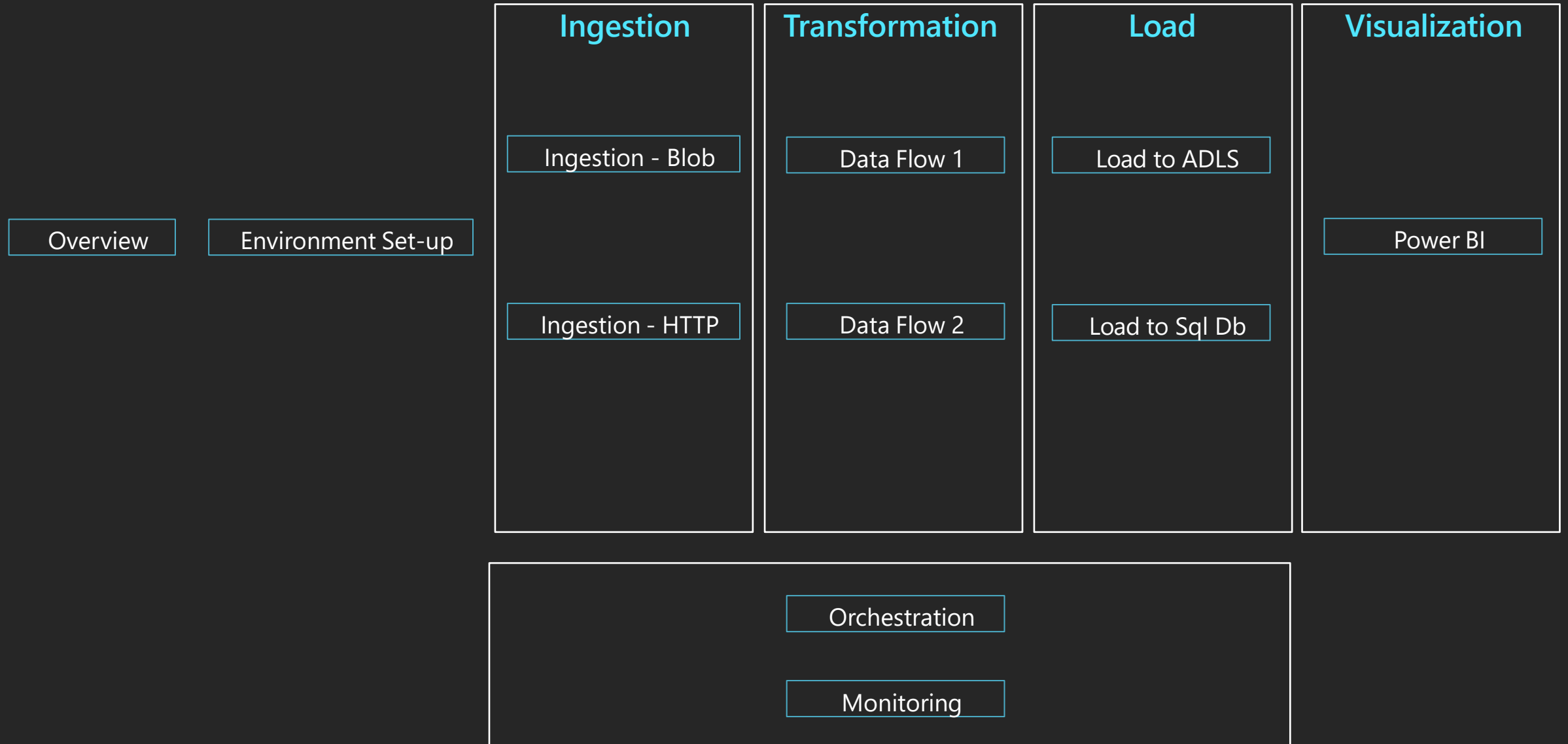


Azure Data Factory

Course Structure



Overview



Azure Data Factory (ADF)

Azure storage solutions



Azure SQL Database



Azure Blob Storage



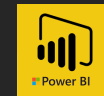
Azure Data Lake Storage Gen2

Other Bigdata Solutions

(continued as part of other modules)



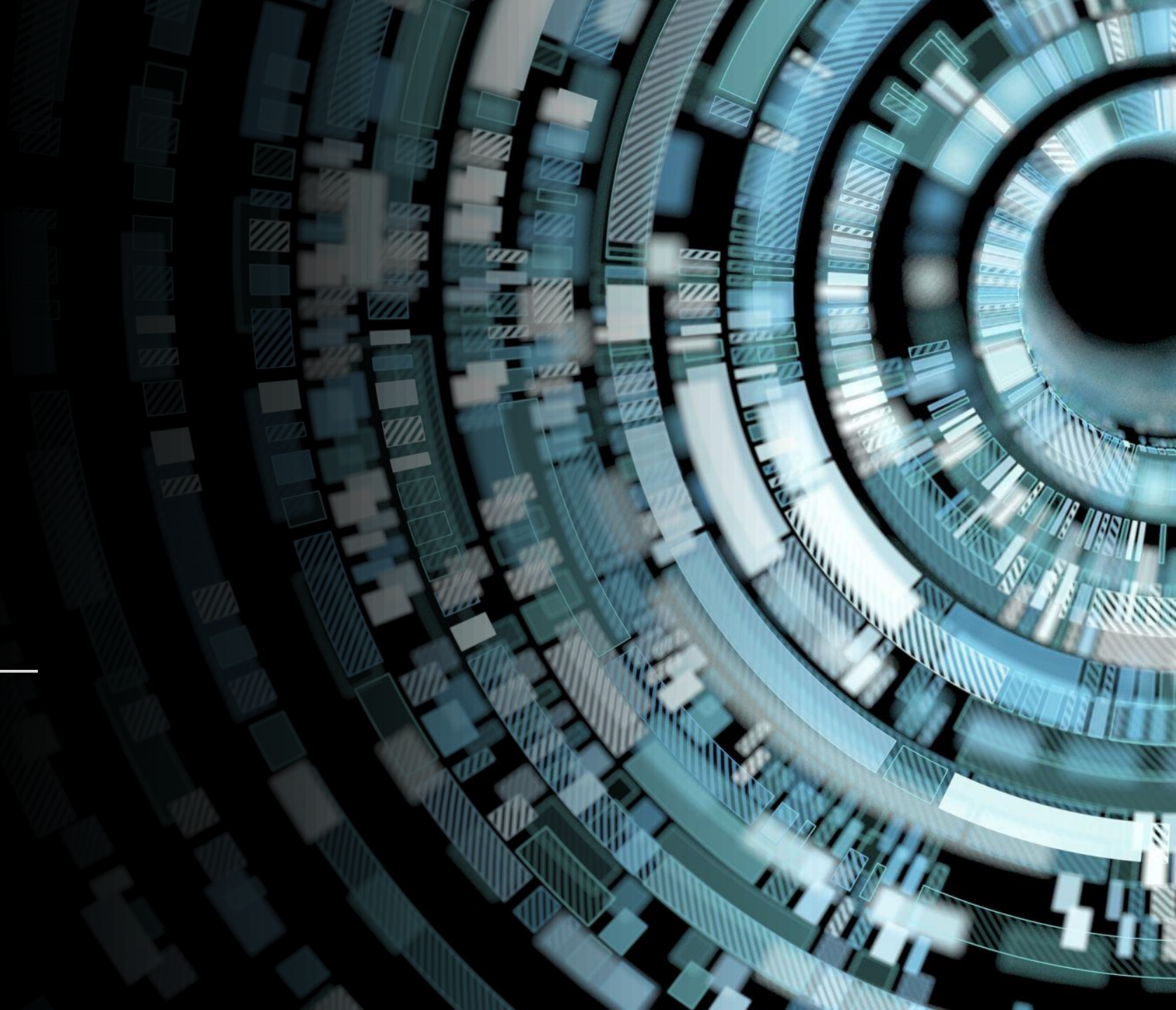
Azure Databricks



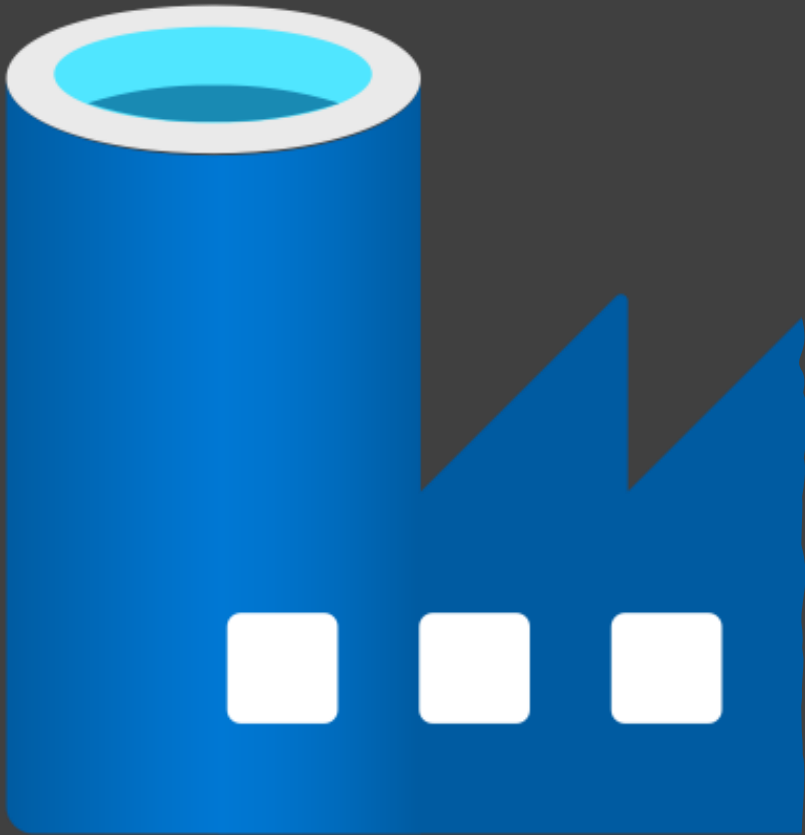
Power BI



Azure Data Factory Overview

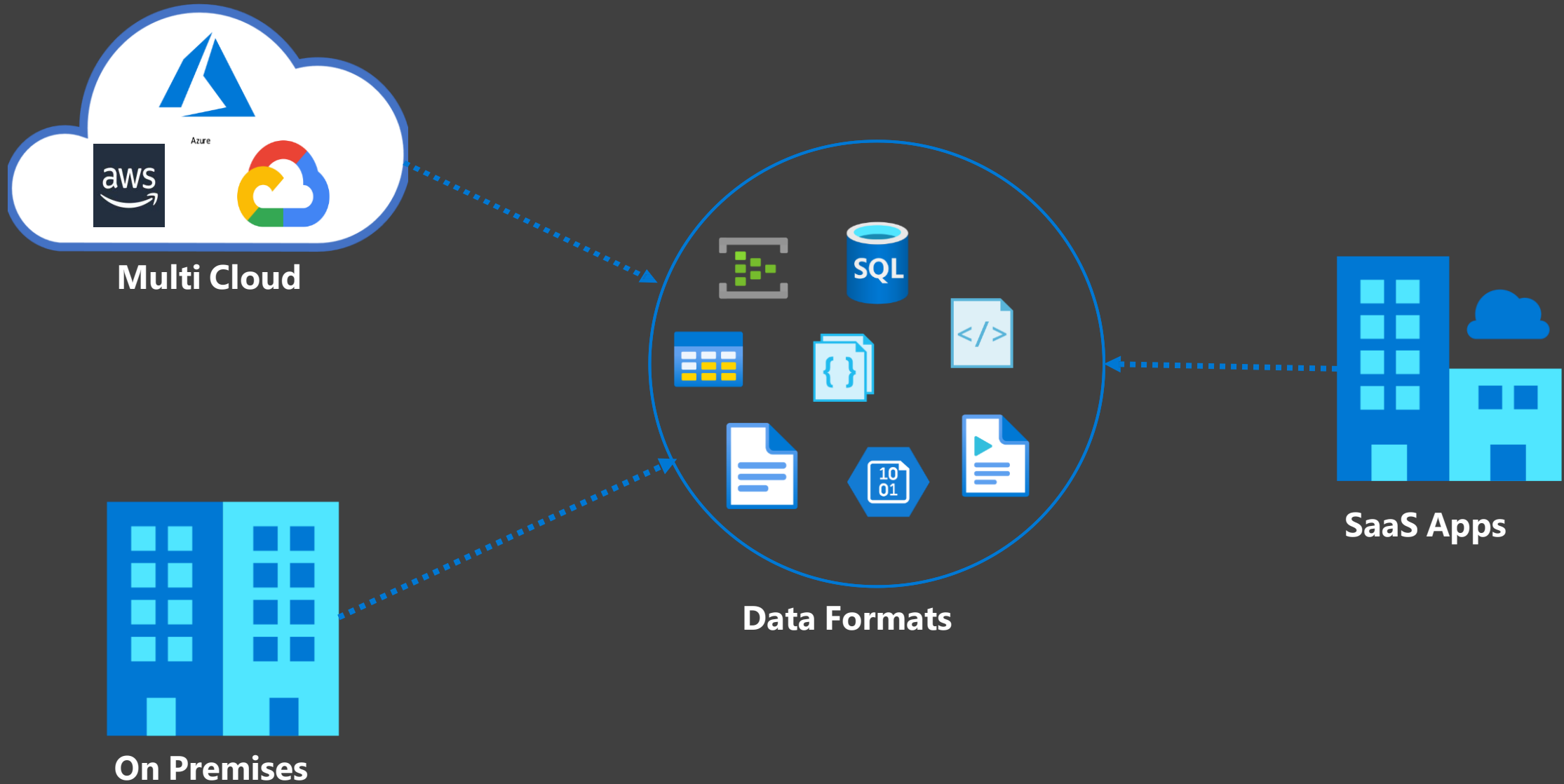


What is Azure Data Factory?

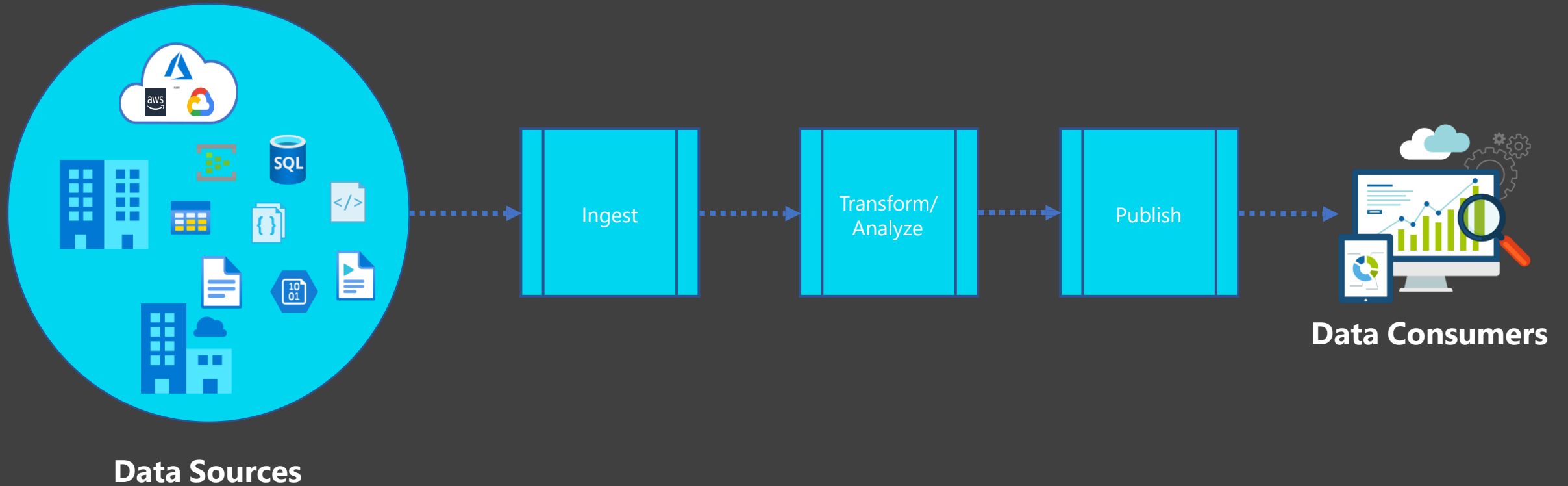


A fully managed, serverless data integration solution for ingesting, preparing and transforming all your data at scale.

