

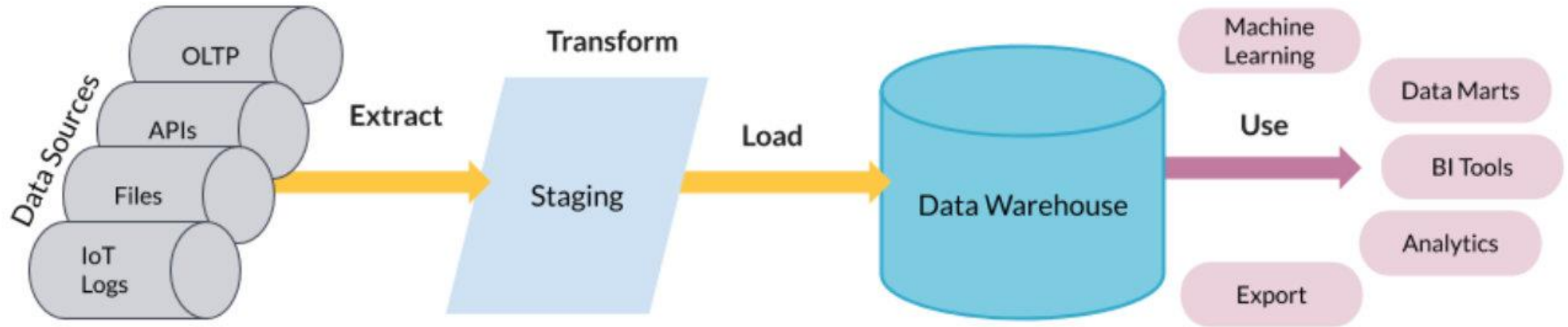
Azure Data Engineering

ADF + Databricks

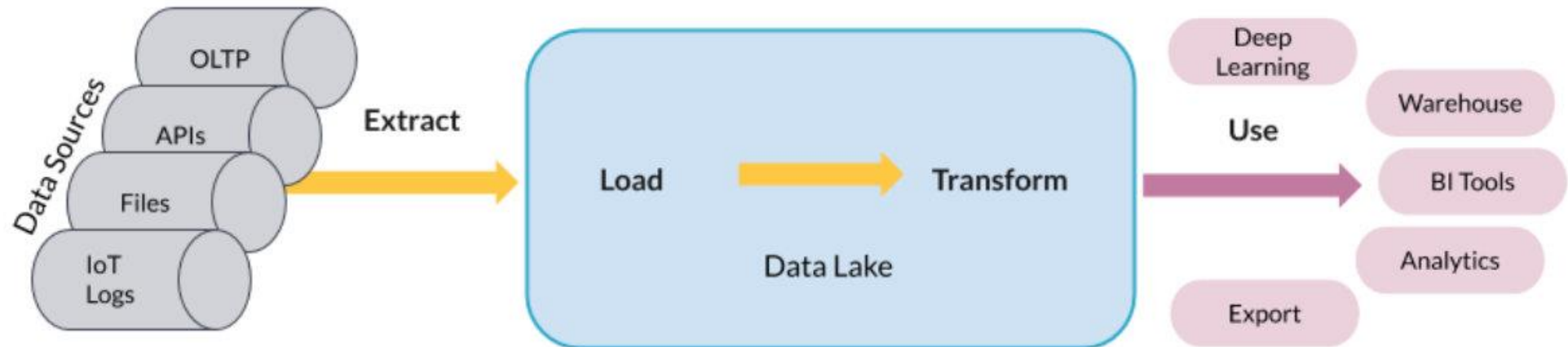
Architecture

ETL Vs ELT

ETL



ELT

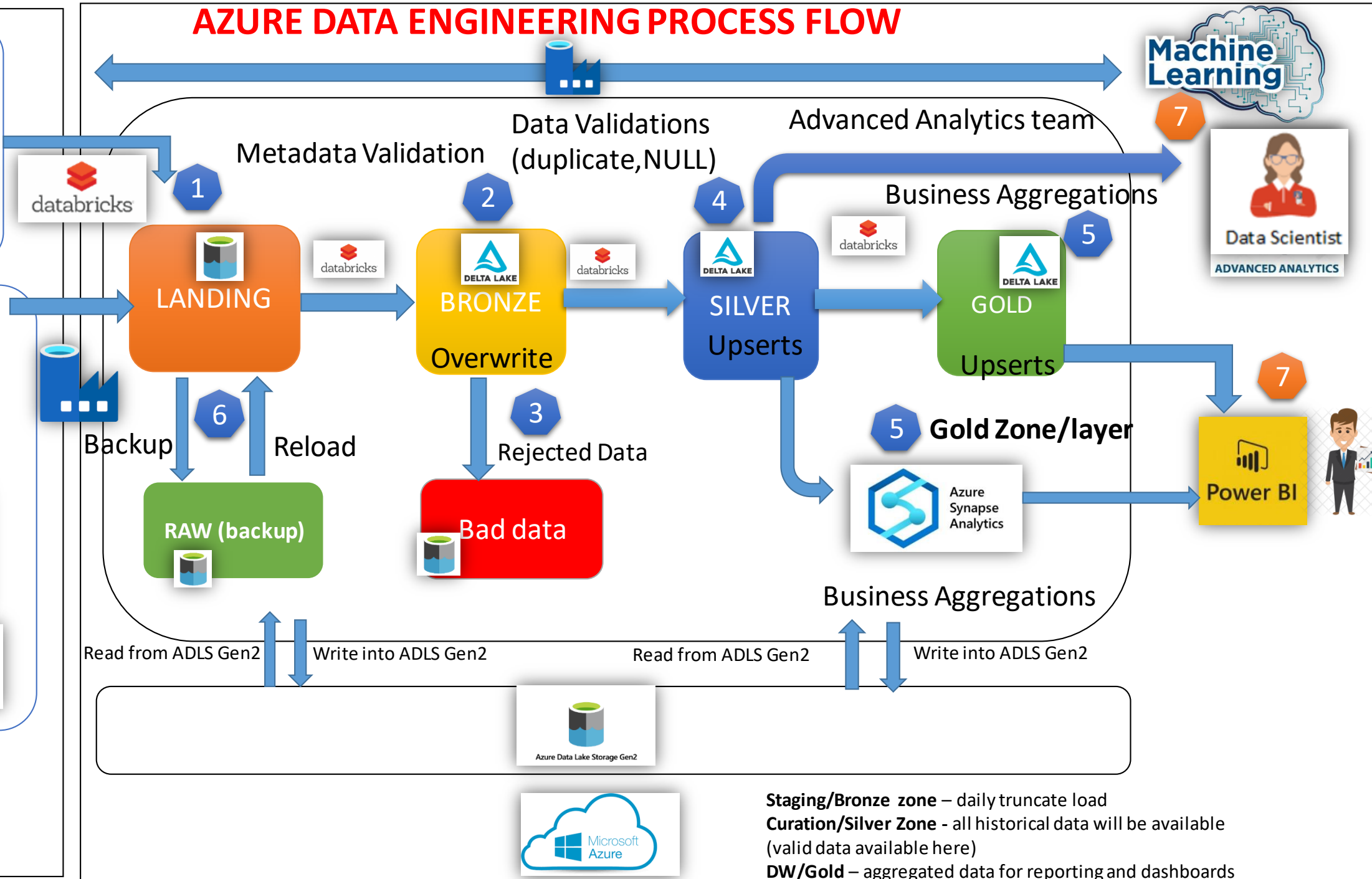


AZURE DATA ENGINEERING PROCESS FLOW

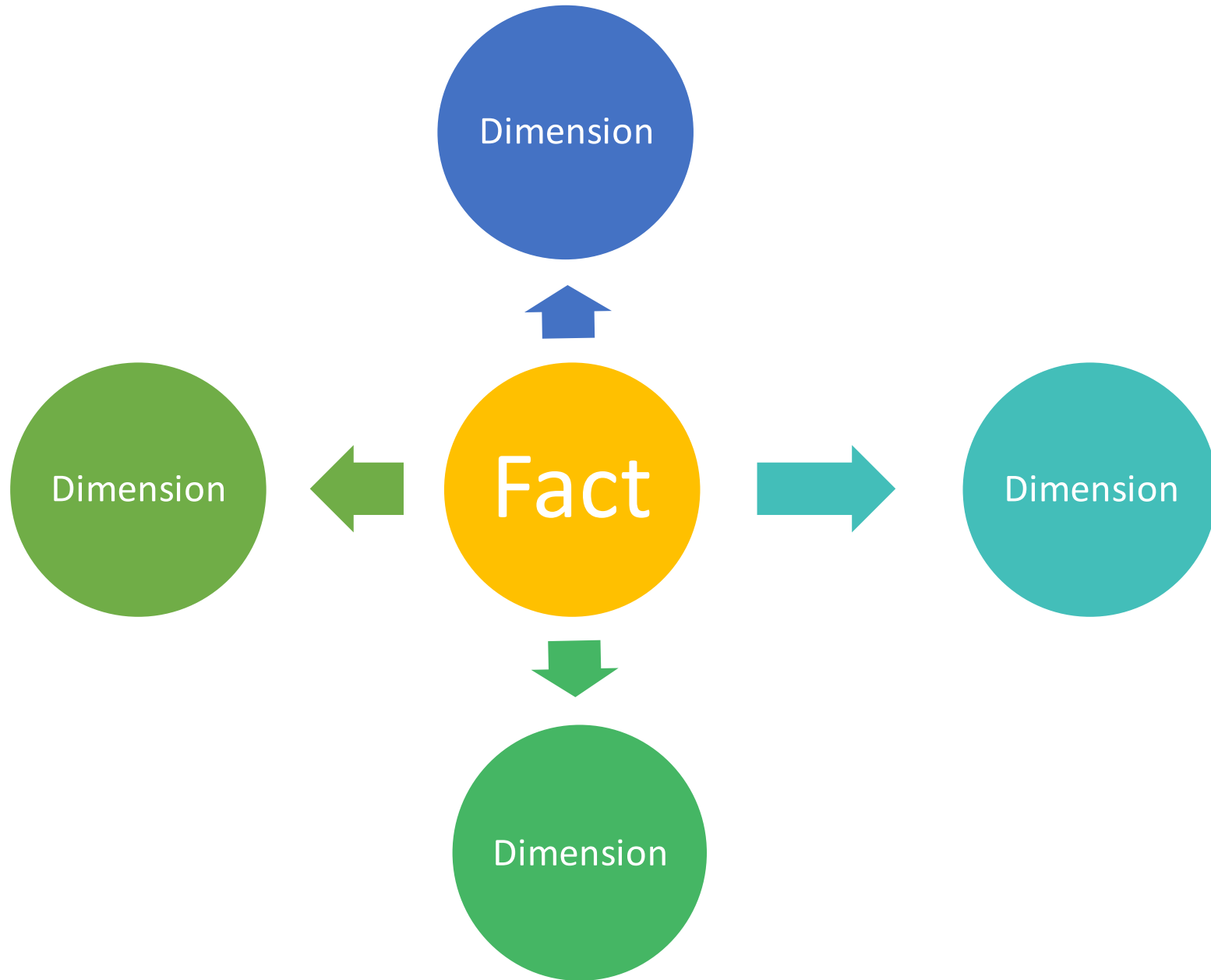
Streaming



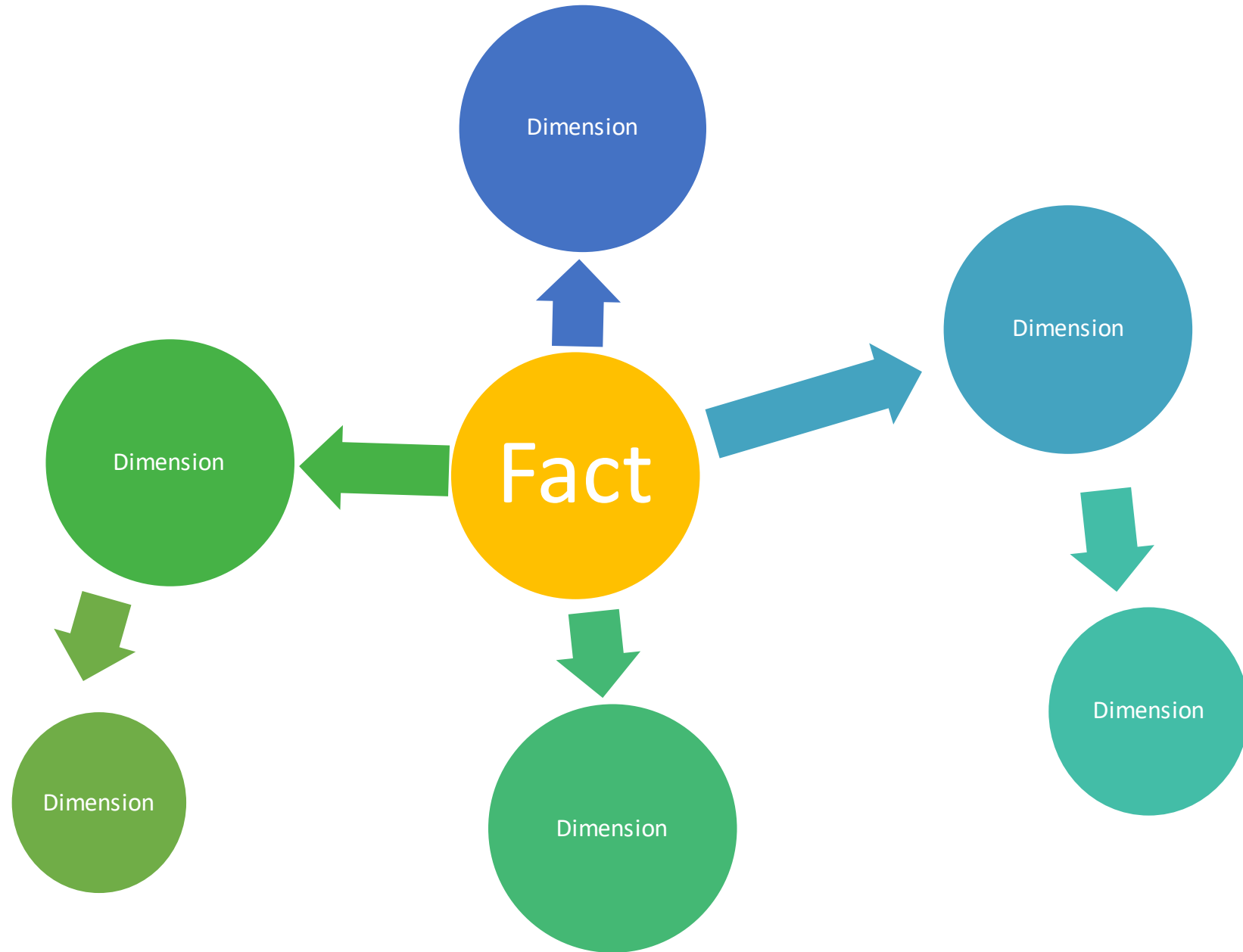
Batch Load



Star Scheme Design



Snowflake Scheme Design



Master Pipeline

pl_master_sales_load x

Activities

Search activities

- Move & transform
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Save as template ✓ Validate ▶ Debug ⚡ Add trigger

foreach for processing one by one job from control table and calling child pipeline

Lookup for job table data

Lookup



Lookup_Getjobdetails

ForEach



ForEach_Eachjob

Activities

1 activities

Execute Pipeline



pl_alerts

Sending Email Alerts after processing

Parameters Variables Settings Output

Pipeline run ID: [icon] [refresh] [info]

Name

Type

Run start

Duration

Status

Integr

Lookup Activity

pl_master_sales_load

Activities

Search activities

- Move & transform
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Save as template ✓ Validate ▶ Debug ⚡ Add trigger

Lookup

Lookup_Getjobdetails

ForEach

ForEach_Eachjob

Activities
1 activities

Execute Pipeline

pl_alerts

General **Settings** User properties

Source dataset * ds_deltalake_lookup

Use query ☐ Table ☒ Query

Query *
select * from jobs.JOB_LIST_DYNAMIC
order by job_id

First row only ☐

ForEach Activity

pl_master_sales_load

Activities

Search activities

- Move & transform
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Save as template ✓ Validate ▶ Debug ⚡ Add trigger

Lookup

Lookup_Getjobdetails

ForEach

ForEach_Eachjob

