Data Curation

Class 4 | 6th July 2024

Electronic Health Records and Data Structures

Stephen Blackwelder, PhD

Welcome Back!

Questions or Concerns?

Course Topic Outline

Getting Data In

1. EHRs and Clinical Records

   Origin and relevant history of the medical record; contemporary promise and problems

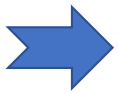
2. Clinical Decision Support

   Uses of medical record data to drive patient care

Storing, Finding, Retrieving

1. Data Structures and Liquidity

   Relational database structures and effective use of EHR data



2. Data Curation

   Raw data, refined data, and patient data from non-EHR sources

Turning Data into Insight

1. Leading Innovation

   Analytics strategy in healthcare organizations

2. Systemic Analytical Decision Making

   Designing an environment compatible with data-driven decision making

Individual Essay - What Did You Learn?

Taking a Practical Approach…

- How does business get ”done”?
- Your most persuasive statements are supported by *all the things you know, but do not say.*
- Leadership in technical fields requires *stronger* communication skills than in “pure” business fields.
- Capturing and holding audience attention is as important as your message.

Group Data Analytics Project

- Dataset has been provided
- Due July 31st; 30% of final grade
- Teams will share or divide up among yourselves these "research publication project" tasks:

  - Research question identification
  - Study design
  - Data management
  - Data analysis
  - Written Findings (2-5 pages)

- Should have begun data management already
- Email instructor if questions/concerns

Data Curation

Class 4 | 6th July 2024

Class 4 Learning Objectives

- Understand how data benefit from transformation, and where data management practices come from
- Understand the difference between structured and unstructured data, and why the difference matters.
- Consider how healthcare differs from other industries in the ability to use analytics on the data available (different challenges?)
- Explore motivations behind Quantified Self and understand how this movement has influenced healthcare, and vice-versa. Understand the challenges encountered in leveraging IoT data.
- Be able to speak to several challenges in data management along the continuum from "raw" EHR data to "live" IoT data to highly curated claims and data warehouse datasets.

Raw vs Curated Data

- *Understand how data benefit from transformation, and where data management practices come from*

❖ What insights have you gained from working with the synthetic data for your group project?

Structured vs Unstructured

- *Understand the difference between structured and unstructured data, and why the difference matters.*

❖ Any group adding in structured or unstructured data to their project?

❖ Other experience with unstructured data outside class?

- *Consider how healthcare differs from other industries in the ability to use analytics*

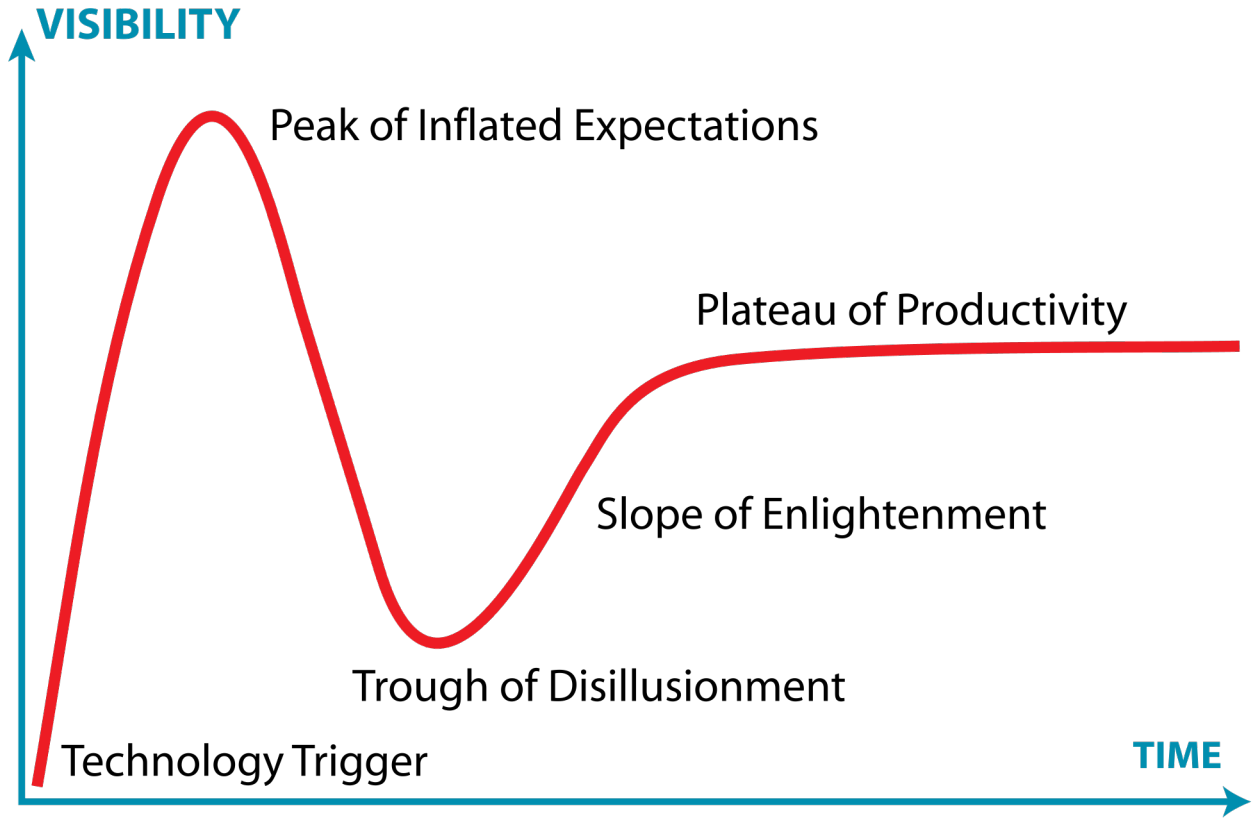
”Patient Wearables” and Other IOT-Generated Data

- Explore motivations behind Quantified Self and understand how this movement has influenced healthcare, and vice-versa.

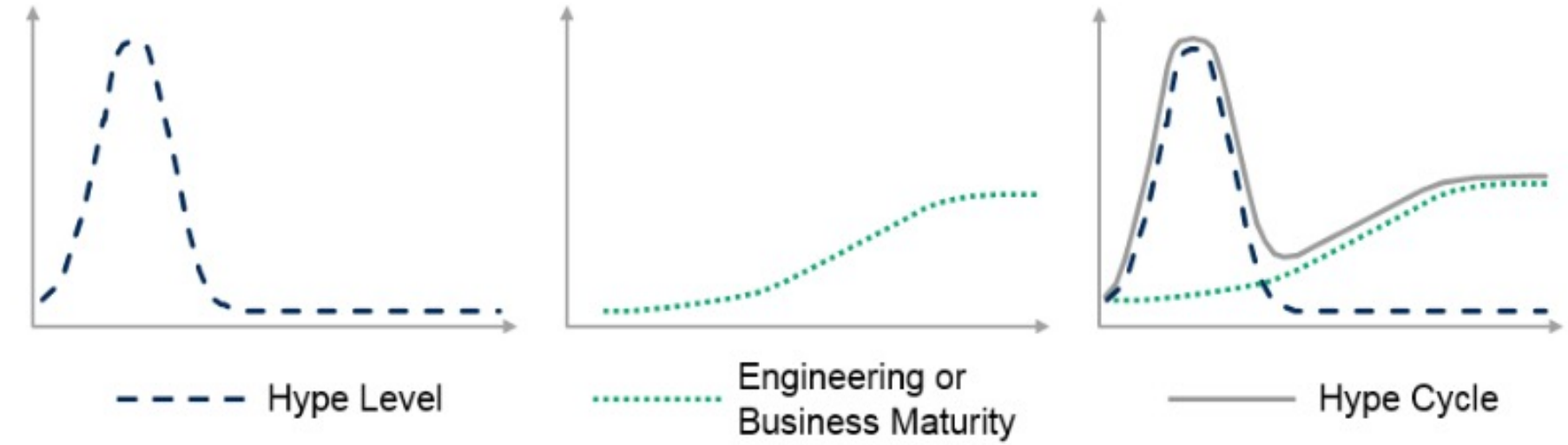
❖ What challenges are encountered in leveraging IoT data?

❖ Is the Mayo Data Trust a persuasive solution to these challenges?