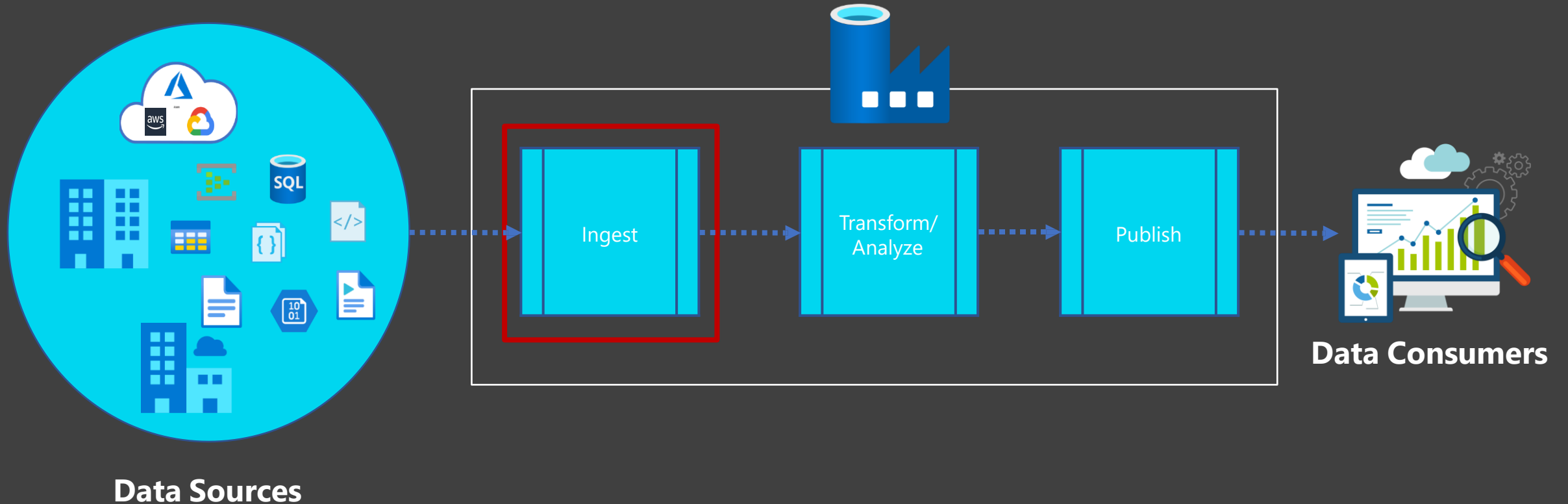
The Data Problem



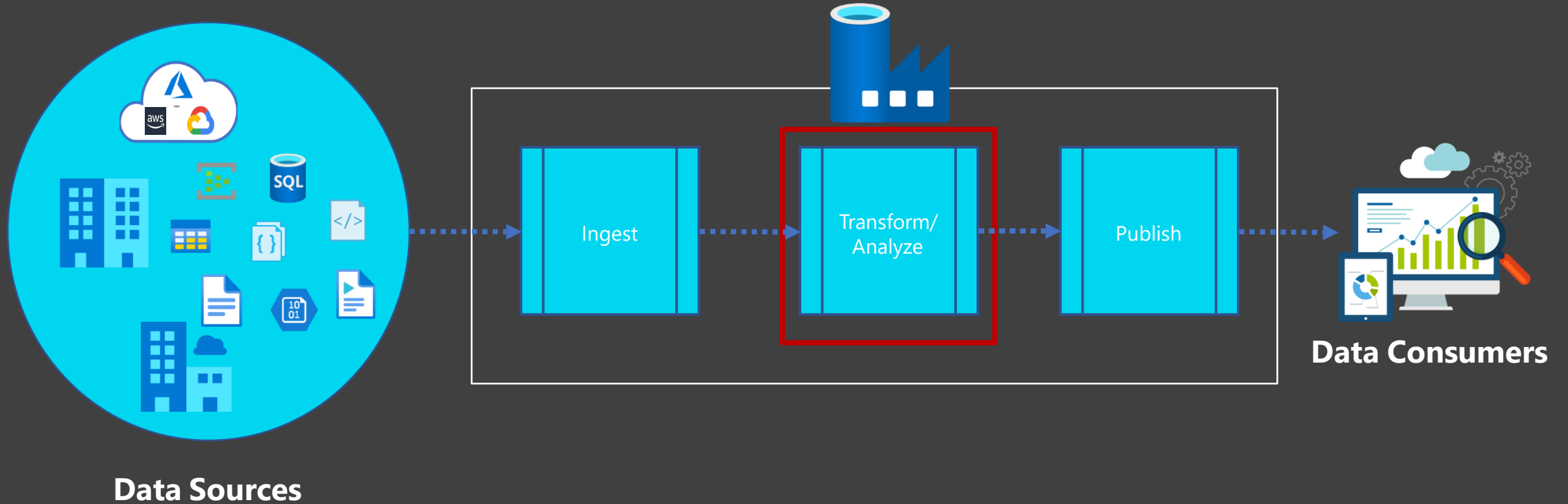
The Data Problem



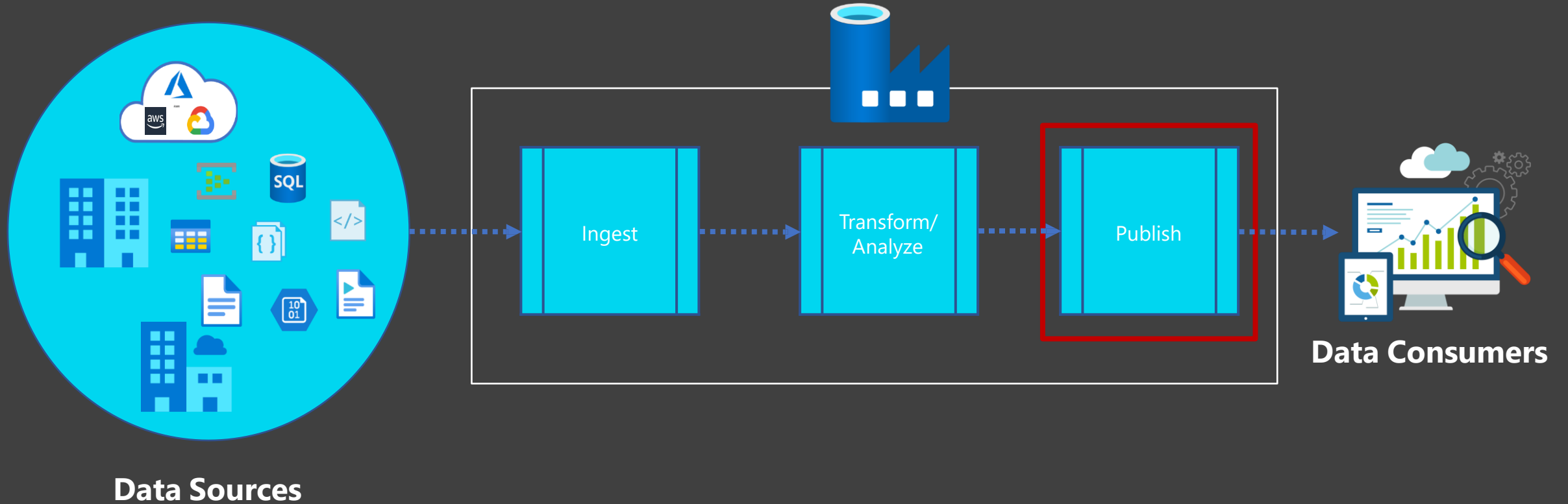
The Data Problem



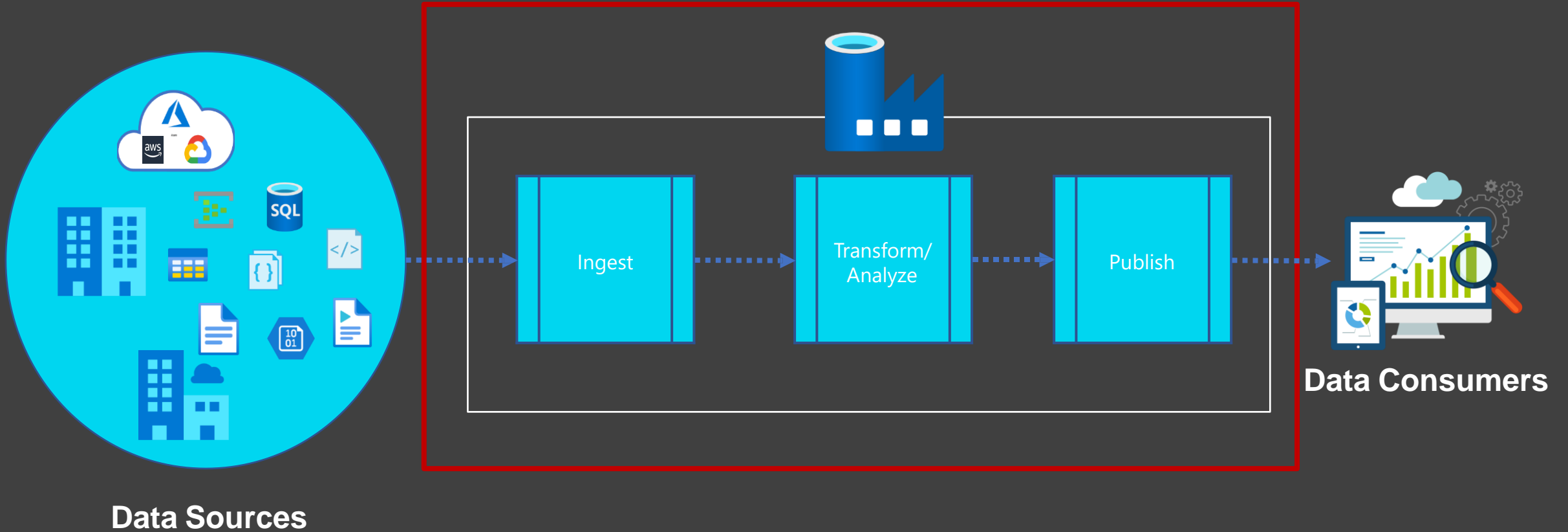
The Data Problem



The Data Problem



The Data Problem



What is Azure Data Factory?



Fully Managed Service

Serverless

Data Integration Service

Data Transformation Service

Data Orchestration Service

A fully managed, serverless data integration solution for ingesting, preparing and transforming all of your data at scale.

What Azure Data Factory Is Not



Data Migration Tool

Data Streaming Service

Suitable for Complex Data Transformations

Data Storage Service

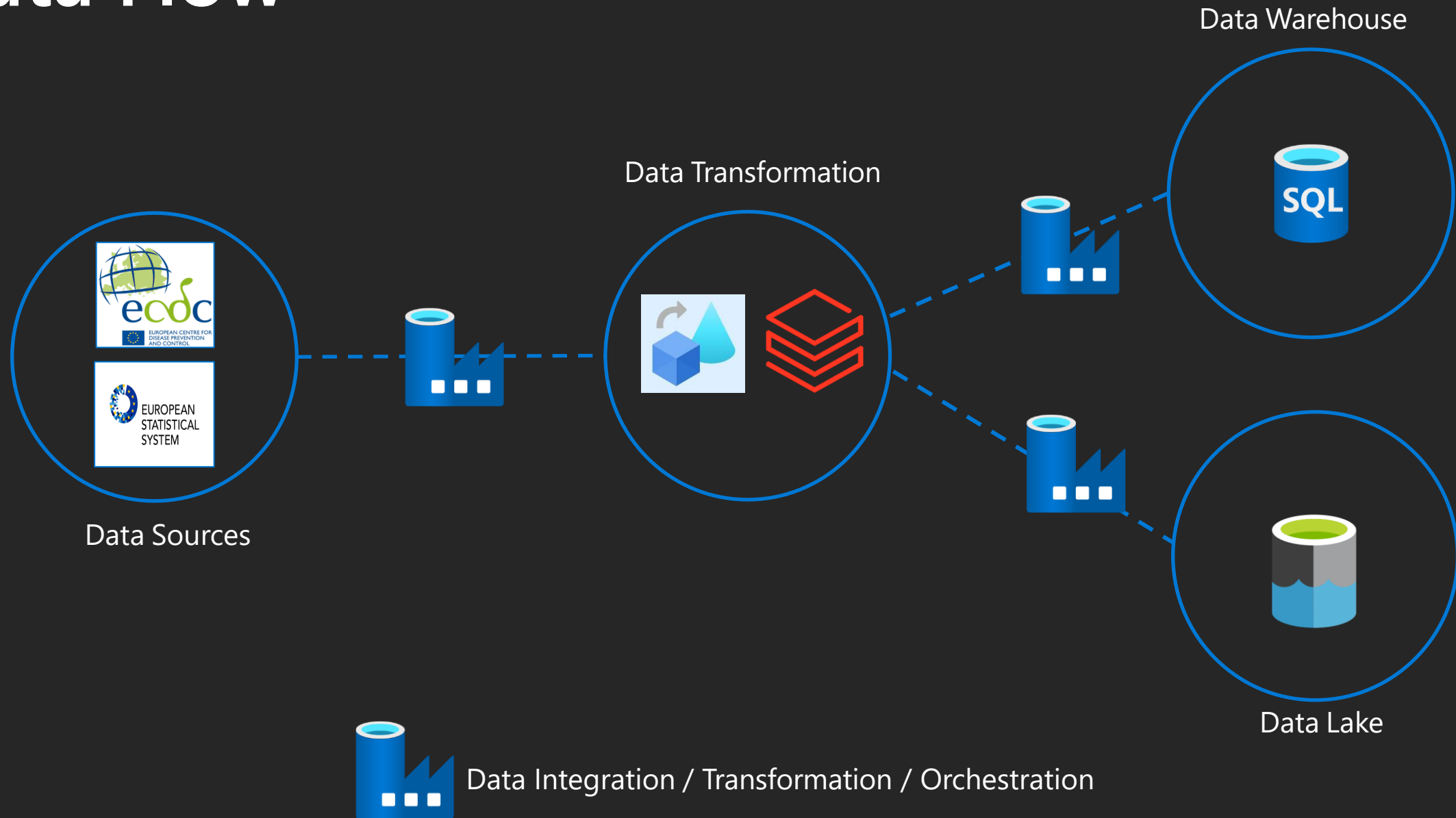


Project Overview



CORONAVIRUS
[COVID-19]

Data Flow



Covid-19 Sample Reporting

Covid-19 Cases EU/EEA & UK

Total Confirmed Cases

1,438,022

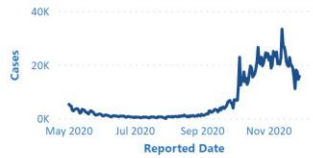
Total Deaths

31,981

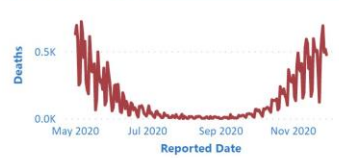
Reported Date (From - To)

01/05/2020 29/11/2020

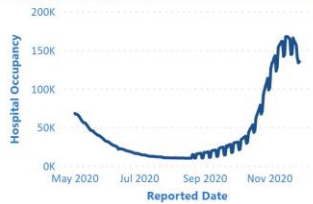
Total Cases Trend



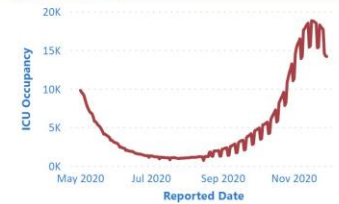
Total Deaths Trend



Hospital Occupancy Trend



ICU Occupancy Trend



Country

- ☐ Albania
- ☐ Andorra
- ☐ Armenia
- ☐ Austria
- ☐ Azerbaijan
- ☐ Belarus
- ☐ Belgium
- ☐ Bosnia and Herzegovina
- ☐ Bulgaria
- ☐ Croatia
- ☐ Cyprus
- ☐ Czechia
- ☐ Denmark
- ☐ Estonia
- ☐ Faroes
- ☐ Finland
- ☐ France
- ☐ Georgia
- ☐ Germany
- ☐ Gibraltar
- ☐ Greece
- ☐ Guernsey