Activities

1 activities

Execute Pipeline

pl_alerts

General **Settings** Activities (1) User properties

Sequential ☐ refer lookup activity output in foreach activity input

Batch count ⓘ 5

Items @activity('Lookup_Getjobdetails').outp...

@activity('Lookup_Getjobdetails').output.value

Calling Child pipeline in Master Pipeline ForEach Activity to process one by one

The screenshot displays the Azure Data Factory (ADF) interface for a pipeline named `pl_master_sales_load`. The `Activities` pane on the left lists various activity types, including `Move & transform`, `Azure Data Explorer`, `Azure Function`, `Batch Service`, `Databricks`, `Data Lake Analytics`, `General`, `HDInsight`, `Iteration & conditionals`, `Machine Learning`, and `Power Query`.

The main canvas shows the pipeline design. The `pl_master_sales_load` pipeline contains a `ForEach_Eachjob` activity. Inside this loop, an `Execute Pipeline` activity is configured to call the child pipeline `pl_child_sales_load`. A red box highlights the `Execute Pipeline` activity, with a note: "calling child pipeline in master pipeline inside foreach loop".

The `Settings` tab for the `Execute Pipeline` activity is selected. It shows the `Invoked pipeline *` set to `pl_child_sales_load`. The `Wait on completion` checkbox is checked. The `Parameters` section is expanded, showing a table of parameters to be passed to the child pipeline. A red box highlights this table, with a note: "specify all parameter values in pipeline and get values from foreach @item().parameter".

Name	Type	Value
job_id	string	@item().job_id
job_name	string	@item().job_name
source_folder_name	string	@item().source_folder_name
landing_zone_file_path	string	@item().landing_zone_file_path

