Modern data stack


Supreme
PWS
Regular
93
89
87


HAZET VANADIUM
CHROM HAZET VANADIUM
HAZET VANADIUM

The **modern data stack** is one **framework** used to **conceptualize** how different **data tools** work together to allow a complete **data journey**.

Modern Data Stack

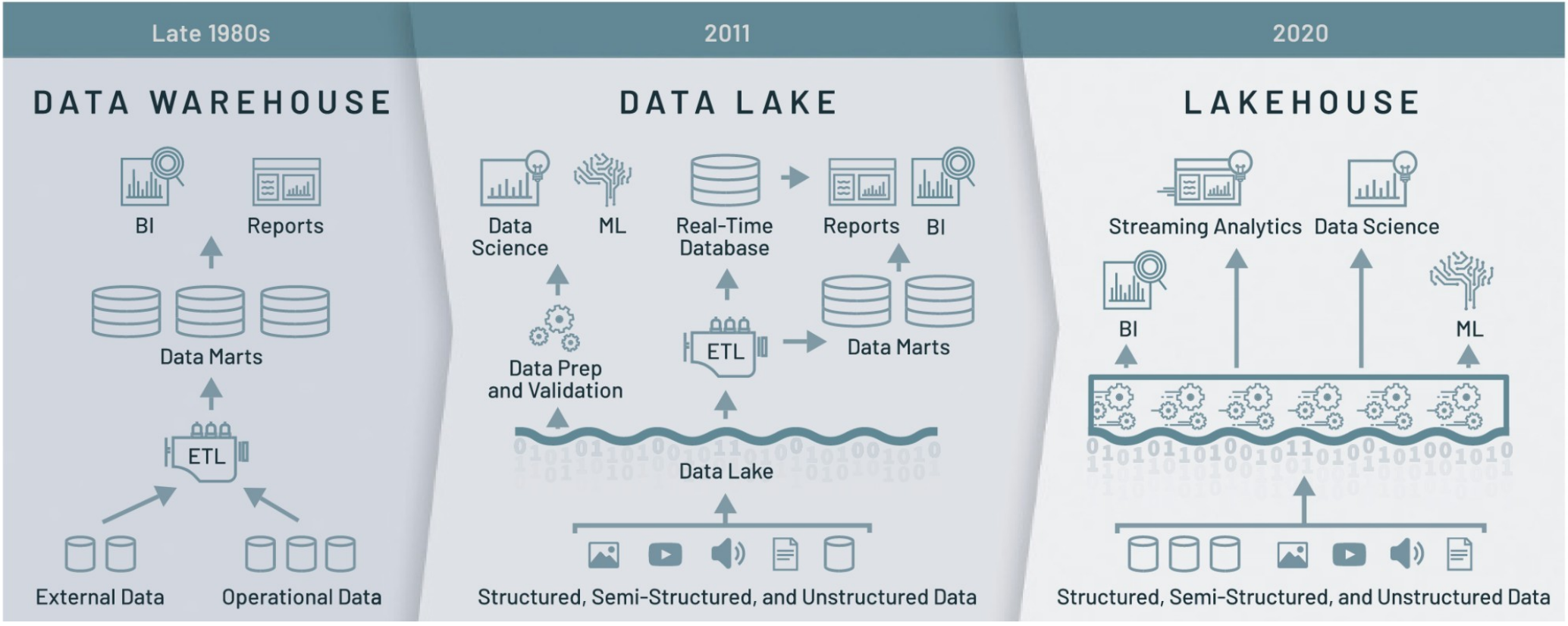
- **Infraestructure as code**


HELM


Build
Release
Deploy
build
release
deploy_produc...
deploy_staging

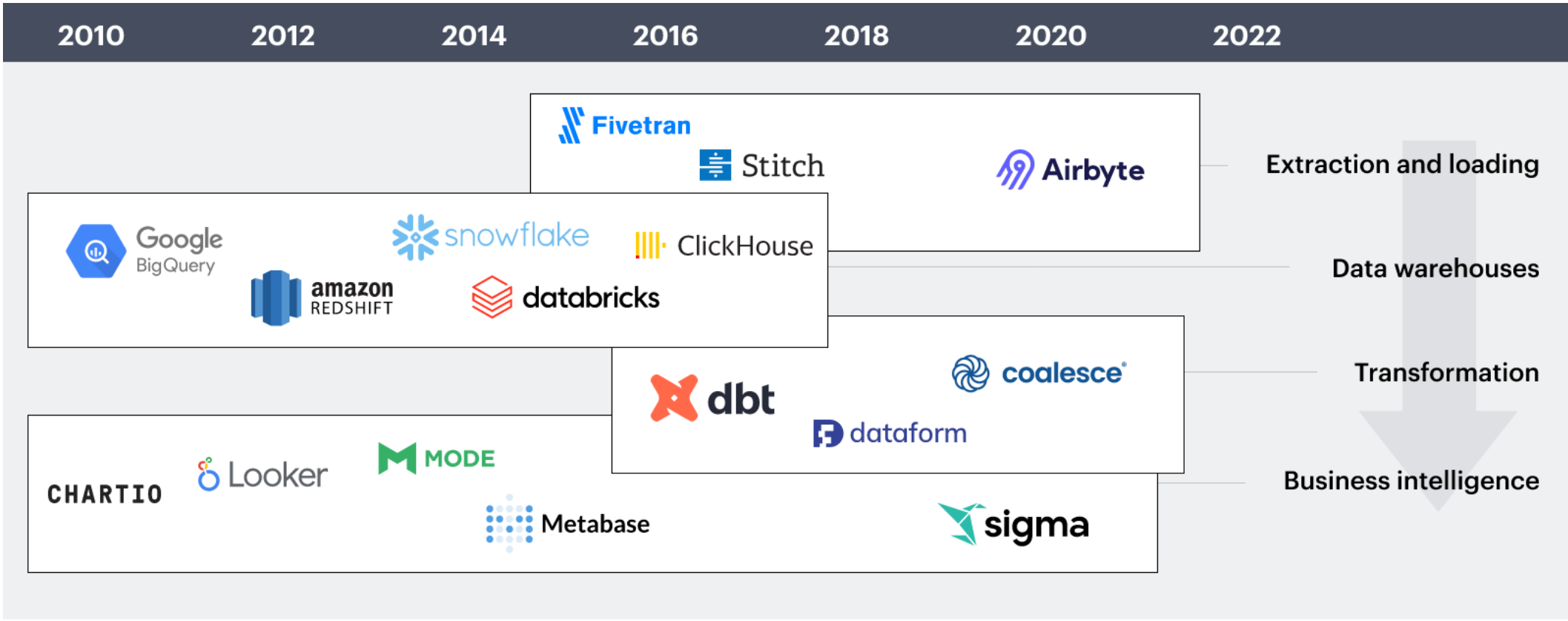
Modern Data Stack

- Infraestructure as code
- **From Datawarehouse to Lakehouse**


Late 1980s
2011
2020
DATA WAREHOUSE
BI
Reports
Data Marts
ETL
External Data
Operational Data
DATA LAKE
Data Science
ML
Real-Time Database
Reports
BI
Data Prep and Validation
ETL
Data Marts
Data Lake
Structured, Semi-Structured, and Unstructured Data
LAKEHOUSE
Streaming Analytics
Data Science
BI
ML
Structured, Semi-Structured, and Unstructured Data

Modern Data Stack

- Infraestructure as code
- From Datawarehouse to Lakehouse
- **From ETL to ELT**