Covid-19 Testing EU/EEA & UK

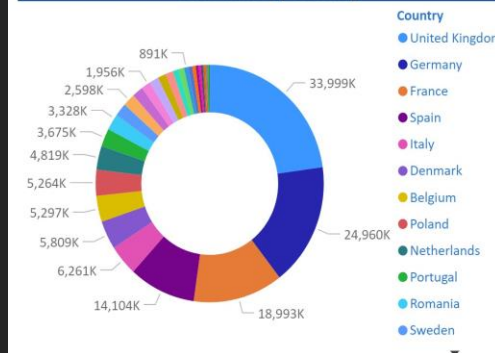
Country

All

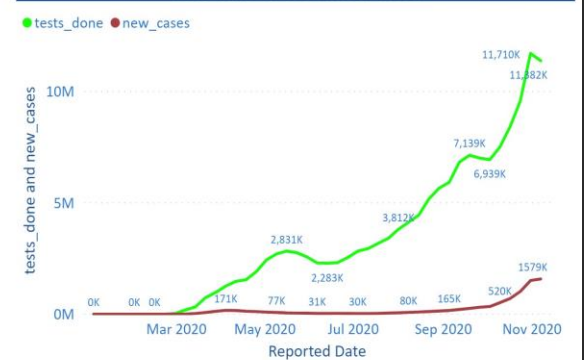
Reported Date (From - To)

04/01/2020 07/11/2020

Tests done by Country



Tests done Vs Confirmed Cases



Data Pipeline Monitoring

Microsoft Azure | Data Factory | covid-reporting-adf

Dashboard

Runs

Pipeline runs

Trigger runs

Runtimes & sessions

Integration runtimes

Data flow debug

Notifications

Alerts & metrics

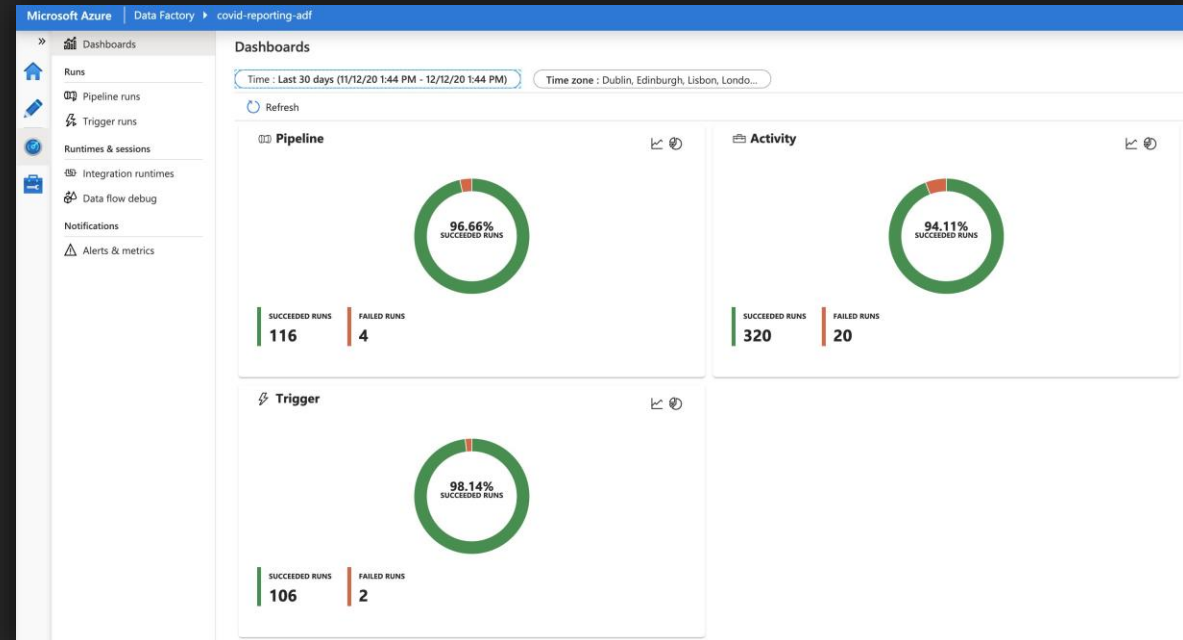
Pipeline runs

Triggered Debug Run Cancel Refresh

Search by run ID or name: Dublin, Edinburgh, L... Last 7 days Pipeline name: All Status: All Runs: Latest runs Add filter Copy filters

Showing 1 - 35 items

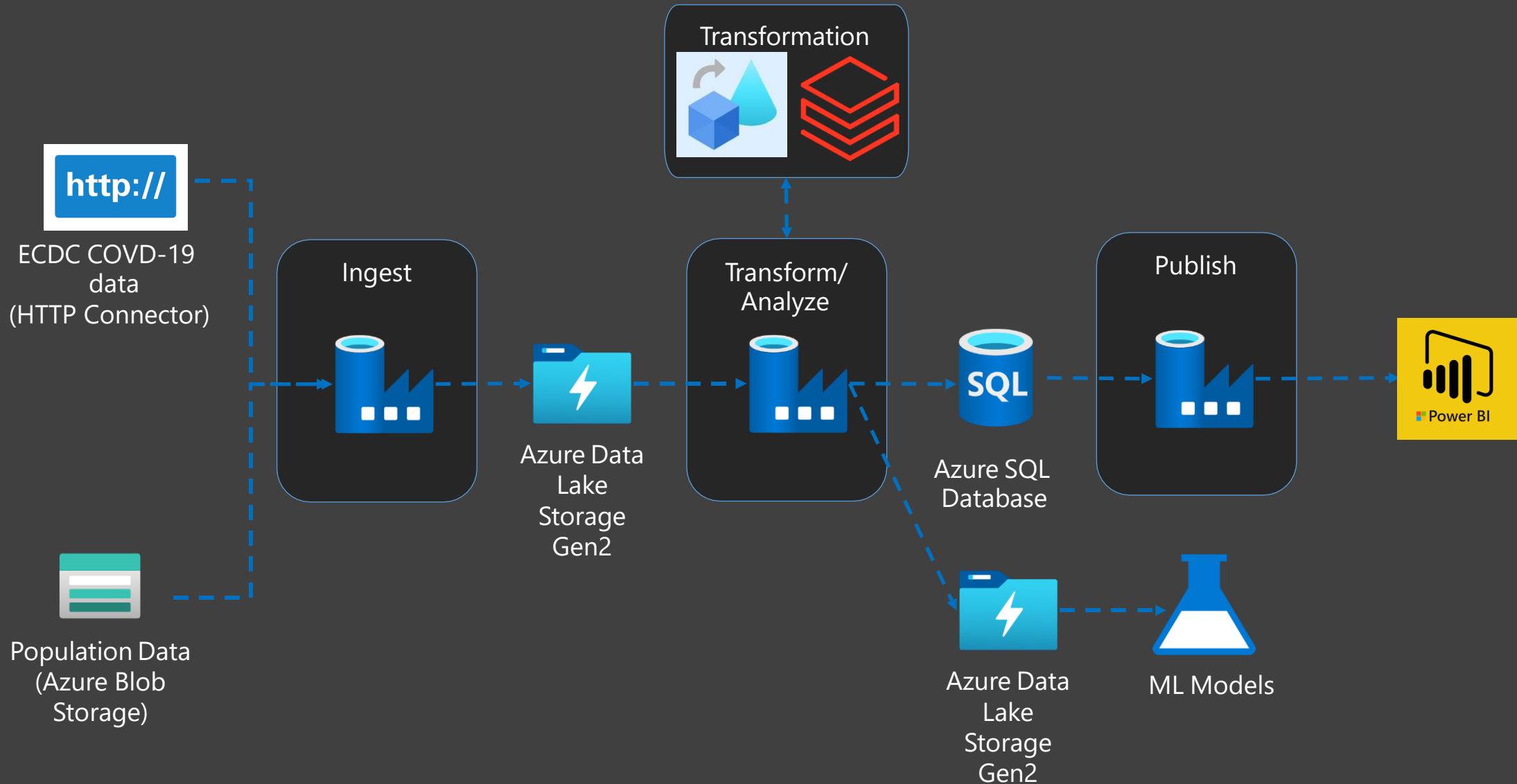
Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters	Annotations	Error	Run ID
pl_sqlize_hospital_admissions...	12/10/20, 12:08:18 AM	12/10/20, 12:08:26 AM	00:00:08	tr_sqlize_hospital_admiss	Succeeded	Original				6c6d4e32-5a16-4f8a-
pl_sqlize_cases_and_deaths_d...	12/10/20, 12:08:05 AM	12/10/20, 12:08:18 AM	00:00:13	tr_sqlize_cases_and_deat	Succeeded	Original				3de4217c-a018-4947-
pl_process_hospital_admission...	12/10/20, 12:01:53 AM	12/10/20, 12:08:06 AM	00:06:13	tr_process_hospital_admi	Succeeded	Original				a216226f-e230-4ac3-
pl_process_cases_and_deaths...	12/10/20, 12:01:47 AM	12/10/20, 12:07:55 AM	00:06:07	tr_process_cases_and_de	Succeeded	Original				38492eb6-383f-4afa-
pl_ingest_esdc_data	12/10/20, 12:00:12 AM	12/10/20, 12:01:30 AM	00:01:17	tr_ingest_esdc_data	Succeeded	Original				d17a5375-1153-4ca2-
pl_sqlize_cases_and_deaths_d...	12/9/20, 12:07:42 AM	12/9/20, 12:07:54 AM	00:00:12	tr_sqlize_cases_and_deat	Succeeded	Original				3d6d1837-0d6f-4796-
pl_sqlize_hospital_admissions...	12/9/20, 12:07:06 AM	12/9/20, 12:07:15 AM	00:00:08	tr_sqlize_hospital_admiss	Succeeded	Original				919741a2-94d9-43fe-
pl_process_cases_and_deaths...	12/9/20, 12:01:22 AM	12/9/20, 12:07:31 AM	00:06:09	tr_process_cases_and_de	Succeeded	Original				1b09a2b6-1260-4f63-
pl_process_hospital_admission...	12/9/20, 12:01:16 AM	12/9/20, 12:06:55 AM	00:05:39	tr_process_hospital_admi	Succeeded	Original				0bdabae9-1dea-42ef-
pl_ingest_esdc_data	12/9/20, 12:00:13 AM	12/9/20, 12:00:59 AM	00:00:45	tr_ingest_esdc_data	Succeeded	Original				30daacfe-6e6a-497e-
pl_sqlize_hospital_admissions...	12/8/20, 12:08:20 AM	12/8/20, 12:08:27 AM	00:00:07	tr_sqlize_hospital_admiss	Succeeded	Original				77ab3327-4a45-406f-
pl_sqlize_cases_and_deaths_d...	12/8/20, 12:07:55 AM	12/8/20, 12:08:08 AM	00:00:13	tr_sqlize_cases_and_deat	Succeeded	Original				e7712d45-5ecd-4646-
pl_process_cases_and_deaths...	12/8/20, 12:01:33 AM	12/8/20, 12:07:44 AM	00:06:10	tr_process_cases_and_de	Succeeded	Original				0c9e0a3d-a8e1-4337-
pl_process_hospital_admission...	12/8/20, 12:01:27 AM	12/8/20, 12:08:10 AM	00:06:43	tr_process_hospital_admi	Succeeded	Original				e44ca7d3-8cd2-4f7b-
pl_ingest_esdc_data	12/8/20, 12:00:12 AM	12/8/20, 12:01:10 AM	00:00:57	tr_ingest_esdc_data	Succeeded	Original				521a9fc7-ba89-4660-
pl_sqlize_cases_and_deaths_d...	12/7/20, 12:08:07 AM	12/7/20, 12:08:20 AM	00:00:13	tr_sqlize_cases_and_deat	Succeeded	Original				8a3572c4-2206-4247-
pl_sqlize_hospital_admissions...	12/7/20, 12:07:34 AM	12/7/20, 12:08:35 AM	00:01:01	tr_sqlize_hospital_admiss	Succeeded	Original				4956c433-7793-427b-
pl_process_cases_and_deaths...	12/7/20, 12:01:16 AM	12/7/20, 12:07:57 AM	00:06:40	tr_process_cases_and_de	Succeeded	Original				69642793-8b0d-4964-
pl_process_hospital_admission...	12/7/20, 12:01:10 AM	12/7/20, 12:07:23 AM	00:06:13	tr_process_hospital_admi	Succeeded	Original				9a08c76e-90bc-464d-
pl_ingest_esdc_data	12/7/20, 12:00:12 AM	12/7/20, 12:00:53 AM	00:00:40	tr_ingest_esdc_data	Succeeded	Original				83268d76-2472-4b0c-



