Sovereign, High-Performance Orchestration of **5 Billion** Healthcare Records using **dbt** and **Dagster**

Transforming 5 Billion Rows of Healthcare Data: A Secure, Locally Deployed Open-Source Pipeline for OMOP CDM

Challenge: Cloud ETL poses **data residency risks** and high costs for large-scale health data.

Goal: A local, open-source pipeline to transform **5B+ rows** into OMOP CDM without compromising institutional sovereignty.

Methods

Architecture

Containerized **Docker** environment using **dbt** for transformations and **Dagster** for orchestration.

Optimization

Custom **dbt materializations** for SQL Server and 6-month **partitioned execution** to minimize memory footprint.

Quality Gates

Mandatory **Achilles** and **DQD** audits before migration to production environment

Local AI

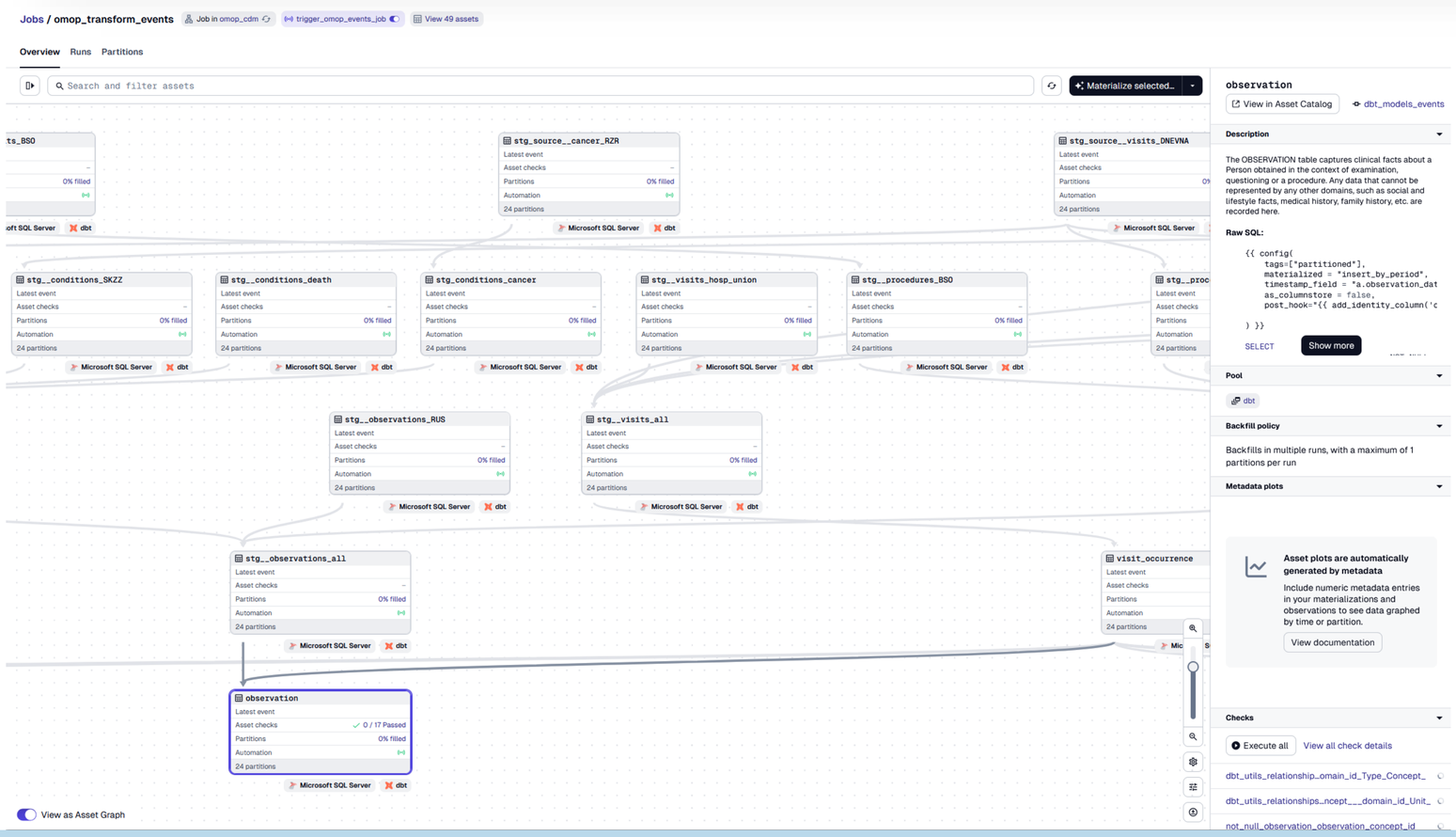
LLM sensors for private, natural-language troubleshooting of SQL logs on-site.

Results

10x Performance: Partitioned execution delivered **a 10x speed increase** in OMOP model updates compared to legacy methods.

Scalability: Proved that **standard on-premises hardware** can process 5B+ rows at enterprise scale.

Transparency: Software-defined assets provide end-to-end **lineage** and **auditability**



HZJZ
HRVATSKI ZAVOD ZA JAVNO ZDRAVSTVO

Karlo Pintarić, MD, et al.

karlo.pintaric@hzjz.hr

EU-HIP
EU INTEROPERABILITY WITH HERA'S IT PLATFORM

OHDSI