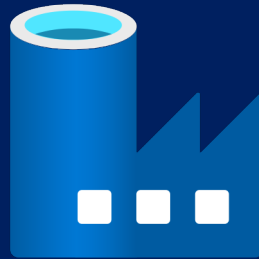


Covid-19 Prediction/ Reporting

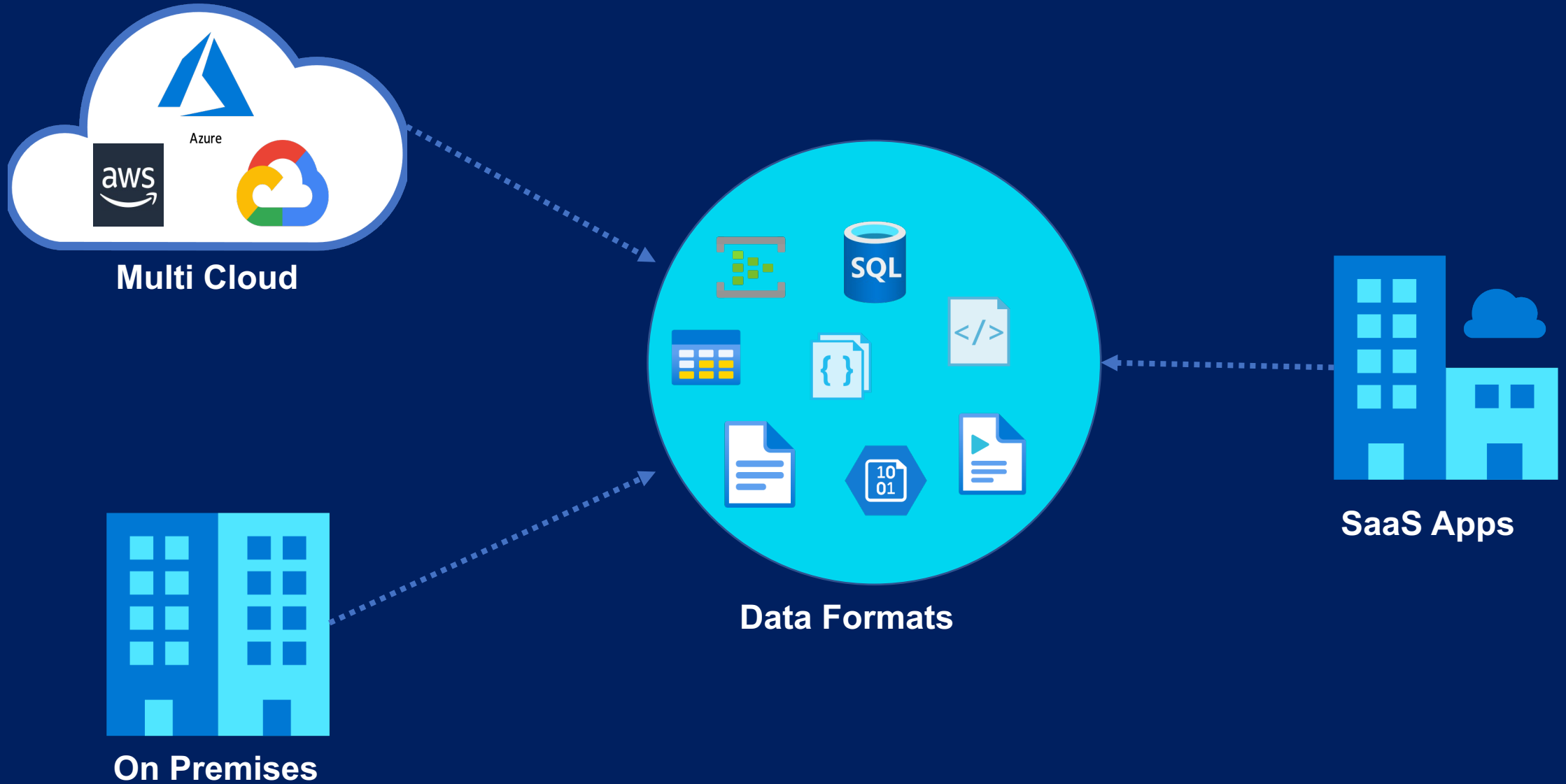
Azure Data Factory Overview

What is Azure Data Factory

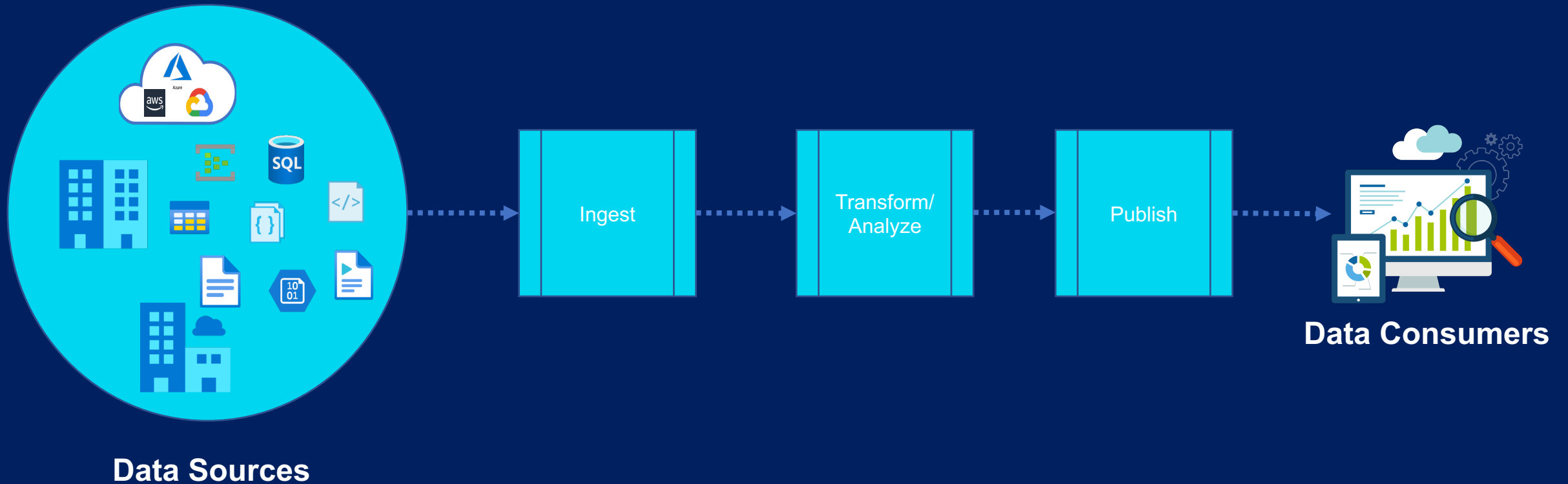


A fully managed, serverless data integration solution for ingesting, preparing and transforming all of your data at scale.