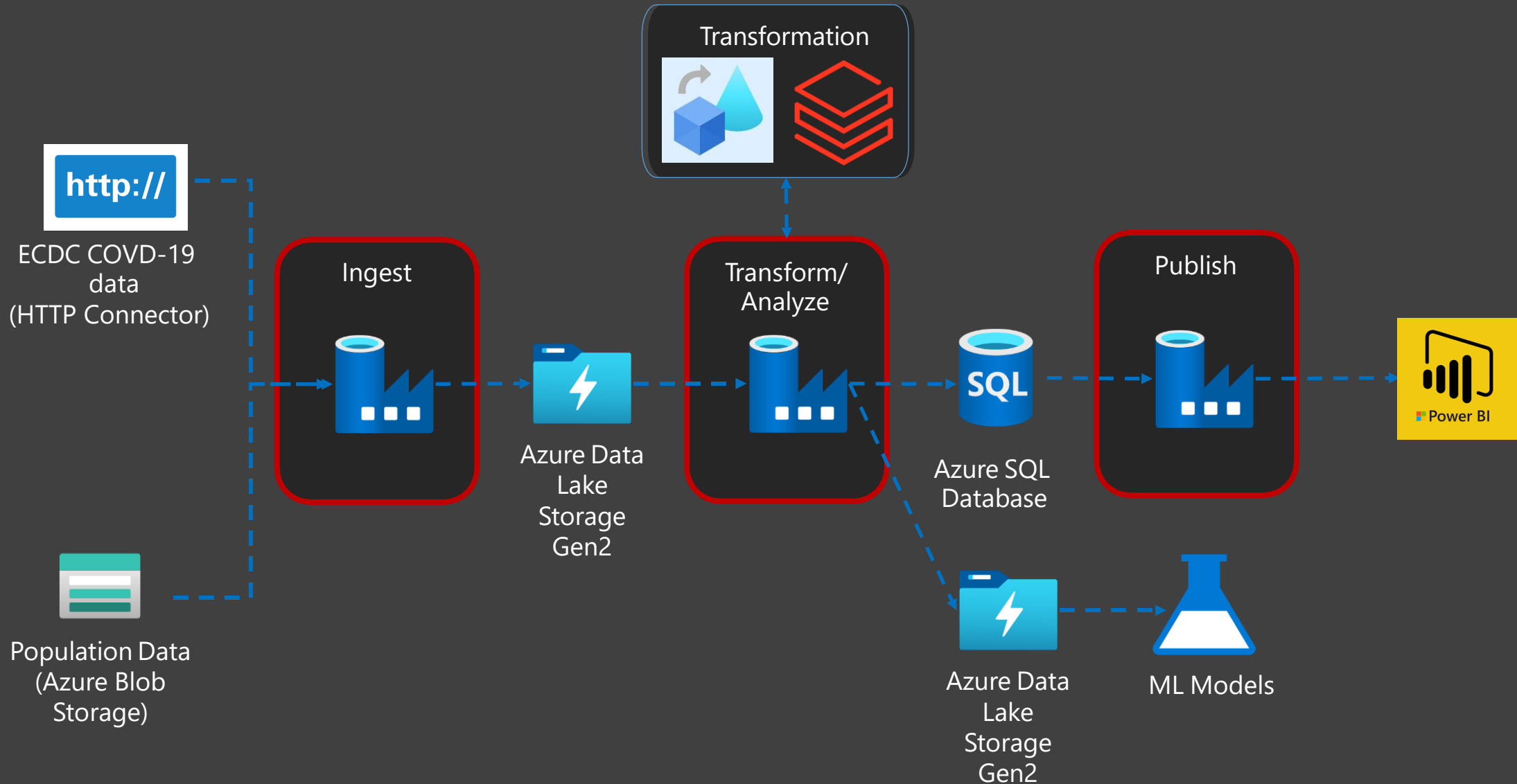


Solution Architecture

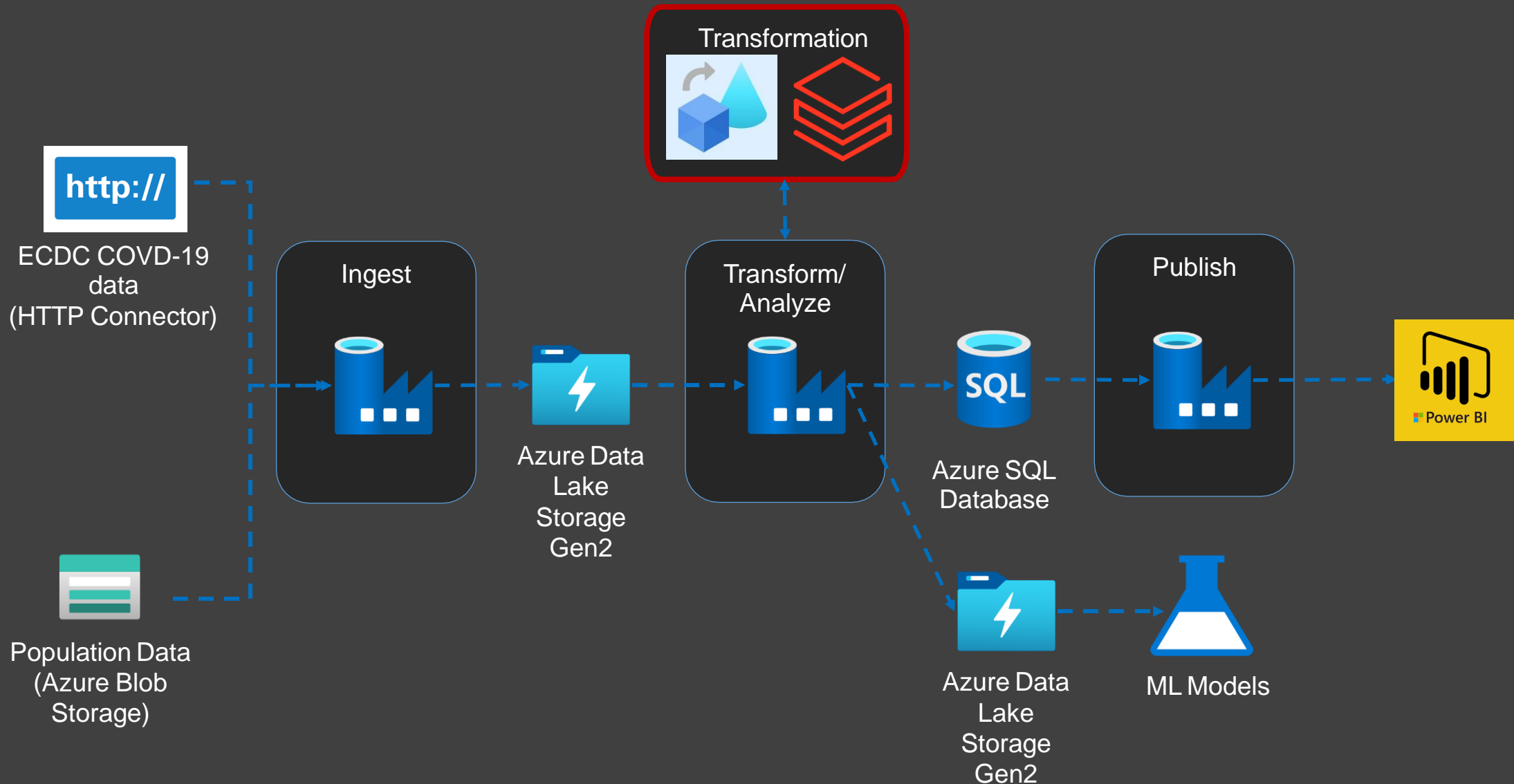
Solution Architecture



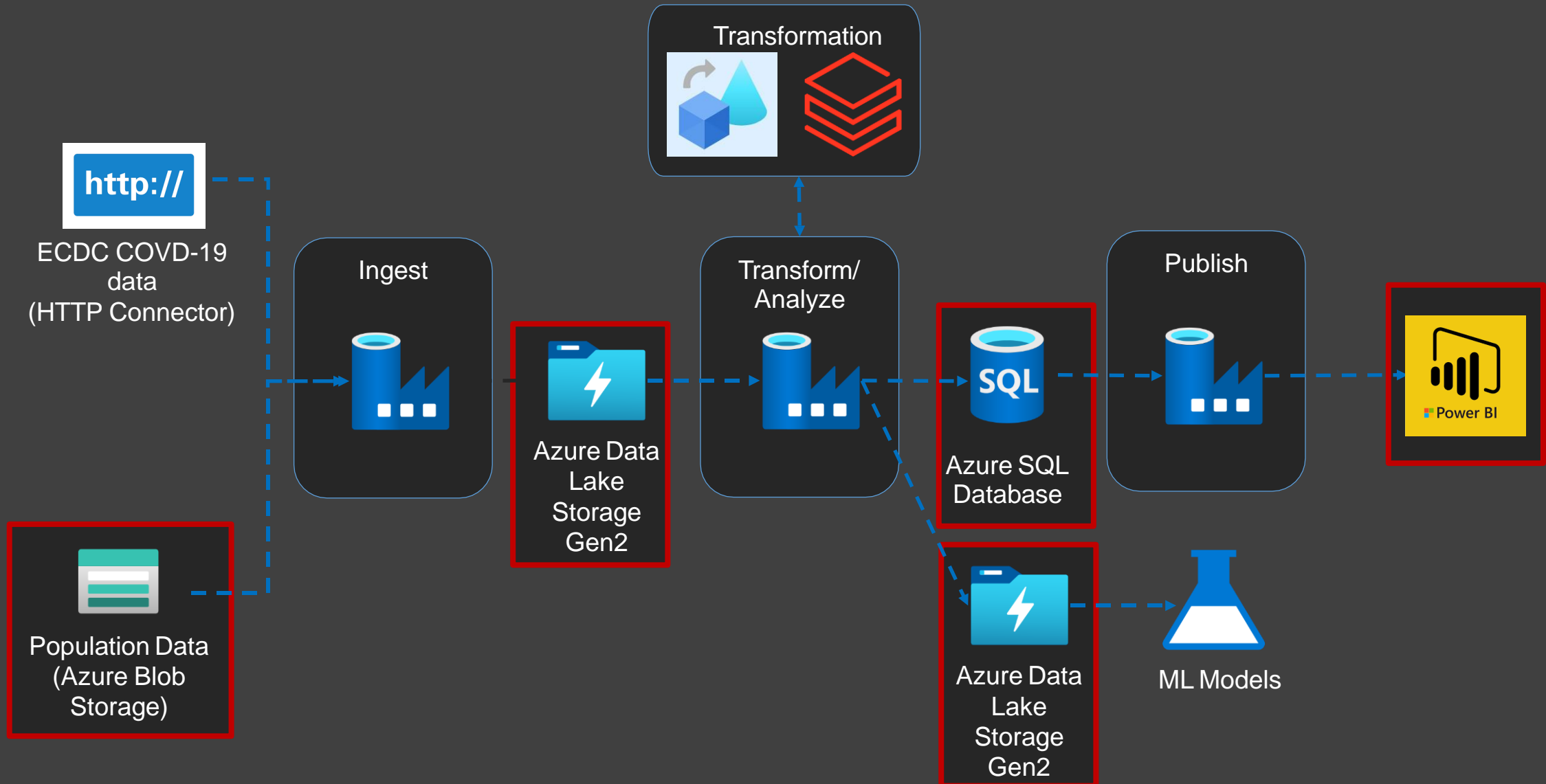
Solution Architecture



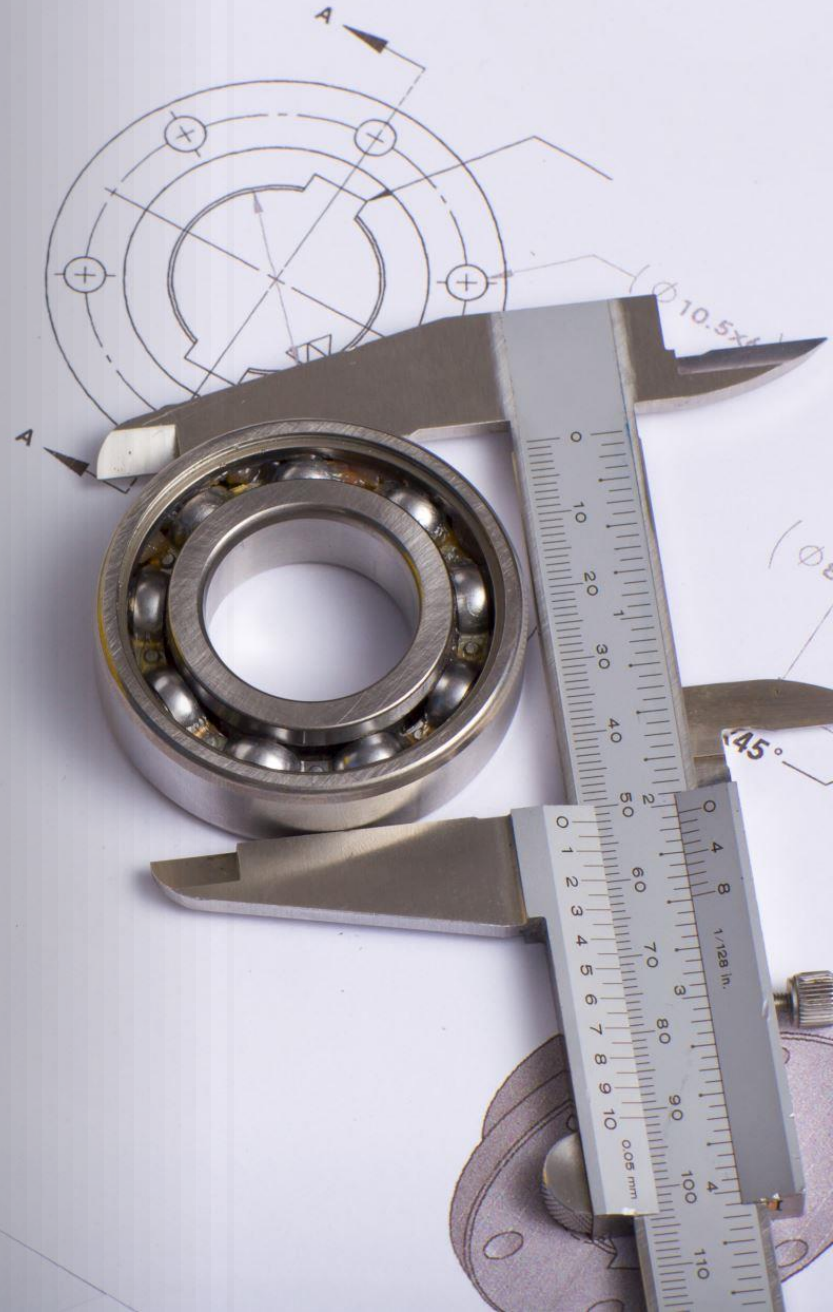
Solution Architecture



Solution Architecture



Detailed Solution Design



Overview

Regional Office



Clusters by Funding Allocation



Clusters by COVID Integrated into 2/4 W



Clusters by GHRP Indicator



Data Sources

- ECDC Website
 - Confirmed cases
 - Mortality
 - Hospitalization/ ICU Cases
 - Testing Numbers
- Eurostat Website
 - Population by age

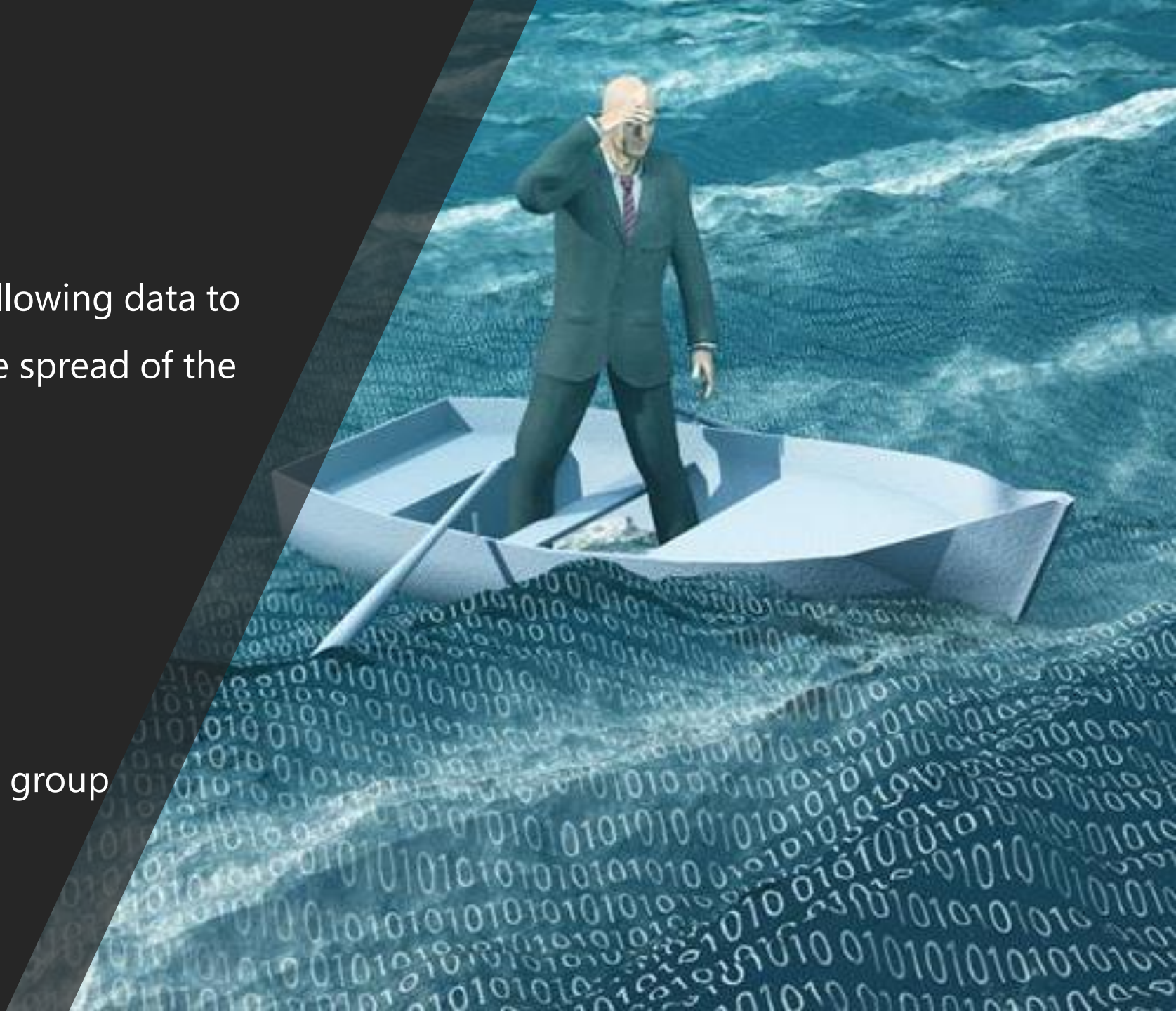
COVID Dashboard: Bulletin	Regional Distribution	CCBF	CIPF	COVID Integrated into 2/4 W	GHRP Indicator Reporting	WHO Monitoring Indicator Reporting
Government Dashboard	EMRO	Yes	Yes	Yes	Yes, in full	Yes, all
EMRO	EMRO	Yes	Yes	Yes	Yes, in full	Yes, all
Government Dashboard	Afro	-	Yes	Yes	Yes, partial	Yes, some
Government Report	Afro	-	Yes	Yes	Yes, partial	Yes, some
Health Cluster Reports	Afro	-	Yes	Yes	Yes, partial	Yes, some
Country/Region/Entity	Afro	-	Yes	Not specified	Not specified	Not specified
Health Cluster Reports	Afro	-	-	In process	Yes, partial	Yes, all
Government Reports	Afro	-	Yes	Separate	Yes	Yes, some
Government of WHO Region	Afro	-	Yes	Separate	Yes, in full	Not specified
Country/Region/Entity	EMRO	-	Yes	Not specified	Not specified	Not specified
Health Cluster Reports	EMRO	-	-	In process	No	No
Government Dashboard	EMRO	-	-	In process	No	No
Health Cluster Reports	EMRO	-	-	In process	No	No
WHO Region	EMRO	-	-	In process	No	No
Health Cluster Reports	EMRO	Yes	-	Not specified	Not specified	Not specified
Health Cluster Reports	EMRO	Yes	-	Separate	No	No
Health Cluster Reports	Afro	Yes	-	In process	In process	Yes, all
Government Dashboard	Afro	Yes	-	In process	In process	Yes, all
Government Reports	Afro	Yes	-	In process	In process	Yes, all
Health Cluster Reports	Afro	-	Yes	In process	No	Yes, all
Government Dashboard	Afro	-	Yes	In process	No	Yes, all
WHO Region	Afro	-	Yes	In process	No	Yes, all
WHO Region	EMRO	-	-	-	-	-
WHO Region	EMRO	-	-	-	-	-