Chile Pipeline with Parameters

pl_master_sales_load

pl_child_sales_load

Save as template

Validate

Debug

Add trigger

Lookup

StartLogging

Get Metadata

GetMetadata_Onprem

Filter

Filter_Files

ForEach

ForEach_CopyFile

Activities
1 activities

Notebook

LandingToStaging

Wait

Wait1

Lookup

LandingToStaging_Log

Notebook

StagingToCuration

Wait

Wait

Lookup

StagingToCuration_Log

If Condition

IFDim

True
2 activities

False
No activities

Switch

SwitchFacts

Default
No activities

sales_transaction
1 activities

costs_transaction
1 activities

Parameters

Variables

Settings

Output

+ New

Delete

pl_child_sales_load

Save as template Validate Debug Add trigger

Parameters

Variables

Settings

Output

<input type="checkbox"/>	Name	Type	Default value
<input type="checkbox"/>	job_id	String	Value
<input type="checkbox"/>	job_name	String	Value
<input type="checkbox"/>	source_folder_name	String	Value
<input type="checkbox"/>	landing_zone_file_path	String	Value
<input type="checkbox"/>	landing_zone_folder_name	String	Value
<input type="checkbox"/>	landing_zone_file_name	String	Value
<input type="checkbox"/>	staging_zone_database_name	String	Value
<input type="checkbox"/>	staging_zone_table_name	String	Value
<input type="checkbox"/>	staging_zone_table_pk_column	String	Value
<input type="checkbox"/>	curation_zone_database_name	String	Value
<input type="checkbox"/>	curation_zone_table_name	String	Value
<input type="checkbox"/>	curation_zone_table_pk_column	String	Value
<input type="checkbox"/>	dw_zone_database_name	String	Value
<input type="checkbox"/>	dw_zone_table_name	String	Value
<input type="checkbox"/>	dw_zone_table_pk_column	String	Value
<input type="checkbox"/>	pyspark_schema	String	Value
<input type="checkbox"/>	table_type	String	Value
<input type="checkbox"/>	raw_zone_file_path	String	Value
<input type="checkbox"/>	raw_zone_folder_name	String	Value
<input type="checkbox"/>	raw_zone_file_name	String	Value

pl_child_sales_load

Save as template Validate Debug Add trigger

Lookup

StartLogging

Get Metadata

GetMetadata_Onprem

Filter

Filter_Files

General Settings User properties

Source dataset * ds_deltalake_lookup Open + New Preview data

Use query ☐ Table ☒ Query

Query insert into JOBS.pipeline_log select sh...