The Hype Cycle

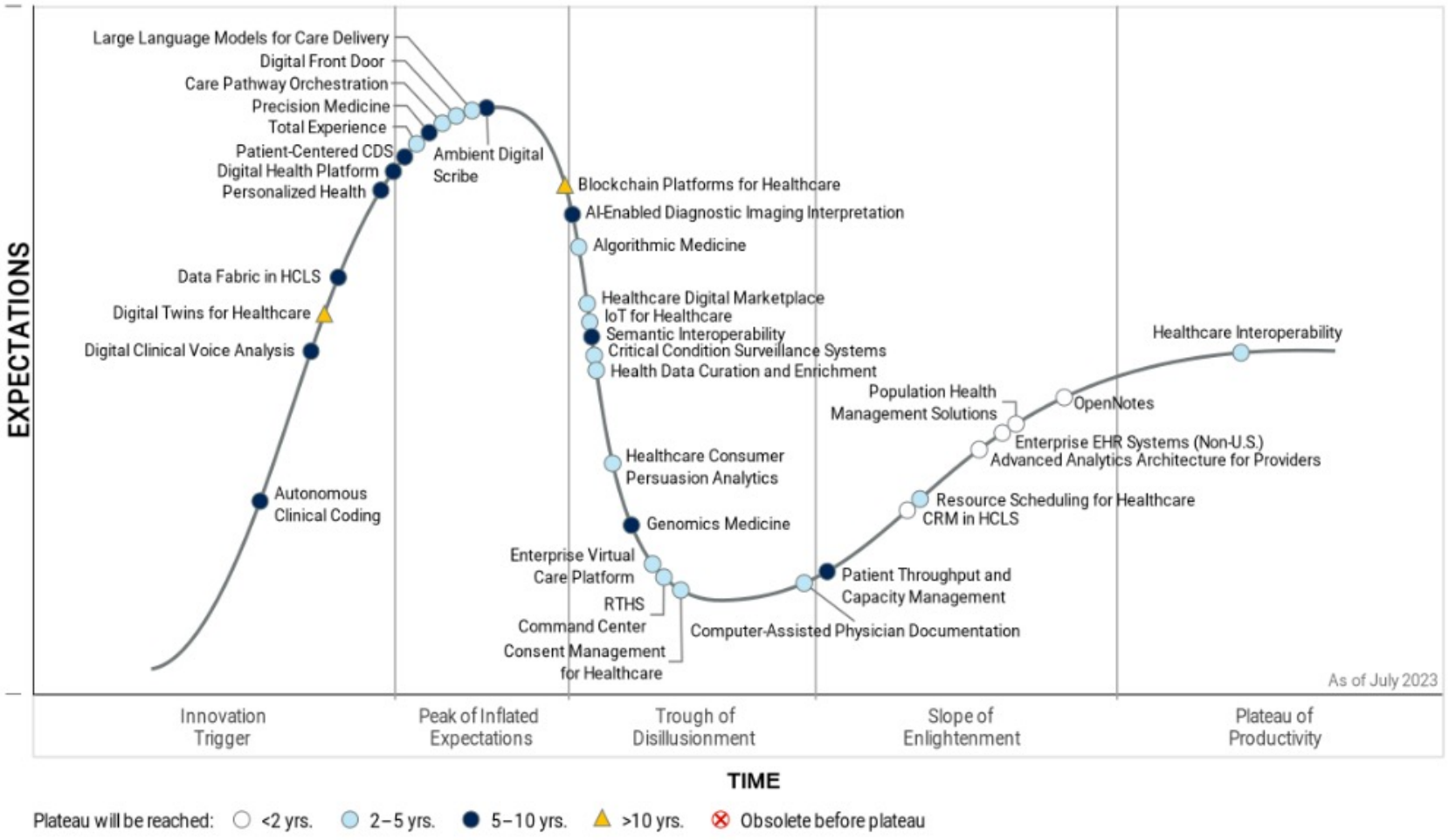


Hype Cycle Components



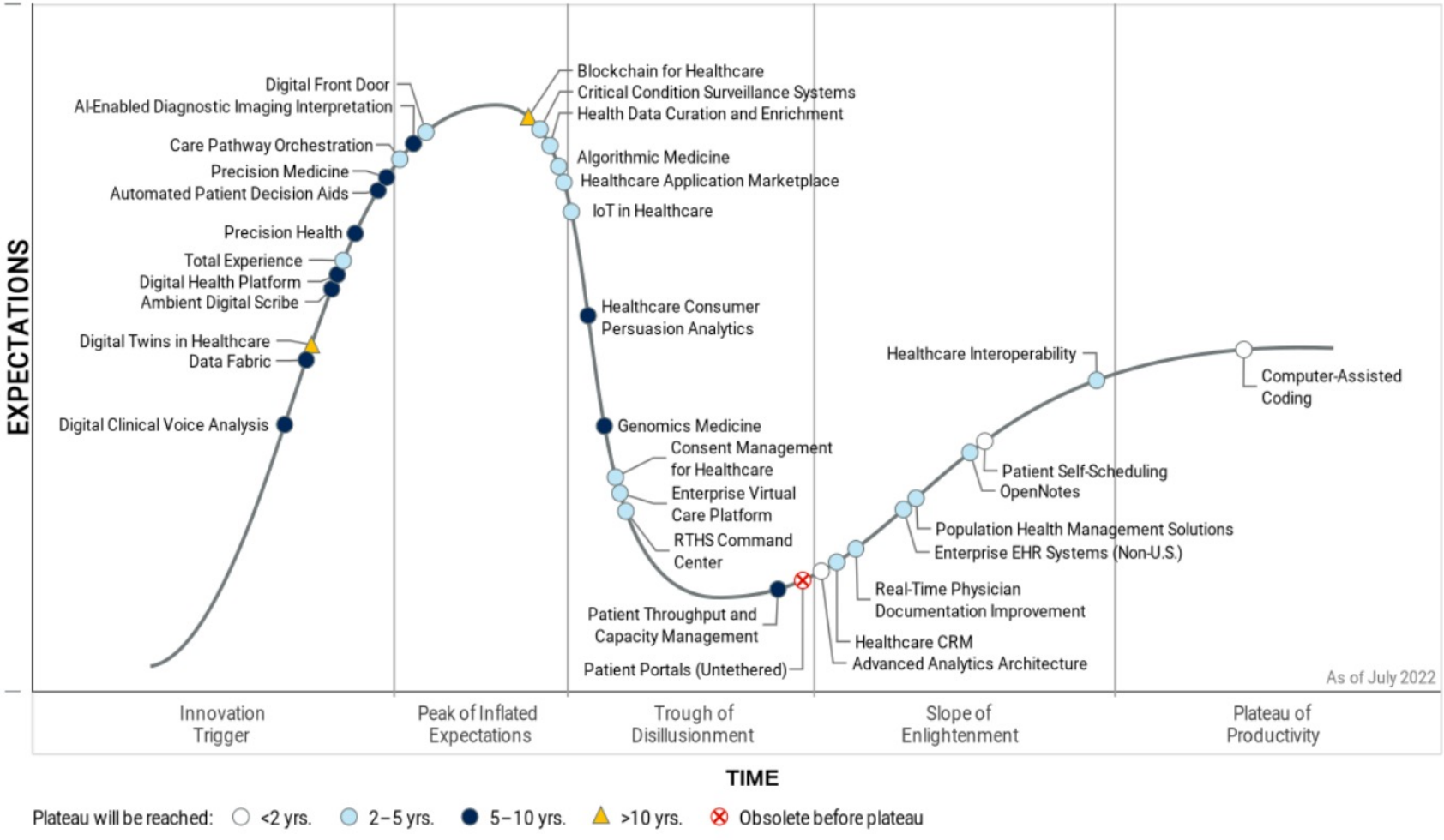
ID: 370163

Hype Cycle for Healthcare Providers, 2023



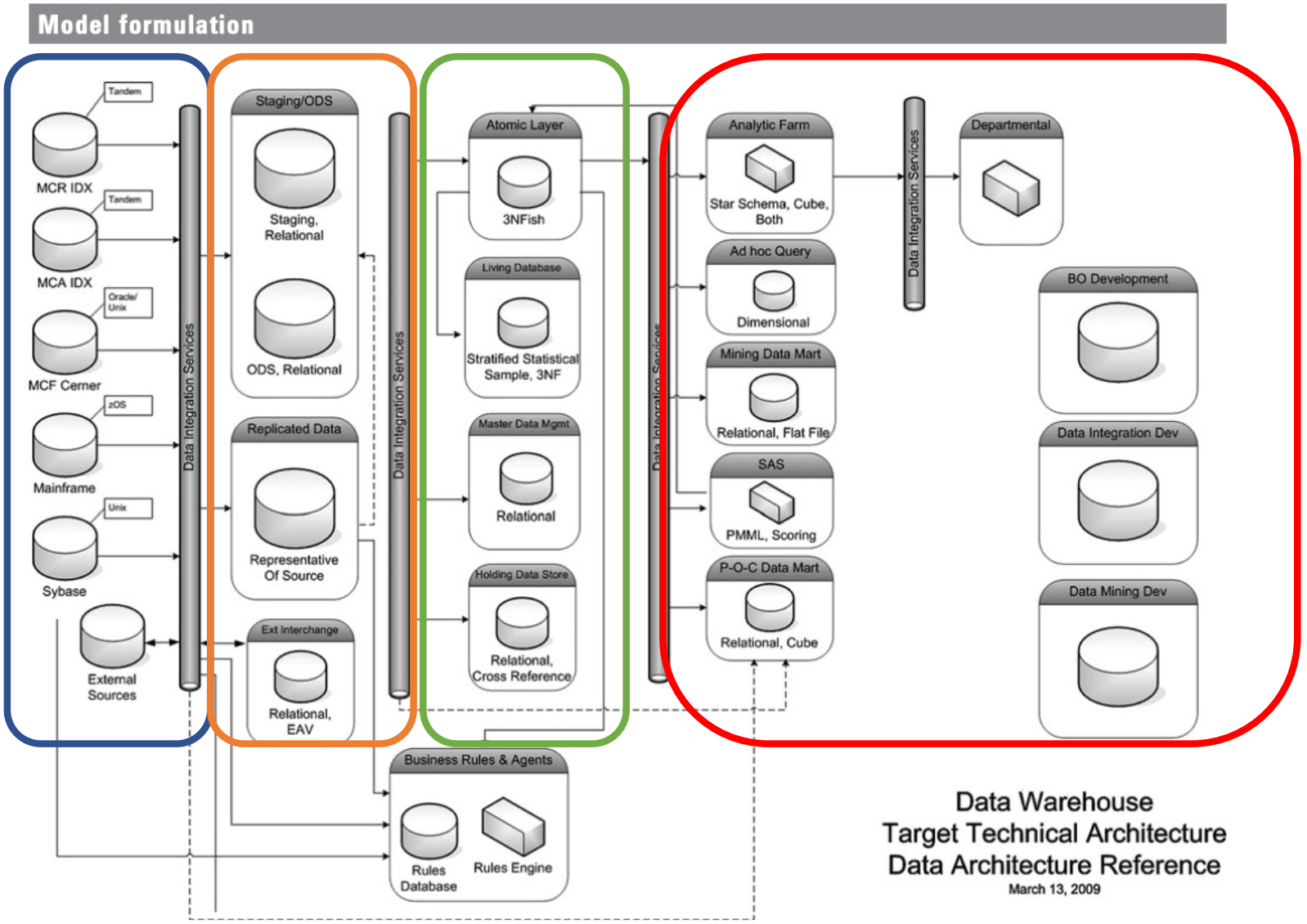
Gartner

Hype Cycle for Healthcare Providers, 2022

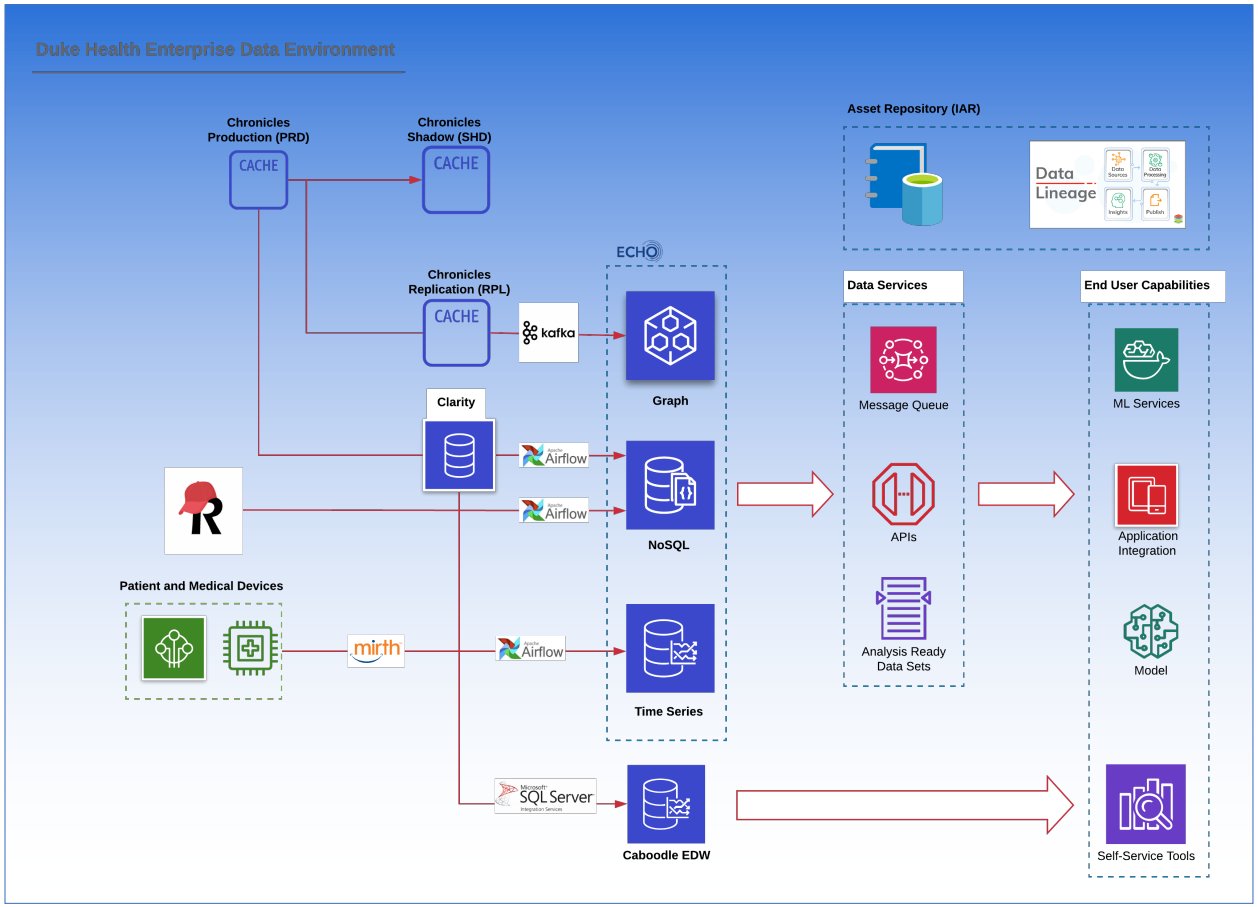


Gartner