Data Lake

Data Lake to be built with the following data to aid Data Scientists to predict the spread of the virus/mortality

- Confirmed cases
- Mortality
- Hospitalization/ ICU Cases
- Testing Numbers
- Country's population by age group



Data Warehouse

Data Warehouse to be built with the following data to aid Reporting on trends

- Confirmed cases
- Mortality rates
- Hospitalization / ICU Cases
- Testing Numbers





Environment set-up

- Azure Subscription
- Azure Data Factory
- Azure Blob Storage Account
- Azure Data Lake Storage Gen2
- Azure SQL Database
- [Azure Databricks](#)



Creating Azure Free Account



Creating Azure Data Factory



Creating Azure Storage Account



Creating Azure Data Lake Gen2



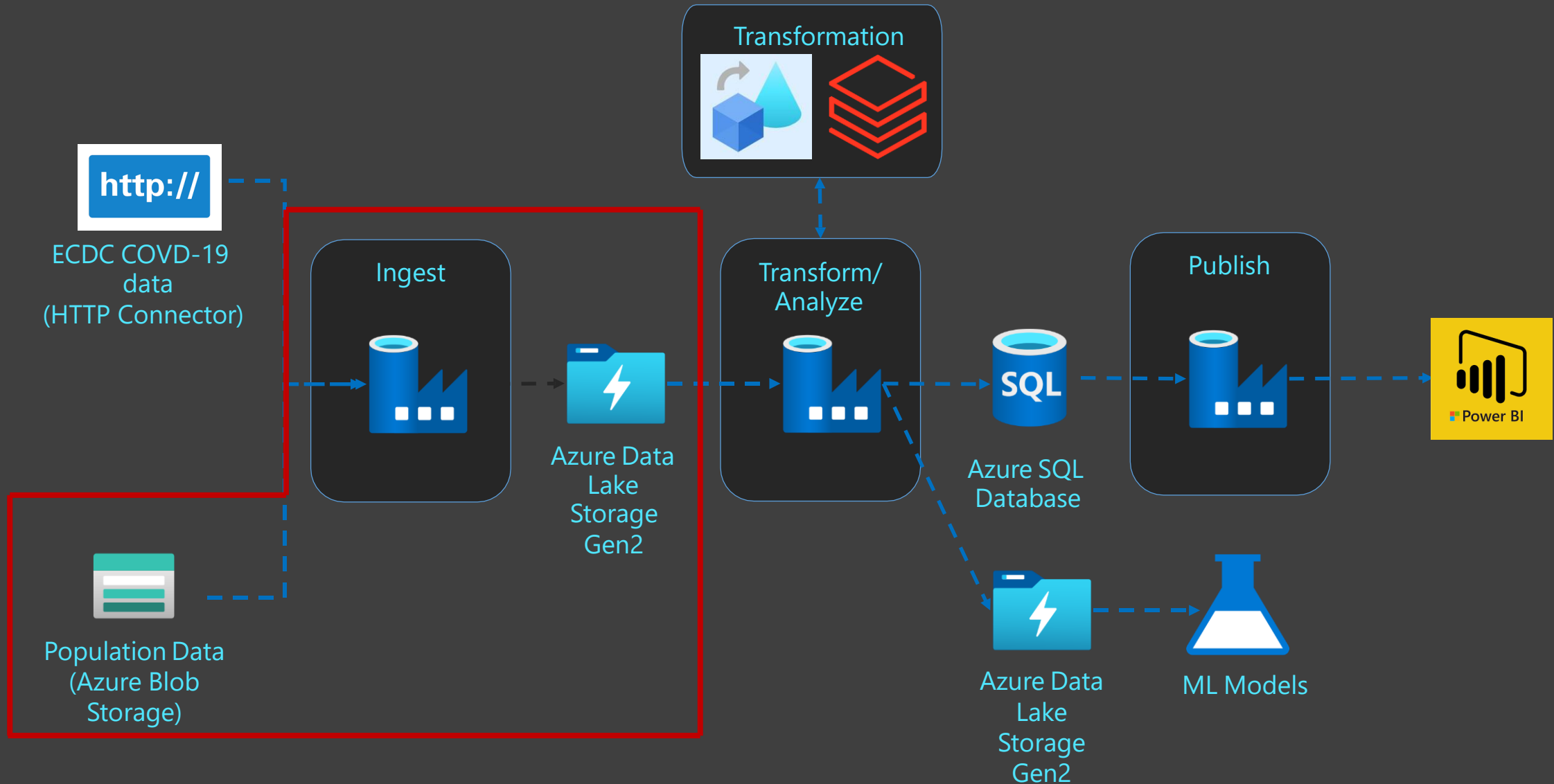
Creating Azure SQL Database



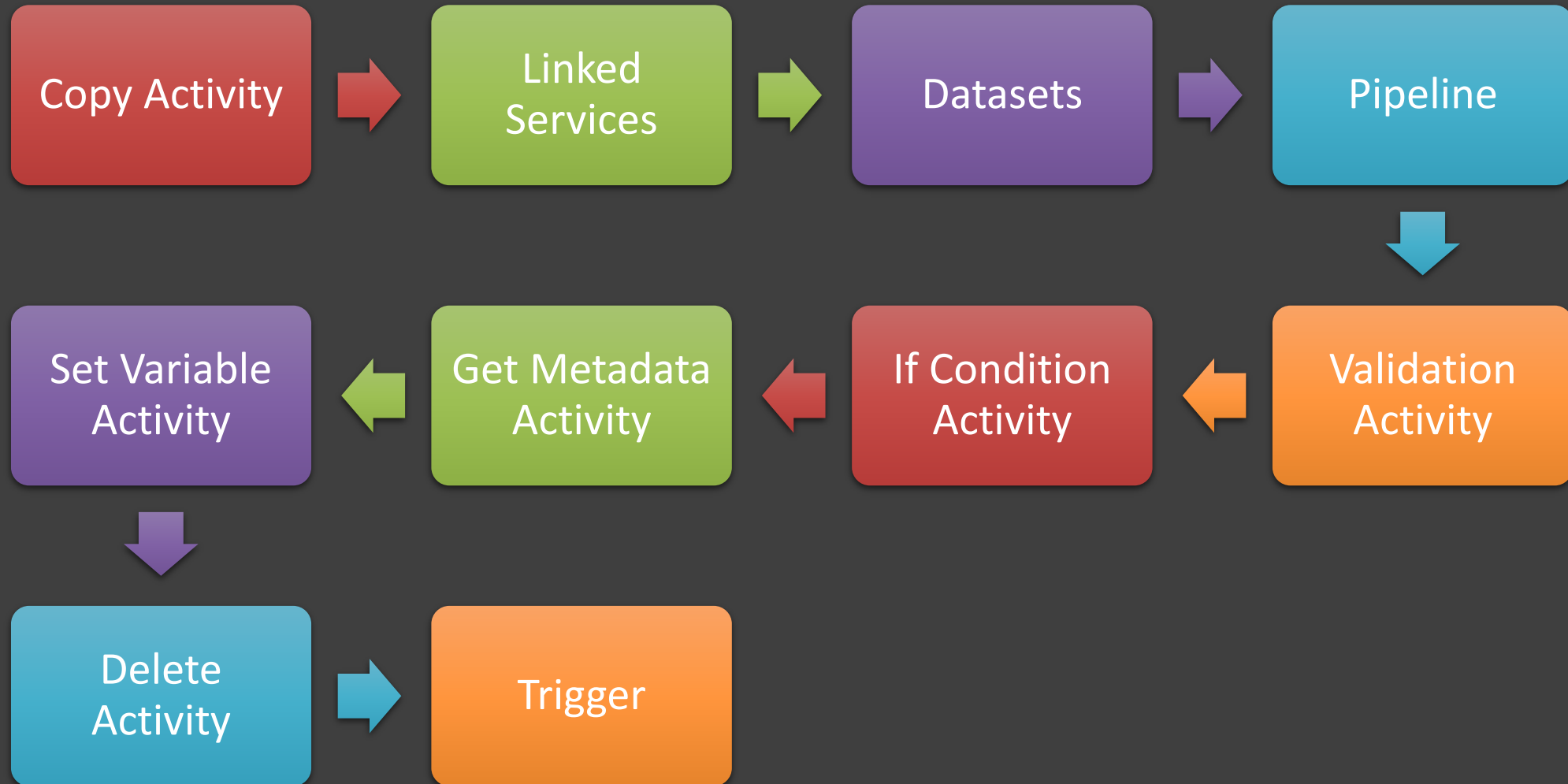
Data Ingestion



Population Dataset Ingestion



Population Dataset Ingestion



Copy Activity

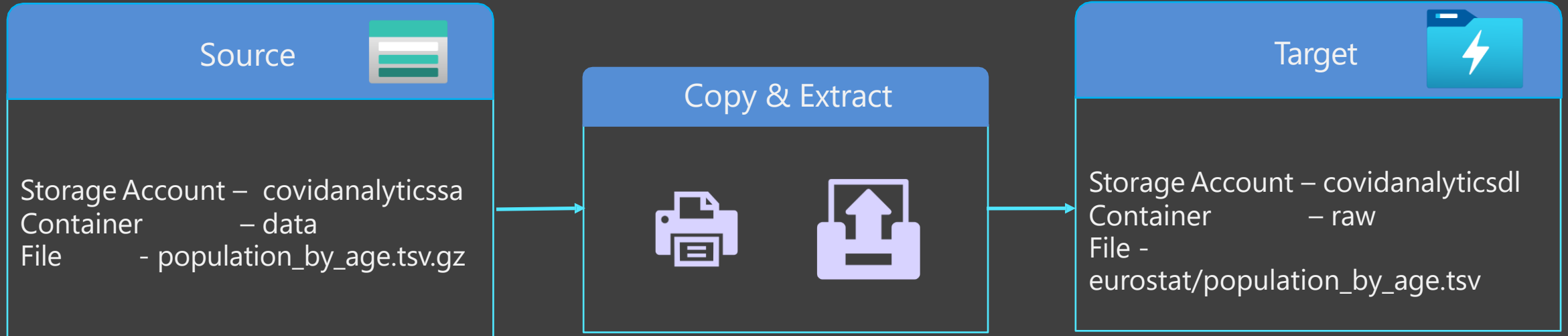
Azure Blob Storage → Azure Data Lake

Copy Activity

Ingest “population by age” for all EU Countries into the Data Lake to support the machine learning models to predict increase in Covid-19 mortality rates



Copy Activity



Copy Activity

Source



Azure Blob
Storage



Source File
(Zipped
TSV)

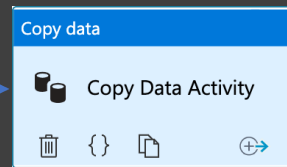
Sink



Azure Data
Lake

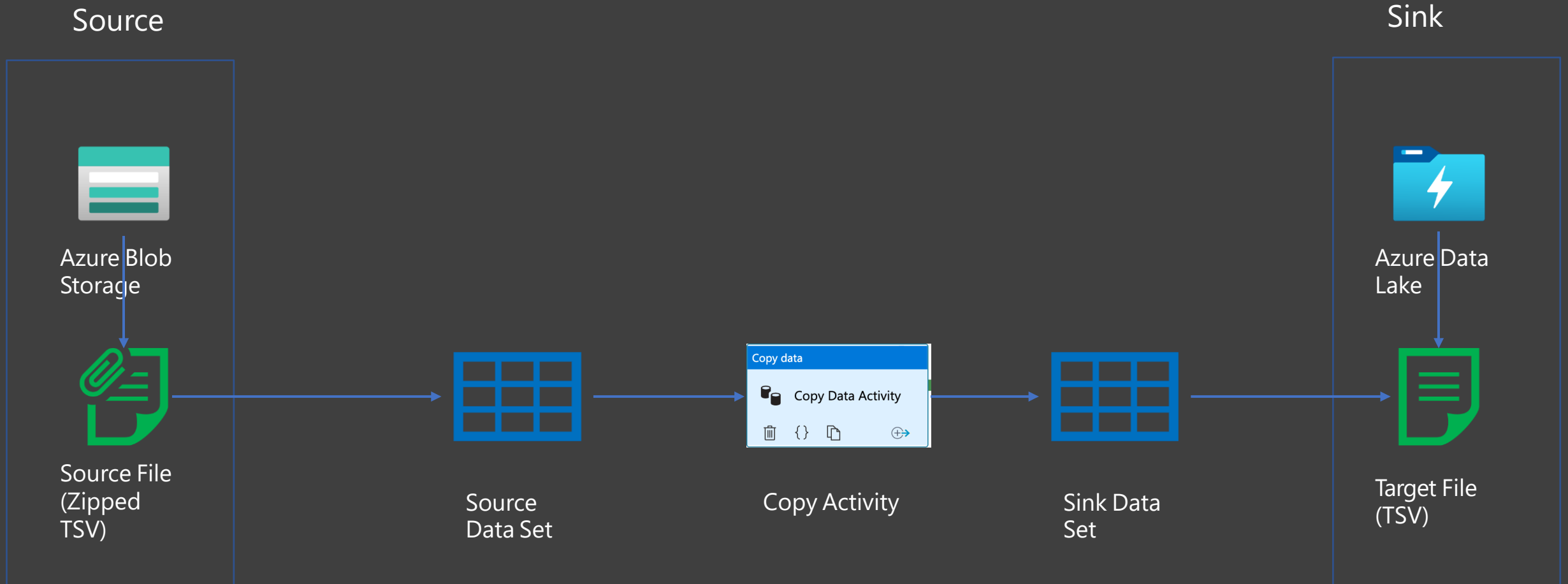


Target File
(TSV)

