



# *End To End Data Engineering Project using Healthcare Data*

25 October 2024

Auni Khairina Azman



# *Problem Statement*

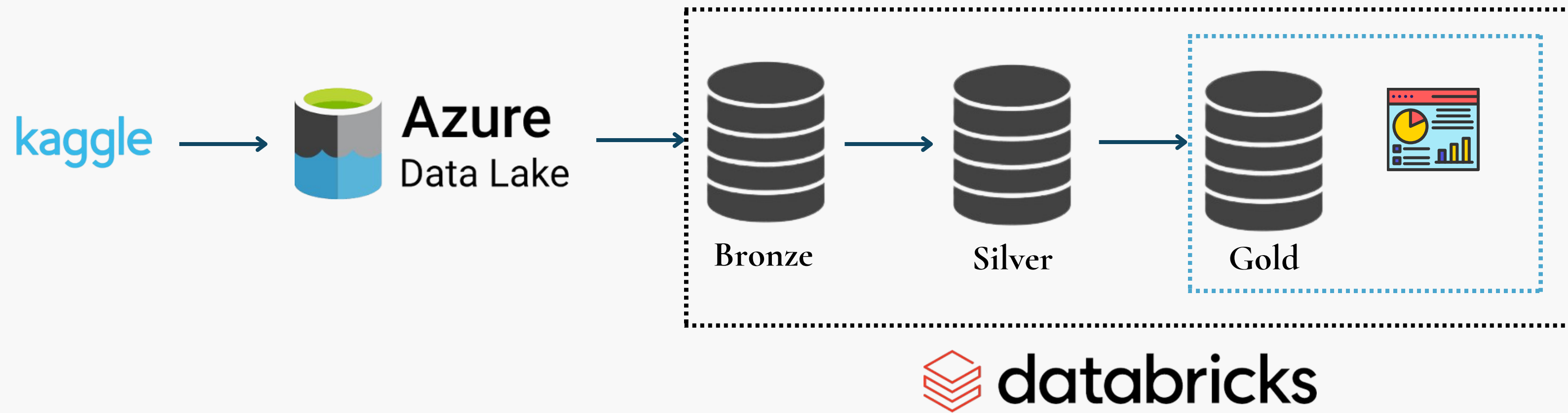
## *Assessing Vaccine Coverage Trends*

To analyze the trends in vaccination coverage for BCG, Hep B, and DTP among children under 1 year old in Malaysia from 2010 to 2018, determining if there has been a consistent increase, decrease, or fluctuation over the years.

Data Source : Kaggle (<https://www.kaggle.com/datasets/thedevastator/who-malaysia-health-indicators>)



# Architecture Diagram



# *Detail explanation in each processing notebook*



## Bronze

- Get adls connection
- Read csv file from adls
- Write data as parquet file in Bronze



## Silver

- Read parquet file from Bronze
- Remove anomaly from dataset (invalid data)
- Remove special characters & white space from variable column
- Convert value from String into Integer
- Select few columns to make analysis
- Write data as parquet file in Silver



## Gold

- Read cleaned BCG, DTaP, Hep B parquet files from Silver
- Combine those 3 files into a dataframe
- Add one more column “Immunization\_Type”
- Filter “year” since 2010 until 2018 only
- Write data as parquet file in Gold
- Create line graph to visualize the data

# ETL Data Pipeline

Workflows > Jobs >

Auni\_ETL\_Data ☆

Send feedback

Run now

RunsTasks

Runs

Start date

< Previous

Next >

Run total duration

1m 34s

47s

Oct 25

Bronze

Silver

Gold

Go to the latest successful run

Cancel runs

| Start time             | Run ID        | Launched | Duration | Status      | Error code | Run parame... |  |
|------------------------|---------------|----------|----------|-------------|------------|---------------|--|
| Oct 25, 2024, 02:51 PM | 5711167443... | Manually | 1m 35s   | ✔ Succeeded |            |               |  |

Job details

Job ID

415619989274559

Creator

Auni Khairina Binti Azman

Run as

Auni Khairina Binti Azman

Tags

Add tag

Description

Add description

Lineage

No lineage information for this job.  
Learn more

Git

Not configured

Add Git settings

Schedules & Triggers

None

Add trigger

# Detail explanation in each task in ETL

Workflows > Jobs > Auni\_ETL\_Data ☆

Runs Tasks

Task name\* ① Bronze

Type\* Notebook ▾

Source\* ① Workspace ▾

Path\* ① /Workspace/Users/khairinaauni@gmail.com/bronze\_child\_health ▾

Cluster\* ① Kotak Sakti's Shared Compute Cluster 42 GB · 12 Cores · DBR 15.4 LTS · Spark 3.5.0 · S... ▾

① Jobs running on all-purpose clusters are considered all-purpose compute. [Learn more](#)

Depends on Select task dependencies... ▾

Dependent libraries ① + Add

Bronze

Workflows > Jobs > Auni\_ETL\_Data ☆

Runs Tasks

Task name\* ① Silver

Type\* Notebook ▾

Source\* ① Workspace ▾

Path\* ① /Workspace/Users/khairinaauni@gmail.com/silver\_child\_health ▾

Cluster\* ① Kotak Sakti's Shared Compute Cluster 42 GB · 12 Cores · DBR 15.4 LTS · Spark 3.5.0 · S... ▾

① Jobs running on all-purpose clusters are considered all-purpose compute. [Learn more](#)

Depends on Bronze X ▾

Run if dependencies ① All succeeded ▾

Silver

Workflows > Jobs > Auni\_ETL\_Data ☆

Runs Tasks

Task name\* ① Gold

Type\* Notebook ▾

Source\* ① Workspace ▾

Path\* ① /Workspace/Users/khairinaauni@gmail.com/gold\_child\_health ▾

Cluster\* ① Kotak Sakti's Shared Compute Cluster 42 GB · 12 Cores · DBR 15.4 LTS · Spark 3.5.0 · S... ▾

① Jobs running on all-purpose clusters are considered all-purpose compute. [Learn more](#)

Depends on Silver X ▾

Run if dependencies ① All succeeded ▾

Gold



*Thank you*

