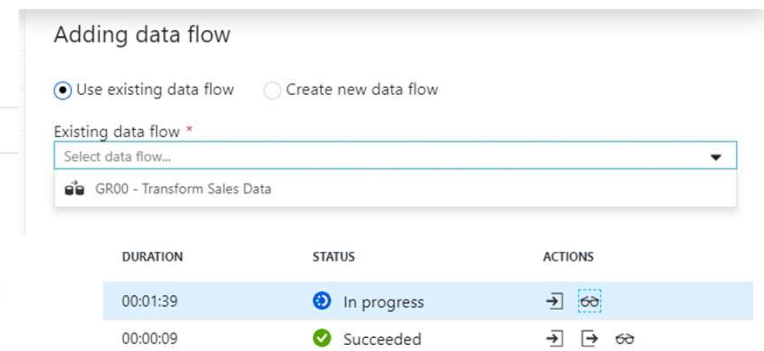
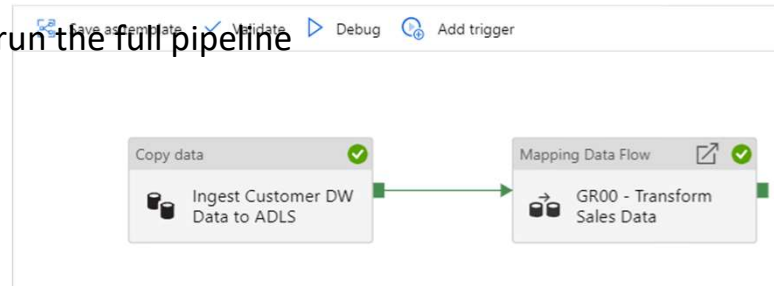
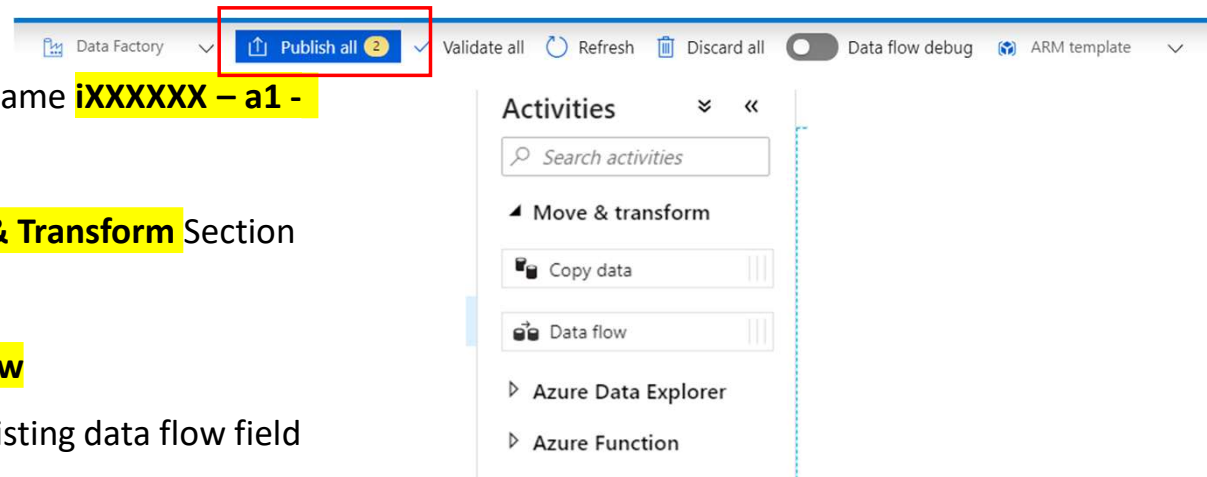
Import schema
☒ From connection/store ☐ From sample file ☐ None



ADF Copy Activity – Deploy Package

In this task, you will operationalize the package to run it

1. Press **Publish All** button
2. Open the Pipeline developed in the previous class name **iXXXXXX – a1 – ingest data**
3. In the Activities pane on the Right open the **Move & Transform** Section
4. Drag the **Data Flow** activity into the main pane
5. In the Adding data Flow select **Use existing data flow**
6. Select the **iXXXXXX_Transform Sales Data** in the existing data flow field
7. Press **OK**
8. Connect the **Copy Activity** to the **Mapping data flow Activity** for precedence
9. Run the **Debug** option to run the full pipeline
10. Use the **Debug** option to follow the pipeline execution



Projecto Final

O trabalho é dividido em 2 partes: A 1ª parte corresponde à elaboração do Dimensional Model e a 2ª ao ETL.

- ~~1ª Parte (40% do trabalho): O enunciado será disponibilizado em 29/06, com data de entrega até 31/07/2022;~~
- 2ª Parte (60% do trabalho): O enunciado será disponibilizado em 19/09, com data de entrega até 19/10.



Executive
Education

www.isegexecutive.education

Rua do Quelhas, 6
1200-781 Lisboa

(+351) 213 922 891
info@executive.education