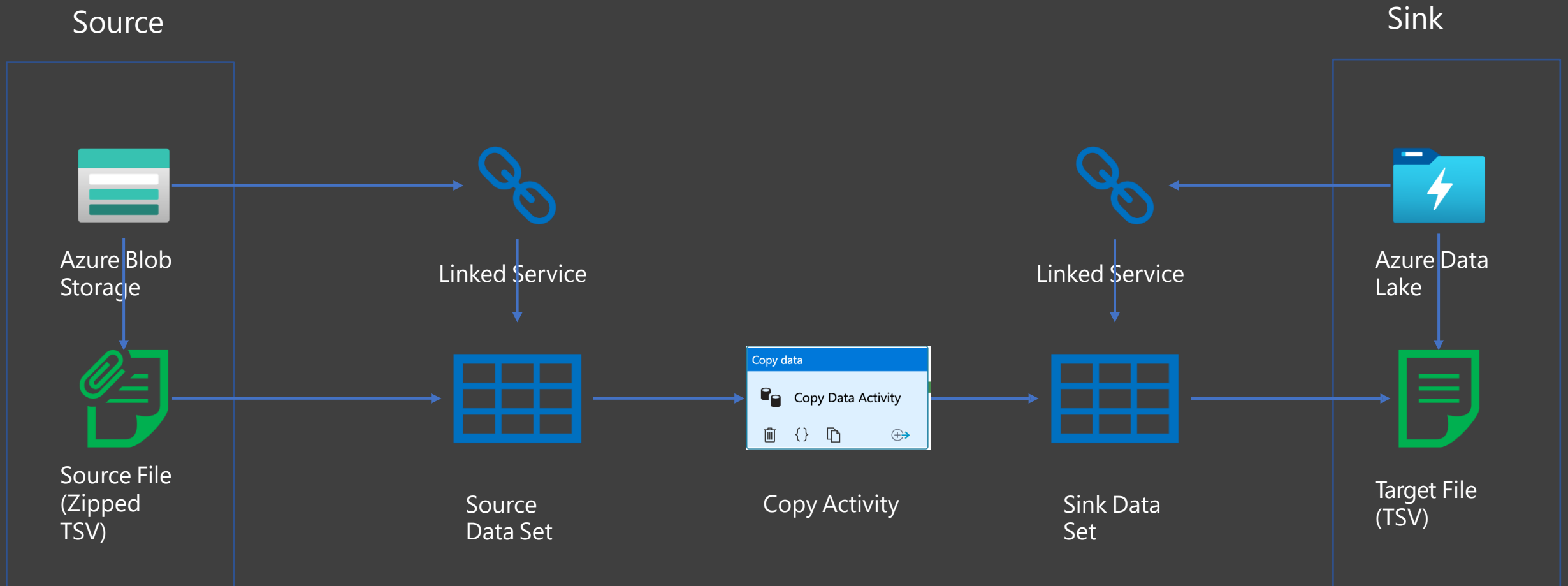


Copy Activity

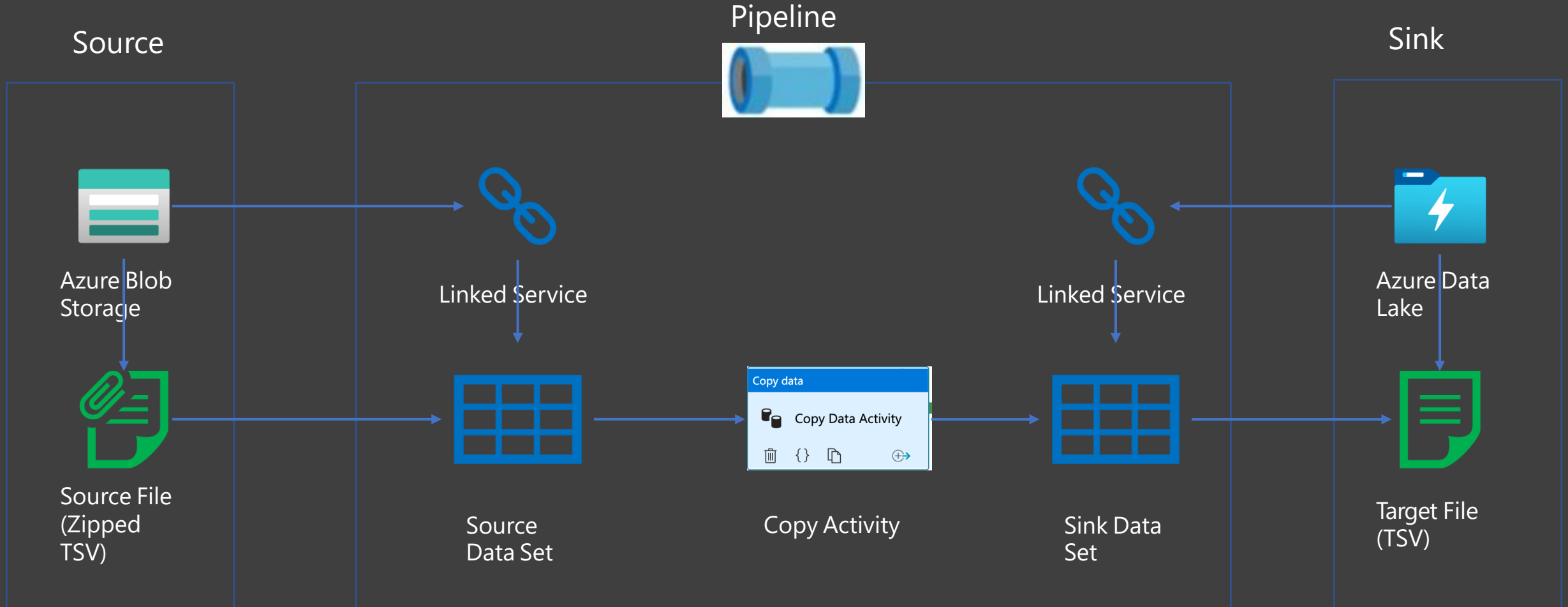
Copy Activity



Copy Activity



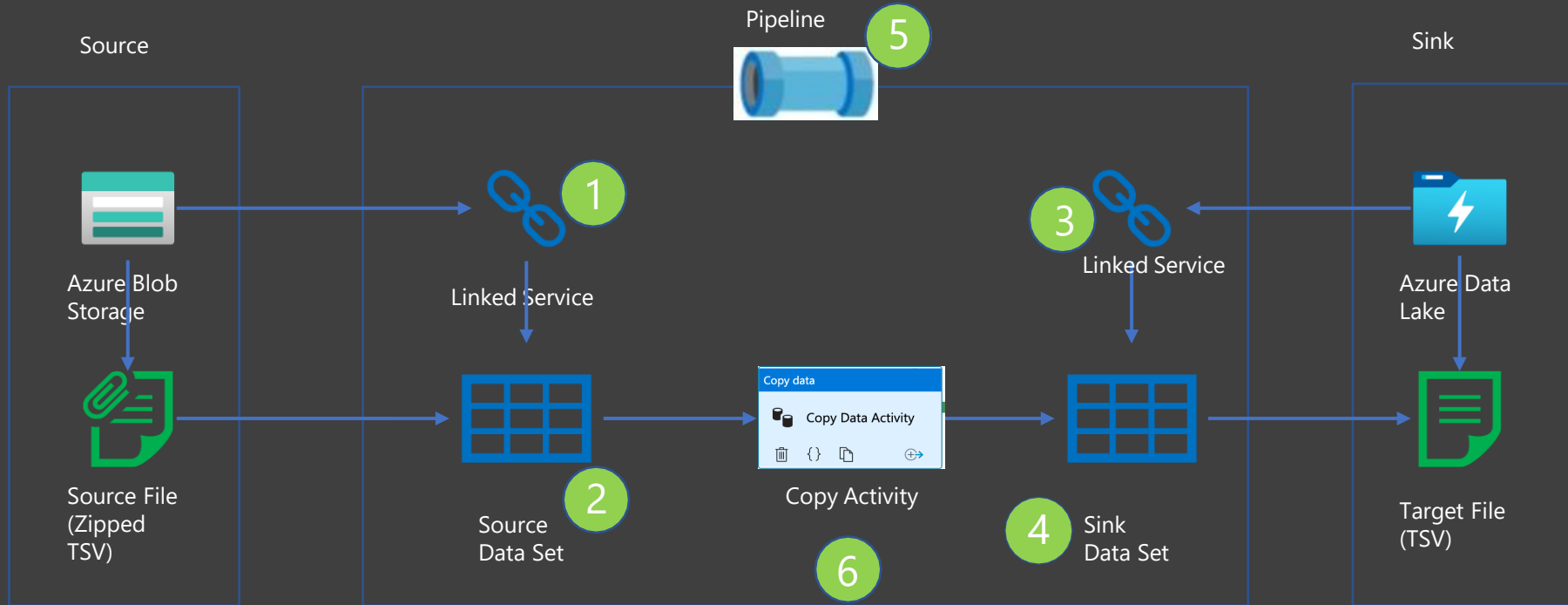
Copy Activity



Copy Activity From Azure Blob Storage



Copy Activity



Storage Account: covidanalyticssa
Container: data
File: population_by_age.tsv.gz

Storage Account: covidanalyticddl
Container: raw
File: eurostat/eu/population_by_age.tsv

- 1 ls_asa_covidreportingsa
- 2 ds_population_raw_gz
- 3 ls_adls_covidreportingdl
- 4 ds_population_raw_tsv
- 5 pl_ingest_population_data
- 6 Copy Population Data

Handling Real World Scenarios



Scenario 1

Execute Copy Activity when the file becomes available



Scenario 2

Execute Copy Activity only if file contents are as expected



Scenario 3

Delete the source file on successful copy



Scheduling Pipeline Execution





Triggers



Schedule Trigger



Tumbling Window Trigger



Event Trigger



Schedule Trigger



Runs on a calendar/ Clock



Supports periodic and specific times



Trigger to Pipeline is Many to Many



Can only be scheduled for a future time to start



Tumbling Window Trigger



Runs at periodic intervals



Windows are fixed sized, non-overlapping



Can be scheduled for the past windows/slices



Trigger to Pipeline is one to one



Storage Events Trigger



Runs in response to events



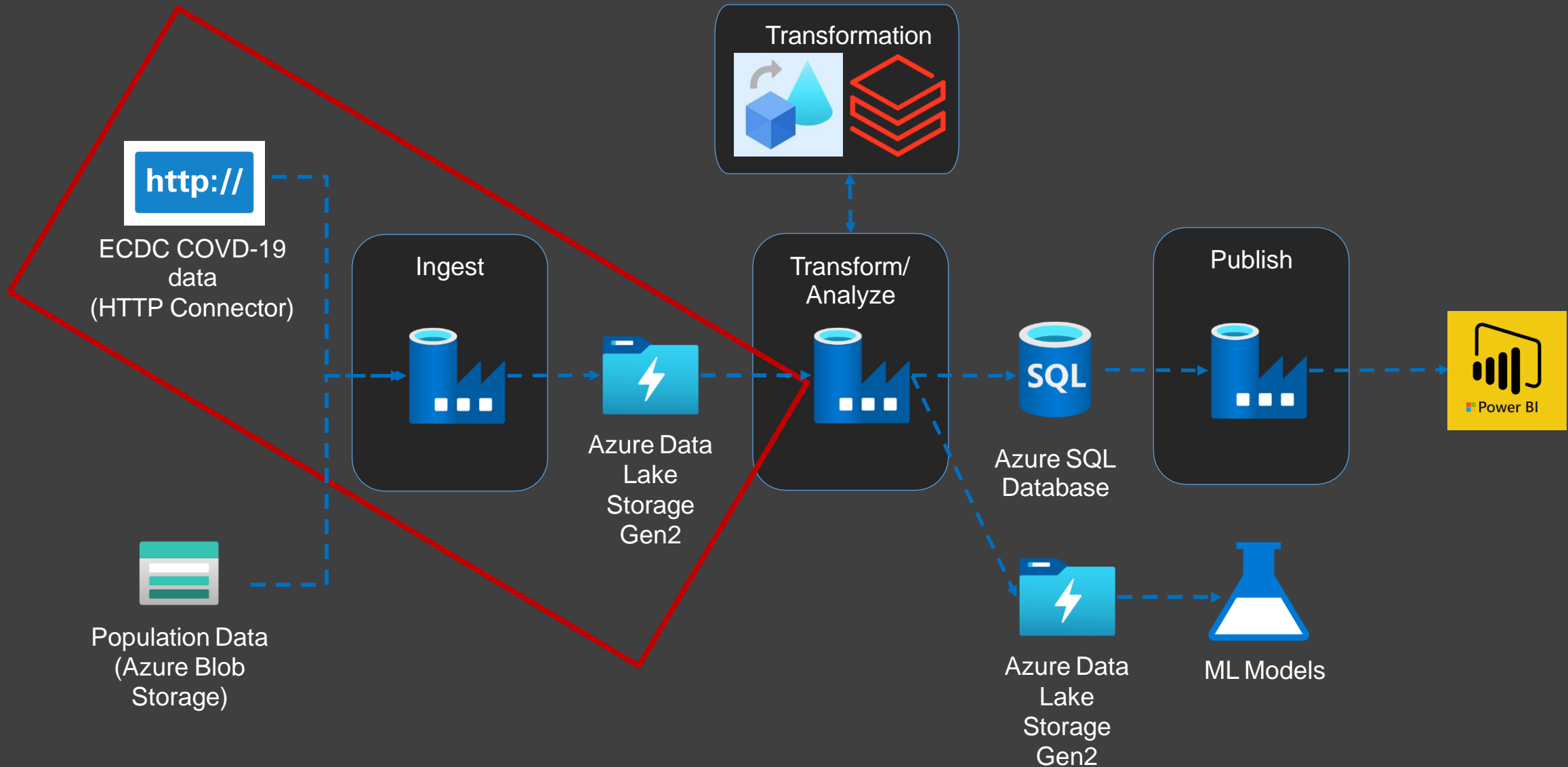
Events can be creation or deletion of Blobs/Files



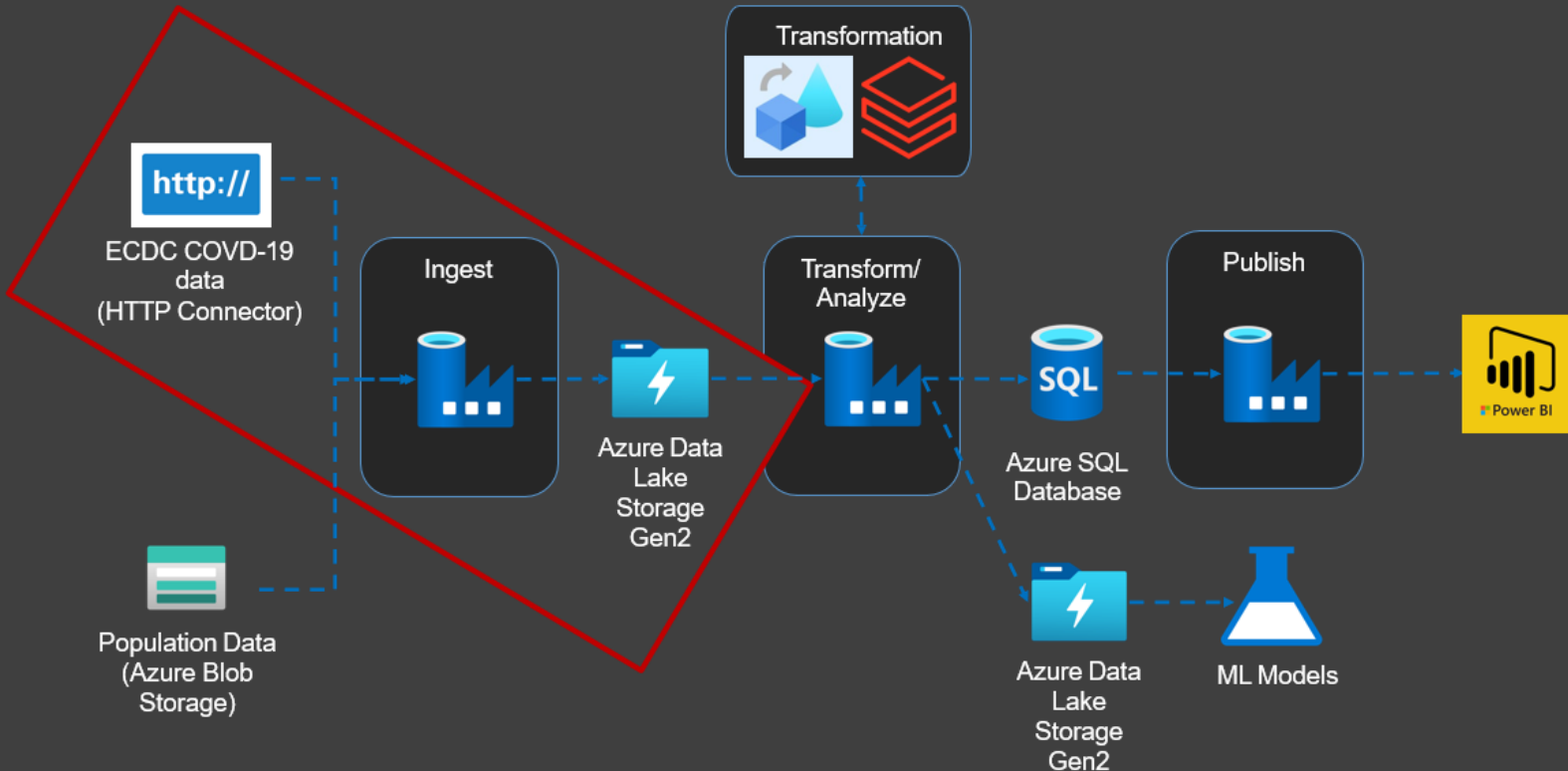
Trigger to Pipeline is Many to Many

Data Ingestion - Module Overview (ECDC Data)

Data Ingestion – ECDC Data



Data Ingestion – ECDC Data



- ECDC Data Overview
- Create Initial Pipeline
- Pipeline Variables
- Pipeline Parameters
- Lookup Activity
- For Each Activity
- Linked Service Parameters
- Metadata driven pipeline

Data Ingestion – ECDC Data

HTTP



Azure Data Lake

Data Ingestion Requirements

- Covid-19 new cases and deaths by Country
- Covid-19 Hospital admissions & ICU cases
- Covid-19 Testing Numbers
- Country Response to Covid-19