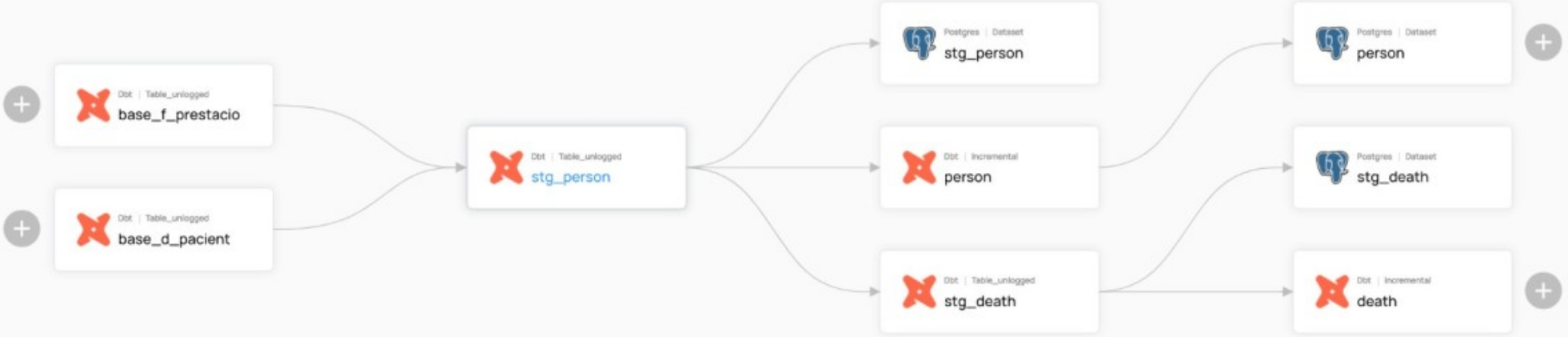


Adoption of the **Modern Data Stack**

Helps data teams pull data from sources (**extraction and loading**) into one place (**warehouses**), transform it into a usable form (**transformation**), and then give access with accessible tools (**business intelligence**).

Modern Data Stack

- Infraestructure as code
- From Datawarehouse to Lakehouse
- From ETL to ELT
- **Data governance and observability**



Accuracy
The degree to which the data correctly represents the entity or attribute being described.

Completeness
The percentage of missing data from a given data set.

Consistency
The absence of difference or contradiction in data irrespective of the data's source.

Validity
Invalid data affects the accuracy and completeness of a data set.

Integrity
The validity of relationships across various data entities.

Uniqueness
Ensures duplicate or overlapping data is identified and marked.

Modern Data Stack

- Infraestructure as code
- From Datawarehouse to Lakehouse
- From ETL to ELT
- Data governance and observability
- **Self-service analytics**

PRODU CTION APPME D

Search by Title...

Recent

PERSON

Person analysis

Domain Visit

DRUGS (QA)

EDA: DRUG_EXPOSURE (H12O)

Favorites ↓U@

No items in this section.