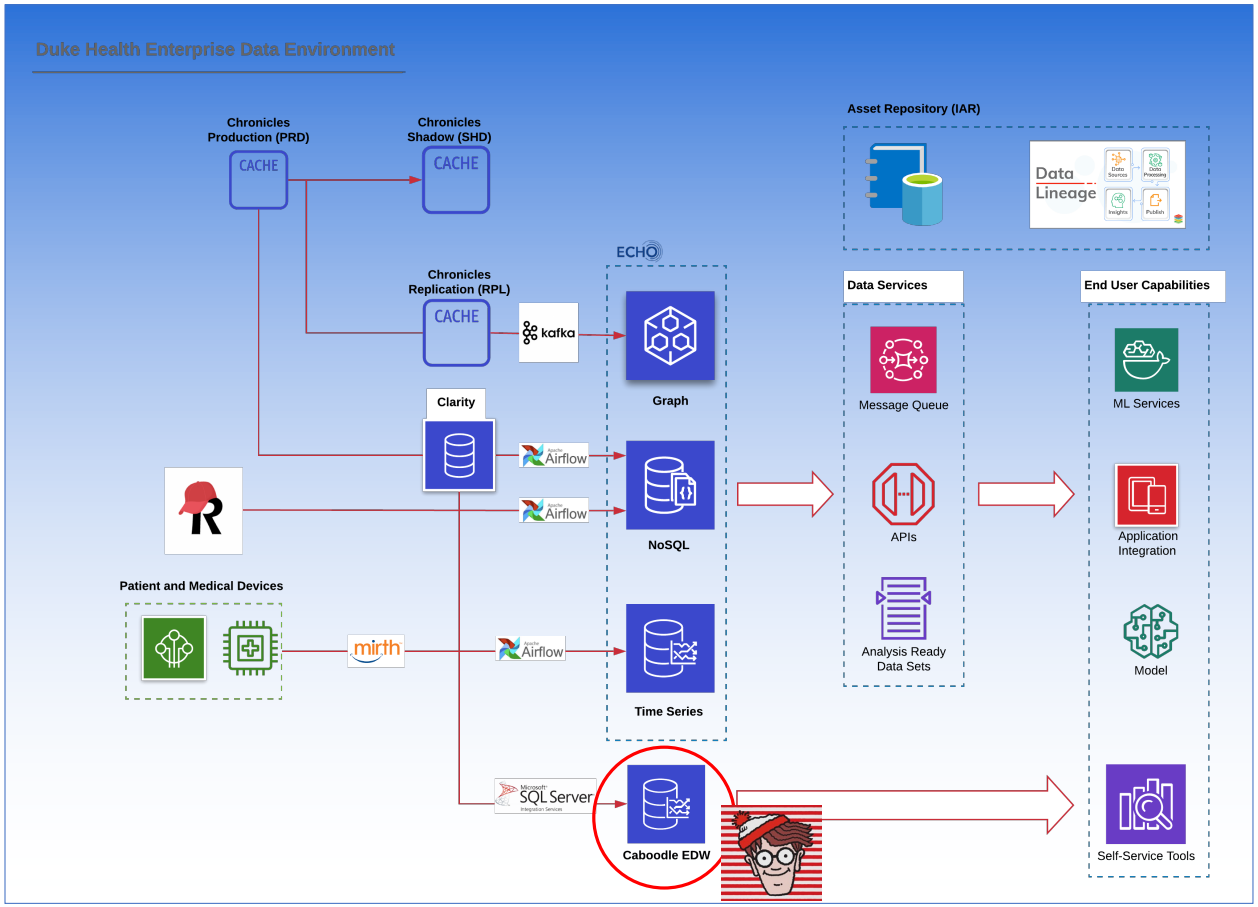
Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *Journal of the American Medical Informatics Association*. 2010;17(2):131-135. doi:10.1136/jamia.2009.002691



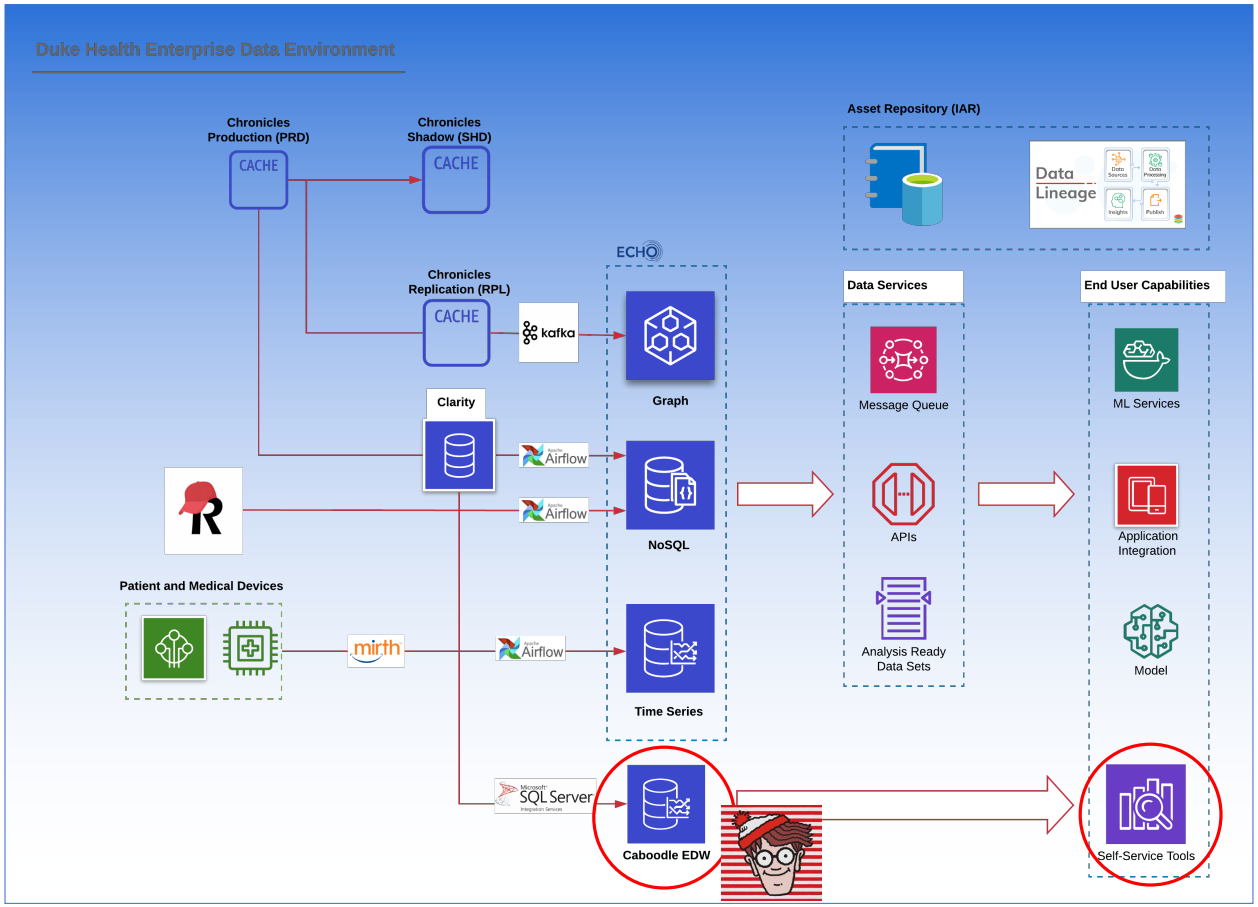
Where's the Warehouse?



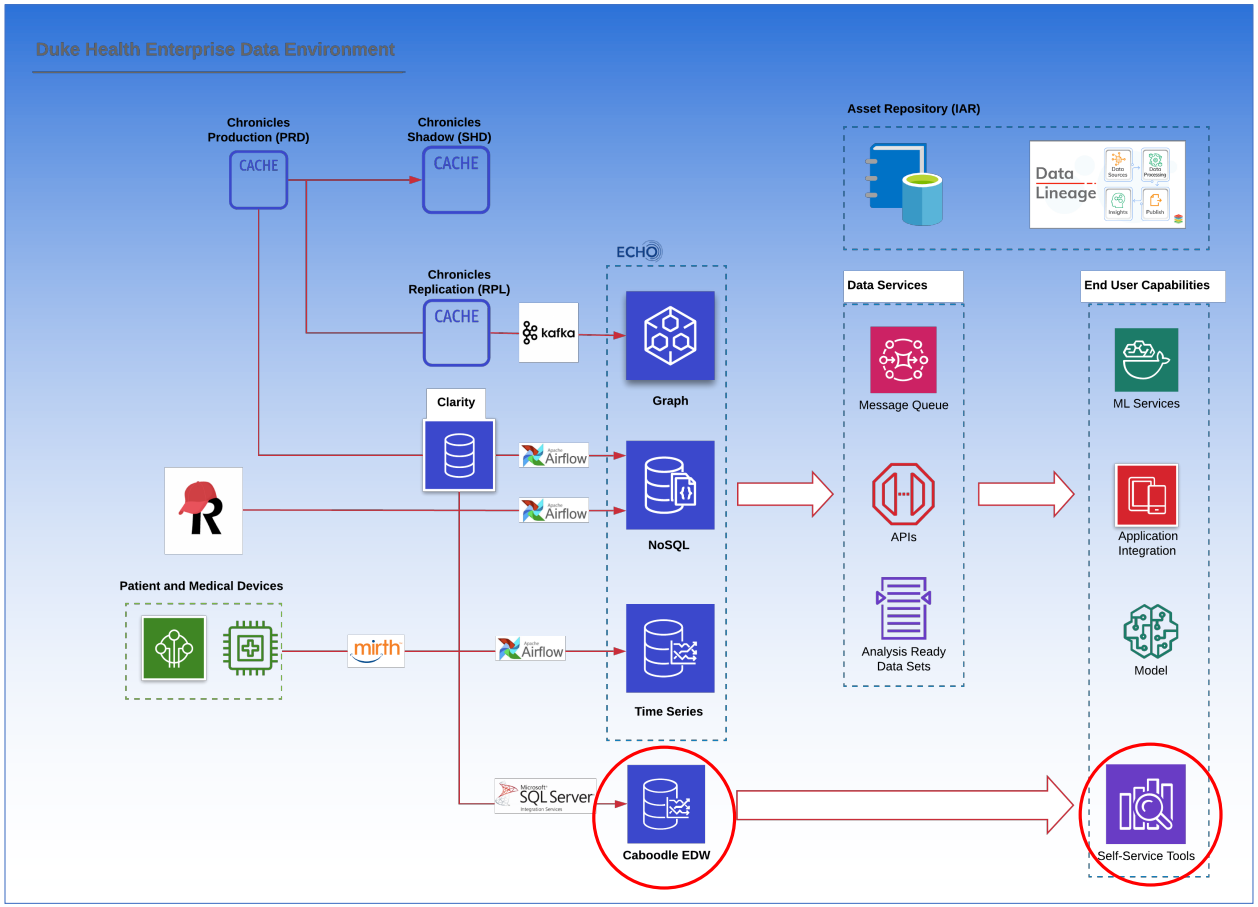
Where's the Warehouse?