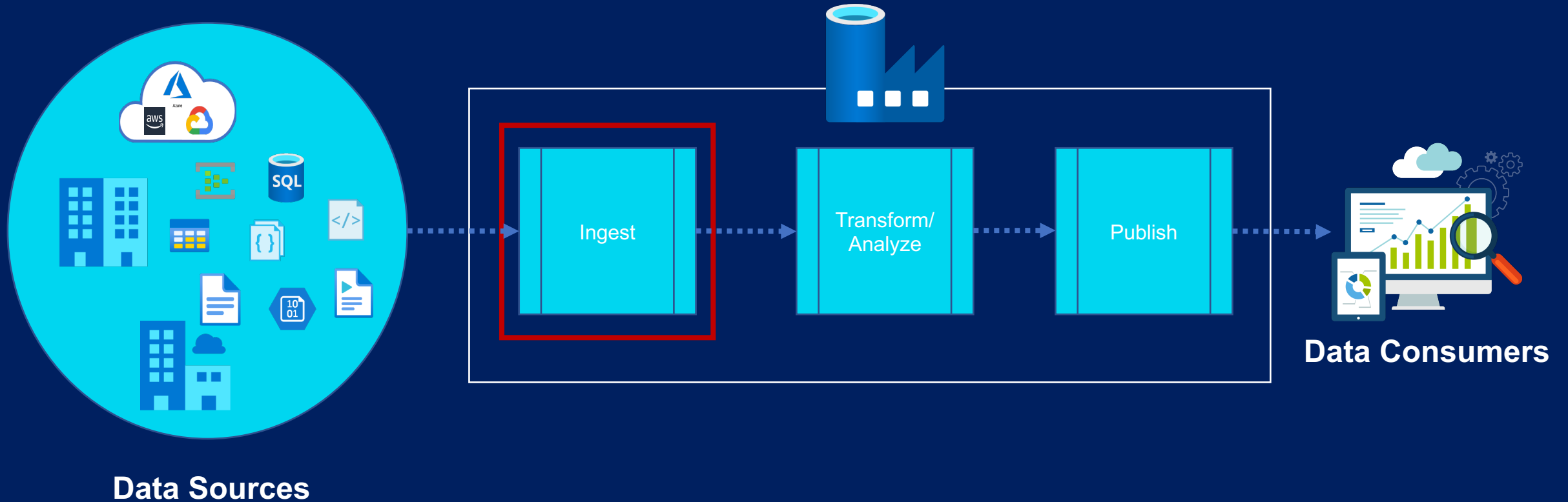
The Data Problem



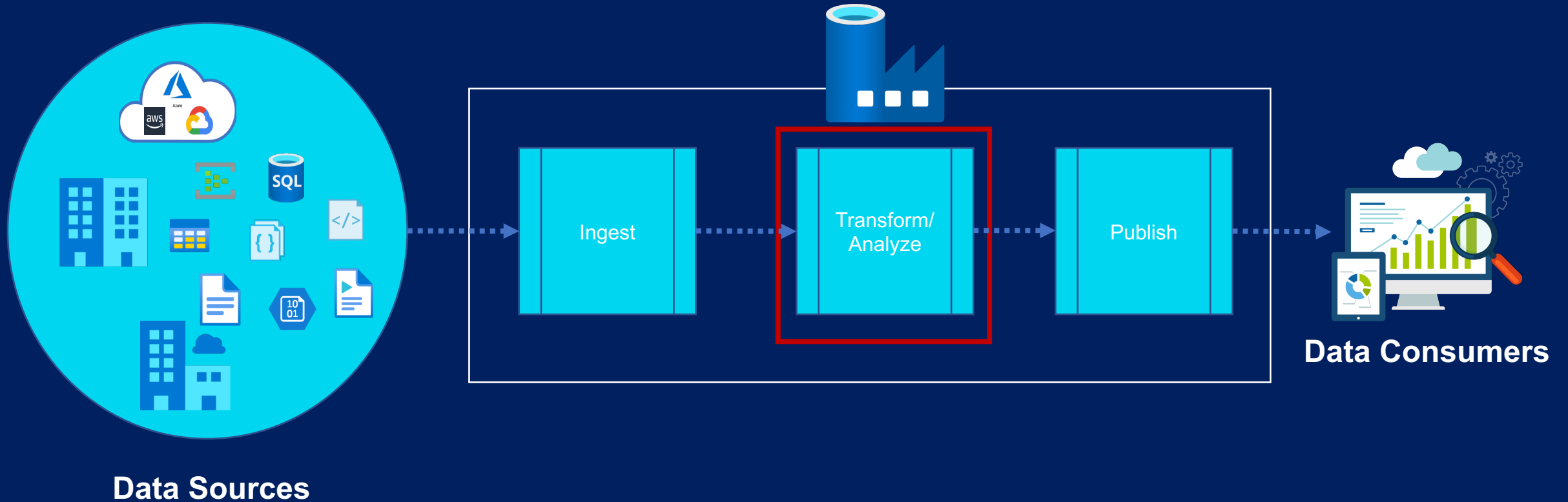
The Data Problem



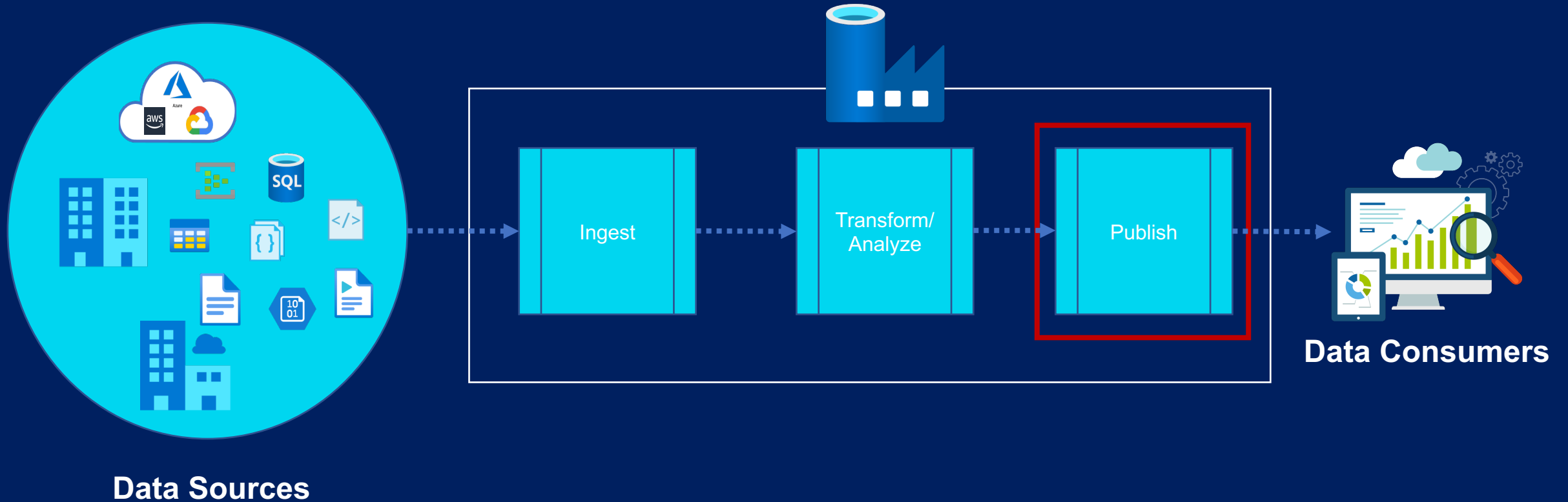
The Data Problem



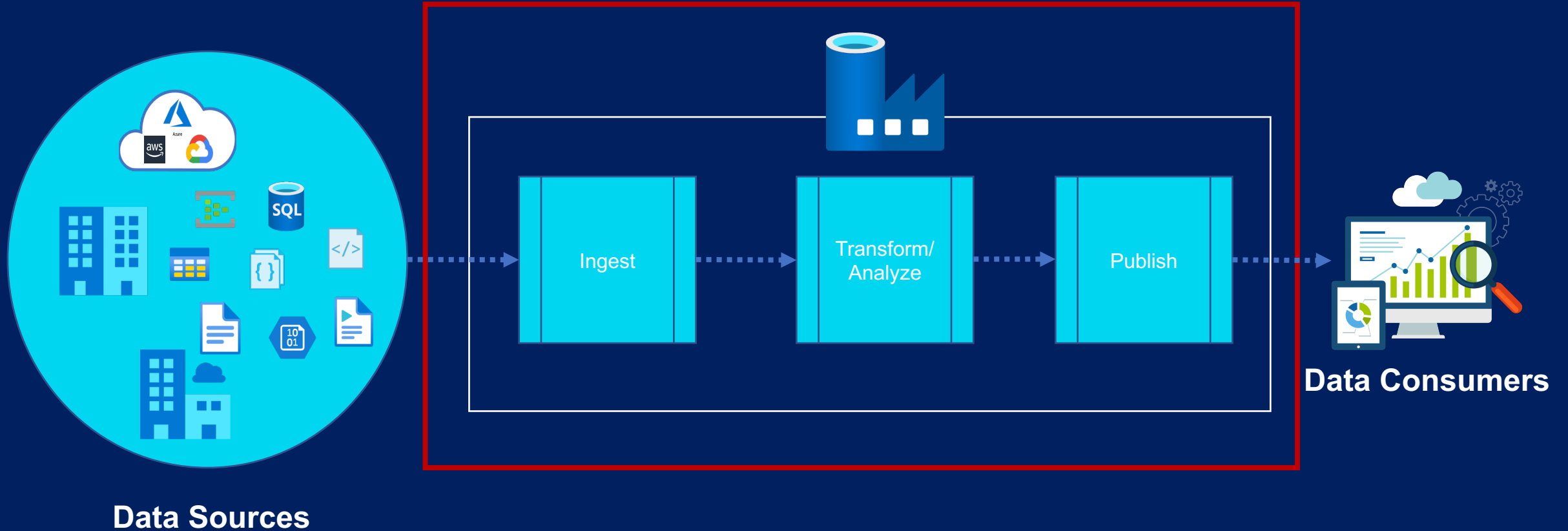
The Data Problem



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



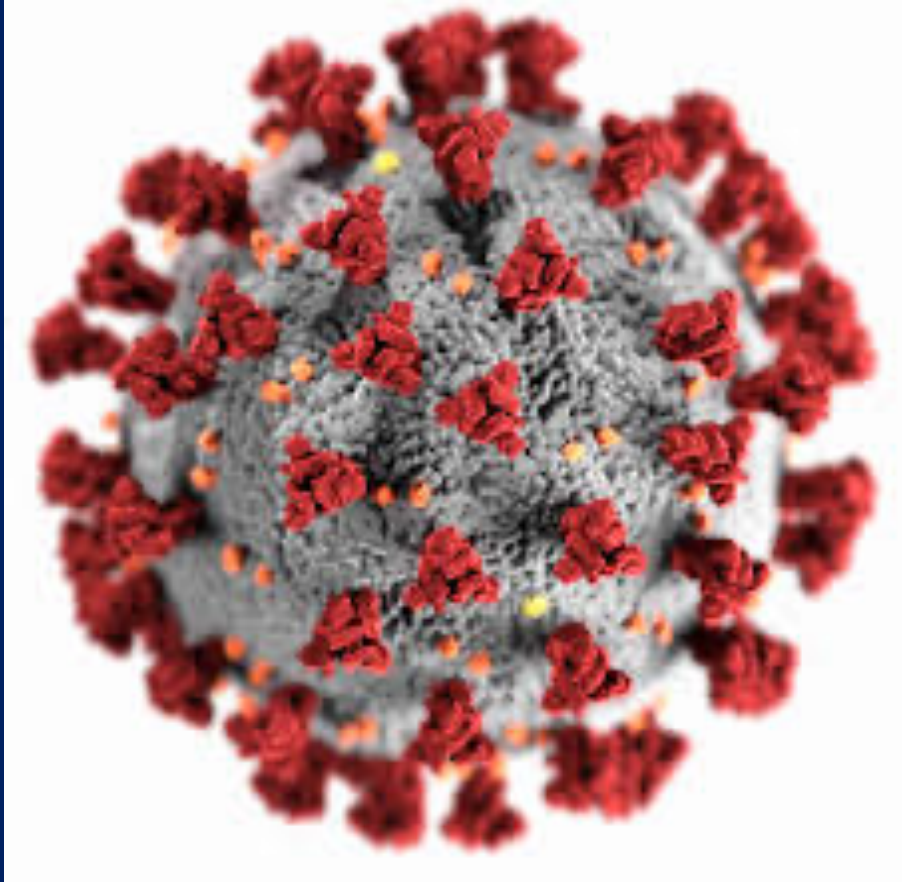
Data Migration Tool

Data Streaming Service

Suitable for Complex Data Transformations

Data Storage Service

Project Overview



Covid-19 Prediction/ Reporting

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
- Hospitalization/ ICU Cases
- Testing Numbers

Data Sources



ECDC Website

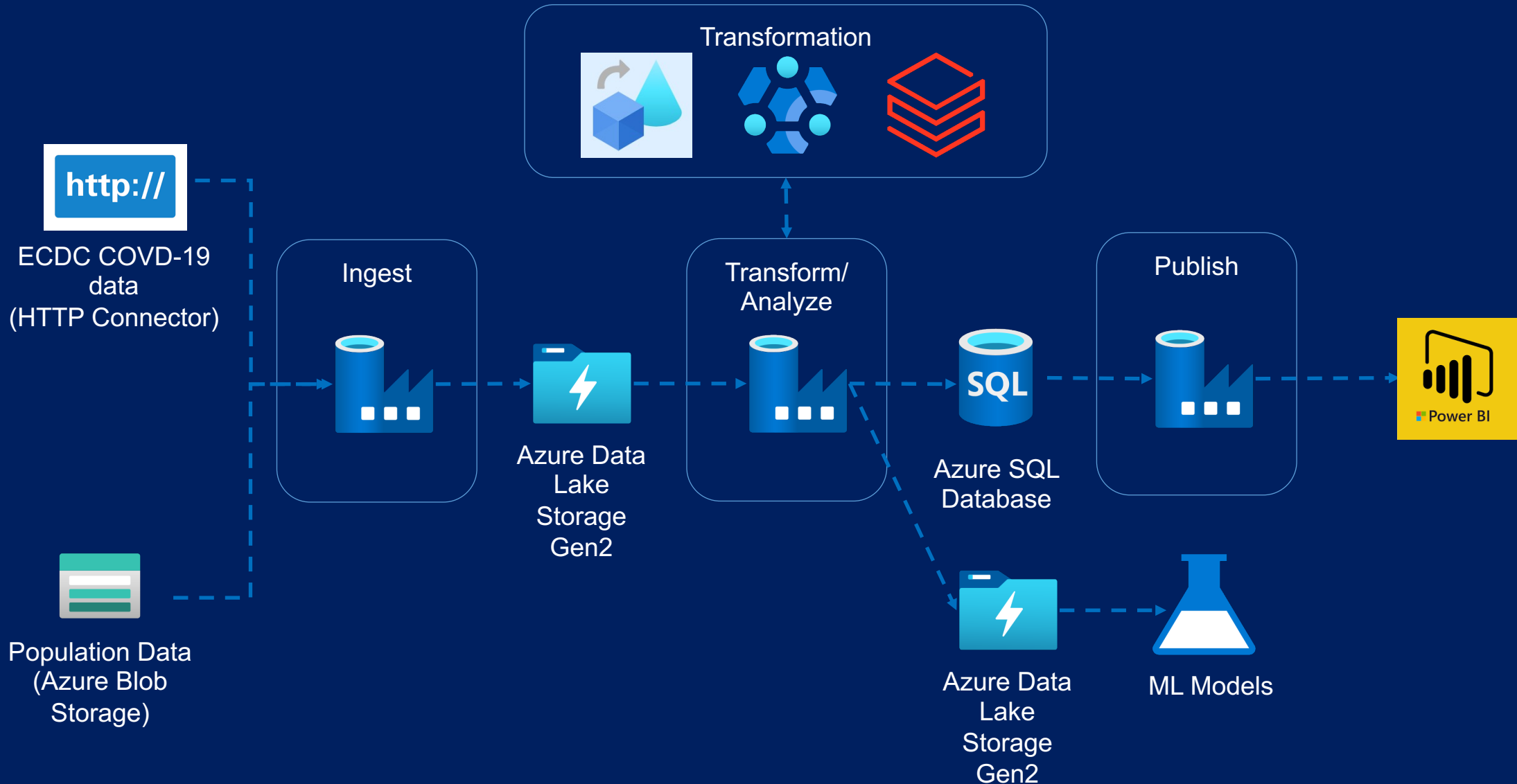
- Confirmed cases
- Mortality
- Hospitalization/ ICU Cases
- Testing Numbers

Eurostat Website

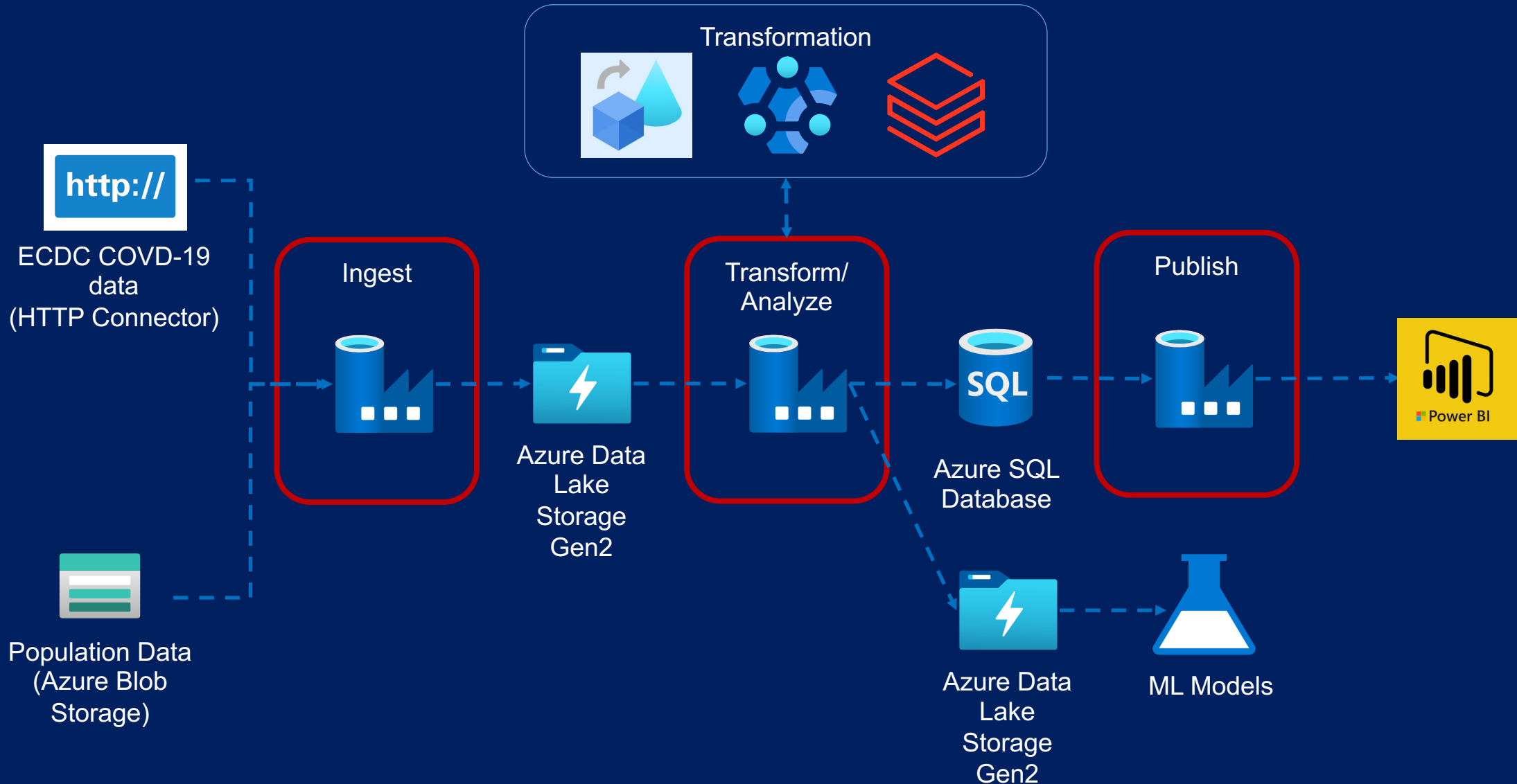
- Population by age

Solution Architecture

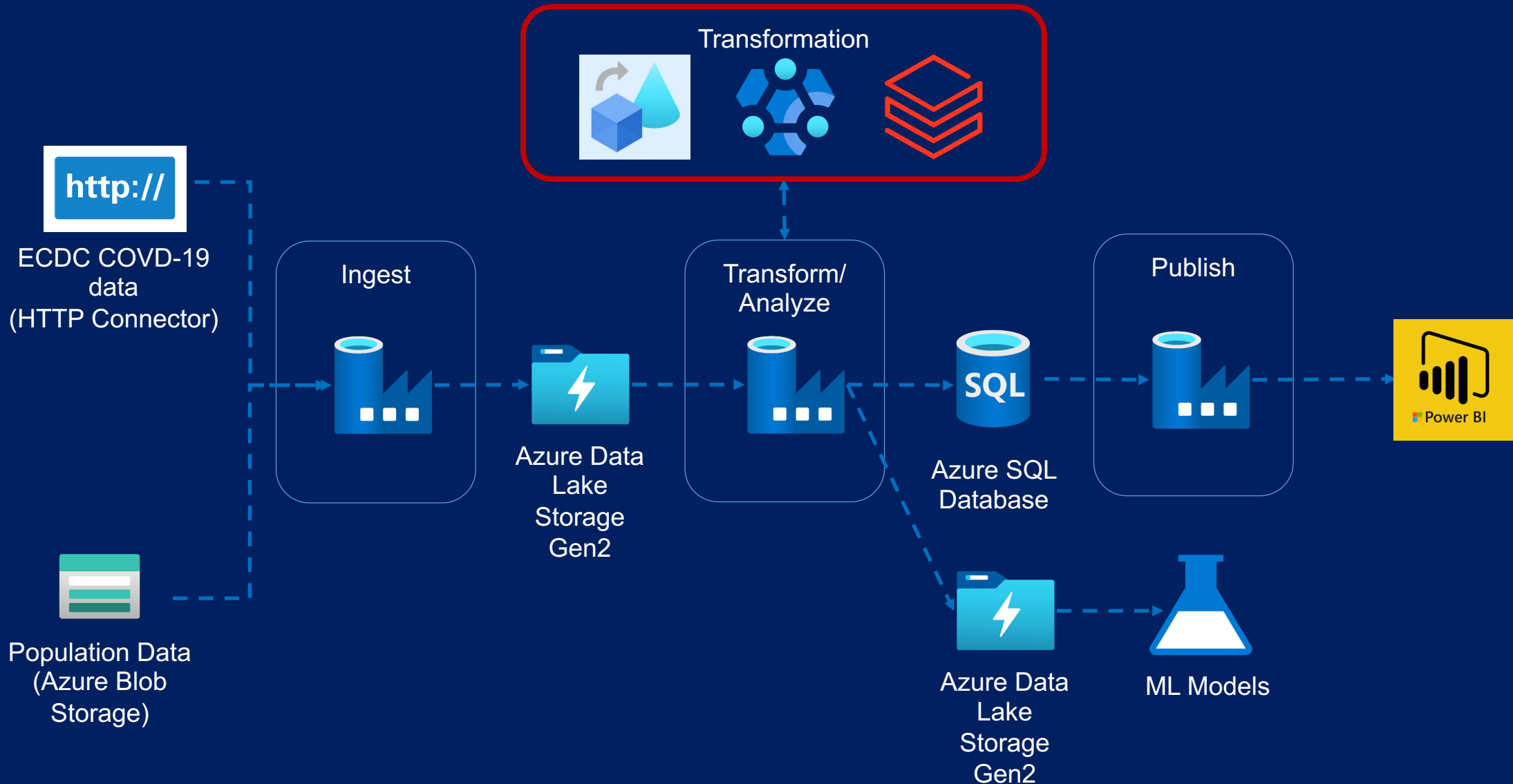
Solution Architecture



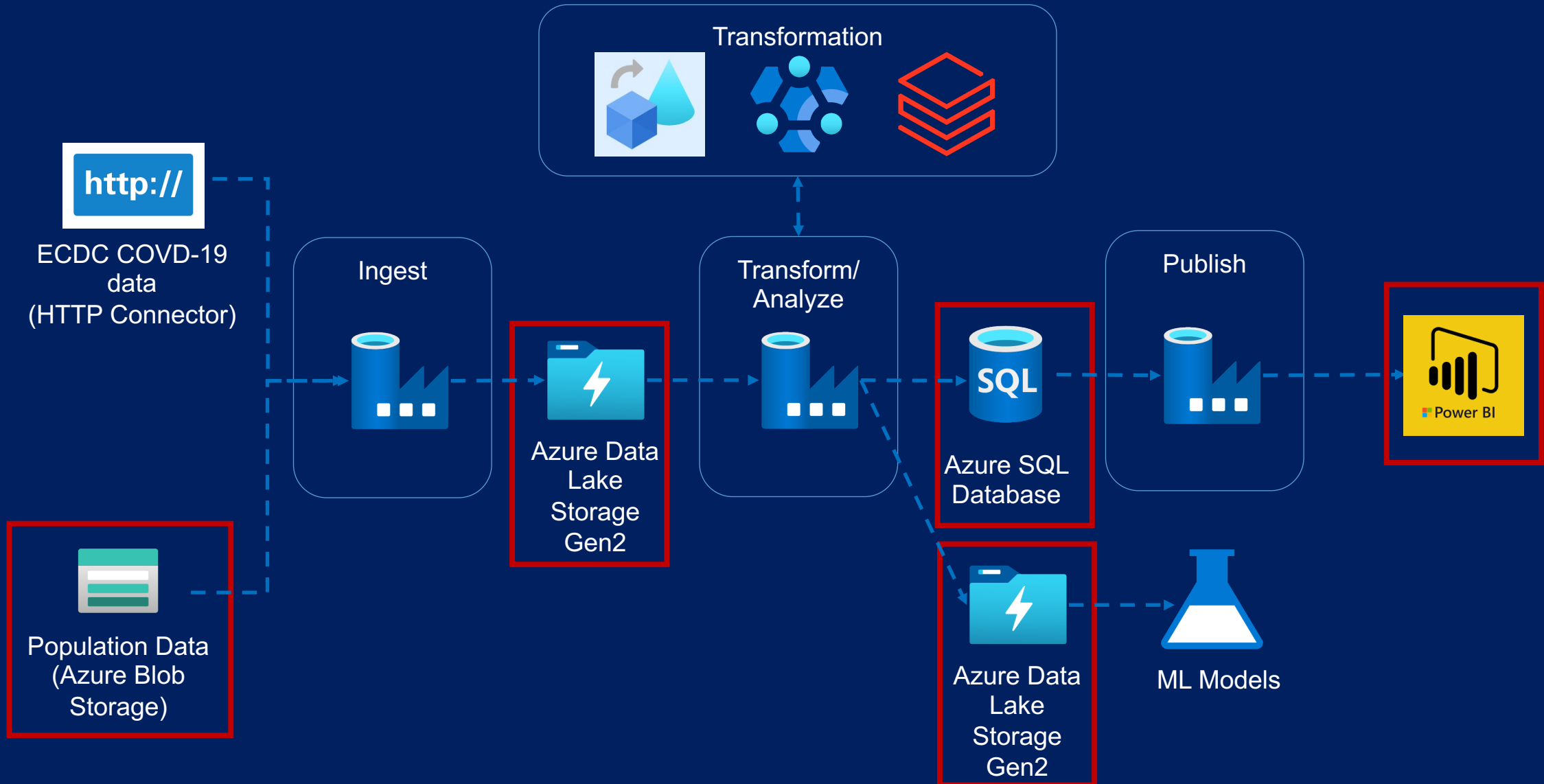
Solution Architecture



Solution Architecture



Solution Architecture



Storage Solutions

Key Factors to Consider

Structure of the data

Structured

Semi-Structured

Unstructured

Operational needs

How often is the data accessed?