First row only ☒

```
insert into JOBS.pipeline_log select  
sha2('@{pipeline().RunId}',256),@{pipeline().parameters.job_id}, '@{pipeline().DataFactory}','@{pipeline().Pipeline}','@{pipeline().RunId}','@{pipeline().TriggerType}','@{pipeline().TriggerId}','  
@{pipeline().TriggerName}','@{pipeline().TriggerTime}',,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

Child pipeline – GetMetadata Activity to get files from Onprem

pl_child_sales_load

Save as template Validate Debug Add trigger

Lookup StartLogging

Get Metadata GetMetadata_Onprem

Filter Filter_Files

General Dataset User properties

Dataset * ds_onprem_file_getmetadata Open + New Learn more

Dataset properties

Name	Value
ds_folder_name	@pipeline().parameters.source_folder_...

Filter by last modified

Start time (UTC) End time (UTC)

Field list + New | Delete

- ☐ Argument
- ☐ Child items
- ☐ Item name

Dataset for GetMetadata from onprem

pl_master_sales_load pl_child_sales_load **ds_onprem_file_get...**

01 Binary
ds_onprem_file_getmetadata

Connection **Parameters**

+ New | Delete

<input type="checkbox"/> Name	Type	Default value
<input type="checkbox"/> ds_folder_name	String	Value

pl_master_sales_load pl_child_sales_load **ds_onprem_file_get...**

01 Binary
ds_onprem_file_getmetadata

Connection Parameters

Linked service * ls_file_onprem_source Test connection Edit + New

Integration runtime * OIRSelfhosted (Unavailable) Edit

File path * E:\files\SH / @dataset().ds_folder_name / File

Compression type None

Filter-Activity For filter only files.

The screenshot displays the Azure Data Factory (ADF) interface for a pipeline named `pl_child_sales_load`. The workflow consists of three activities: `Lookup` (containing `StartLogging`), `Get Metadata` (containing `GetMetadata_Onprem`), and `Filter` (containing `Filter_Files`). The `Filter` activity is highlighted with a red box. Below the workflow, the `Settings` tab for the `Filter` activity is open, also highlighted with a red box. It shows the `Items` property set to `@activity('GetMetadata_Onprem').out...` and the `Condition` property set to `@equals(item().type,'File')`. Both input fields are highlighted with red boxes. Explanatory text in blue states: "previous getmetadata activity output as input to filter activity" and "filter only files based on previous activity out type."

pl_child_sales_load

Save as template Validate Debug Add trigger

Lookup
StartLogging

Get Metadata
GetMetadata_Onprem

Filter
Filter_Files

General **Settings** User properties

Items
`@activity('GetMetadata_Onprem').out...` previous getmetadata activity output as input to filter activity

Condition
`@equals(item().type,'File')` filter only files based on previous activity out type.

Foreach Activity

Calling COPY Activity to copy files from on-prem to azure

The screenshot displays the Azure Data Factory designer interface for a pipeline named `pl_child_sales_load`. The workflow consists of three activities: `Filter_Files` (Filter), `Foreach_CopyFile` (Foreach), and `LandingToStaging` (Notebook). The `Foreach` activity is highlighted with a red border and contains one activity, `Foreach_CopyFile`. The settings pane at the bottom shows the `Settings` tab selected, with the `Items` field set to `@activity('Filter_Files').output.value`. The `Batch count` field is empty, and the `Sequential` checkbox is unchecked.

foreach activity to copy files from onprem to azure

General **Settings** Activities (1) User properties

Sequential ☐

Batch count ⓘ

Items

previous filter activity output to Foreach activity input

Copy- Activity

pl_child_sales_load ×

Save as template ✓ Validate ✓ Validate copy runtime ▶ Debug ⚡ Add trigger

pl_child_sales_load > ForEach_CopyFile

Copy activity for processing files from onprem to azure.
Source: On-prem file
Target: azure Data lake adls gen2

Copy data