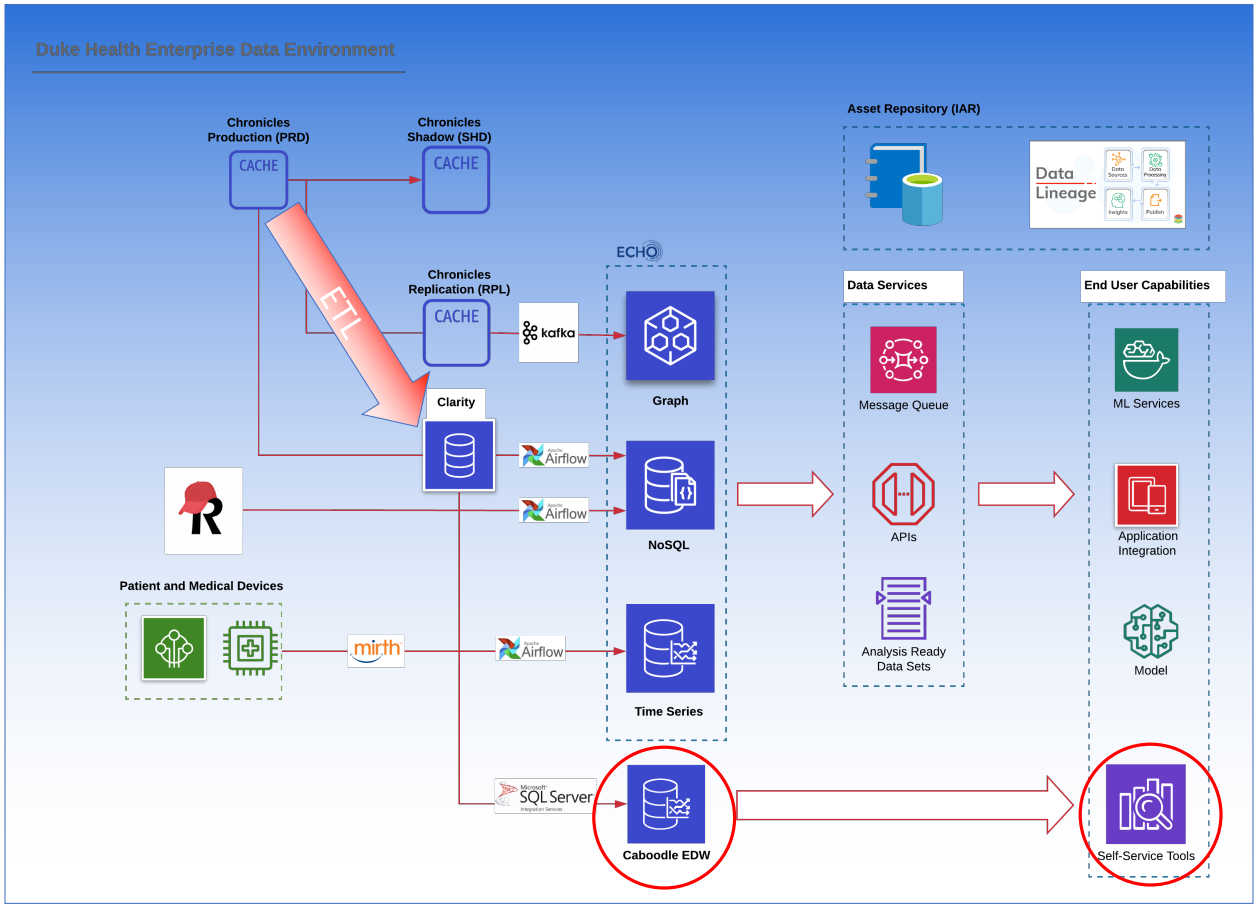
Where's the Warehouse?



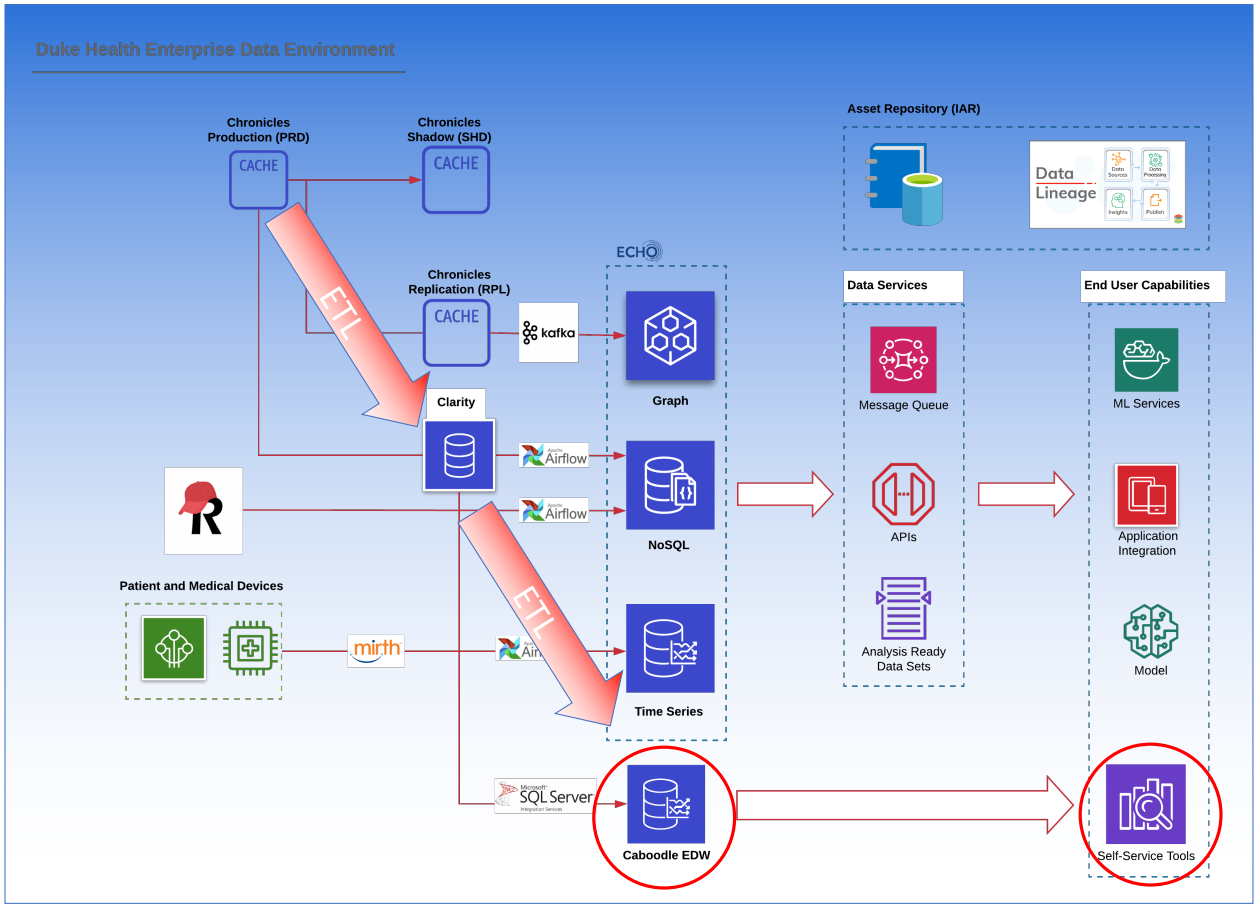
Where's the Warehouse?



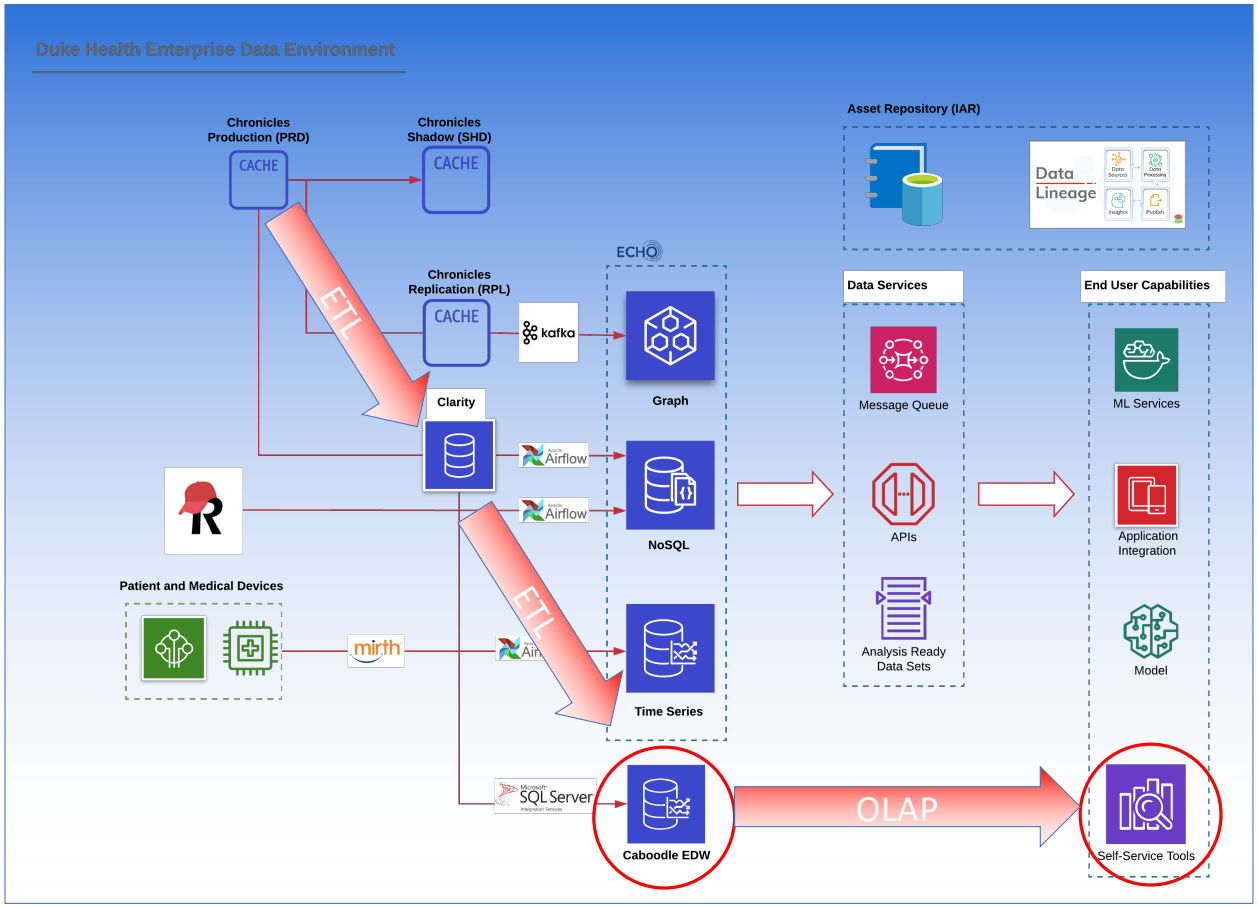
Where's the Warehouse?