Lists

OMOP

All My Docs ↓U@

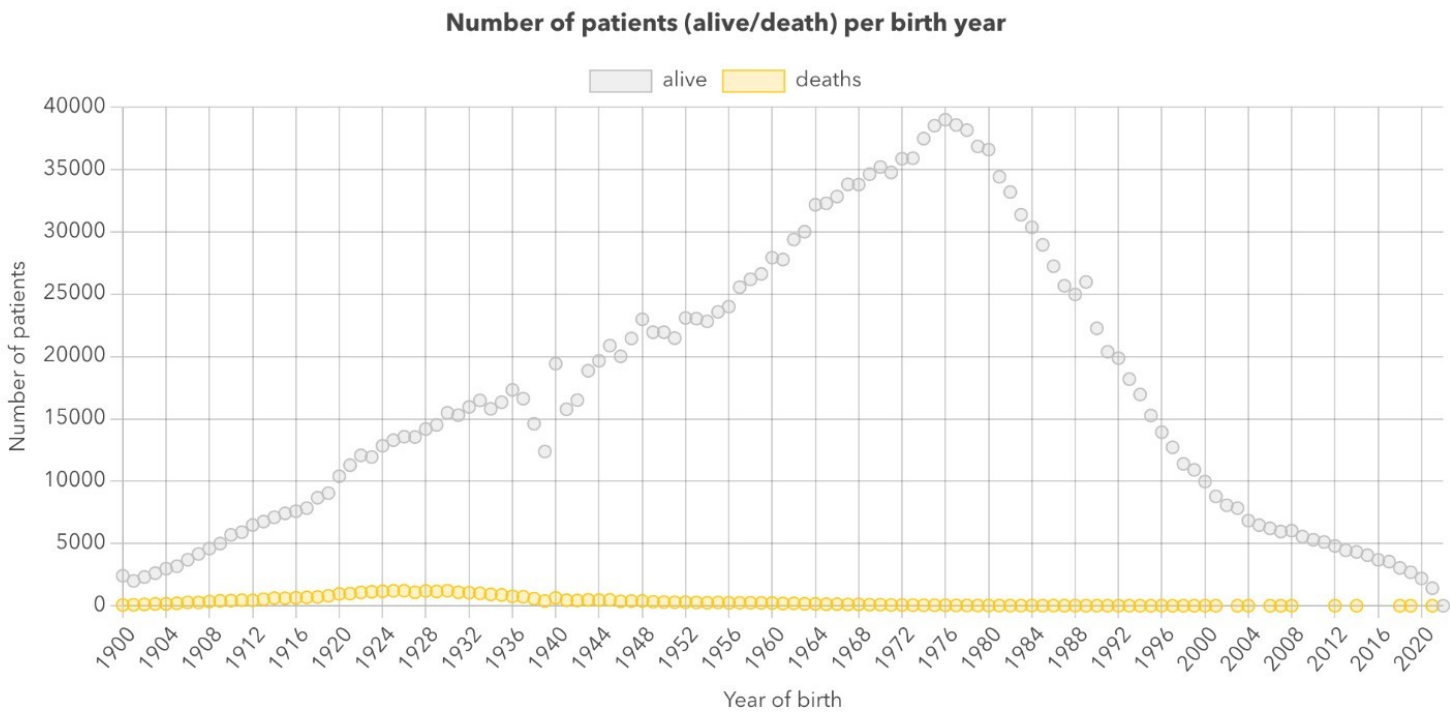
Untitled

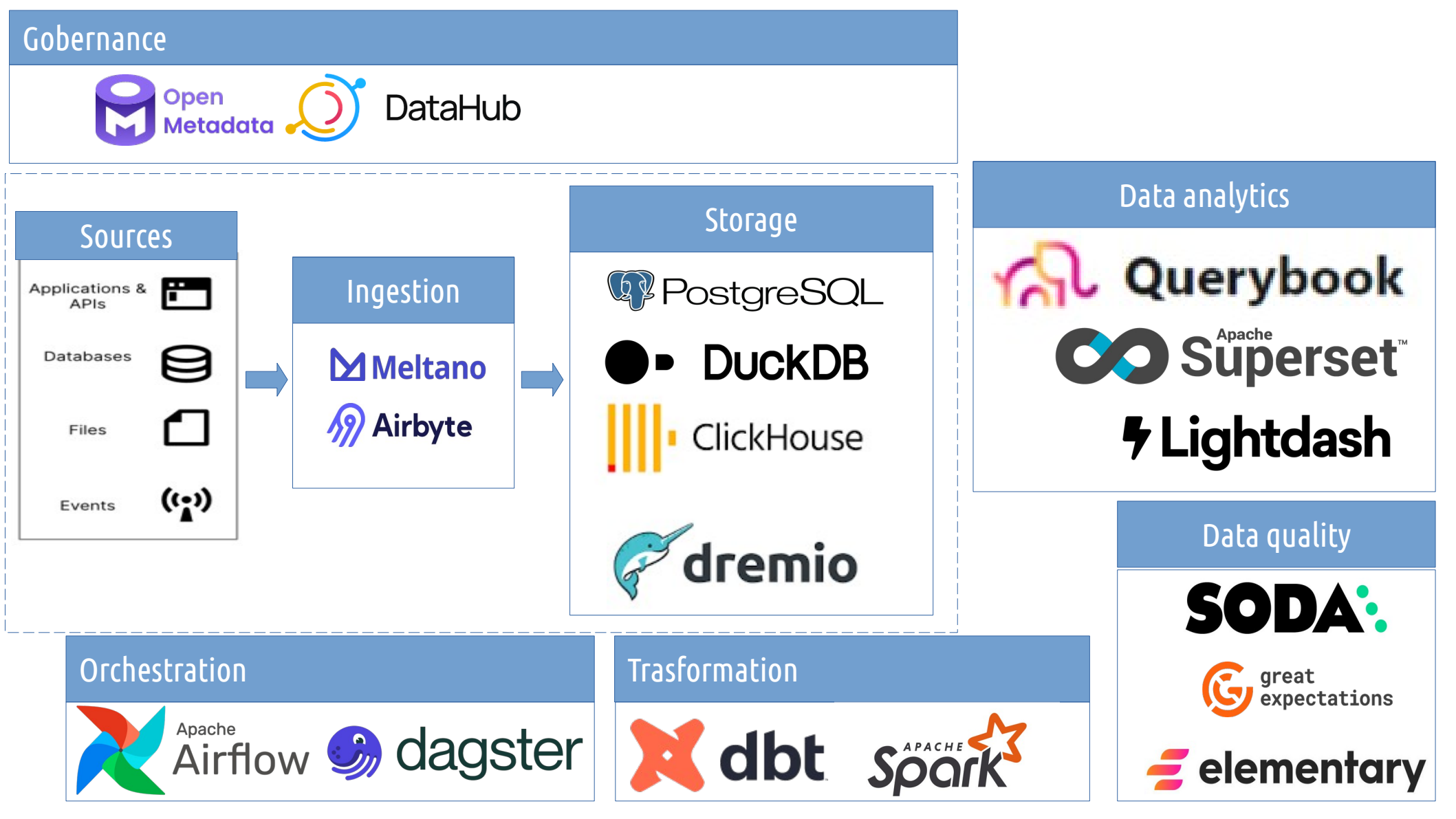
Adhoc Docs Tables Snippets Runs Search

EXECUTION 90

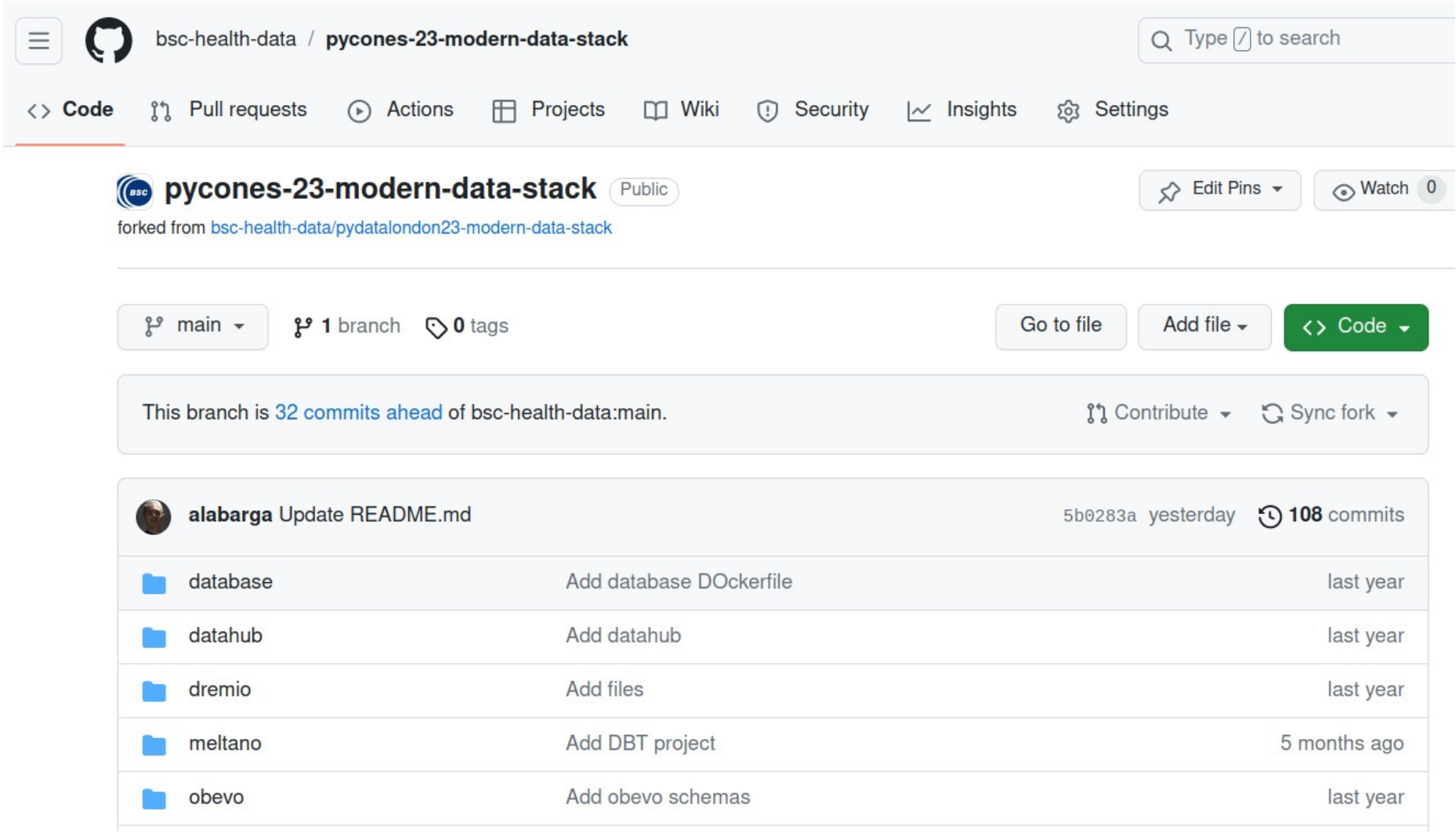
STATEMENT 1 OUT OF 1

CONFIG CHART




Gobernance
Open Metadata
DataHub
Sources
Applications & APIs
Databases
Files
Events
Ingestion
Meltano
Airbyte
Storage
PostgreSQL
DuckDB
ClickHouse
dremio
Data analytics