How quickly do we need to serve?

Need to run simple queries?

Need to run heavy analytical workload?

Accessed from multiple regions?

Azure Databases



Azure SQL Database



Azure Database for MySQL



Azure Database for PostgreSQL



Azure Database for MariaDB



VM Images with Oracle, SQL Server etc.

Azure Storage Account



Blob Storage



File Storage



Disk Storage

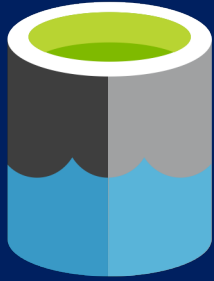


Table Storage



Queue Storage

Azure Data Lake



Azure Data Lake Storage Gen2

Enhance Performance

Better Security

Enhance Management

Azure Cosmos DB



Globally distributed

Multi Model

High Throughput

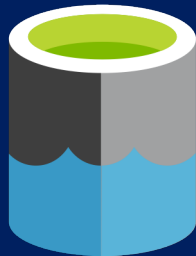
Storage solutions used in this course



Azure SQL Database



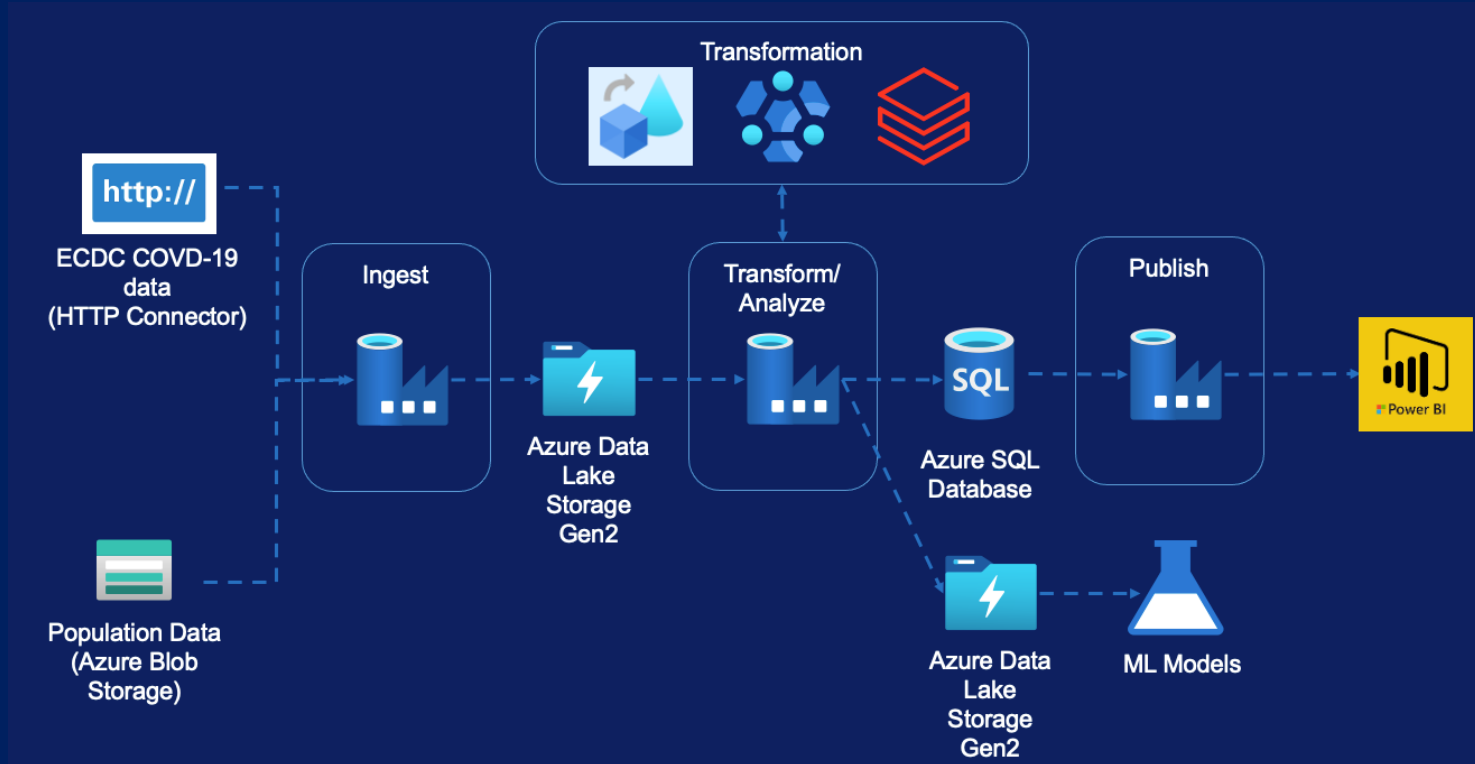
Azure Blob Storage



Azure Data Lake Storage Gen2

Environment set-up

Environment set-up



- Azure Subscription
- Data Factory
- Blob Storage Account
- Data Lake Storage Gen2
- Azure SQL Database
- Azure Databricks Cluster
- HD Insight Cluster

Creating Azure Free Account



Creating Azure Data Factory



Creating Azure Storage Account



Creating Azure Data Lake Gen2



Creating Azure SQL Database

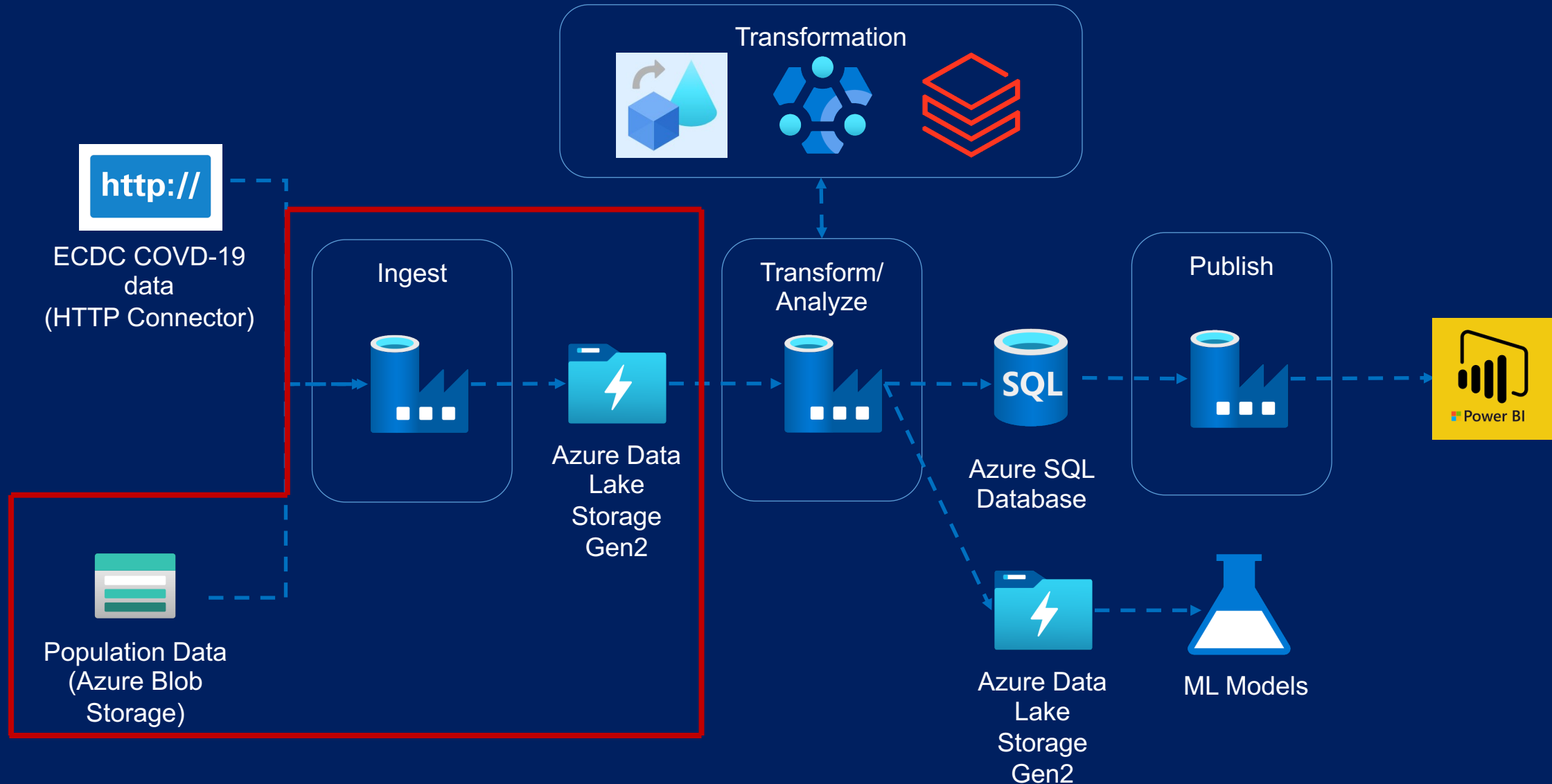


Data Ingestion

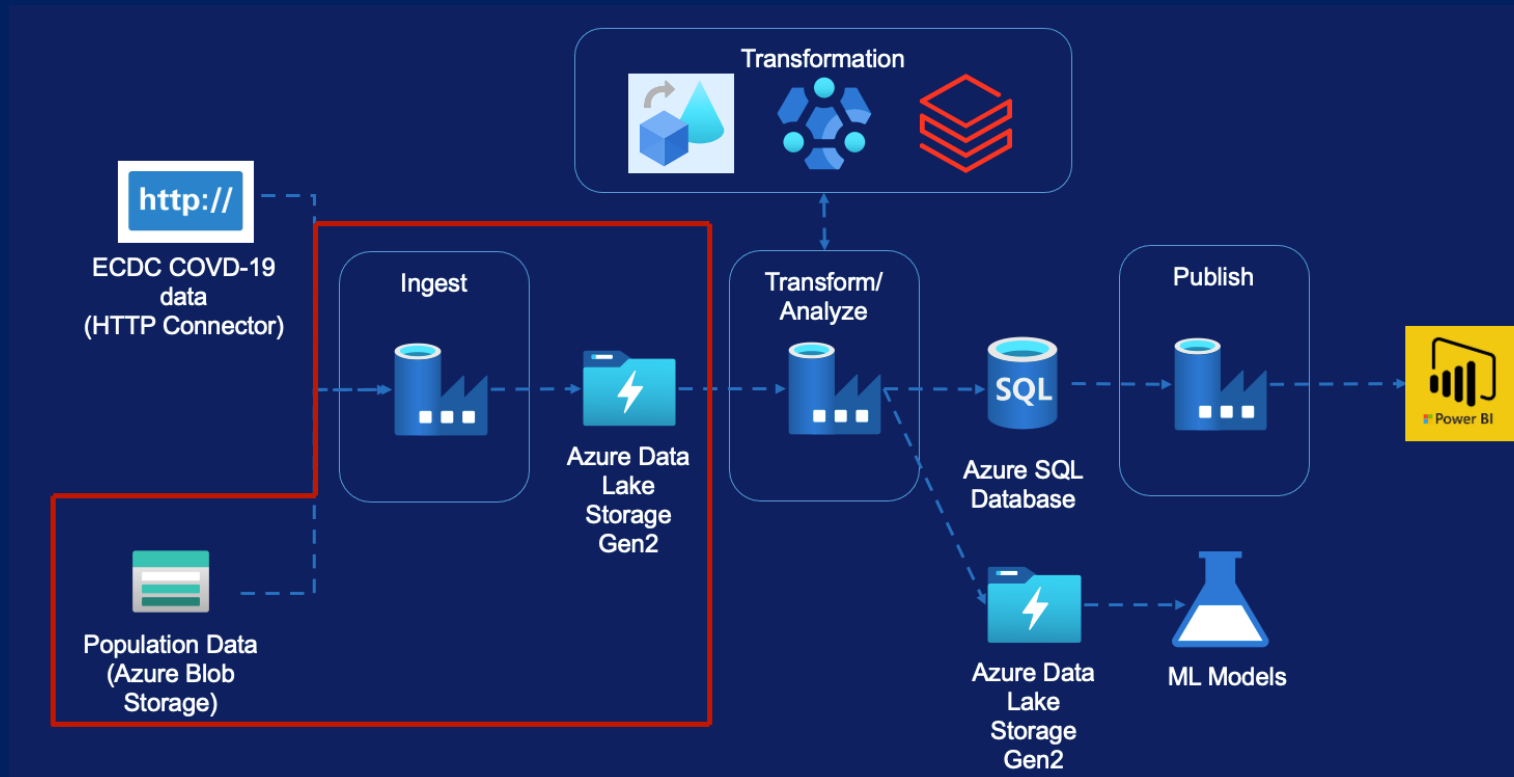
Data Ingestion - Module Overview

(Population by Age)