OnPremToLanding

General **Source** Sink Mapping Settings User properties

Source dataset * ds_onprem_file_binary Open + New Learn more

Dataset properties

Name	Value	Type
ds_folder_name	@pipeline().parameters.source_folder_...	string
ds_file_name	@item().name	string

File path type ☒ File path in dataset ☐ File filter ☐ Wildcard file path ☐ List of files

Start time (UTC) End time (UTC)

Filter by last modified

Recursively ☒

Copy-Activity

Source Dataset - Properties

pl_master_sales_load pl_child_sales_load **ds_onprem_file_binary**

01 Binary
ds_onprem_file_binary

Connection **Parameters**

+ New | Delete

<input type="checkbox"/> Name	Type	Default value
<input type="checkbox"/> ds_folder_name	String	Value
<input type="checkbox"/> ds_file_name	String	Value

pl_master_sales_load pl_child_sales_load **ds_onprem_file_binary**

01 Binary
ds_onprem_file_binary

Connection Parameters

Linked service * ls_file_onprem_source Test connection Edit + New Learn more

Integration runtime * OIRSelfhosted (Unavailable) Edit

File path * E:\files\SH / @dataset().ds_folder_name / @dataset().ds_file_name

Compression type None

Copy-Activity

Sink- Dataset properties

The screenshot shows the configuration of a Copy Activity in Azure Data Factory. The activity is named "pl child sales load". The Sink dataset is configured with the following properties:

Name	Value
ds_folder_name	@pipeline().parameters.landing_zone_...
ds_file_name	@item().name

The "Copy behavior" is set to "None".

Activity Name: pl child sales load

ForEach CopyFile: ForEach CopyFile

Copy data: OnPremToLanding

Sink dataset: ds_adlsgen2_landing_binary

Dataset properties:

- ds_folder_name: @pipeline().parameters.landing_zone_...
- ds_file_name: @item().name

Copy behavior: None

Sink DataSet Parameters and dynamic filepath.

pl_master_sales_load pl_child_sales_load ds_adlsgen2_landing... X

01 Binary
ds_adlsgen2_landing_binary

Connection **Parameters**

+ New | Delete

<input type="checkbox"/> Name	Type	Default value
<input type="checkbox"/> ds_folder_name	String	Value
<input type="checkbox"/> ds_file_name	String	Value

pl_master_sales_load pl_child_sales_load ds_adlsgen2_landing... X

01 Binary
ds_adlsgen2_landing_binary

Connection Parameters

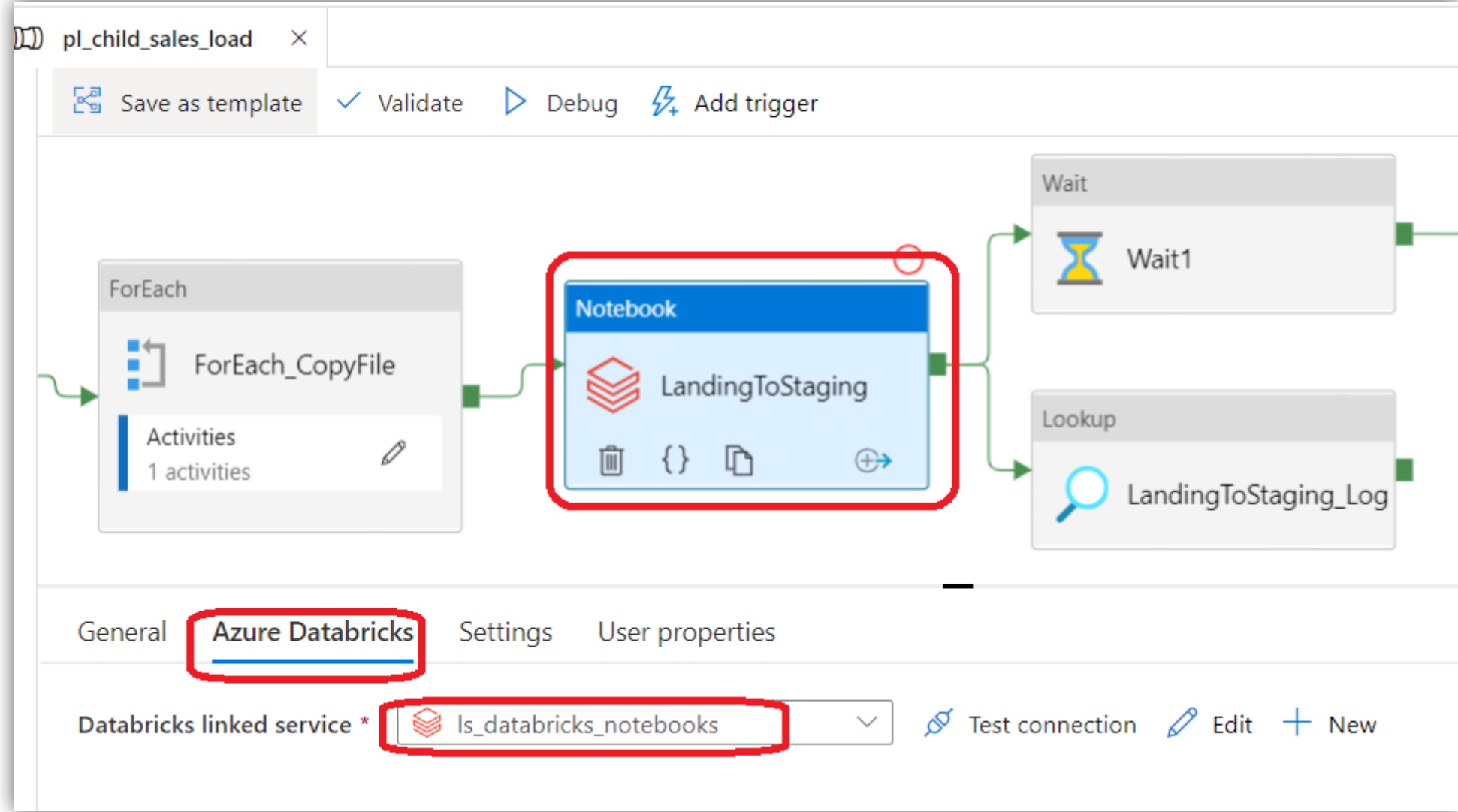
Linked service * ls_adlsgen2_sales Test connection Edit + New Learn more

File path * sales / @concat('landing/sales/',dataset().ds_f... / @dataset().ds_file_name

Compression type None

@concat('landing/sales/',dataset().ds_folder_name)

Calling Databricks Notebook activity for Landing to Staging.



Notebook : Landing to staging widgets parameters

pl_child_sales_load

Save as template ✓ Validate ▶ Debug ⚡ Add trigger

ForEach

ForEach_CopyFile

Activities
1 activities

Notebook

LandingToStaging

Wait

Wait1

Lookup

LandingToStaging_Log

General Azure Databricks **Settings** User properties

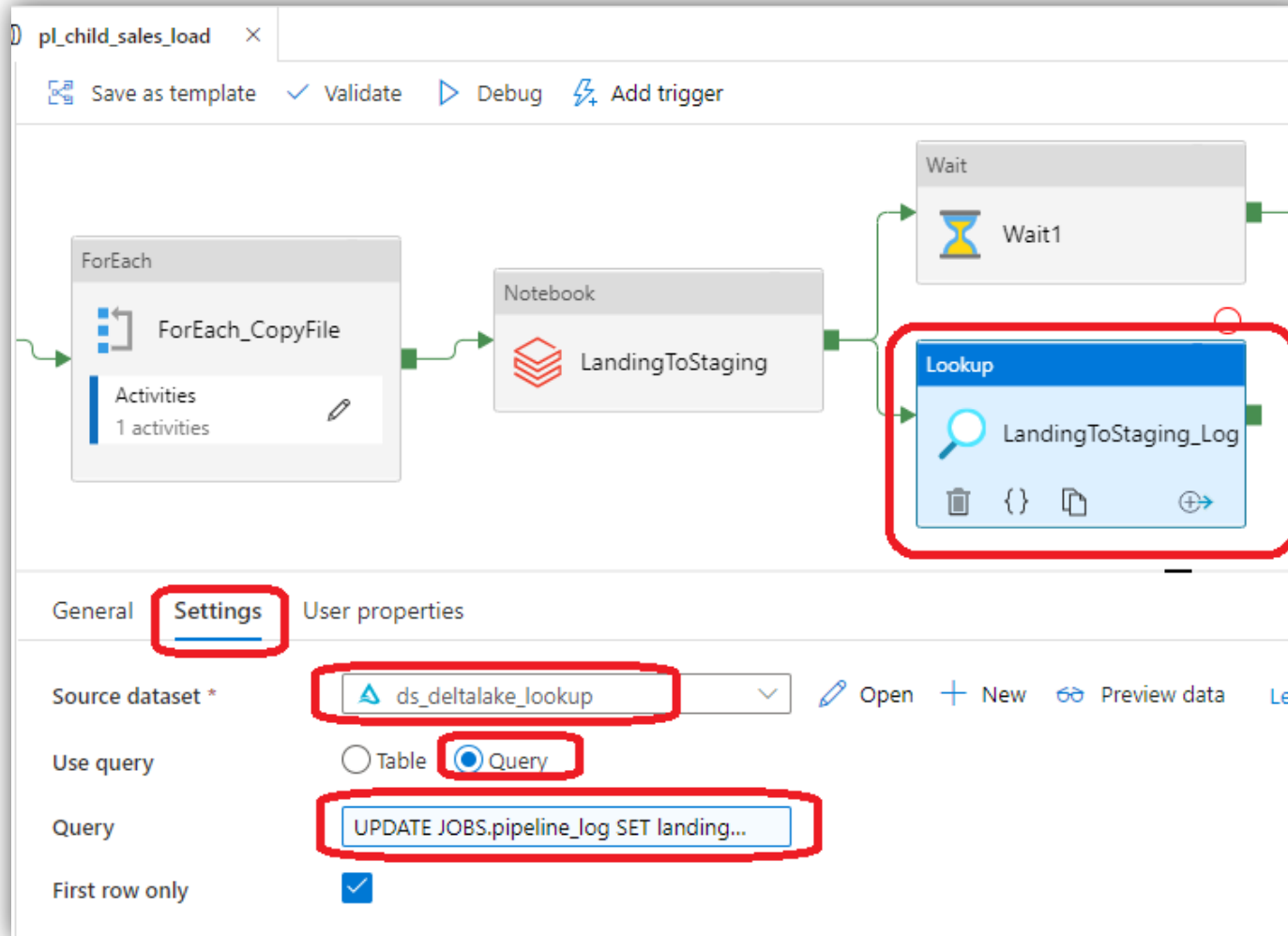
Notebook path * Browse Open

Base parameters

+ New | Delete

<input type="checkbox"/>	Name	Value
<input type="checkbox"/>	landing_zone_file_pat	@pipeline().parameters.landing_zone_...
<input type="checkbox"/>	landing_zone_folder_r	@pipeline().parameters.landing_zone_...
<input type="checkbox"/>	landing_zone_file_nam	@pipeline().parameters.landing_zone_...
<input type="checkbox"/>	staging_zone_databas	@pipeline().parameters.staging_zone_...
<input type="checkbox"/>	staging_zone_table_n	@pipeline().parameters.staging_zone_t...
<input type="checkbox"/>	staging_zone_table_pl	@pipeline().parameters.staging_zone_t...
<input type="checkbox"/>	pyspark_schema	@pipeline().parameters.pyspark_schema
<input type="checkbox"/>	job_name	@pipeline().parameters.job_name
<input type="checkbox"/>	job_id	@pipeline().parameters.job_id

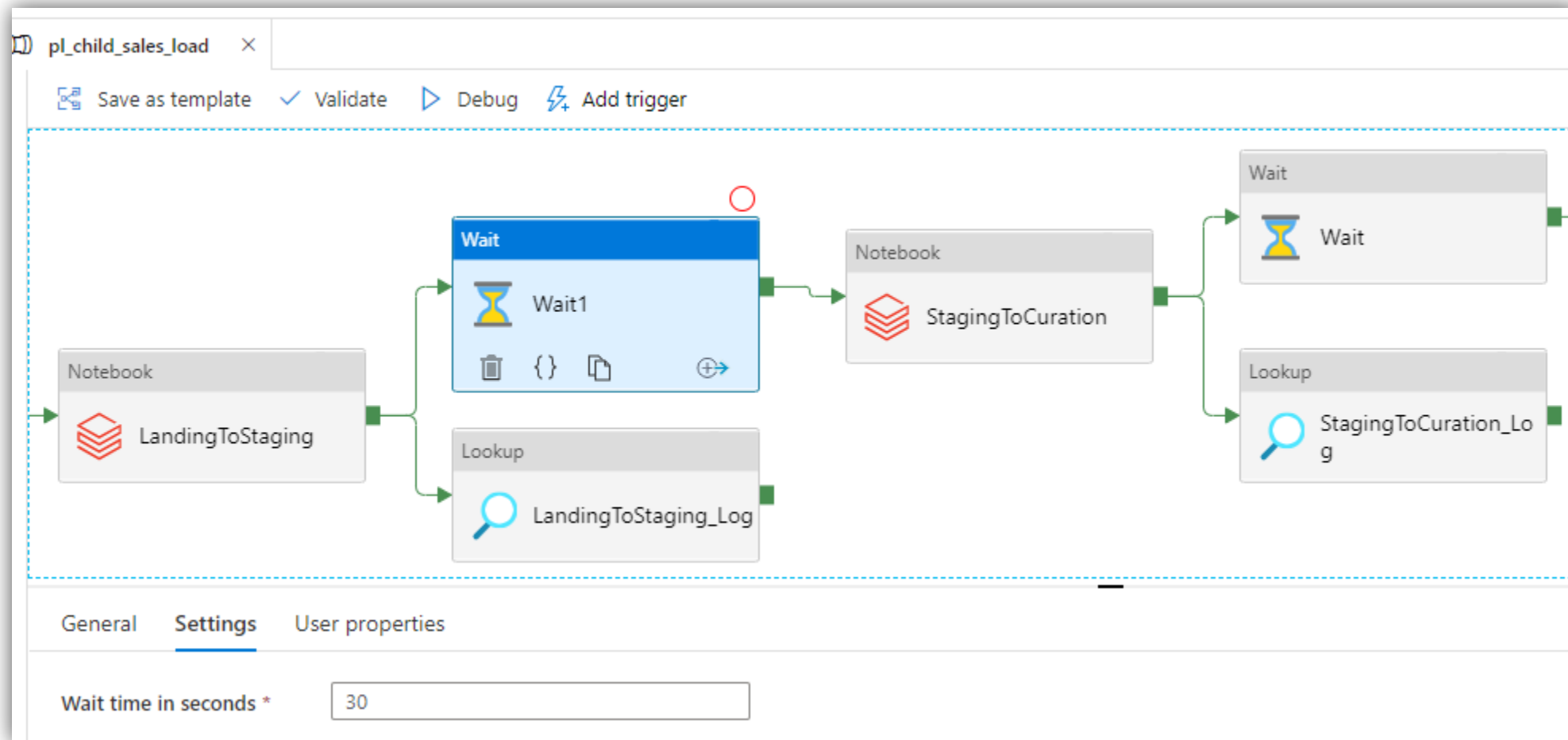
Lookup- Activity for logging into delta lake table