URL - <https://www.ecdc.europa.eu/en/covid-19/data>

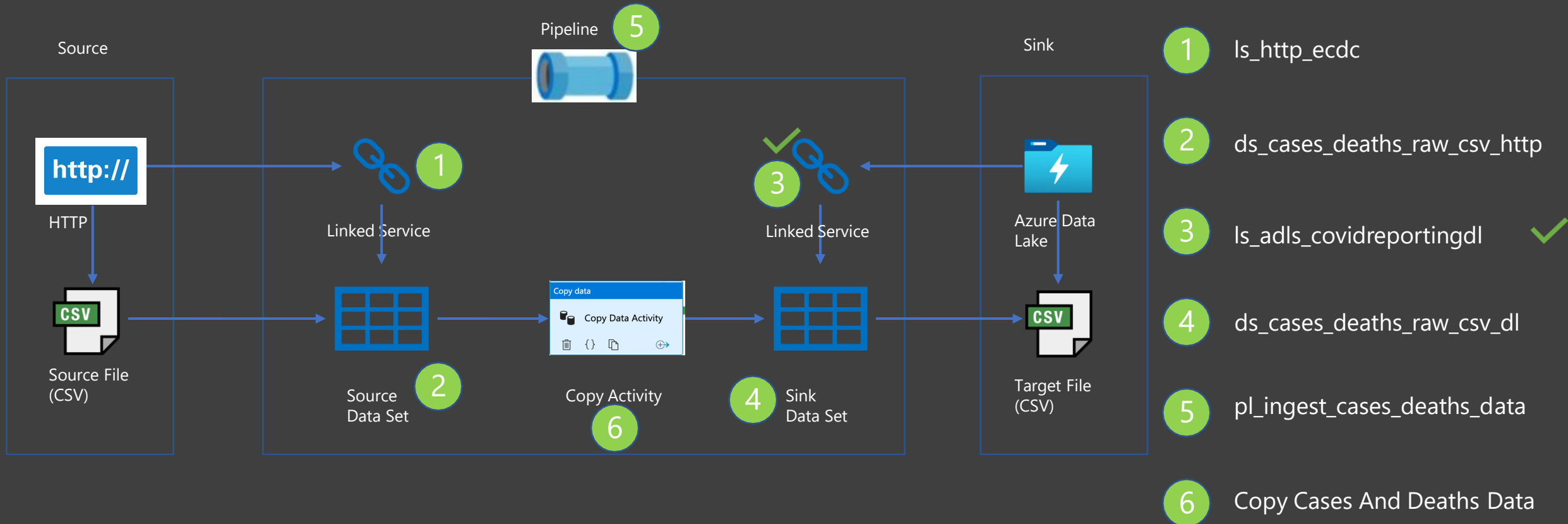


Data Ingestion

Case & Deaths Data

URL - <https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19>

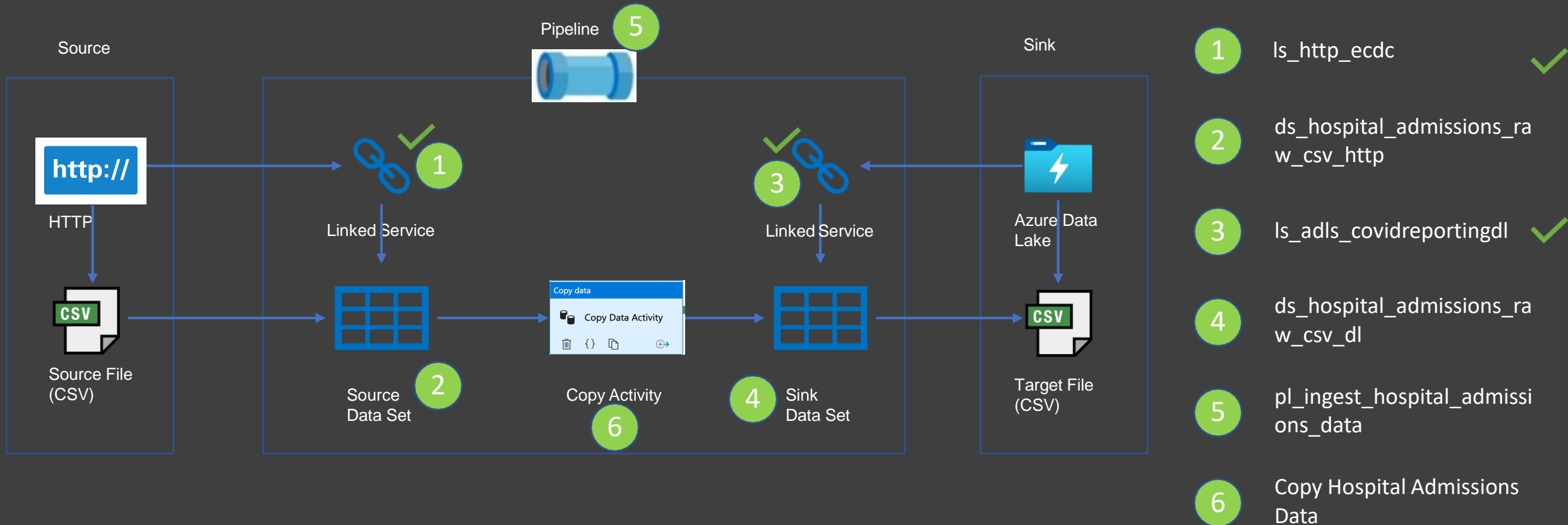
Copy Activity – Case & Deaths Data



URL: <https://opendata.ecdc.europa.eu/covid19/nationalcasedeath/csv/data.csv>

Storage Account: covidreportingdl
Container: raw
File: ecdc/cases_deaths.csv

Copy Activity – Hospital Admission Data



URL: <https://opendata.ecdc.europa.eu/covid19/hospitalicuadmissionrates/csv/data.csv>

Storage Account: covidreportingdl
Container: raw
File: ecdc/hospital_admissions.csv

Parameters & Variables

Parameters are external values passed into pipelines, datasets or linked services. The value cannot be changed inside a pipeline.

Variables are internal values set inside a pipeline. The value can be changed inside the pipeline using Set Variable or Append Variable Activity

Differences

Source

<https://opendata.ecdc.europa.eu/covid19/nationalcasedeath/csv/data.csv>

<https://opendata.ecdc.europa.eu/covid19/hospitalicuadmissionrates/csv/data.csv>

<https://opendata.ecdc.europa.eu/covid19/testing/csv/data.csv>

https://www.ecdc.europa.eu/sites/default/files/documents/response_graphs_data_2021-08-26.csv

Sink

raw/ecdc/case_distribution.csv

raw/ecdc/hospital_admission.csv

raw/ecdc/testing.csv

raw/ecdc/country_response.csv



Scenario 1

Use Variables to Parameterize the Pipeline

Scenario 2

Use Pipeline Parameters to make
the Pipeline generic





Scenario 3

Schedule the Pipeline and pass Trigger Parameters

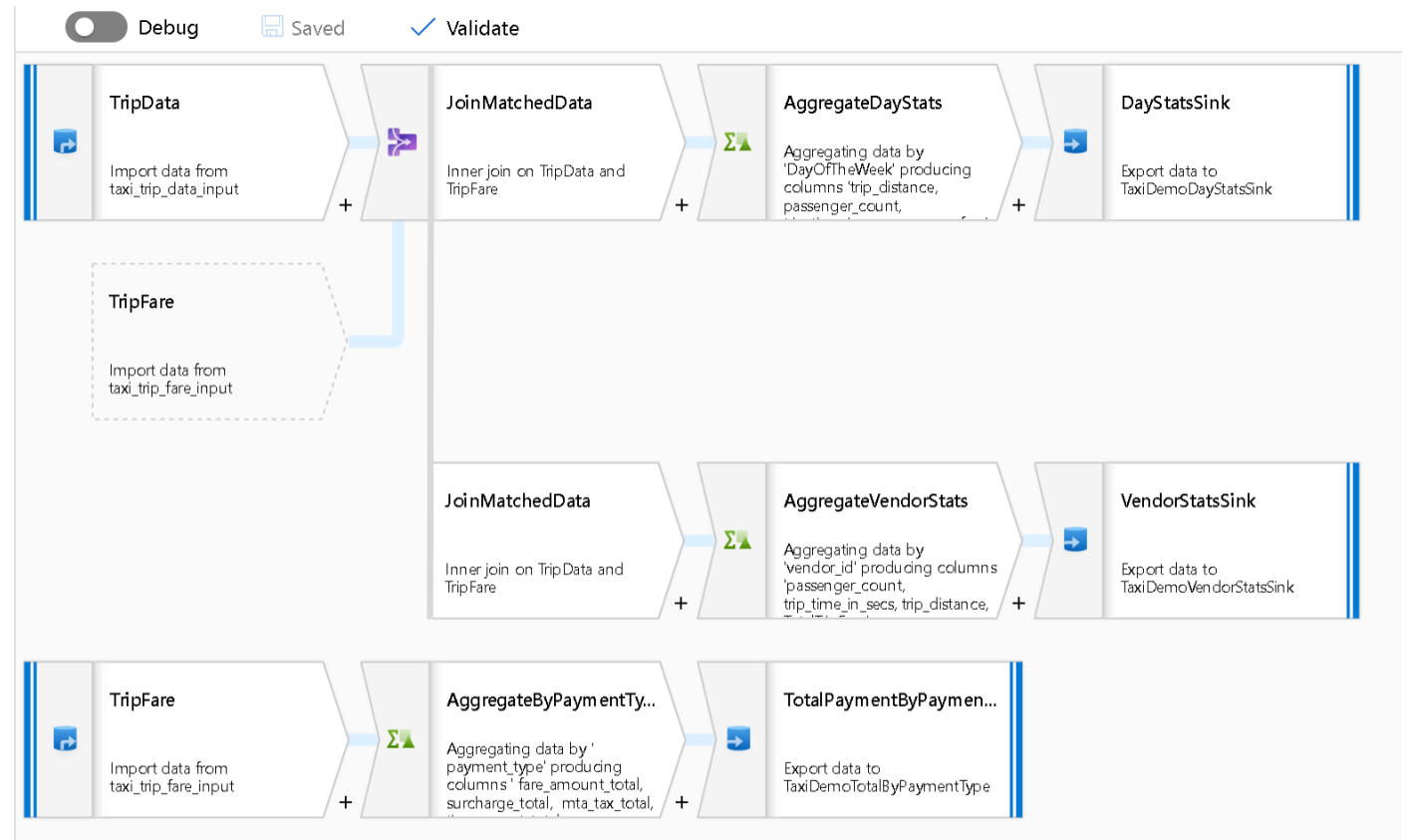
Scenario 4

Use a config file and further
optimize the pipeline design

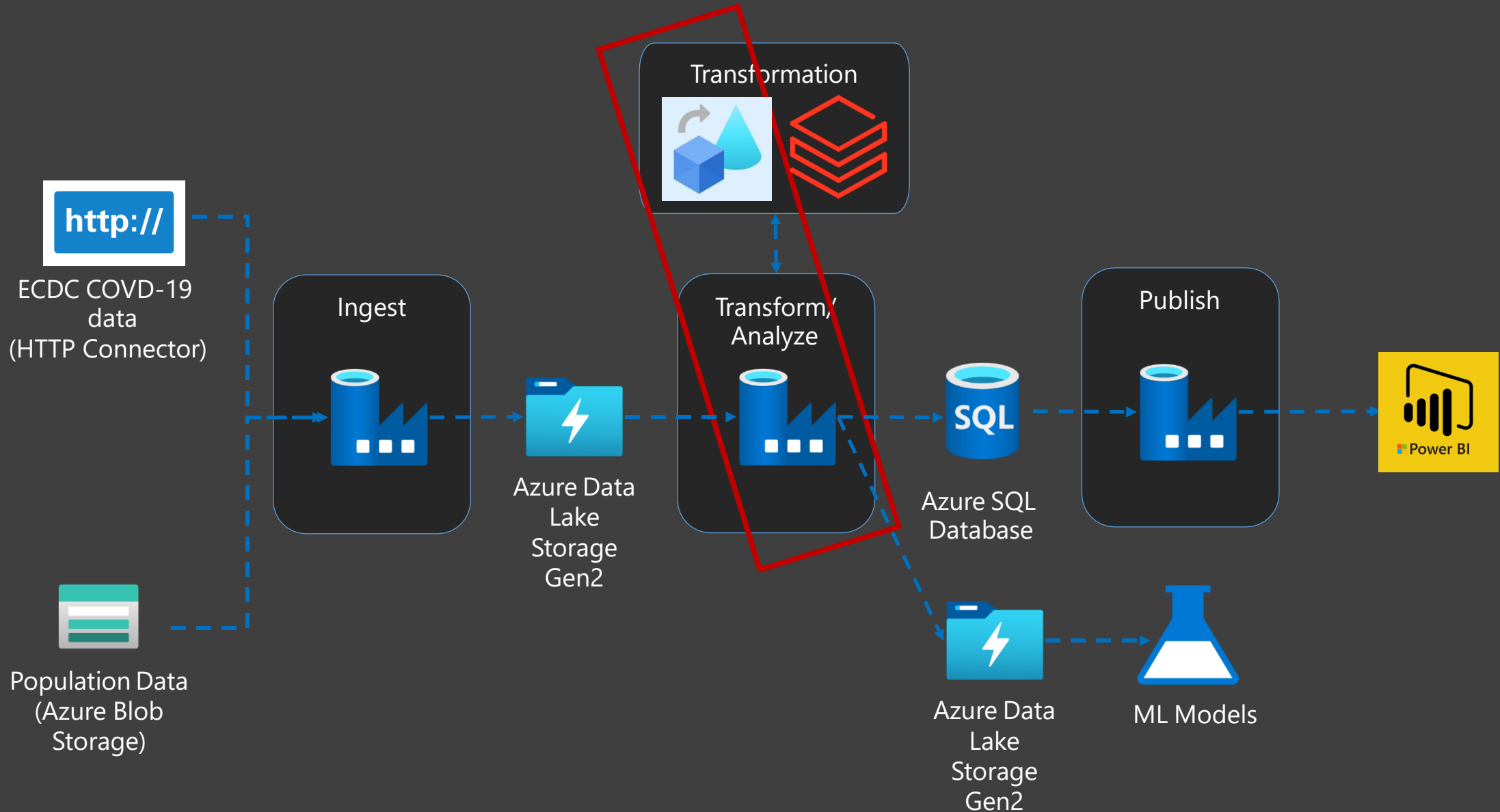


Data Flows

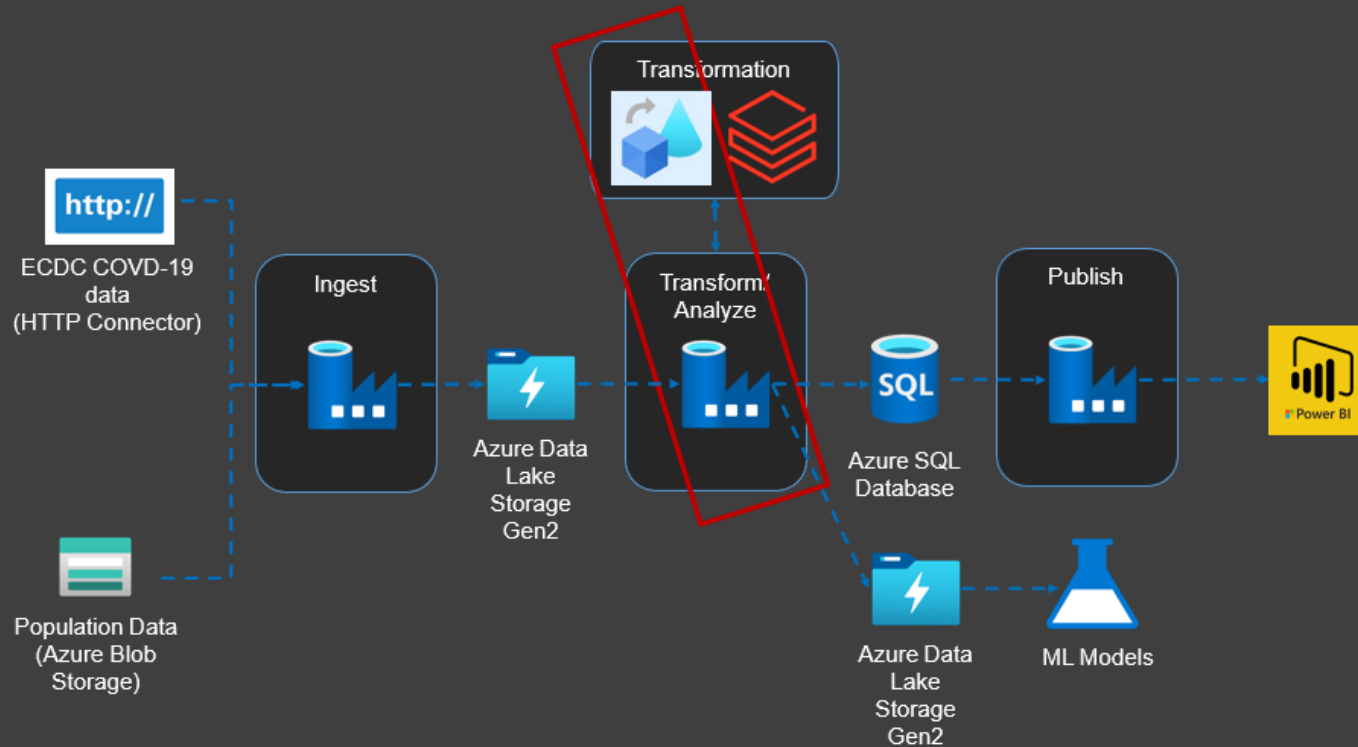
Cases & Deaths File



Data Flow – Cases & Deaths Data



Data Flow – Cases & Deaths Data



- Data Flow Overview
- Requirements
- Source Transformation
- Filter Transformation
- Select Transformation
- Pivot Transformation
- Lookup Transformation
- Sink Transformation
- Create Pipeline

Data Flows

Data Flows

Features

- Code free data transformations
- Executed on Data Factory managed Spark clusters
- Benefits from Data factory scheduling and monitoring capabilities.