Data Ingestion – Population Data



Data Ingestion – Population Data



Copy Activity

Linked Services

Datasets

Pipeline

Validation Activity

If Condition Activity

Web Activity

Get Metadata Activity

Delete Activity

Trigger

Copy Activity

Azure Blob Storage → Azure Data Lake

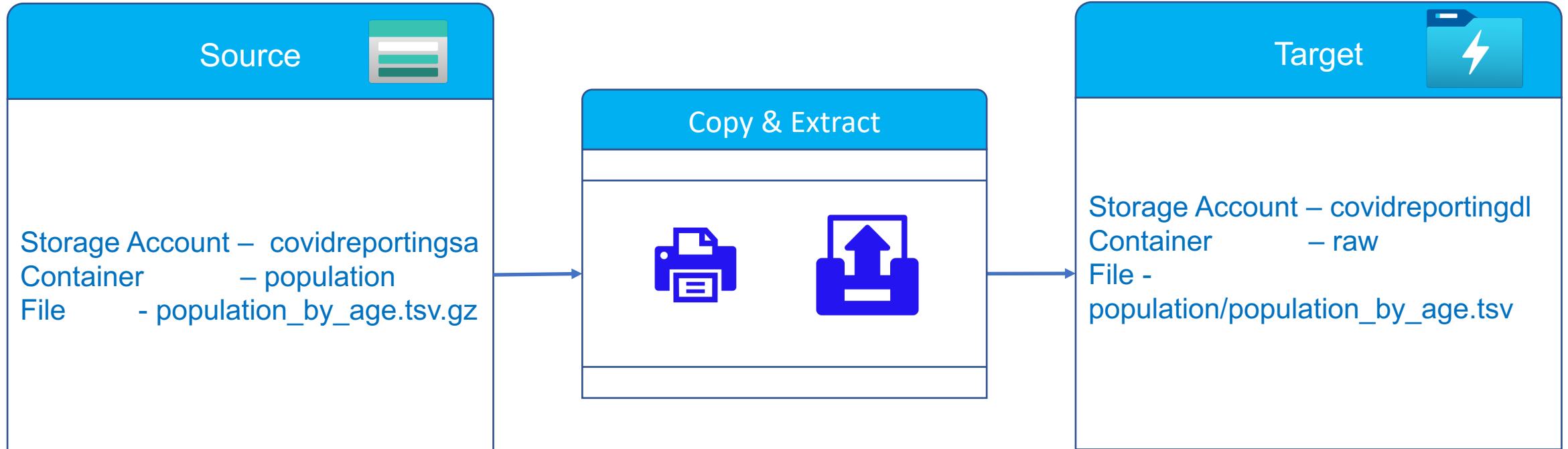
Copy Activity



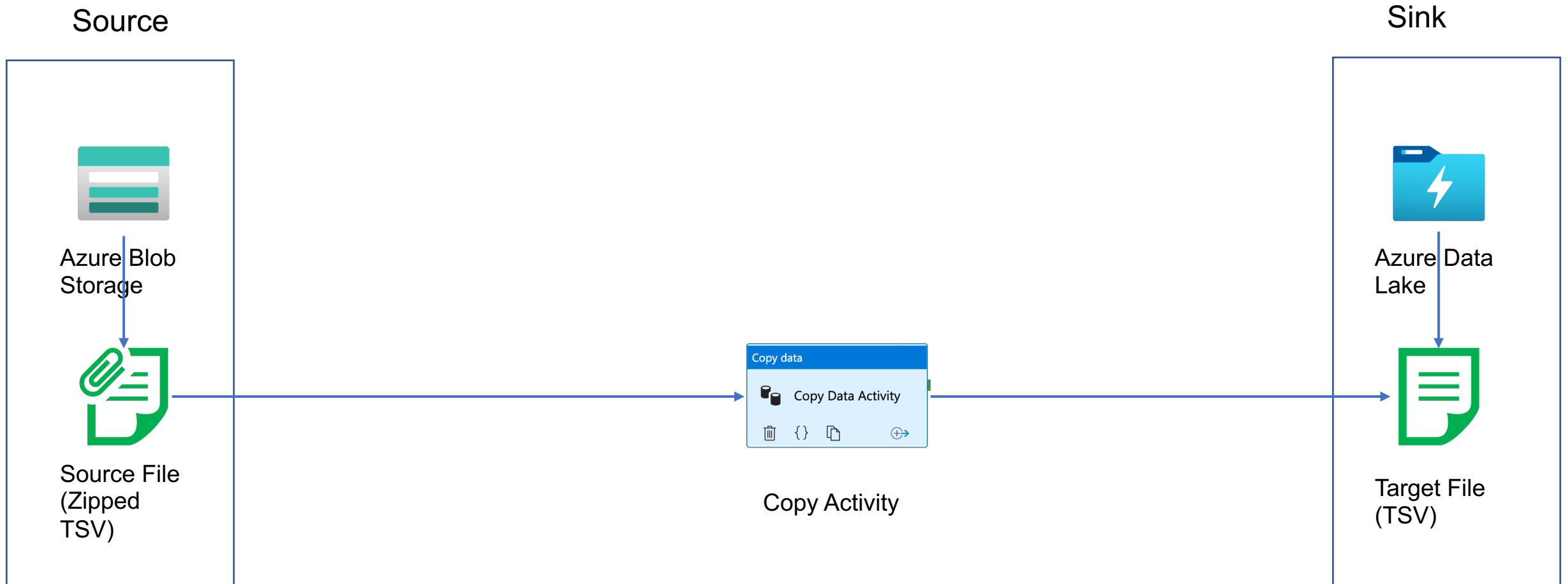
Ingest "population by age" for all EU Countries into the Data Lake to support the machine learning models to predict increase in Covid-19 mortality rates



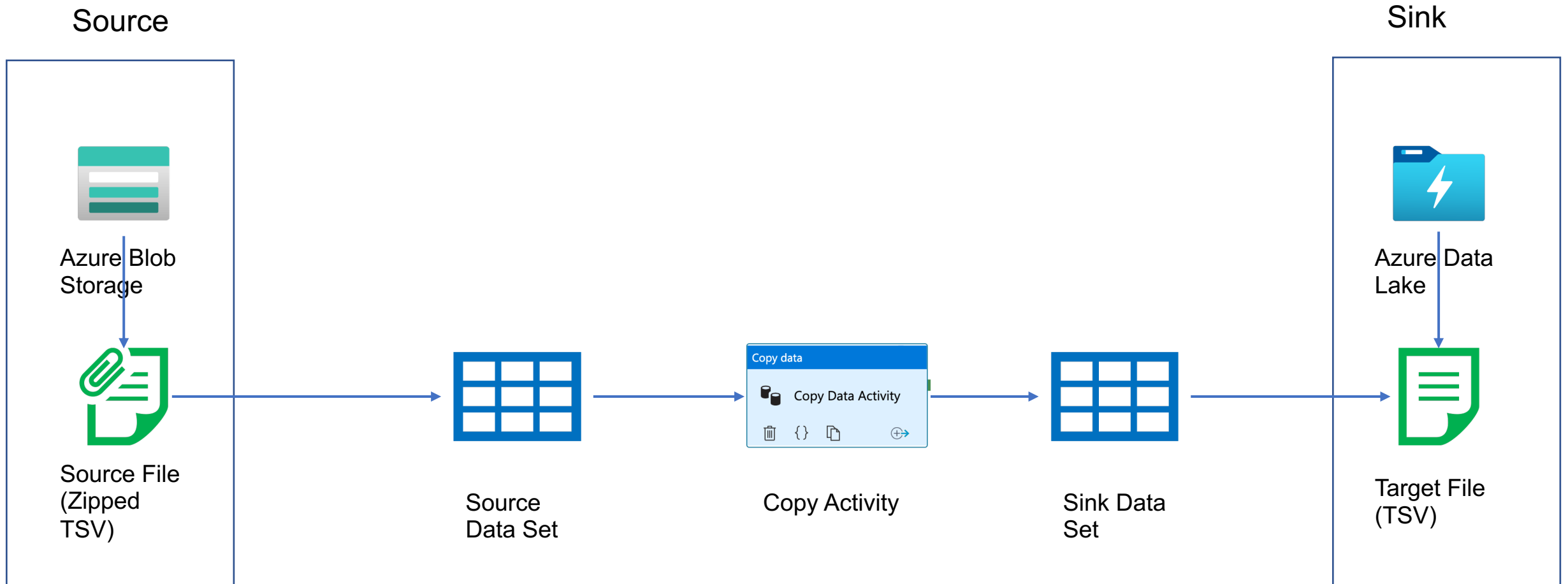
Copy Activity



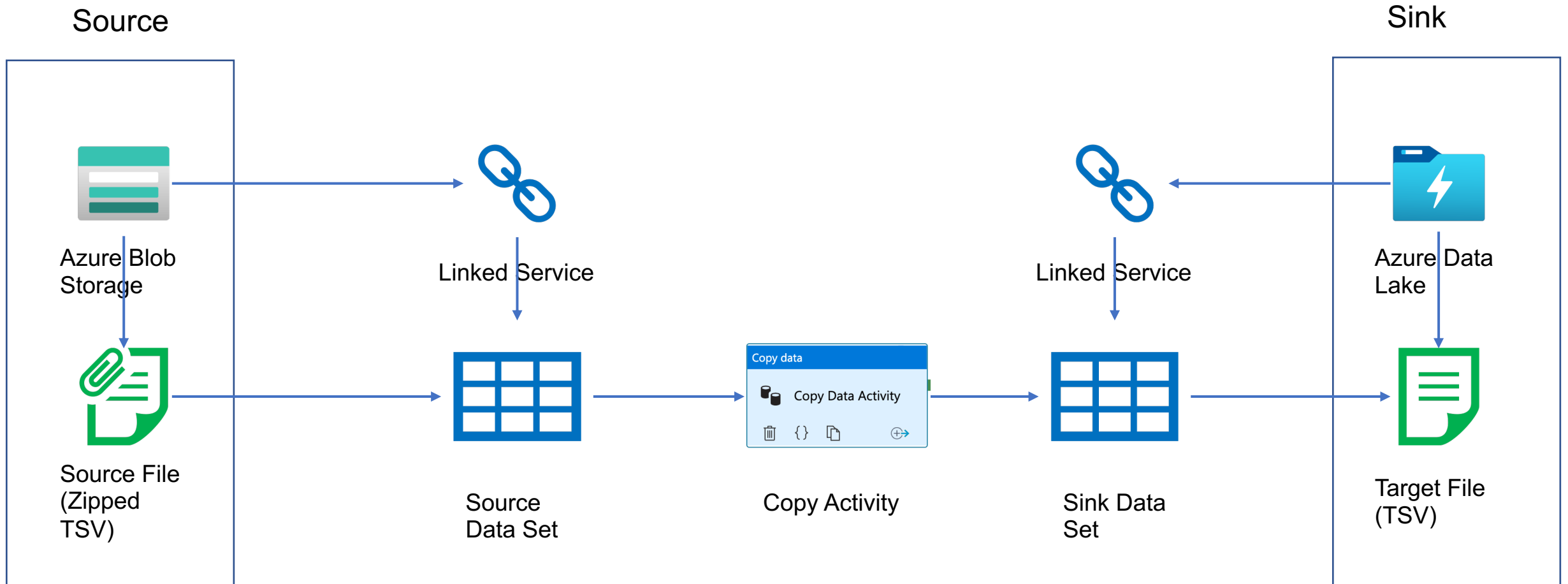
Copy Activity



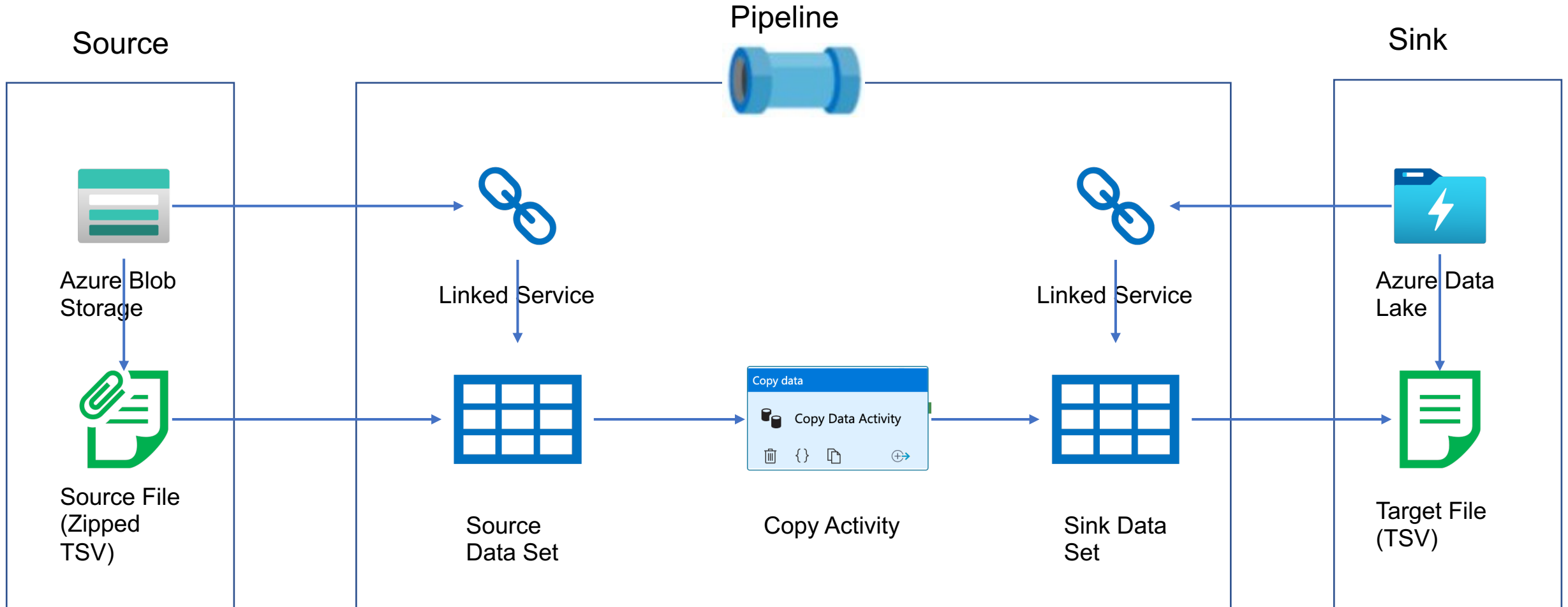
Copy Activity



Copy Activity



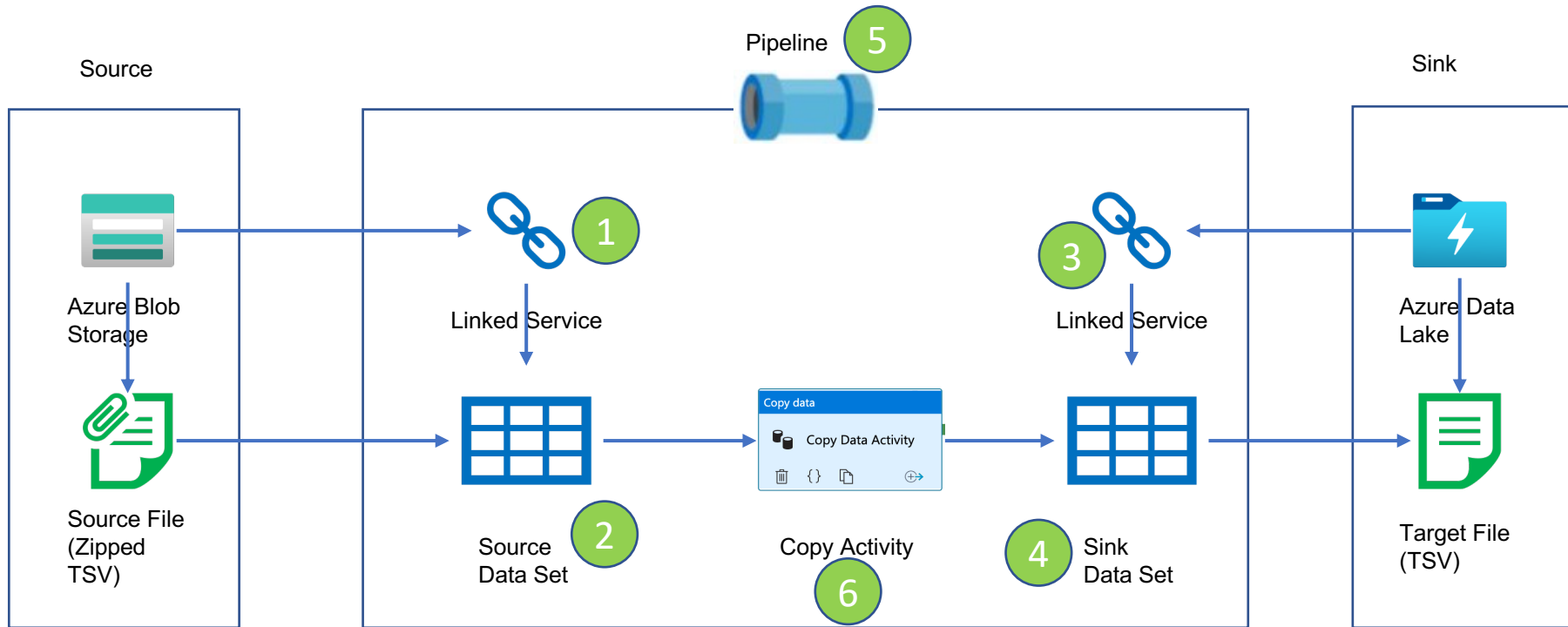
Copy Activity



Copy Activity From Azure Blob Storage



Copy Activity

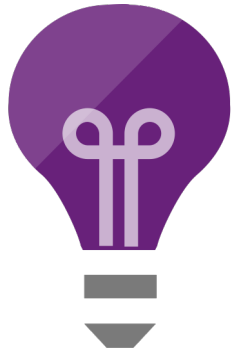


Storage Account: covidreportingsa
Container: population
File: population_by_age.tsv.gz

Storage Account: covidreportingdl
Container: raw
File: population/population_by_age.tsv

- 1 ls_ablob_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





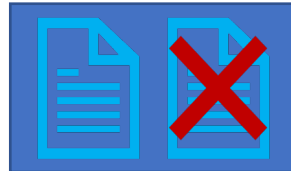
Triggers



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



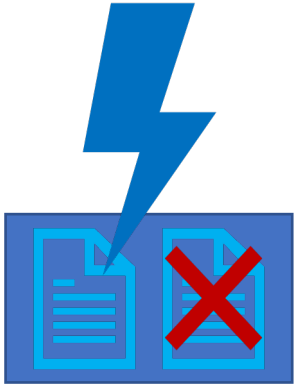
Windows are fixed sized, non-overlapping