```
UPDATE JOBS.pipeline_log SET  
landing_rows=@{activity('LandingToStaging').output.runOutput.landingRows},rejected_rows=@{activity('LandingToStaging').output.runOutput.rejectedRows},  
staging_rows=@{activity('LandingToStaging').output.runOutput.stagingRows} WHERE pipeline_runid='@{pipeline().RunId}'
```

Wait- Activity

Wait activity to wait job execution few seconds (30 sec) before staging next notebook



Notebook : Activity for Staging to Curation

pl_child_sales_load x

Save as template ✓ Validate ▶ Debug ⚡ Add trigger

Wait
Wait1

Lookup
LandingToStaging_Log

Notebook
StagingToCuration

Wait
Wait

Lookup
StagingToCuration_Log

General Azure Databricks **Settings** User properties

Notebook path * /Shared/Pyspark_Project_Dynamic/NB_sale: Browse Open

▲ Base parameters

+ New | Delete

<input type="checkbox"/>	Name	Value
<input type="checkbox"/>	staging_zone_databas	@pipeline().parameters.staging_zone_...
<input type="checkbox"/>	staging_zone_table_n:	@pipeline().parameters.staging_zone_t...
<input type="checkbox"/>	staging_zone_table_pl	@pipeline().parameters.staging_zone_t...
<input type="checkbox"/>	job_name	@pipeline().parameters.job_name
<input type="checkbox"/>	job_id	@pipeline().parameters.job_id
<input type="checkbox"/>	curation_zone_databa	@pipeline().parameters.curation_zone...
<input type="checkbox"/>	curation_zone_table_n	@pipeline().parameters.curation_zone...
<input type="checkbox"/>	curation_zone_table_p	@pipeline().parameters.curation_zone...

Lookup- Activity for updating log

The screenshot displays the Azure Data Factory pipeline editor for a pipeline named 'pl_child_sales_load'. The pipeline flow includes a 'Wait' activity (Wait1), a 'Lookup' activity (LandingToStaging_Log), a 'Notebook' activity (StagingToCuration), and another 'Wait' activity. A 'Lookup' activity (StagingToCuration_Log) is highlighted with a red box. Below the pipeline canvas, the 'Settings' tab is selected, and the configuration for the 'StagingToCuration_Log' activity is shown. The 'Source dataset' is set to 'ds_deltalake_lookup'. The 'Use query' option is selected, and the query text is 'UPDATE JOBS.pipeline_log SET curatio...'. The 'First row only' checkbox is checked.

pl_child_sales_load

Save as template Validate Debug Add trigger

Wait

Wait1

Lookup

LandingToStaging_Log

Notebook

StagingToCuration

Wait

Wait

Lookup

StagingToCuration_Log

General Settings User properties

Source dataset * ds_deltalake_lookup

Use query Table Query

Query UPDATE JOBS.pipeline_log SET curatio...

First row only