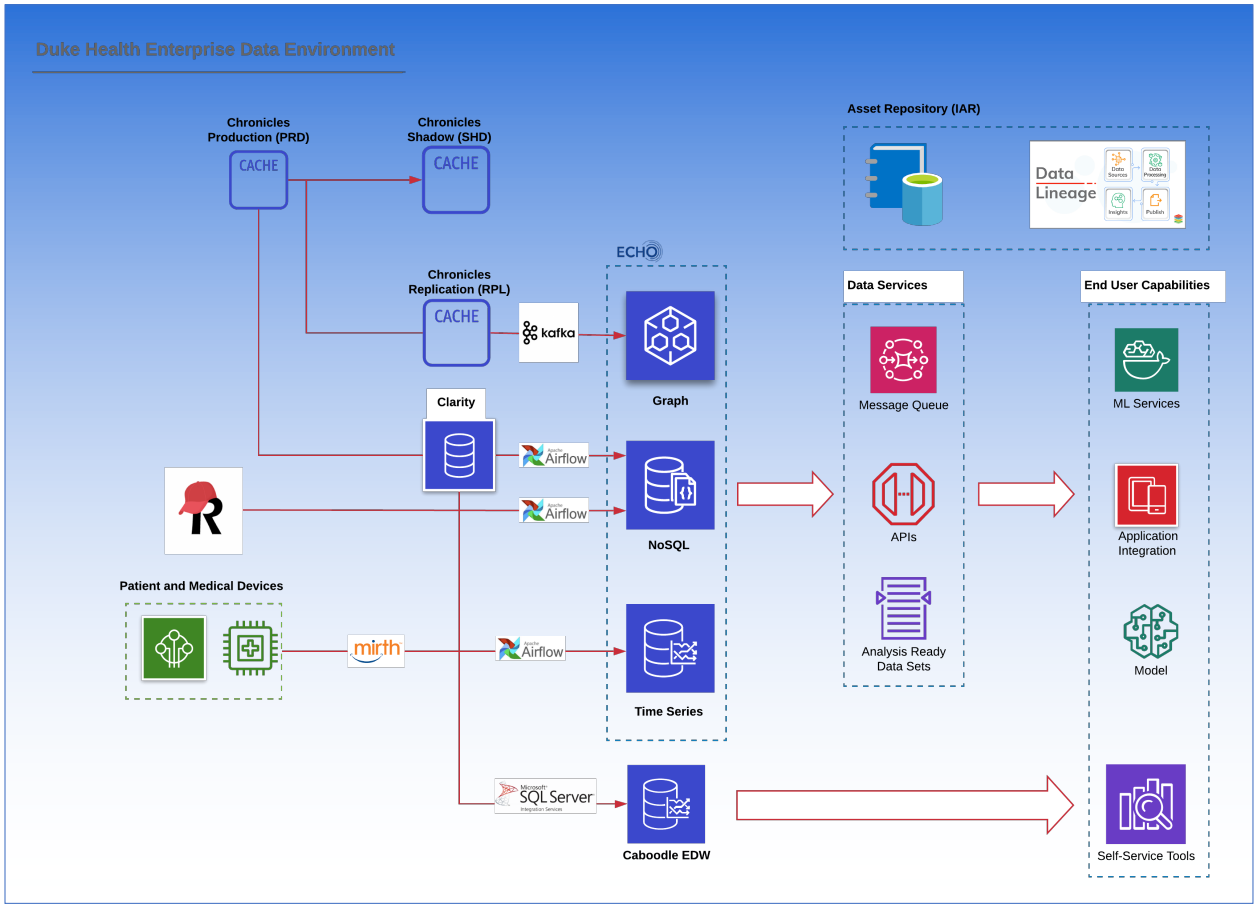


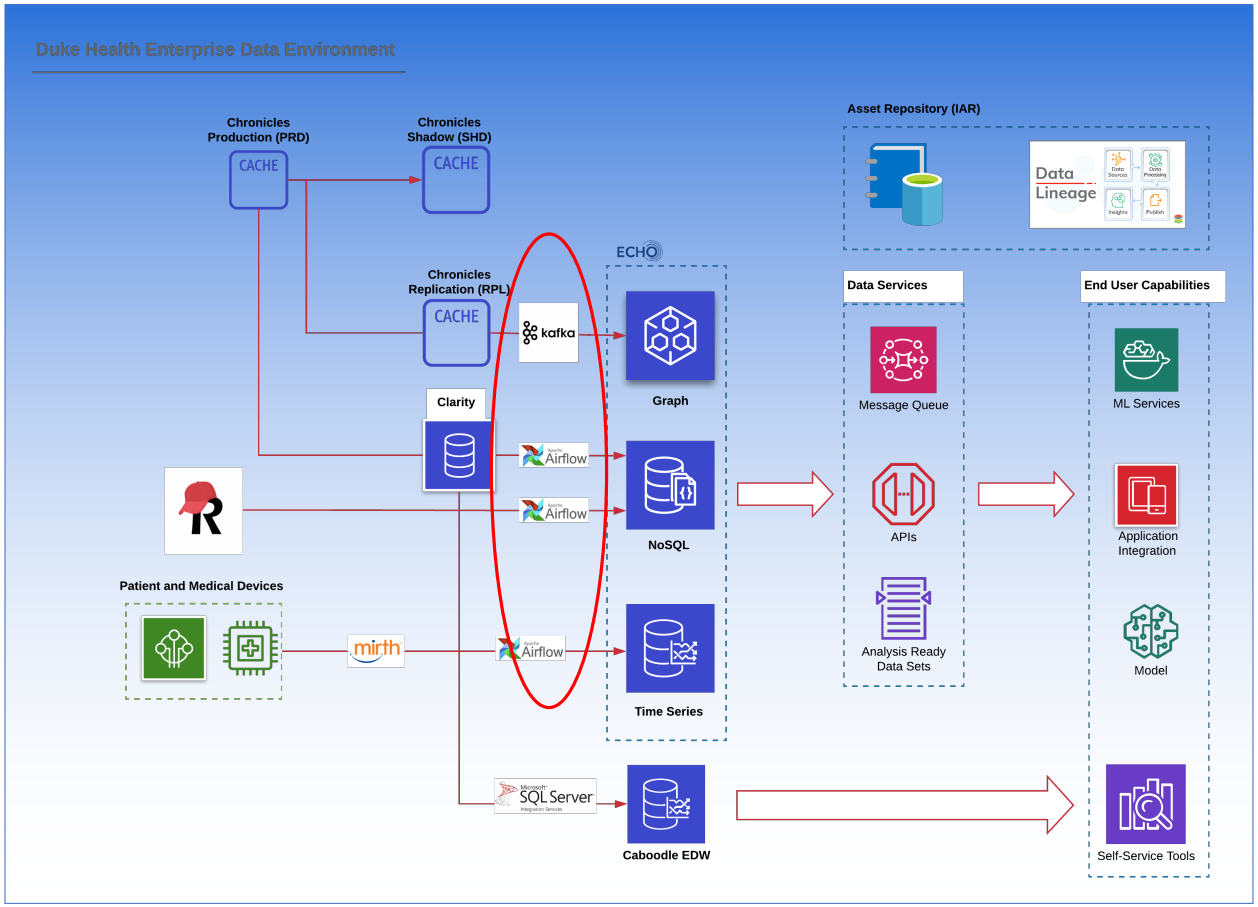
Where's the Warehouse?



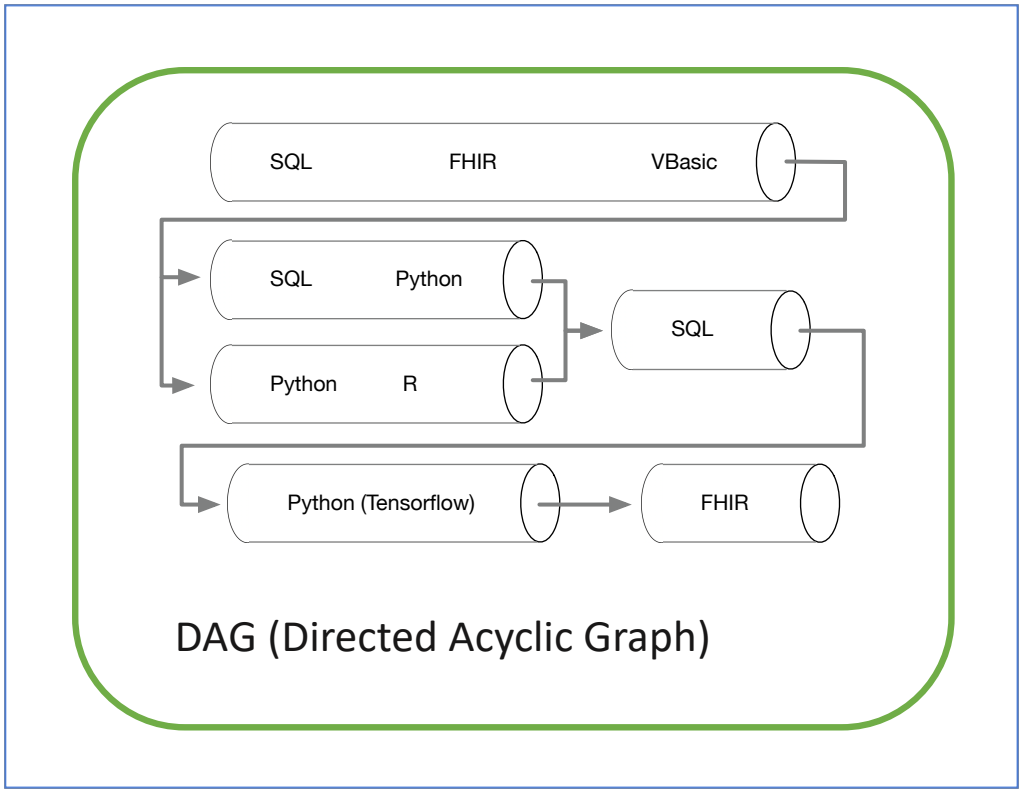
Where's the Warehouse?




Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Chronicles Replication (RPL)
CACHE
kafka
Clarity
Airflow
Airflow
Patient and Medical Devices
mirth
Airflow
Microsoft SQL Server
ECHO
Graph
NoSQL
Time Series
Caboodle EDW
Asset Repository (IAR)
Data Lineage
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools

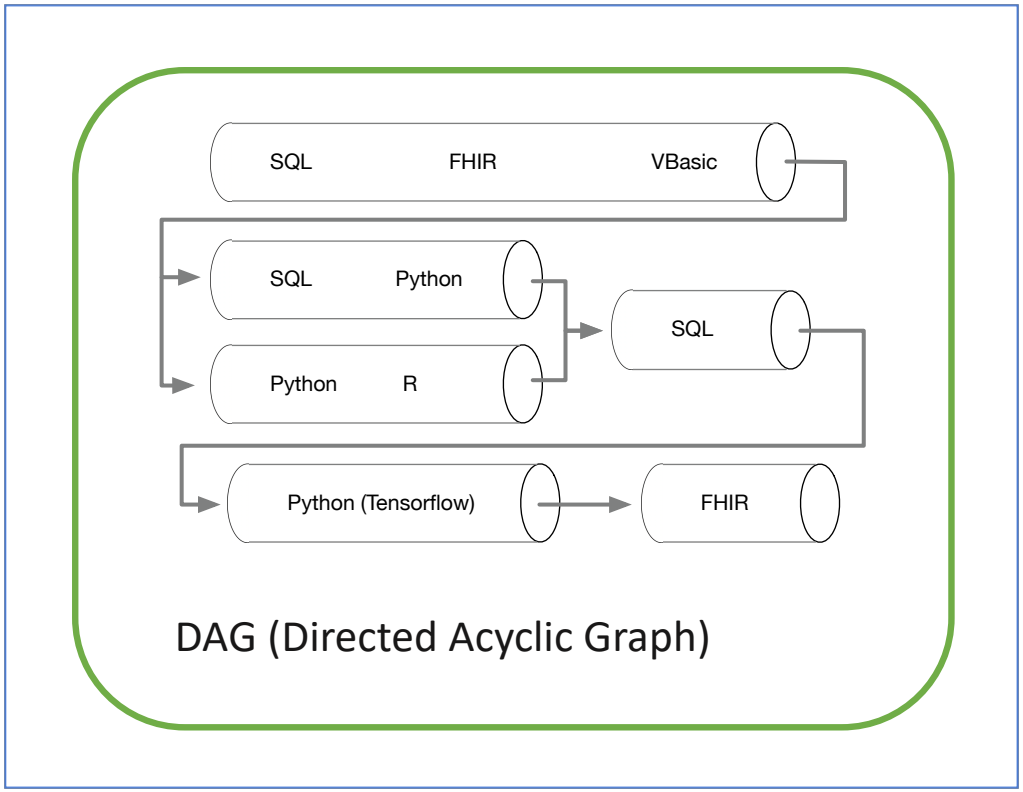

Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Chronicles Replication (RPL)
CACHE
kafka
Clarity
Airflow
Airflow
Airflow
Patient and Medical Devices
mirth
Microsoft SQL Server
ECHO
Graph
NoSQL
Time Series
Caboodle EDW
Asset Repository (IAR)
Data Lineage
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools

Anatomy of a “Data Pipeline”



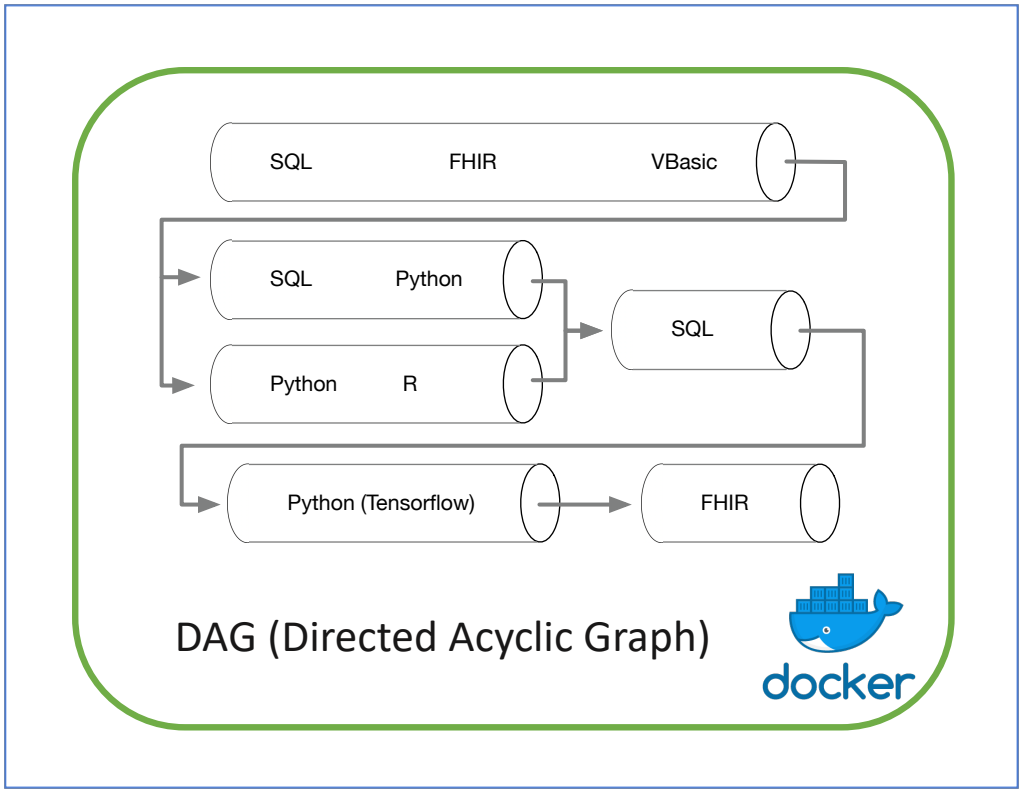
Anatomy of a “Data Pipeline”

Airflow



Anatomy of a “Data Pipeline”

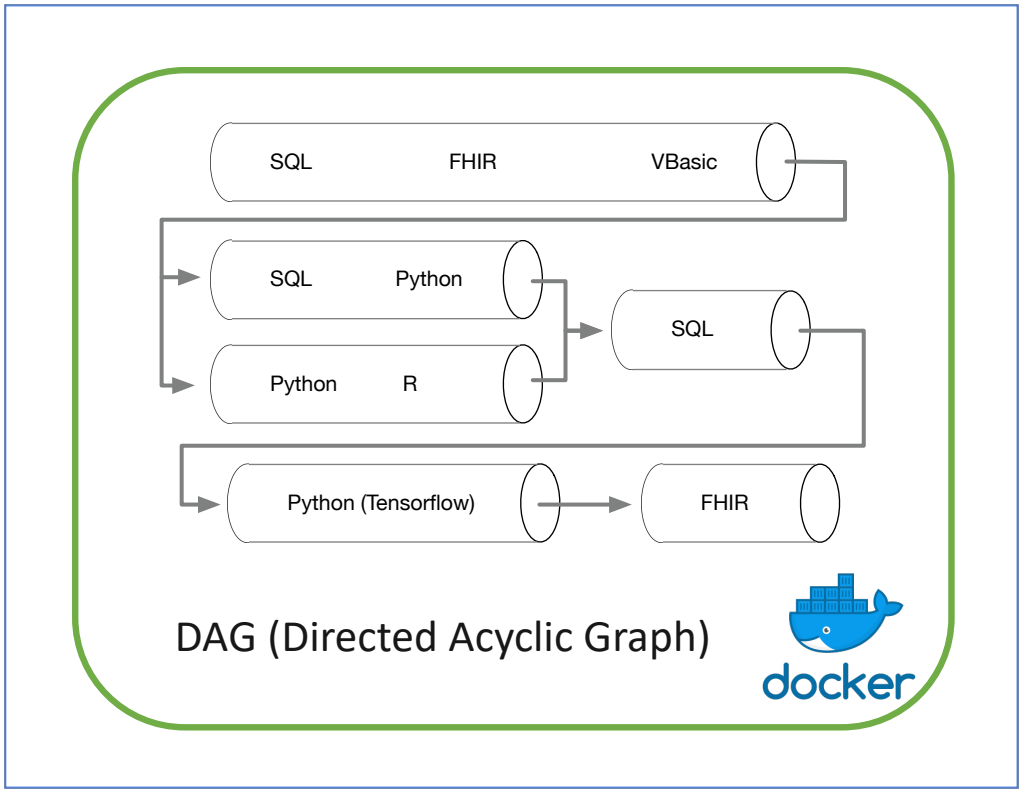
Airflow



Anatomy of a “Data Pipeline”