Can be scheduled for the past windows/
slices



Trigger to Pipeline is one to one



Event Trigger



Runs in response to events



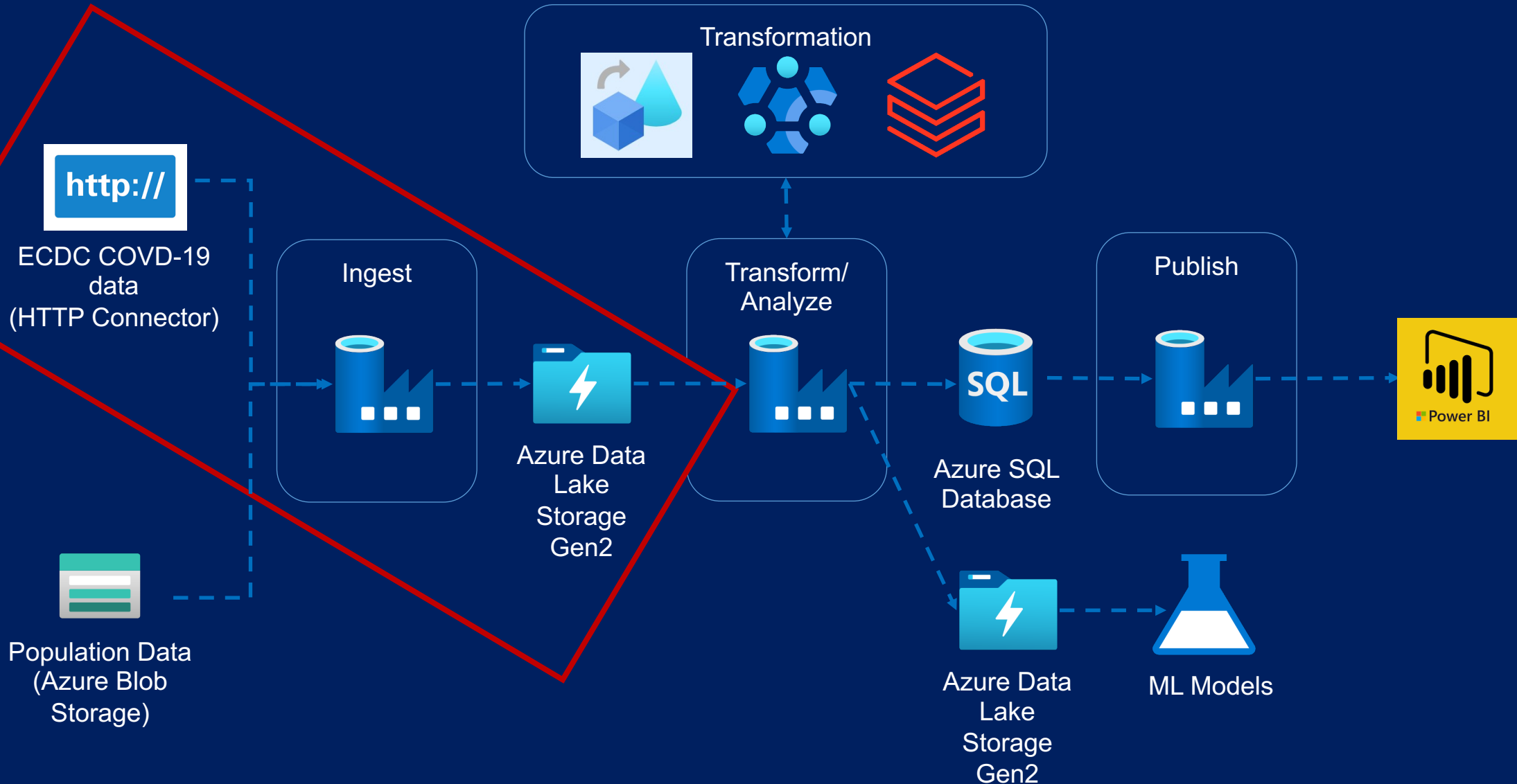
Events can be creation or deletion of Blobs/
Files



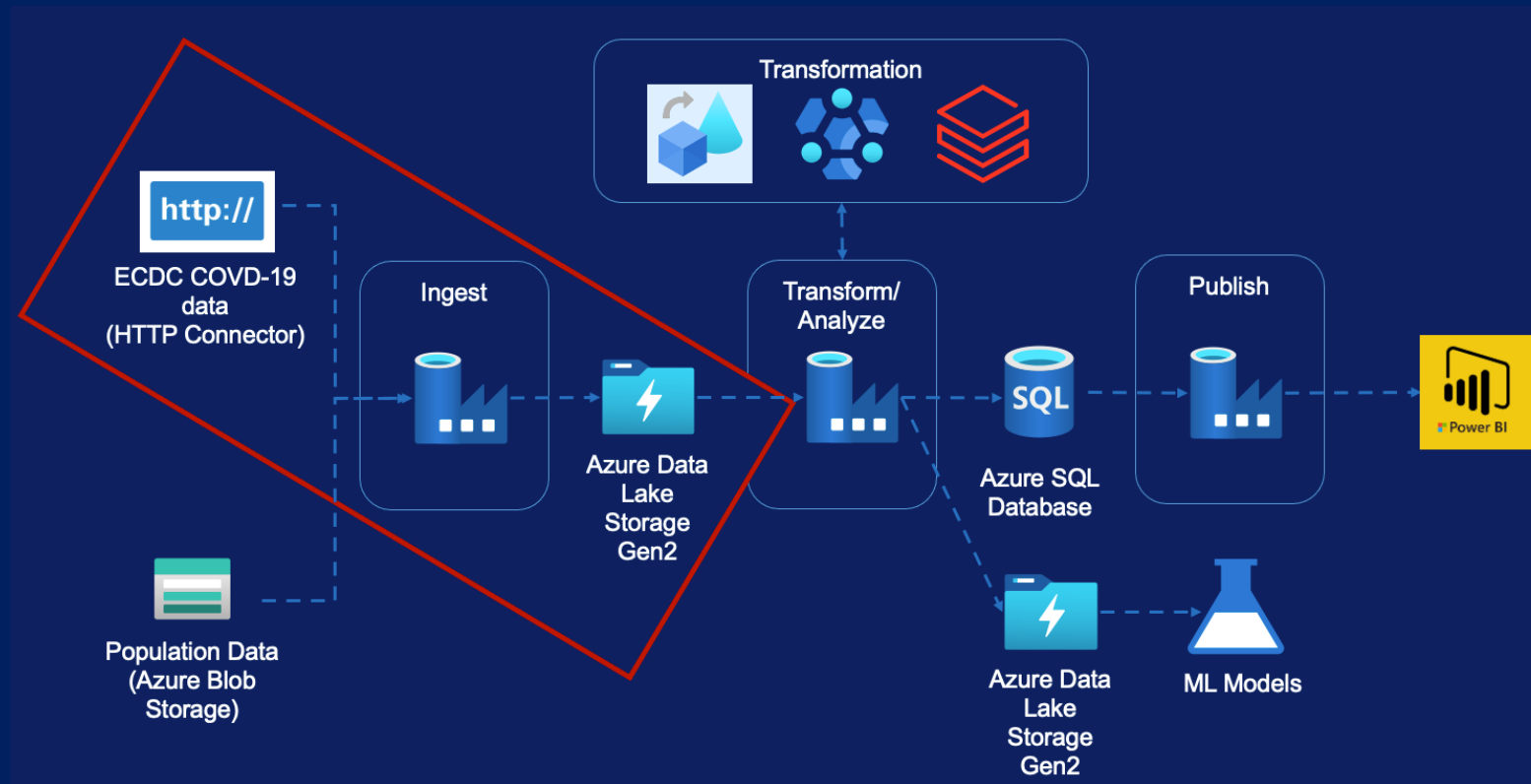
Trigger to Pipeline is Many to Many

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

Recent Changes to ECDC Data

Recent Changes to ECDC Data

Download COVID-19 datasets



ECDC switched to a weekly reporting schedule for the COVID-19 situation worldwide and in the EU/EEA and the UK on 17 December 2020. Hence, all daily updates have been discontinued from 14 December. ECDC will publish updates on the number of cases and deaths reported worldwide and aggregated by week every Thursday. The weekly data will be available as downloadable files in the following formats: XLSX, CSV, JSON and XML. As an exception, the weekly updates for the end-of-year festive season will be published on 23 December and 30 December 2020.

With the switch from daily to weekly reporting, ECDC will shift its Epidemic Intelligence (EI) resources from case counting to signal/event detection and resume its regular EI activities, which will include COVID-19 signal and event detection and analysis but also other potential threats.

- Granularity of the data changed from daily to weekly

- File structure is also different as a result

- Use GIT Repo - <https://github.com/cloudboxacademy/covid19>

Data Ingestion

HTTP

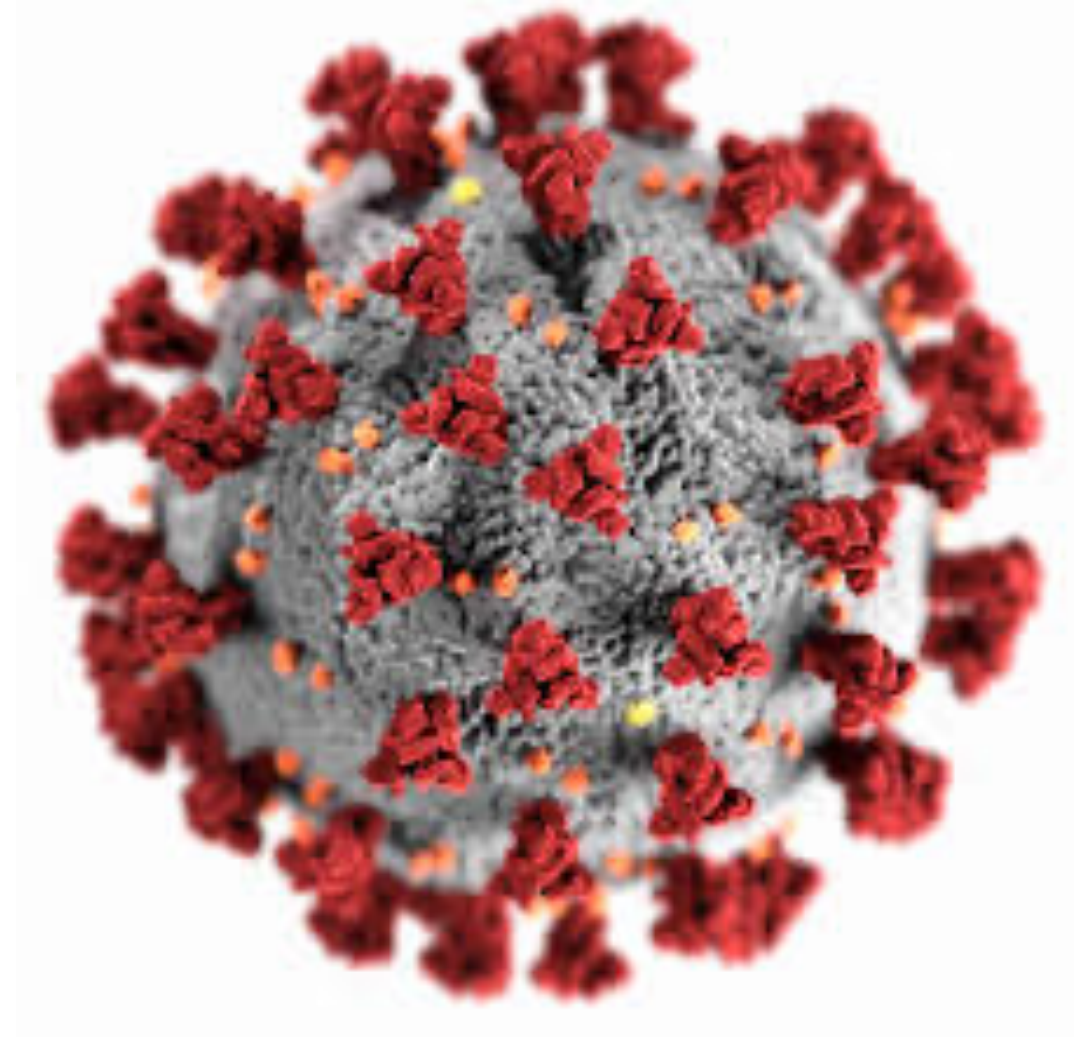


Azure Data Lake

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>

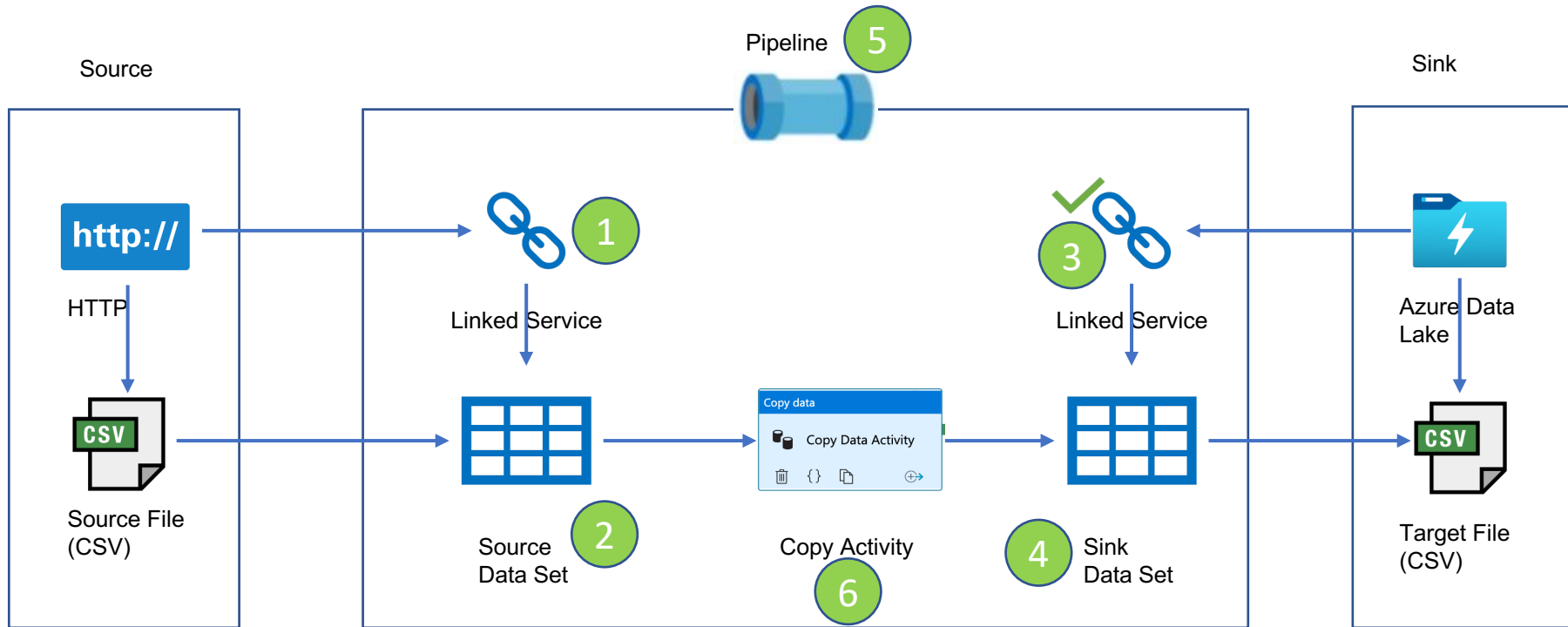


Data Ingestion

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

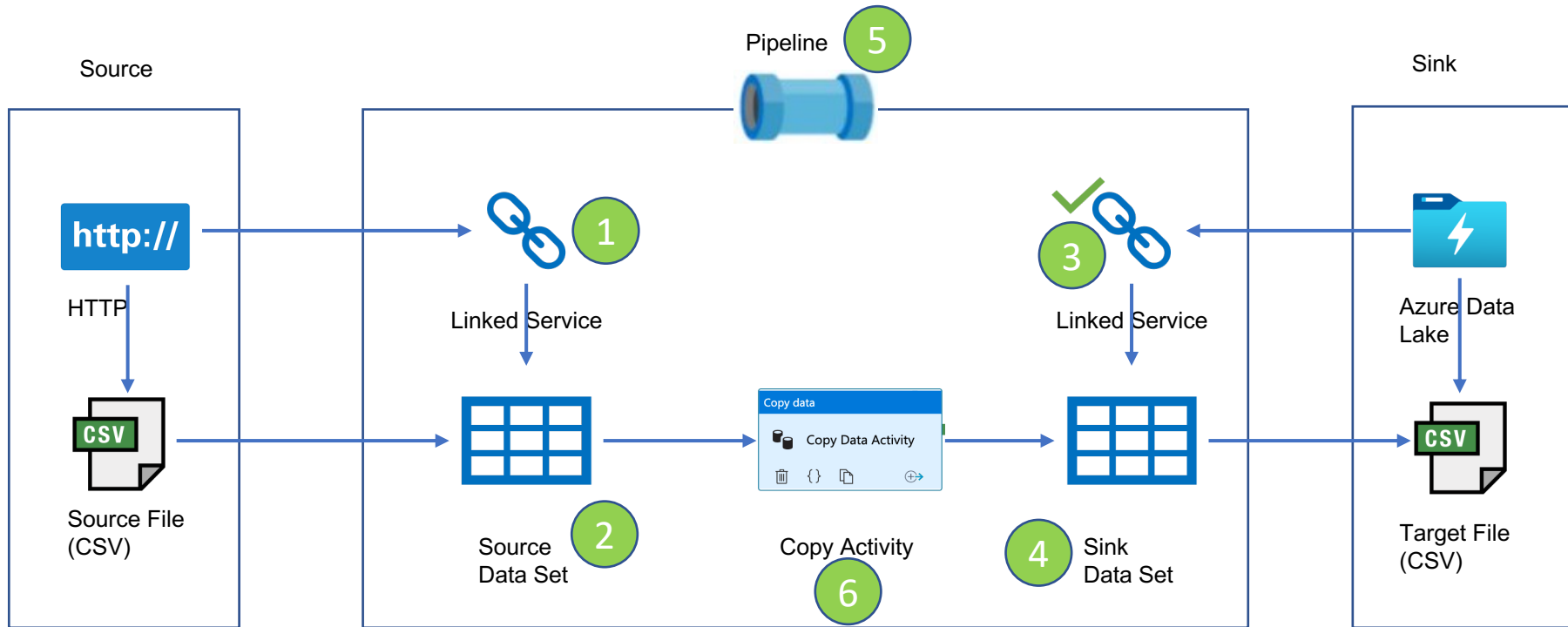


- 1 ls_http_opendata_ecdc_europa_eu
- 2 ds_cases_deaths_raw_csv_http
- 3 ls_adls_covidreportingdl ✓
- 4 ds_cases_deaths_raw_csv_dl
- 5 pl_ingest_cases_deaths_data
- 6 Copy Cases And Deaths Data