Querybook
Apache Superset
Lightdash
Data quality
SODA
great expectations
elementary
Orchestration
Apache Airflow
dagster
Trasformation
dbt
APACHE Spark

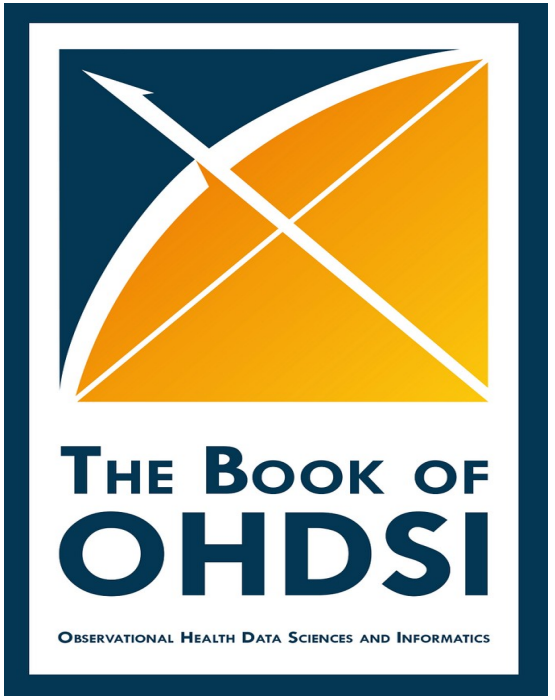

Practice time?



https://github.com/bsc-health-data/pycones-23-modern-data-stack/

Recursos

- Sitio web: https://github.com/alabarga/omop-cdm-course
- Diapositivas
- El Libro de OHDSI
- EHDEN academy
- OHDSI workshops
- Atlas tutorials
- Entorno local (docker)



You are free to:

Share — copy and redistribute the material in any medium or format

Under the following terms:

Attribution - You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

NonCommercial - You may not use the material for commercial purposes .

NoDerivatives - If you remix, transform, or build upon the material, you may not distribute the modified material.