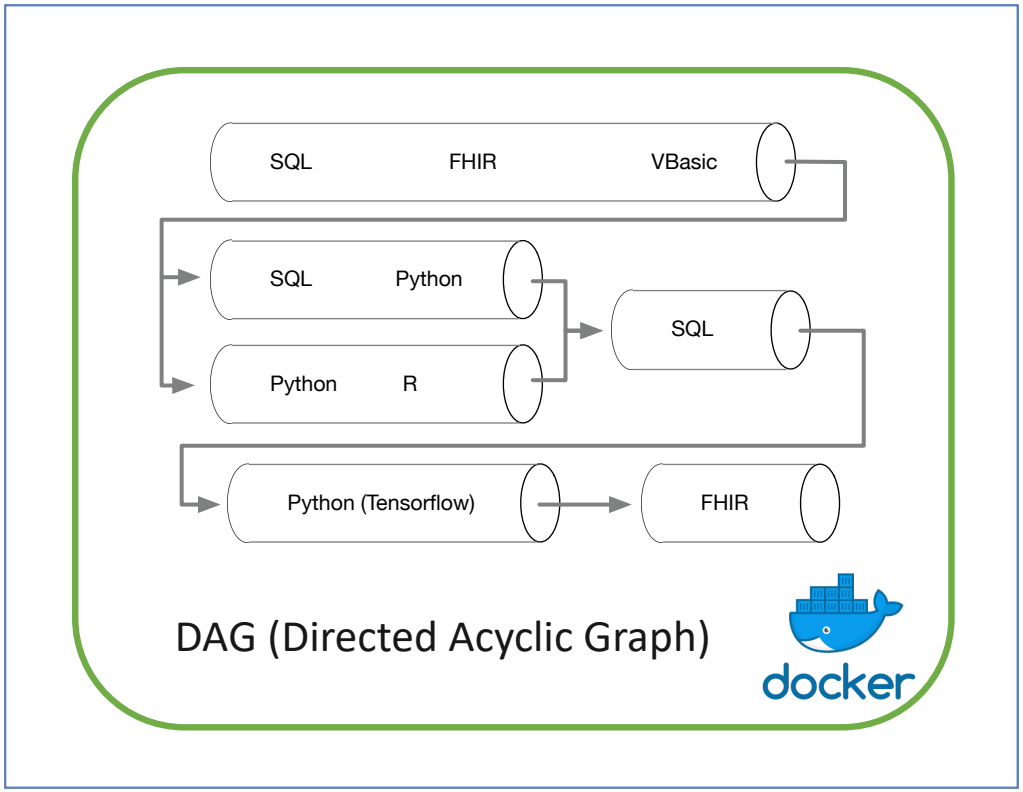
Airflow

kubernetes



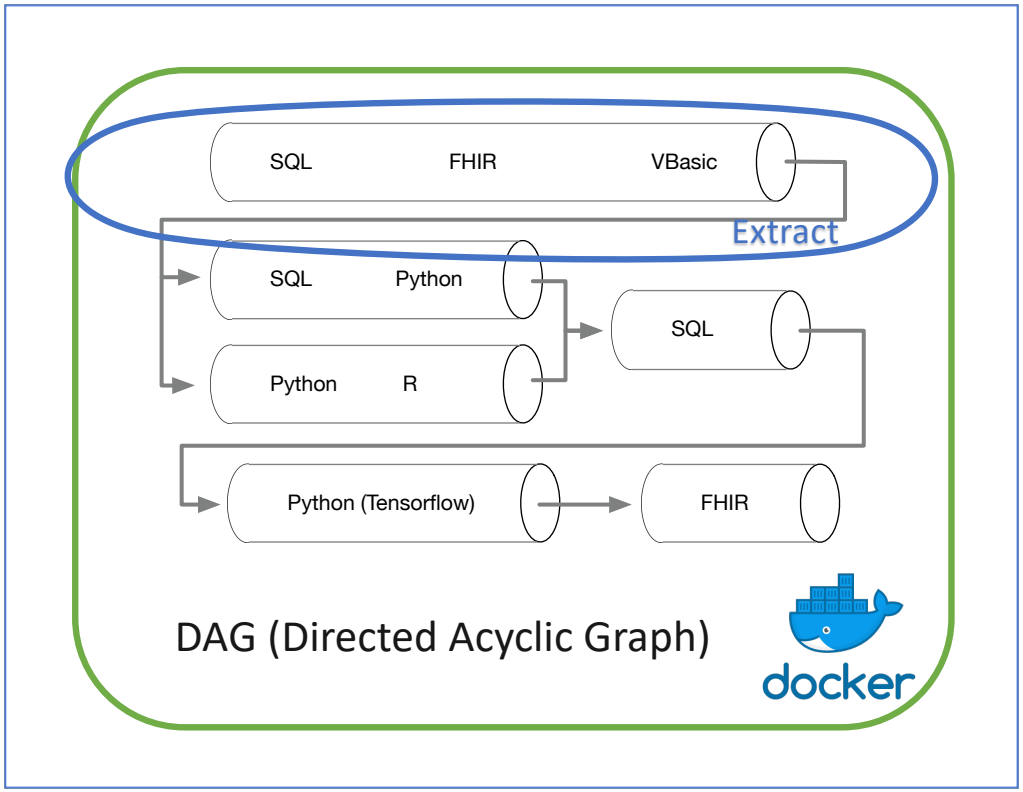
Anatomy of a “Data Pipeline”

Metric definition can be customized right up to the last moment before delivery.



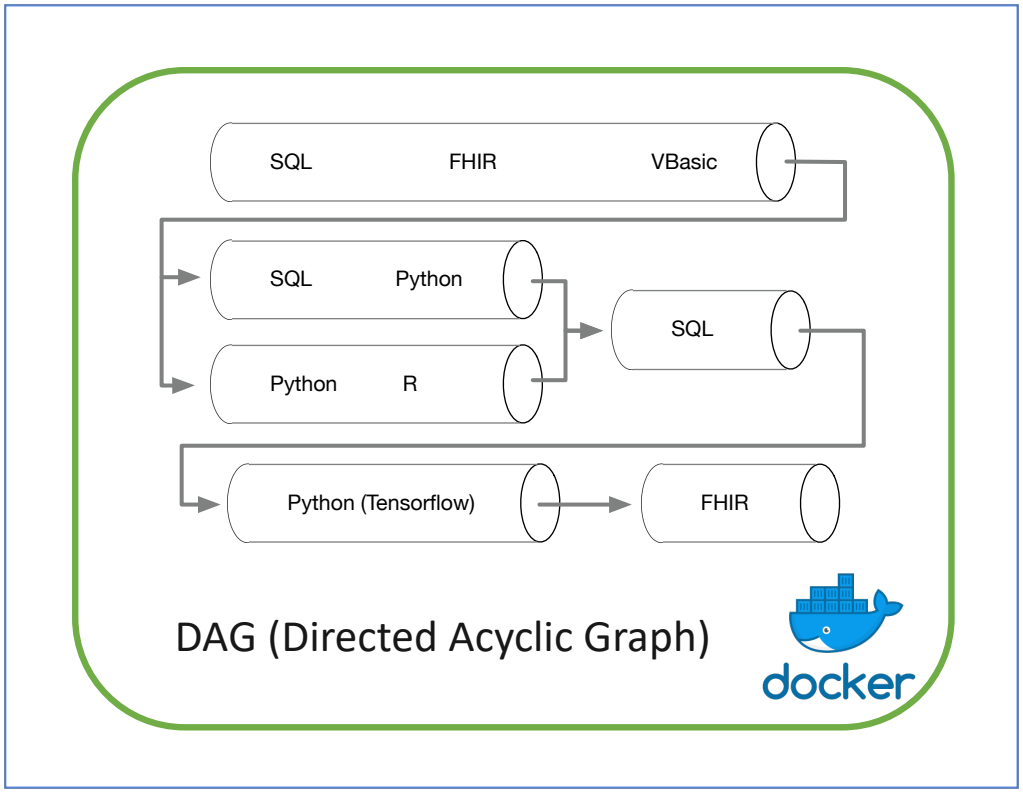
Anatomy of a “Data Pipeline”

Metric definition can be customized right up to the last moment before delivery.



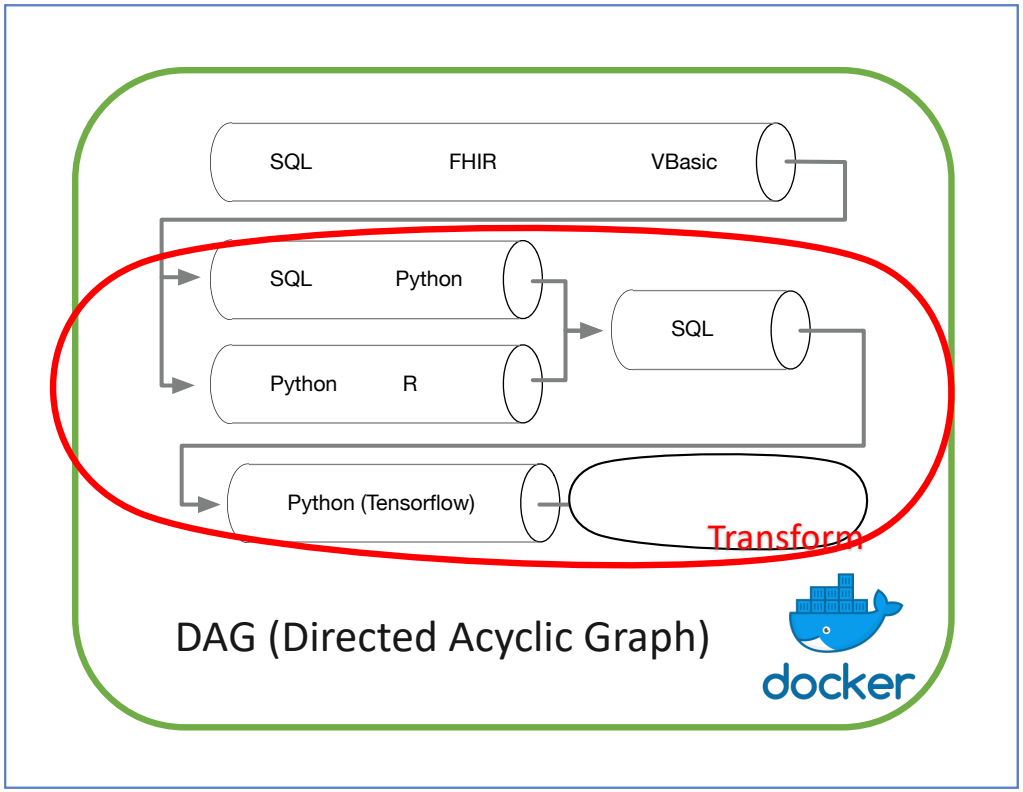
Anatomy of a “Data Pipeline”

Metric definition can be customized right up to the last moment before delivery.



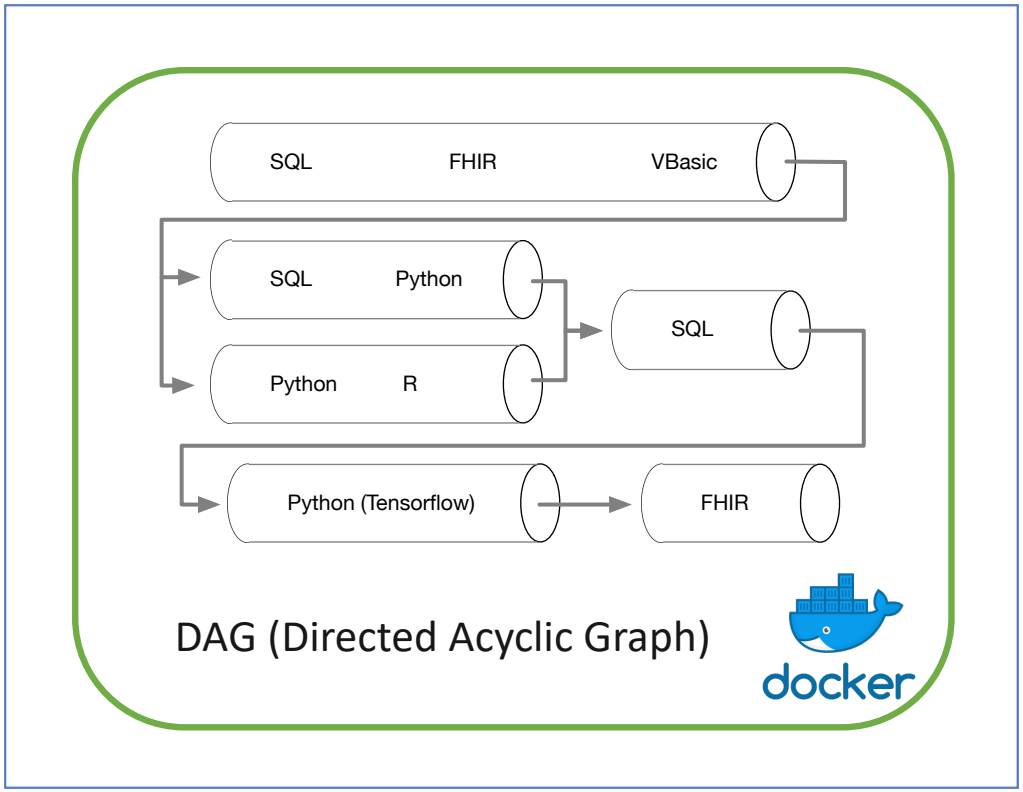
Anatomy of a “Data Pipeline”

Metric definition can be customized right up to the last moment before delivery.