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

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

Copy Activity – Case & Deaths Data

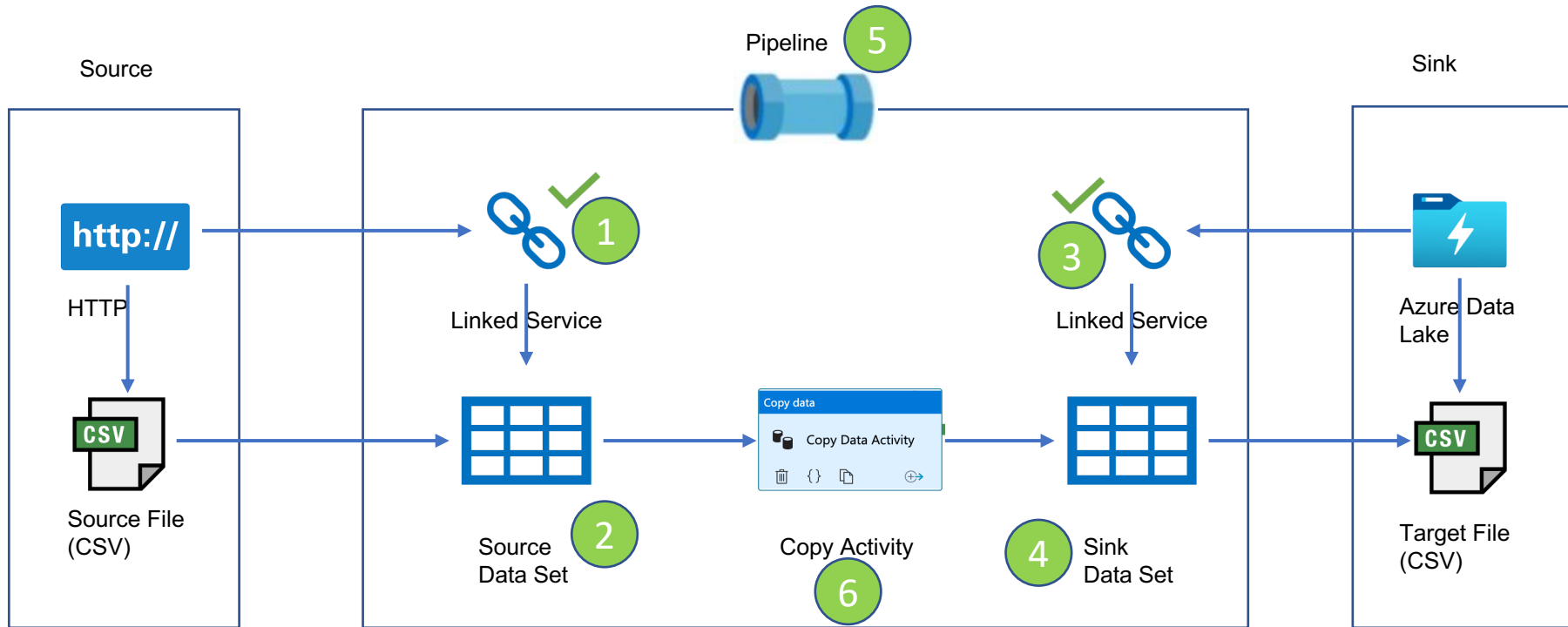


- 1 ls_http_opendata_ecdc_europa_eu
- 2 ds_cases_deaths_raw_csv_http
- 3 ls_adls_covidreportingdl ✓
- 4 ds_cases_deaths_raw_csv_dl
- 5 pl_ingest_cases_deaths_data
- 6 Copy Cases And Deaths Data

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

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

Copy Activity – Hospital Admission Data



- 1 ls_http_opendata_ecdc_eu
ropa_eu ✓
- 2 ds_hospital_admissions_ra
w_csv_http
- 3 ls_adls_covidreportingdl ✓
- 4 ds_hospital_admissions_ra
w_csv_dl
- 5 pl_ingest_hospital_admissi
ons_data
- 6 Copy Hospital Admissions
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>

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

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

https://www.ecdc.europa.eu/sites/default/files/documents/data_response_graphs_0.csv

Sink

raw/ecdc/case_distribution.csv

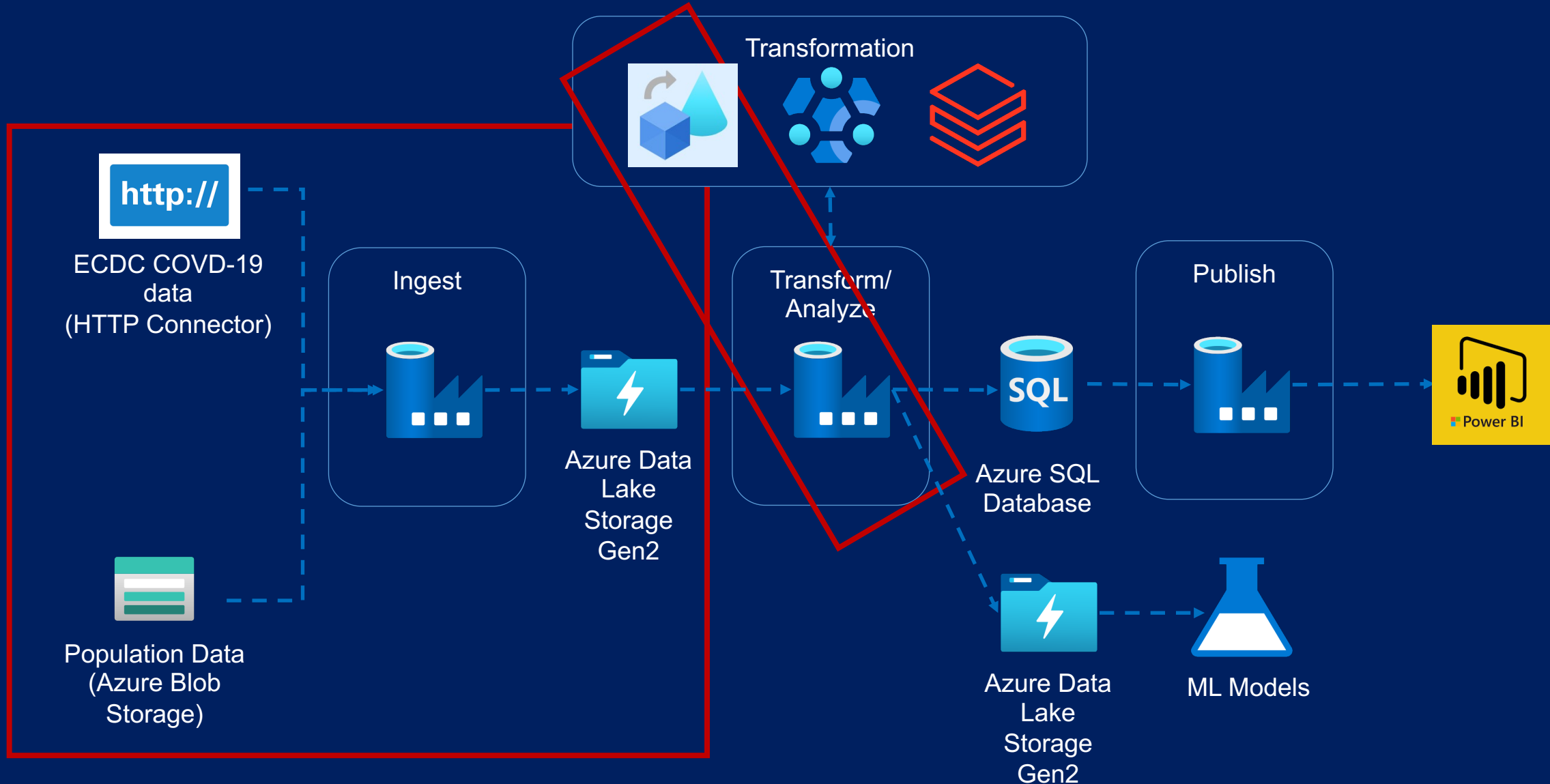
raw/ecdc/hospital_admission.csv

raw/ecdc/testing.csv

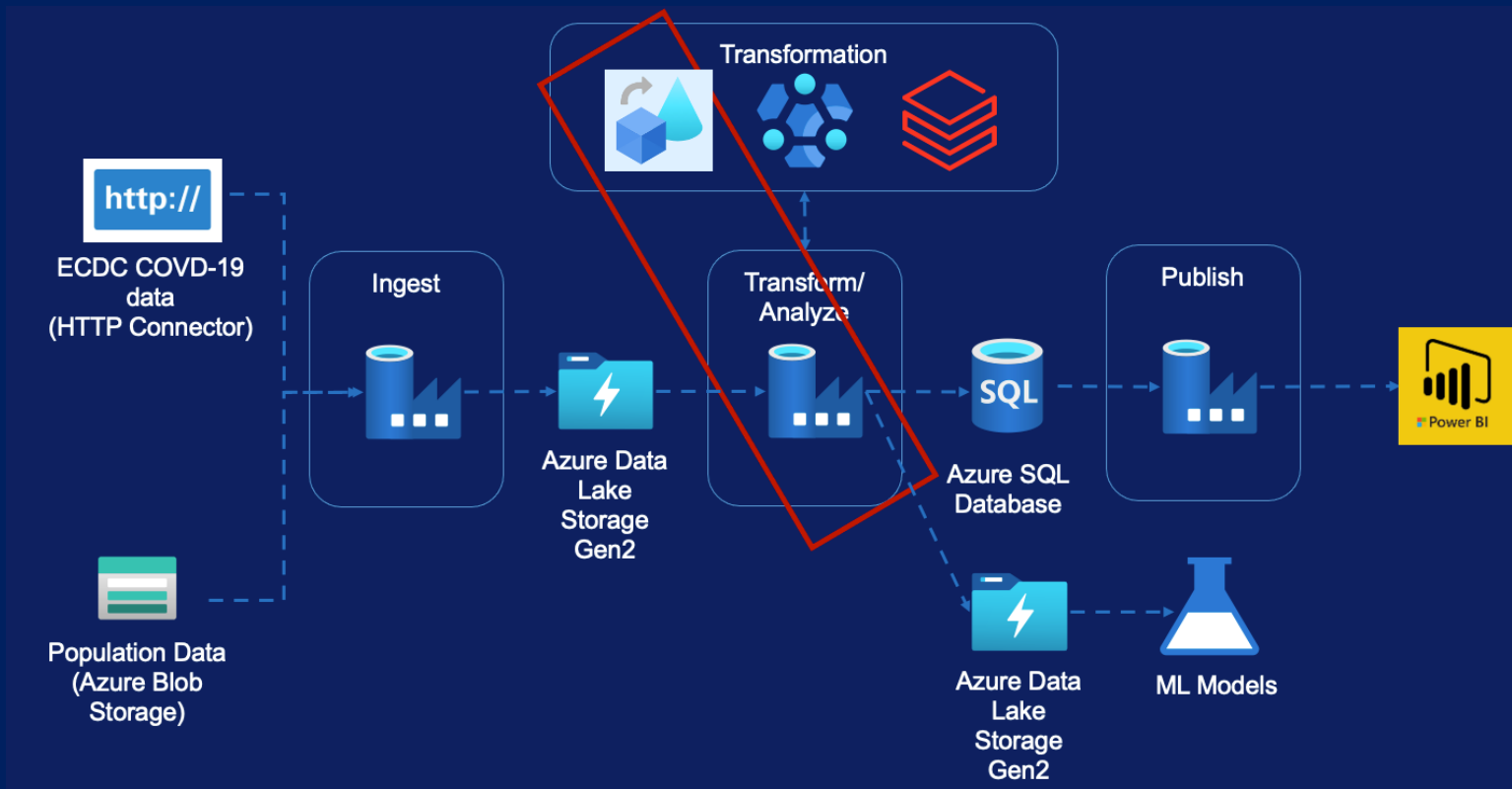
raw/ecdc/country_response.csv

Data Flows (1) - Module Overview (Cases & Deaths File)

Data Flow – Cases & Deaths Data



Data Flow – Cases & Deaths Data



Data Flow Overview

Requirement

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 Databricks 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

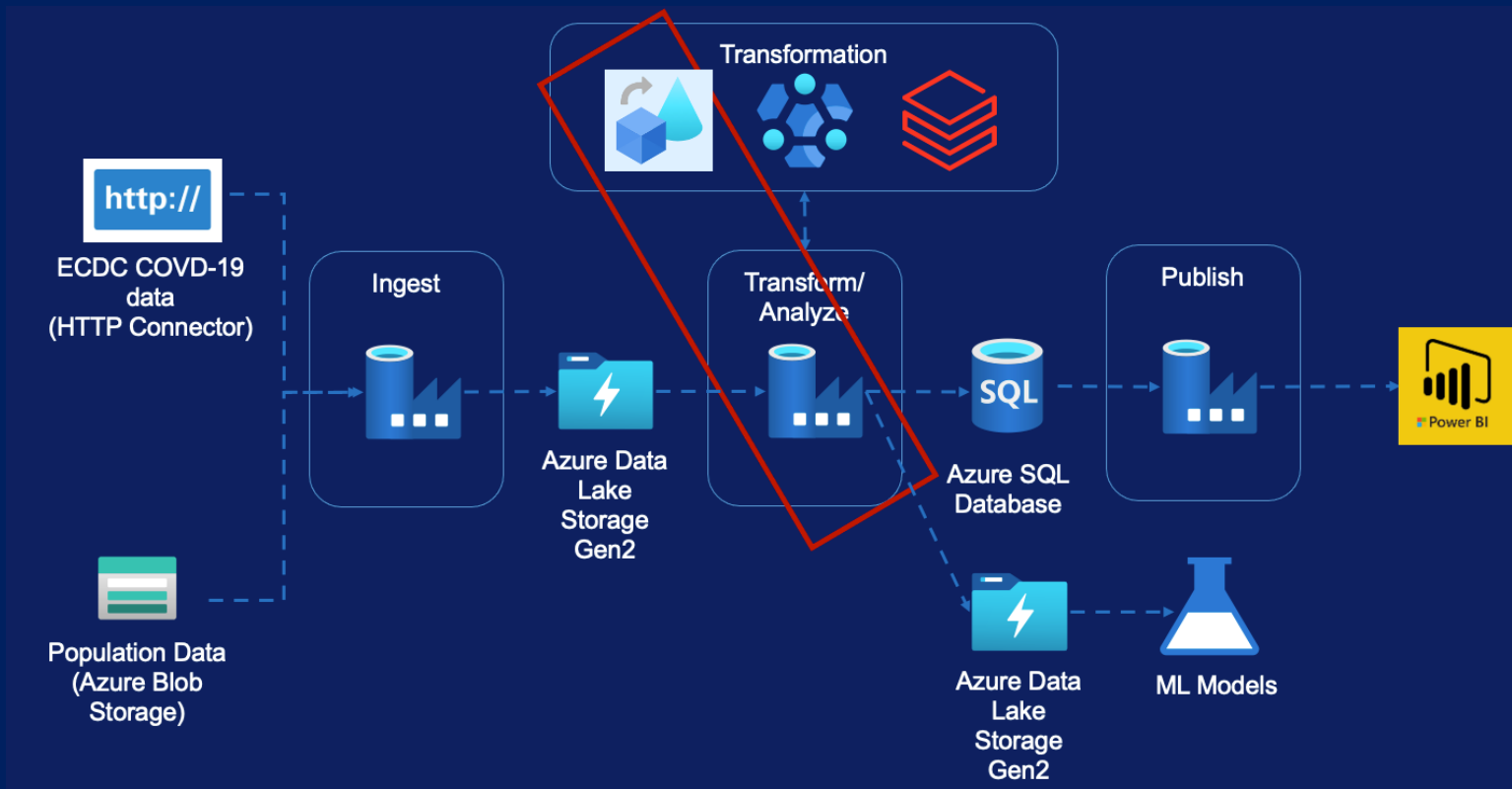


Transformed File

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

Data Flows (2) - Module Overview (Hospital Admissions File)