Data Flows

Types



Data flow

Code free data transformation at scale



Wrangling Data Flow (Preview)

Code free data preparation at scale

Data Flows

Limitations

- Only available in some regions

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-overview#available-regions>

- Limited set of connectors available

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-source#supported-sources>

- Not suitable for very complex logic

Data Flows



Transform Cases & Deaths Data



Transform Cases & Deaths Data

Raw File from ECDC

Column Name
country
country_code
continent
population
indicator
daily_count
date
rate_14_day
source

Europe
Only

Transformed File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit
population
cases_count
deaths_count
reported_date
source

Transform Cases & Deaths Data

Raw File from ECDC

Column Name
country
country_code
continent ✓
population
indicator
daily_count
date
rate_14_day ✓
source

Europe
Only



Transformed File

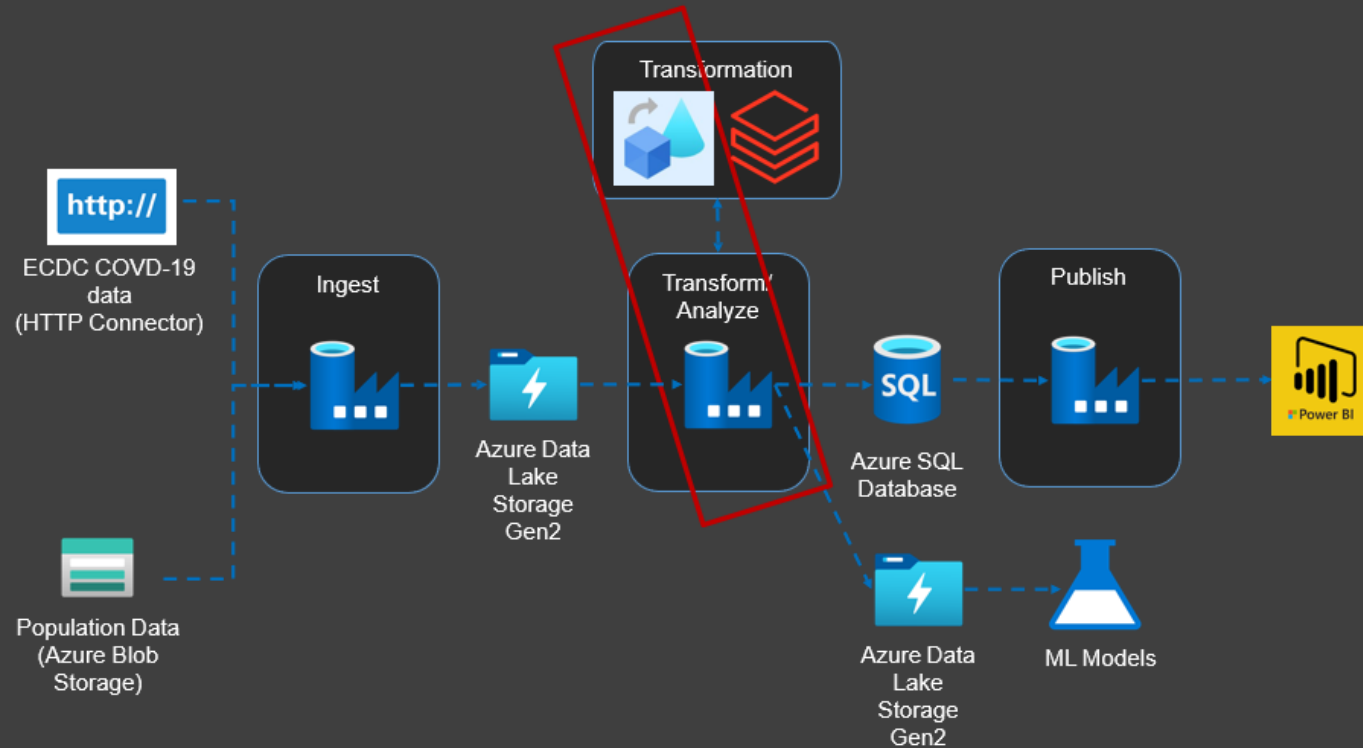
Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit
population
cases_count ✓
deaths_count ✓
reported_date(Rename) ✓
source



Hospital Admissions File

Data Flow 2

Data Flow – Hospital Admission Data



- Requirement
- Source Transformation
- Select Transformation
- Lookup Transformation
- Pivot Transformation
- Sink Transformation
- Conditional Split Transformation
- Derived Column Transformation
- Aggregate Transformation
- Sort Transformation
- Join Transformation
- Create Pipeline

Hospital Admissions Data



Hospital Admissions Data

Raw File from ECDC

Column Name
country
indicator
date
year_week
value
source
url

Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source

Hospital Admissions Data

Raw File from ECDC

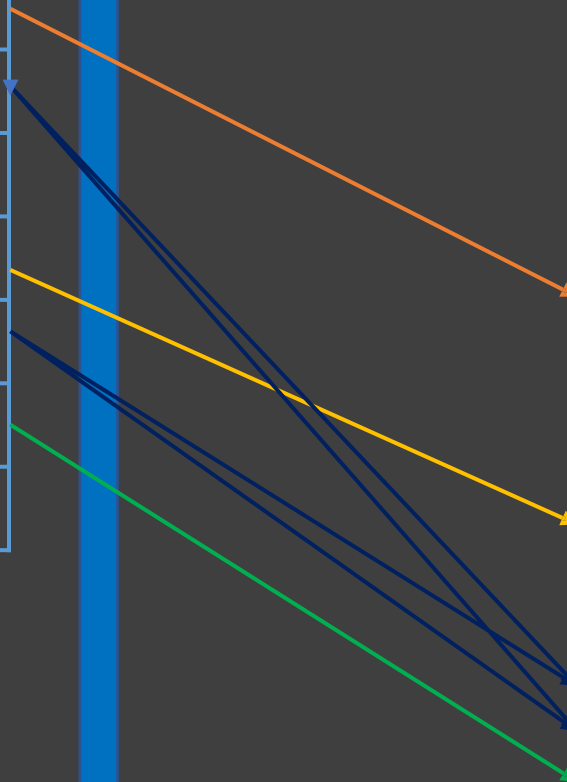
Column Name
country
indicator
date
year_week
value
source
url

Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source



Source Transformation

Assignment



Select Transformation

Assignment



Remove url



Rename date to reported_date



Rename year_week to reported_year_week

Lookup Transformation

Assignment



- Lookup country file
- Select only required fields (i.e. remove additional fields from lookup)

Pivot Transformation

Assignment



Hospital Admissions Data

Raw File from ECDC

Column Name
country
indicator
date
year_week
value
source
url

Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source

Hospital Admissions Data

Raw File from ECDC

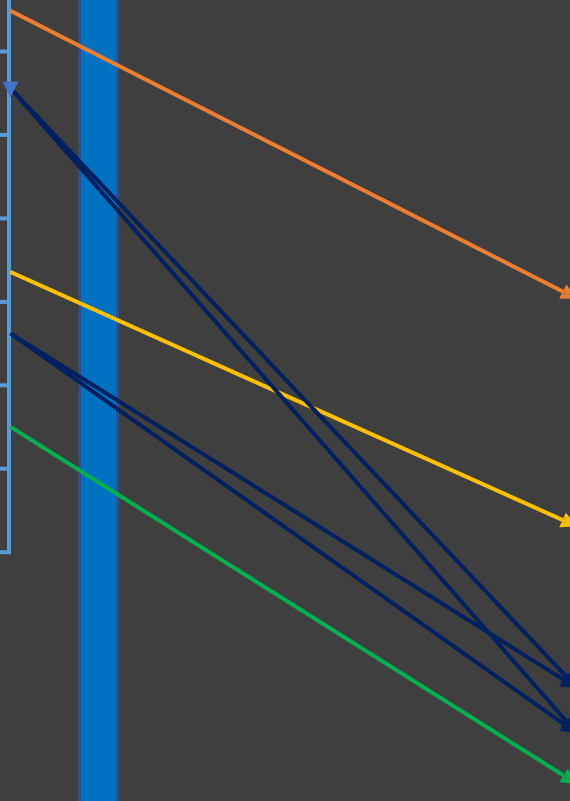
Column Name
country
indicator
date
year_week
value
source
url

Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source



Select & Sink Transformation

Assignment



Hospital Admissions Data

Raw File from ECDC

Column Name
country
indicator
date
year_week
value
source
url

Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source

Hospital Admissions Data

Raw File from ECDC

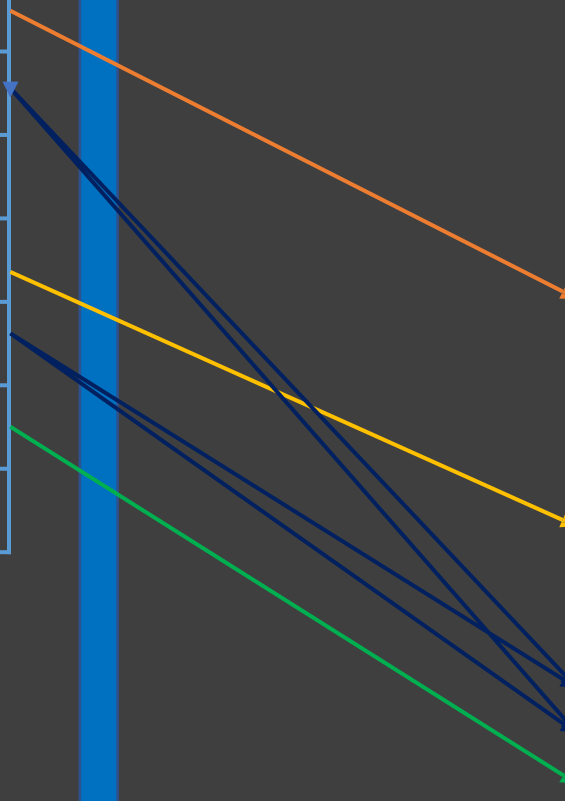
Column Name
country
indicator
date
year_week
value
source
url

Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source



Data Flow Execution

Assignment



Data Orchestration



Data Orchestration Requirements

- Pipeline executions are full automated
- Pipelines run at regular intervals or on an event occurring
- Activities only run once the upstream dependency has been satisfied
- Easier to monitor for execution progress and issues

Data Factory Capability

- Dependency between activities inside a pipeline
- Dependency between pipelines within a parent pipeline
- Dependency between triggers [Only tumbling window triggers]
- Custom-made Solution

Data Orchestration

Option 1 – Parent Pipeline



Data Orchestration

Option 2 – Trigger Dependency



Azure Data Factory - Monitoring

Azure Data Factory - Monitoring




- What to Monitor
- Data Factory
- Monitoring Creating
- Alerts
- Recovery From Failure
- Reporting on Metrics
- Azure Monitor
- Introduction Log Analytics

Monitoring

What do we want to monitor




Azure Data Factory Resource




Integration runtime



Trigger runs



Pipeline runs



Activity runs

Data Factory Monitor

- Ability to monitor status of pipeline/ triggers
- Can be used to re-run failed pipelines/ triggers
- Ability to send alerts from base level metrics
- Provides base level metrics and logs
- Pipeline runs are stored only for 45 days

Azure Monitor

- Ability to route the diagnostic data to other storage solutions
- Provides richer diagnostic data
- Ability to write complex queries and custom reporting
- Ability to report across multiple data factories

Data Factory Monitor



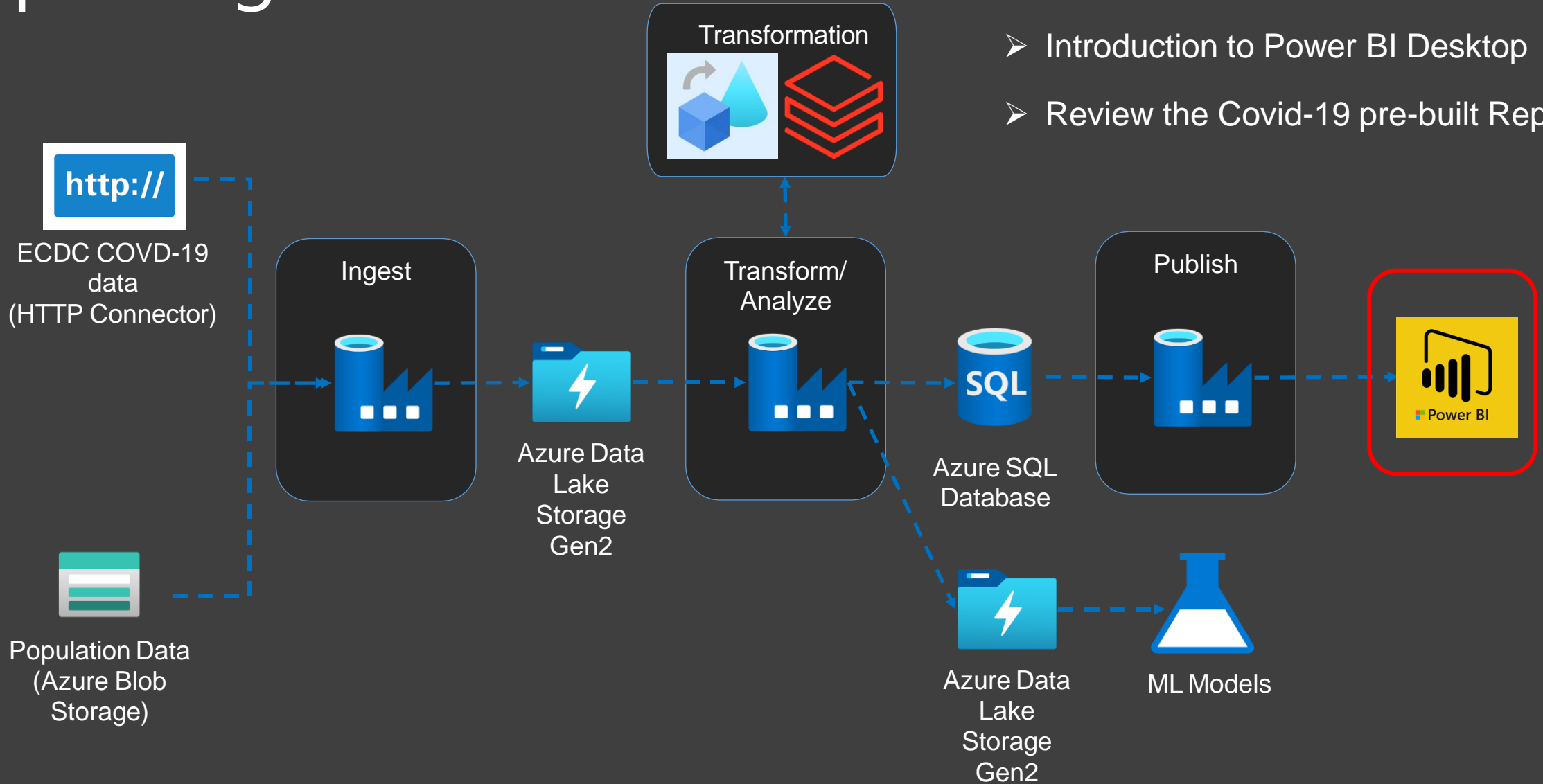
Azure Monitor



Reporting via Power BI

Reporting via Power BI

- Introduction to Power BI Desktop
- Review the Covid-19 pre-built Report



Power BI Desktop Overview