```
UPDATE JOBS.pipeline_log SET
curation_read_rows=@{activity('StagingToCuration').output.runOutput.curationRows['num_affected_rows']},curation_inserted_rows=@{activity('StagingToCuration').output.runOutput.curationRows['num_inserted_rows']},curation_updated_rows=@{activity('StagingToCuration').output.runOutput.curationRows['num_updated_rows']} WHERE
pipeline_runid=@{pipeline().RunId}'
```

IF-Condition : for processing Dimensions based on condition (table_type='DIM')

The screenshot displays the SAP Data Builder interface for a pipeline named 'pl_child_sales_load'. The pipeline flow includes a 'Wait' activity, a 'Lookup' activity named 'StagingToCuration_Loading', and an 'If Condition' activity. The 'If Condition' activity is highlighted with a red box and contains a table with two cases: 'True' (2 activities) and 'False' (No activities). The 'True' case is further detailed in the 'Activities (2)' tab at the bottom, showing a table with columns 'Case', 'Activity', and the condition '@equals(pipeline().parameters.table_type, 'dim')'. The 'False' case is listed as 'No activities'. To the right of the 'If Condition' activity is a 'Switch' activity named 'SwitchFacts' with three cases: 'Default' (No activities), 'sales_transaction' (1 activities), and 'costs_transaction' (1 activities).

pl_child_sales_load

Save as template Validate Debug Add trigger

Wait

Wait

Lookup

StagingToCuration_Loading

If Condition

IFDim

Case	Activity
True	2 activities
False	No activities

Switch

SwitchFacts

Case	Activity
Default	No activities
sales_transaction	1 activities
costs_transaction	1 activities

General Activities (2) User properties

Expression

@equals(pipeline().parameters.table_t...

Case	Activity	@equals(pipeline().parameters.table_type, 'dim')
True	2 Activities	
False	No activities	

Notebook: Curation To DWH processing inside if condition

pl_child_sales_load

Save as template Validate Debug Add trigger

pl_child_sales_load > IFDim > True activities

Notebook

CurationToDWHDim

Lookup

CurationToDWHDim_Log

General Azure Databricks Settings User properties

Notebook path */ /Shared/Pyspark_Project_Dynamic/NB_sale: Browse Open

Base parameters

+ New | Delete

<input type="checkbox"/>	Name	Value
<input type="checkbox"/>	job_name	@pipeline().parameters.job_name
<input type="checkbox"/>	job_id	@pipeline().parameters.job_id
<input type="checkbox"/>	curation_zone_databa	@pipeline().parameters.curation_zone...
<input type="checkbox"/>	curation_zone_table_n	@pipeline().parameters.curation_zone...
<input type="checkbox"/>	curation_zone_table_p	@pipeline().parameters.curation_zone...
<input type="checkbox"/>	dw_zone_database_na	@pipeline().parameters.dw_zone_data...
<input type="checkbox"/>	dw_zone_table_name	@pipeline().parameters.dw_zone_table...
<input type="checkbox"/>	dw_zone_table_pk_col	@pipeline().parameters.dw_zone_table...

Updating Log in log table

The screenshot shows the Databricks IDE interface for a pipeline named 'pl_child_sales_load'. The pipeline is in the 'True activities' state. A 'Notebook' activity named 'CurationToDWHDim' is connected to a 'Lookup' activity. The 'Lookup' activity is highlighted with a red box. Below the canvas, the 'Settings' tab is selected and highlighted with a red box. The 'Source dataset' is 'ds_deltalake_lookup'. The 'Use query' radio button is selected. The 'Query' text box contains the SQL statement: 'UPDATE JOBS.pipeline_log SET executi...'. The 'First row only' checkbox is checked.