Data Flow – Cases & Deaths 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

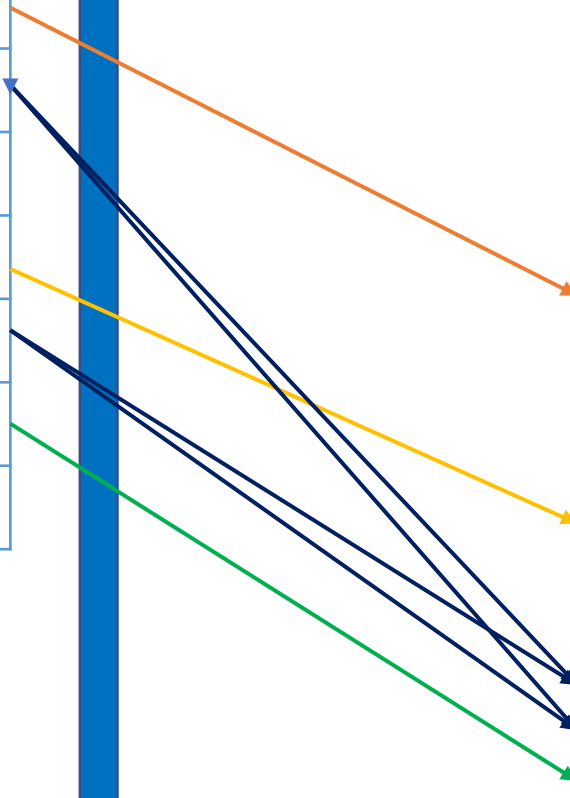
Column Name
country
indicator
date
year_week
value
source
url

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

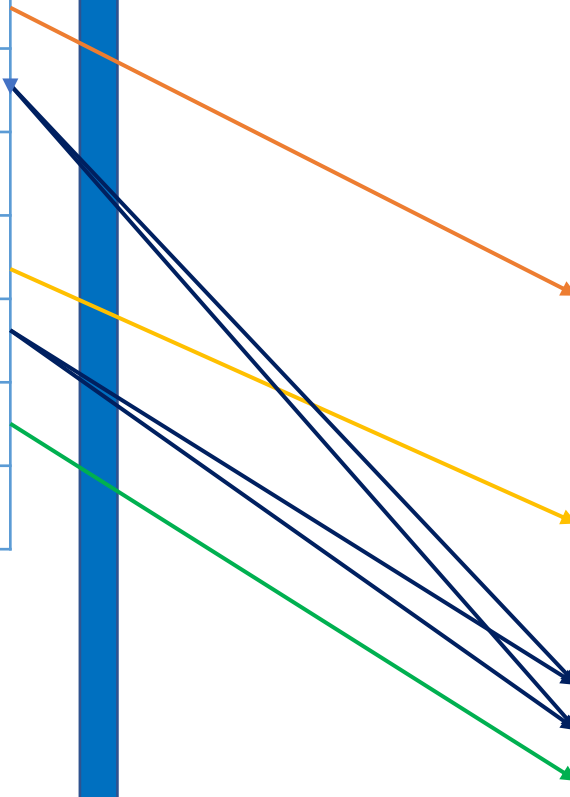
Column Name
country
indicator
date
year_week
value
source
url

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

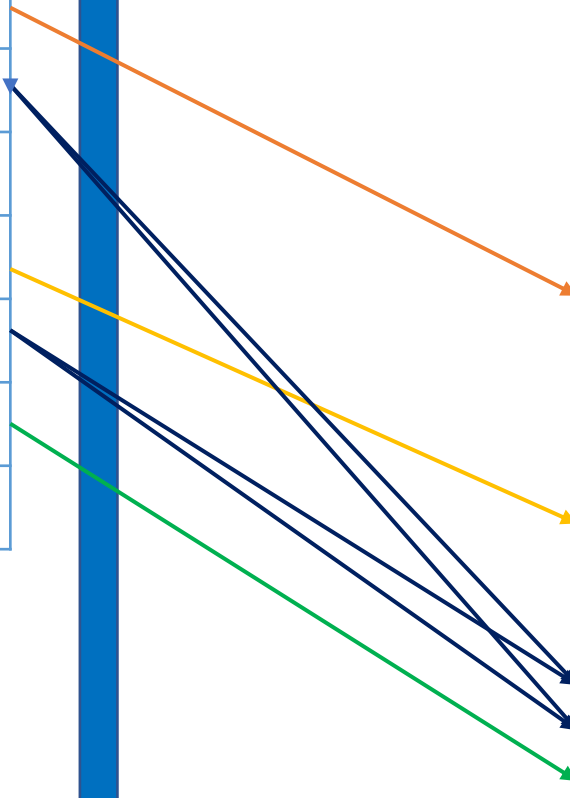
Column Name
country
indicator
date
year_week
value
source
url

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



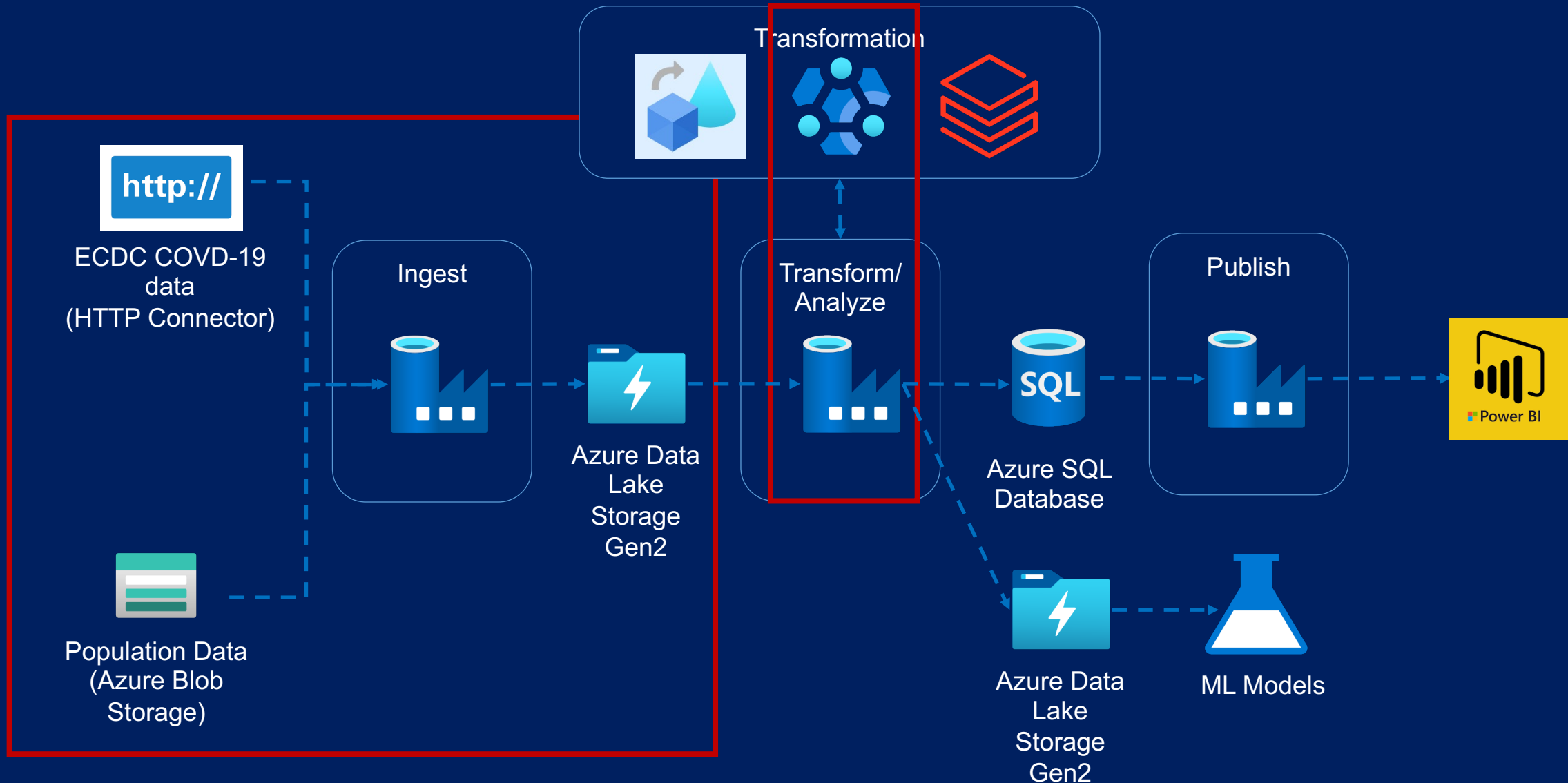
Data Flow Execution

Assignment

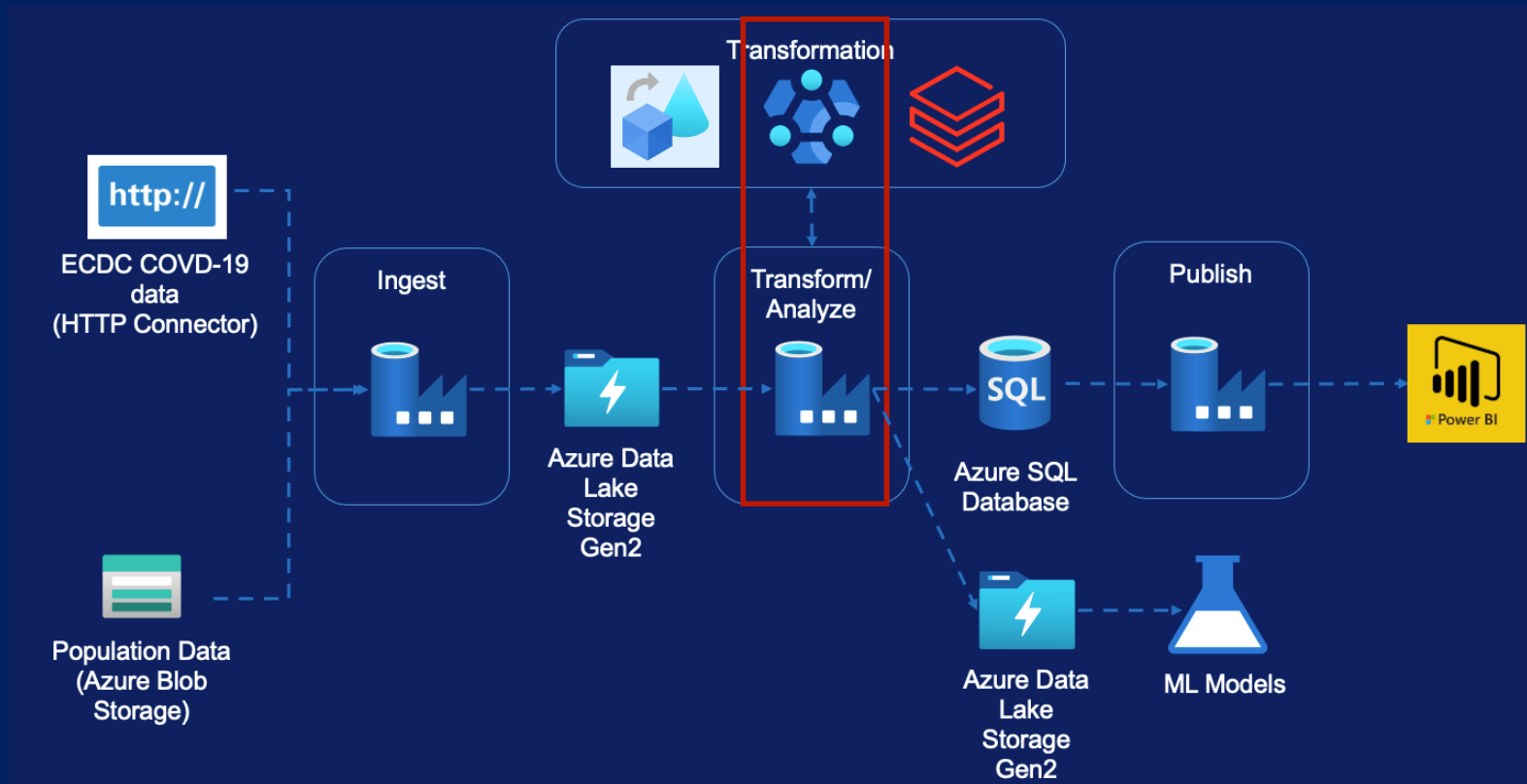


HDInsight Activity - Module Overview (Testing File)

HDInsight Activity – Testing File



HDInsight Activity – Testing File



Creating HDInsight Cluster

HDInsight UI Overview

Transformation Requirement

Hive Script Walk-through

Creating Pipeline

Delete HDInsight Cluster

Creating HDInsight Cluster



Testing Data



Testing Data

Raw File from ECDC

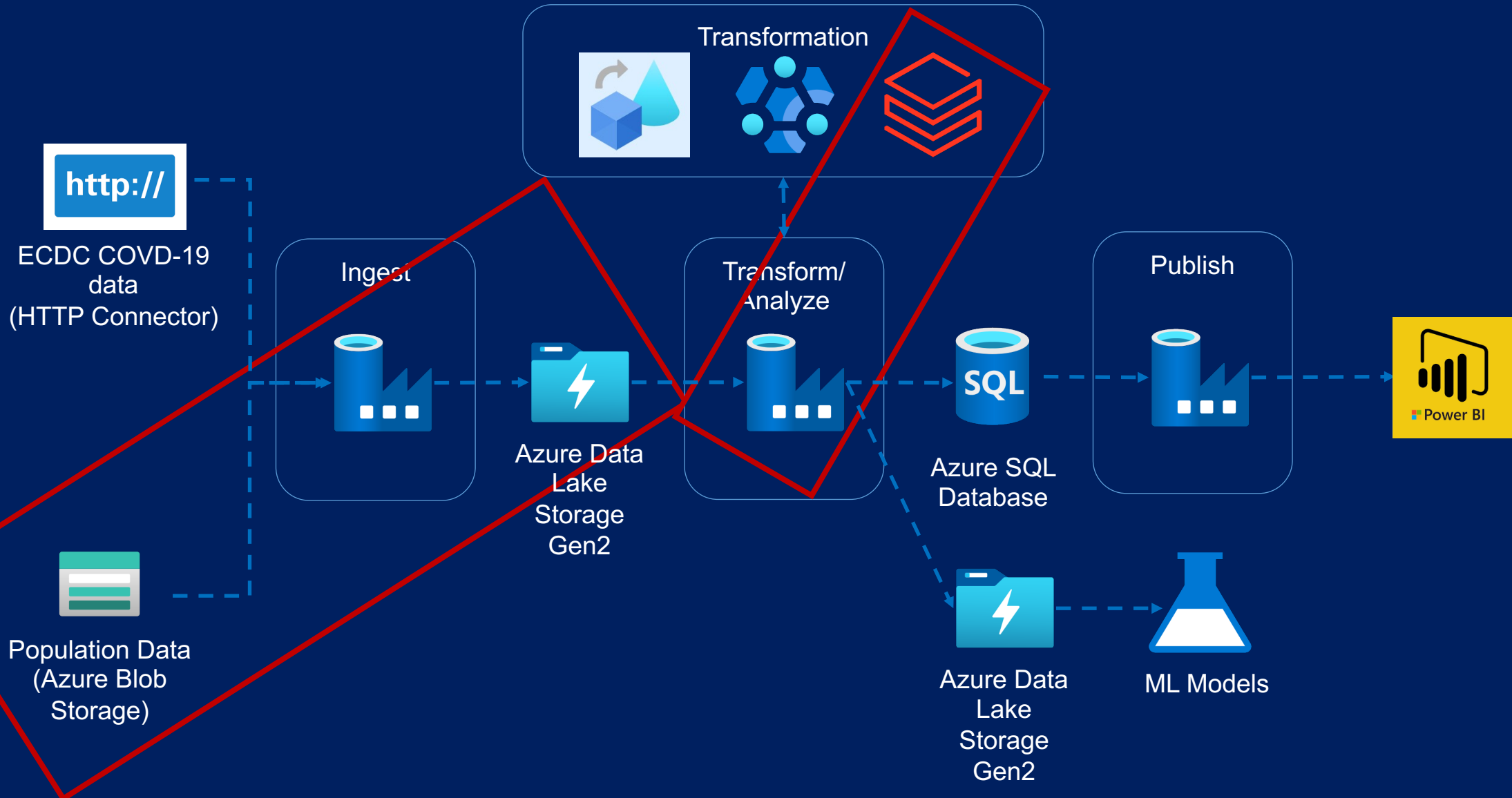
Column Name
country
country_code (Remove)
Year_week
new_cases
test_done
population
testing_rate
positivity_rate
testing_data_source

Transformed File

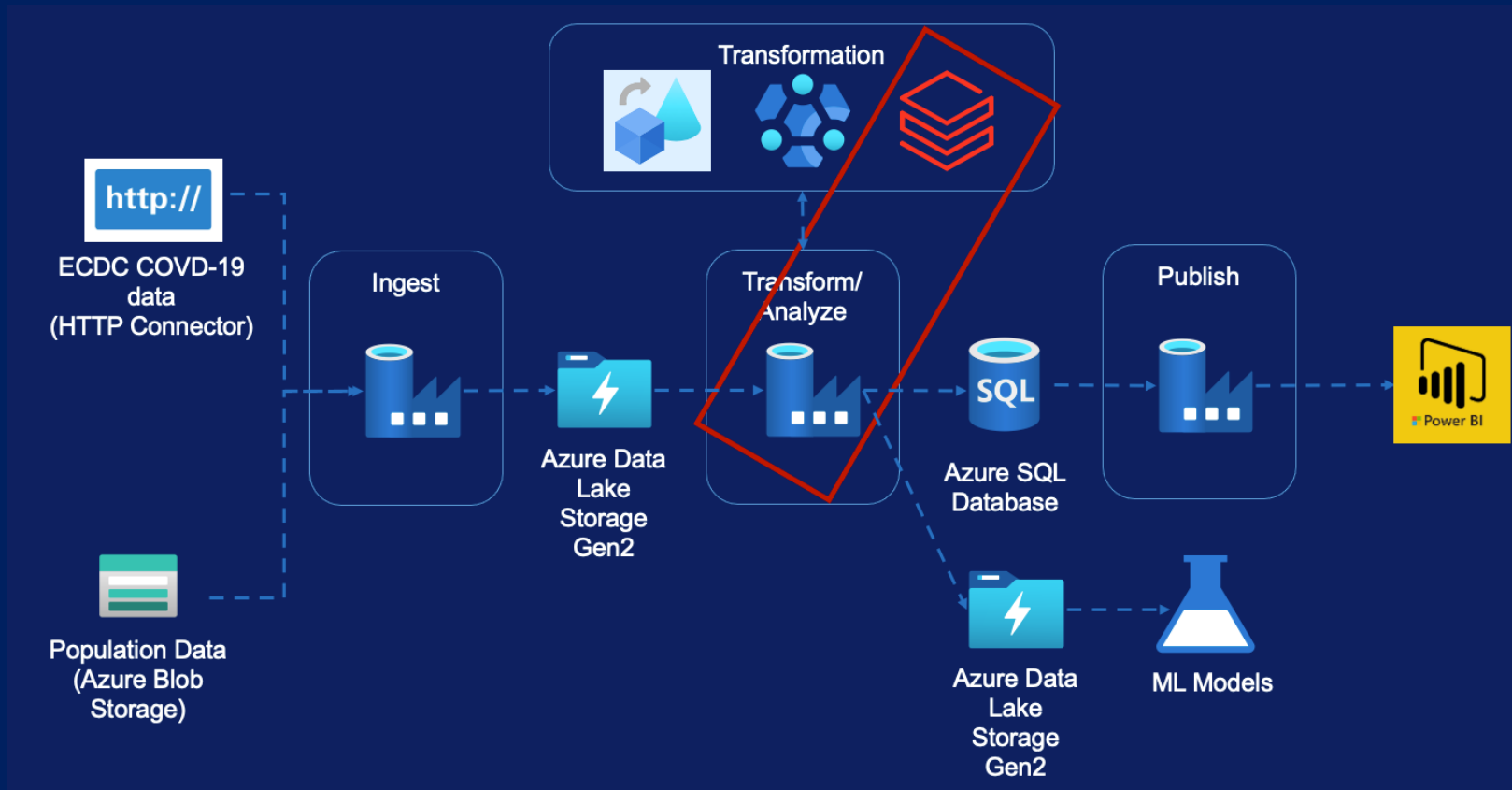
Column Name
country
country_code_2_digit (lookup)
country_code_3_digit(lookup)
reported_year_week
reported_week_start_date(lookup)
reported_week_end_date(lookup)
new_cases
test_done
population
testing_rate
positivity_rate
testing_data_source

Databricks Activity - Module Overview (Population File)

Databricks Activity – Population File



Databricks Activity – Population File



Create Databricks Service

Create Databricks Cluster

Mount Storage Accounts

Transformation Requirements

Creating Pipeline

Databricks Environment Set-up



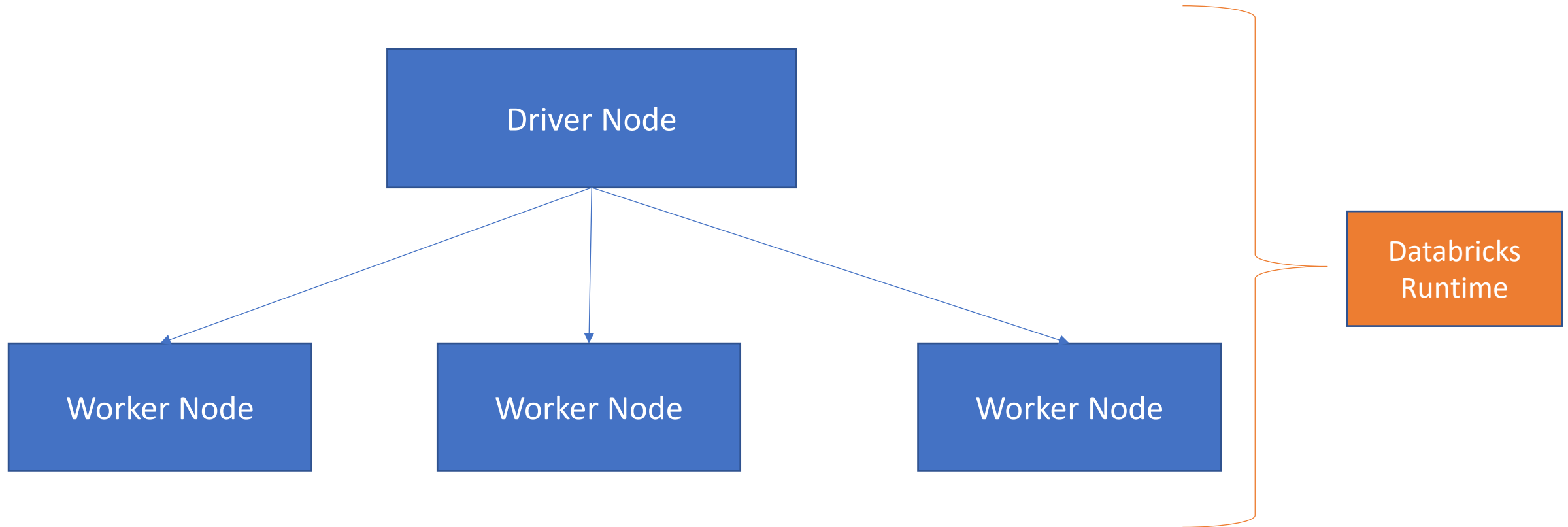
Creating Databricks Service



Creating Databricks Cluster



What is a cluster?



Cluster Types

All Purpose/ Interactive
Clusters

Job Clusters

Mounting Data Lake Storage



Mounting Data Lake Storage

- Create Azure Service Principal
- Grant access for data lake to Azure Service Principal
- Create the mount in databricks using Service Principal

Transform Population By Age Data



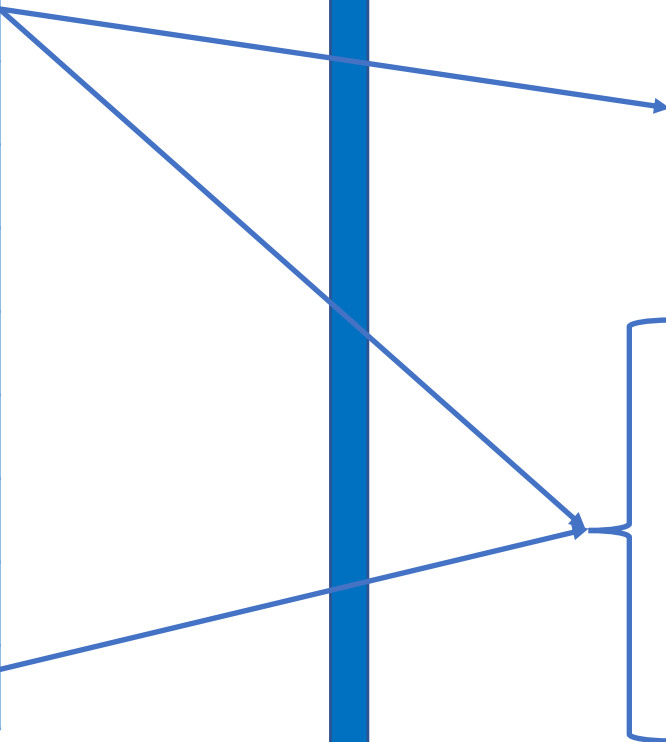
Transform Population By Age Data

Raw File

Column Name
indic_de,geo\time
2008
2009
2010
2011
...
....
2018
2019

Transformed File

Column Name
Country (Lookup)
country_code_2_digit(Substr)
country_code_3_digit(Lookup)
population(Lookup)
age_group_0_14
age_group_25_49
age_group_50_64
age_group_65_79
age_group_80_max



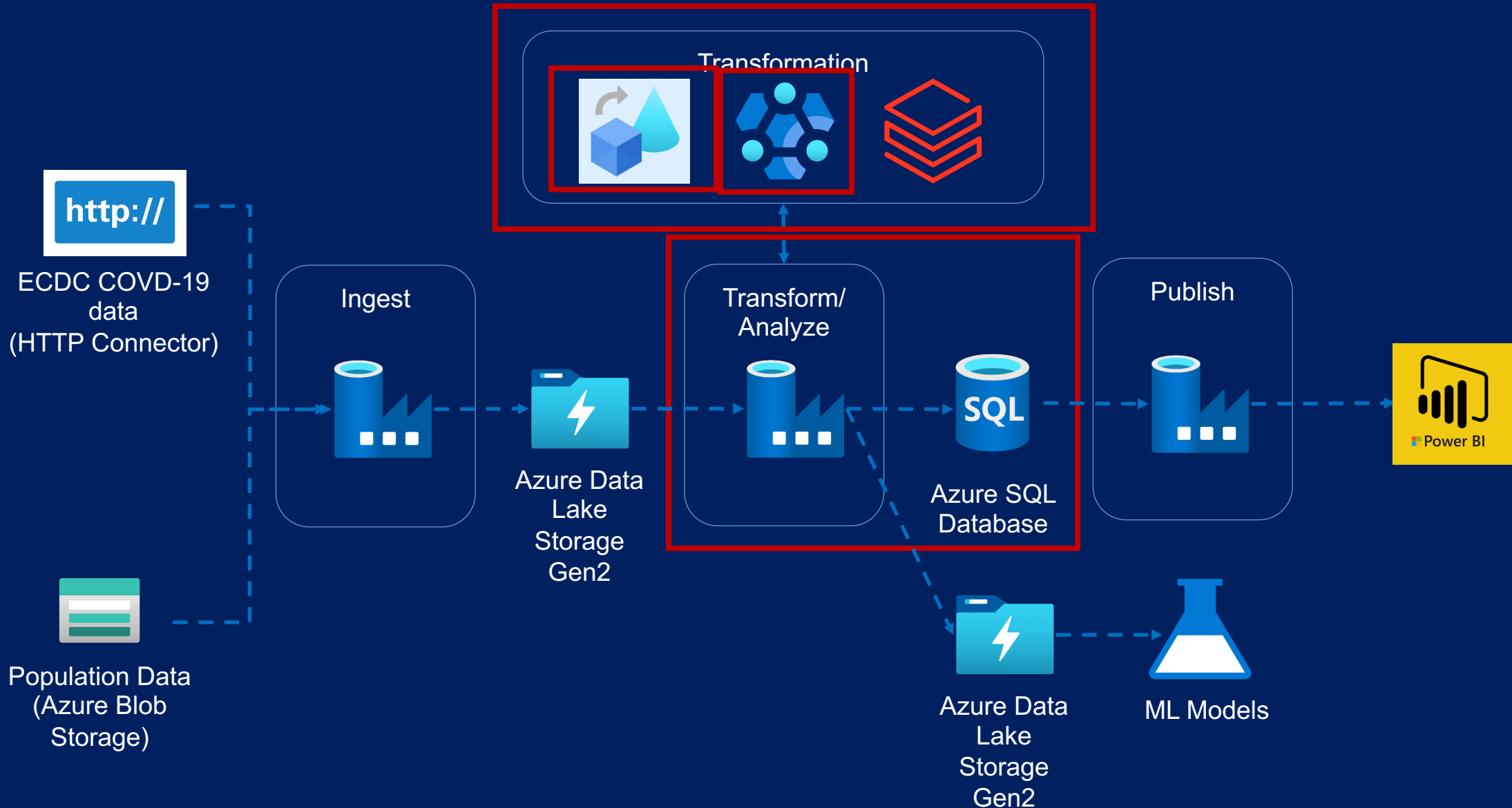
Transform Population By Age Data

Data Factory Pipeline

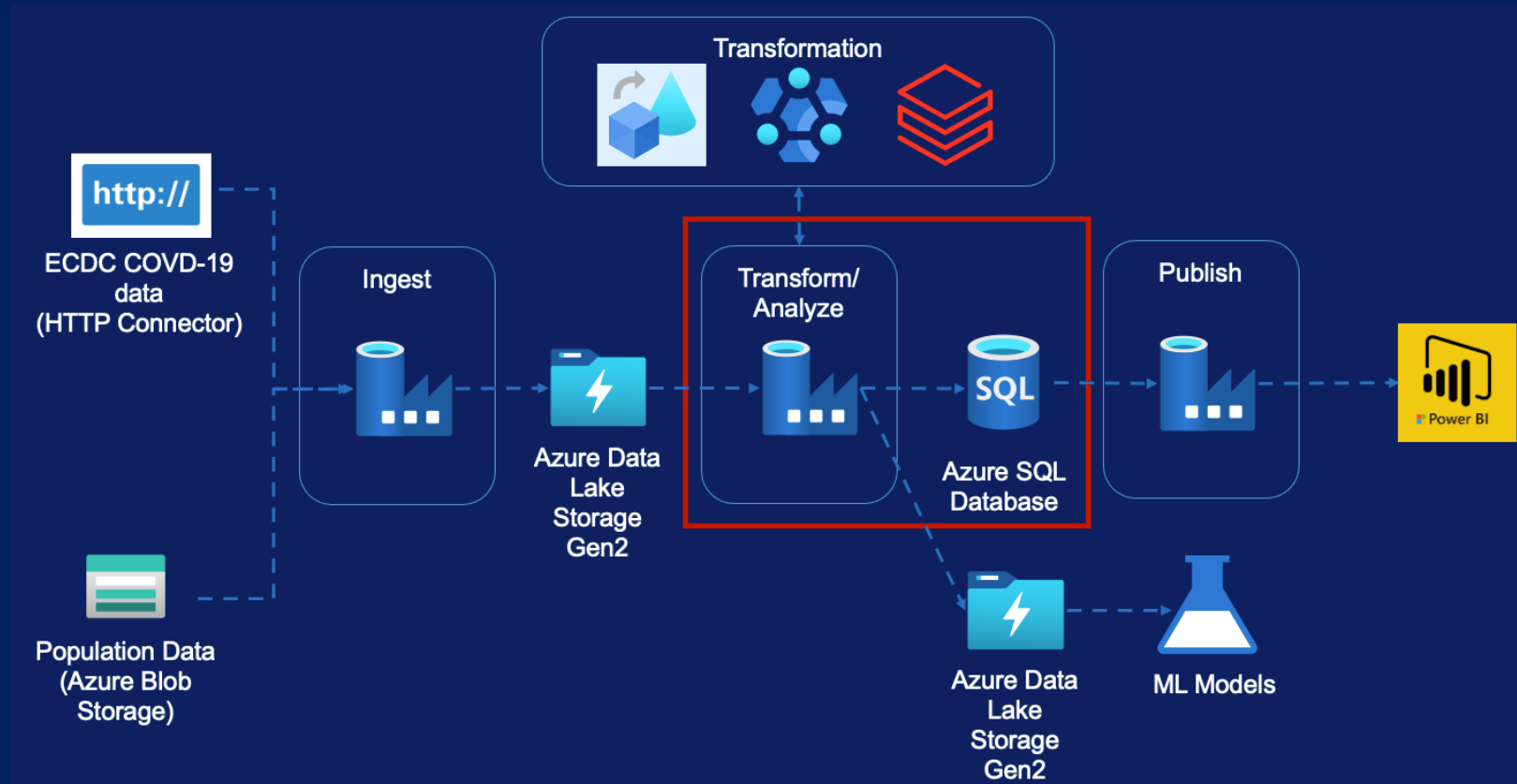


Copy Data to Azure SQL

Copy Data to SQL



Copy Data to SQL



- Copy Cases & Deaths
- Copy Hospital Admissions
- Copy Testing

Copy Activity – Data Lake to SQL

Cases and Deaths Data