```
UPDATE JOBS.pipeline_log SET
execution_status='SUCCESS',dwh_read_rows=@{activity('CurationToDWHDim').output.runOutput.dwhRows['num_affected_rows']],dwh_inserted_rows=@{activity('CurationToDWHDim').output.runOutput.dwhRows['num_inserted_rows']],dwh_updated_rows=@{activity('CurationToDWHDim').output.runOutput.dwhRows['num_updated_rows']] WHERE pipeline_runid='@{pipeline().RunId}'
```

Switch – Case : for processing Fact data loads based on condition.

pl_child_sales_load

Save as template Validate Debug Add trigger

General **Activities (2)** User properties

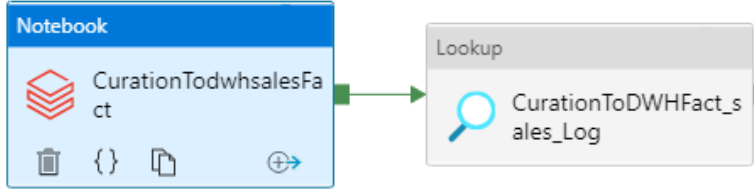
Expression: @pipeline().parameters.job_name

+ Add case

Case	Activity
Default	No activities
sales_transaction	1 Activity
costs_transaction	1 Activity

Processing both Facts inside Switch case

pl_child_sales_load > SwitchFacts - salestransaction



General Azure Databricks **Settings** User properties

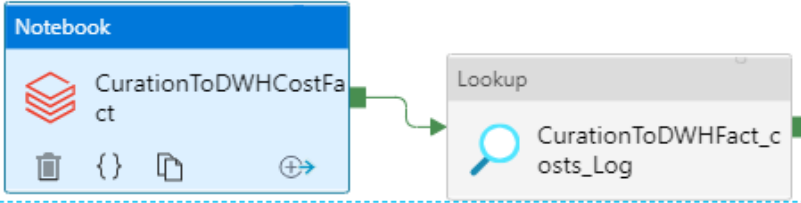
Notebook path * /Shared/Pyspark_Project_Dynamic/NB_sale: Browse Open

Base parameters

+ New | Delete

<input type="checkbox"/> Name	Value
<input type="checkbox"/> job_name	@pipeline().parameters.job_name
<input type="checkbox"/> job_id	@pipeline().parameters.job_id
<input type="checkbox"/> curation_zone_databa	@pipeline().parameters.curation_zone...
<input type="checkbox"/> curation_zone_table_n	@pipeline().parameters.curation_zone...
<input type="checkbox"/> curation_zone_table_p	@pipeline().parameters.curation_zone...
<input type="checkbox"/> dw_zone_database_na	@pipeline().parameters.dw_zone_data...
<input type="checkbox"/> dw_zone_table_name	@pipeline().parameters.dw_zone_table...
<input type="checkbox"/> dw_zone_table_pk_col	@pipeline().parameters.dw_zone_table...
<input type="checkbox"/> staging_zone_databas	@pipeline().parameters.staging_zone...
<input type="checkbox"/> staging_zone_table_n	@pipeline().parameters.staging_zone_t...
<input type="checkbox"/> staging_zone_table_pl	@pipeline().parameters.staging_zone_t...

pl_child_sales_load > SwitchFacts - coststransaction



General Azure Databricks **Settings** User properties

Notebook path * /Shared/Pyspark_Project_Dynamic/NB_cost: Browse Open

Base parameters

+ New | Delete

<input type="checkbox"/> Name	Value
<input type="checkbox"/> job_name	@pipeline().parameters.job_name
<input type="checkbox"/> job_id	@pipeline().parameters.job_id
<input type="checkbox"/> curation_zone_databa	@pipeline().parameters.curation_zone...
<input type="checkbox"/> curation_zone_table_n	@pipeline().parameters.curation_zone...
<input type="checkbox"/> curation_zone_table_p	@pipeline().parameters.curation_zone...
<input type="checkbox"/> dw_zone_database_na	@pipeline().parameters.dw_zone_data...
<input type="checkbox"/> dw_zone_table_name	@pipeline().parameters.dw_zone_table...
<input type="checkbox"/> dw_zone_table_pk_col	@pipeline().parameters.dw_zone_table...
<input type="checkbox"/> staging_zone_databas	@pipeline().parameters.staging_zone...
<input type="checkbox"/> staging_zone_table_n	@pipeline().parameters.staging_zone_t...
<input type="checkbox"/> staging_zone_table_pl	@pipeline().parameters.staging_zone_t...

All The Best 😊