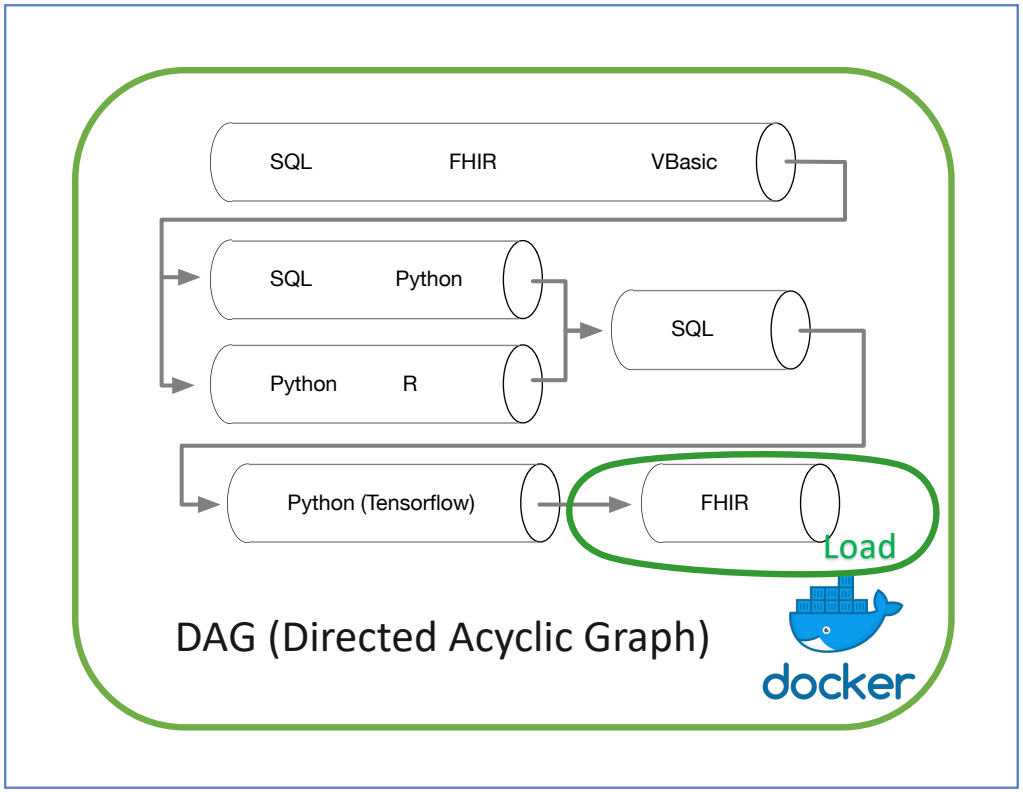
Anatomy of a “Data Pipeline”

Metric definition can be customized right up to the last moment before delivery.



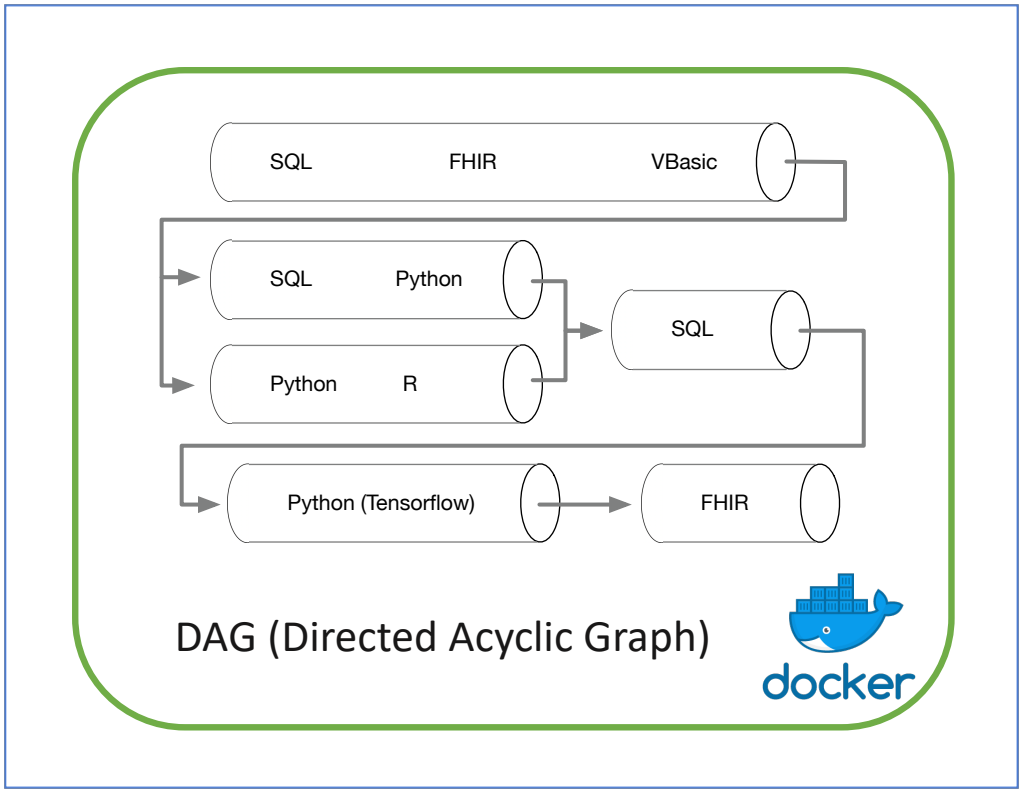
Anatomy of a “Data Pipeline”

Metric definition can be customized right up to the last moment before delivery.



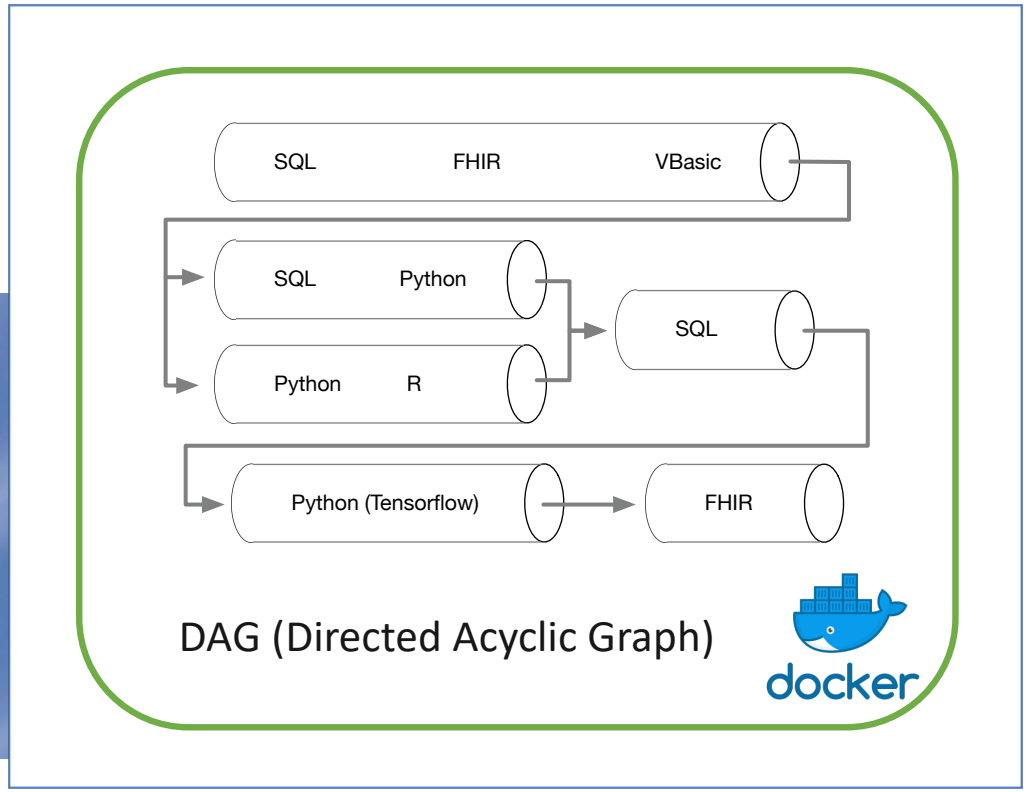
Anatomy of a “Data Pipeline”

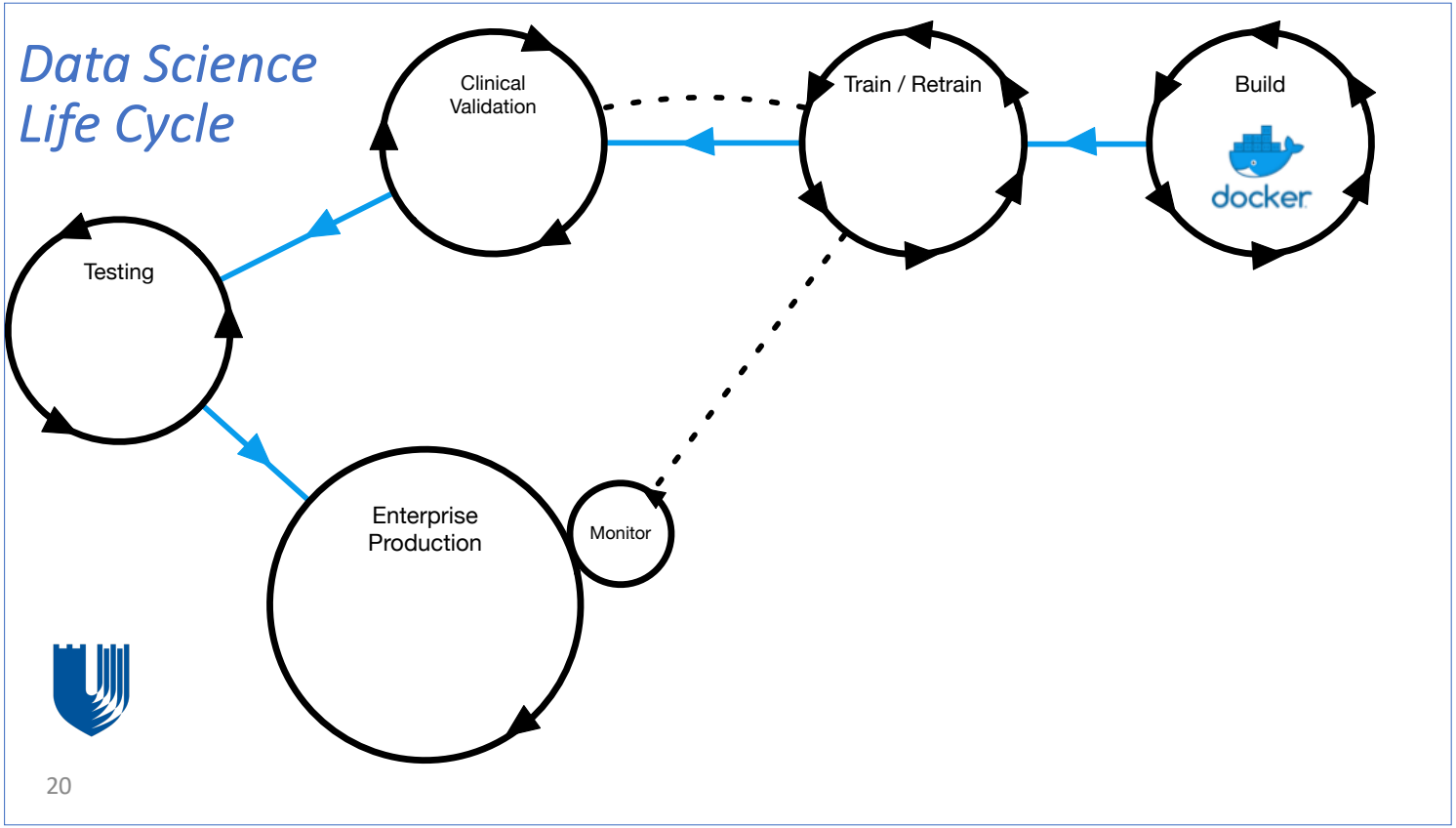
Metric definition can be customized right up to the last moment before delivery.

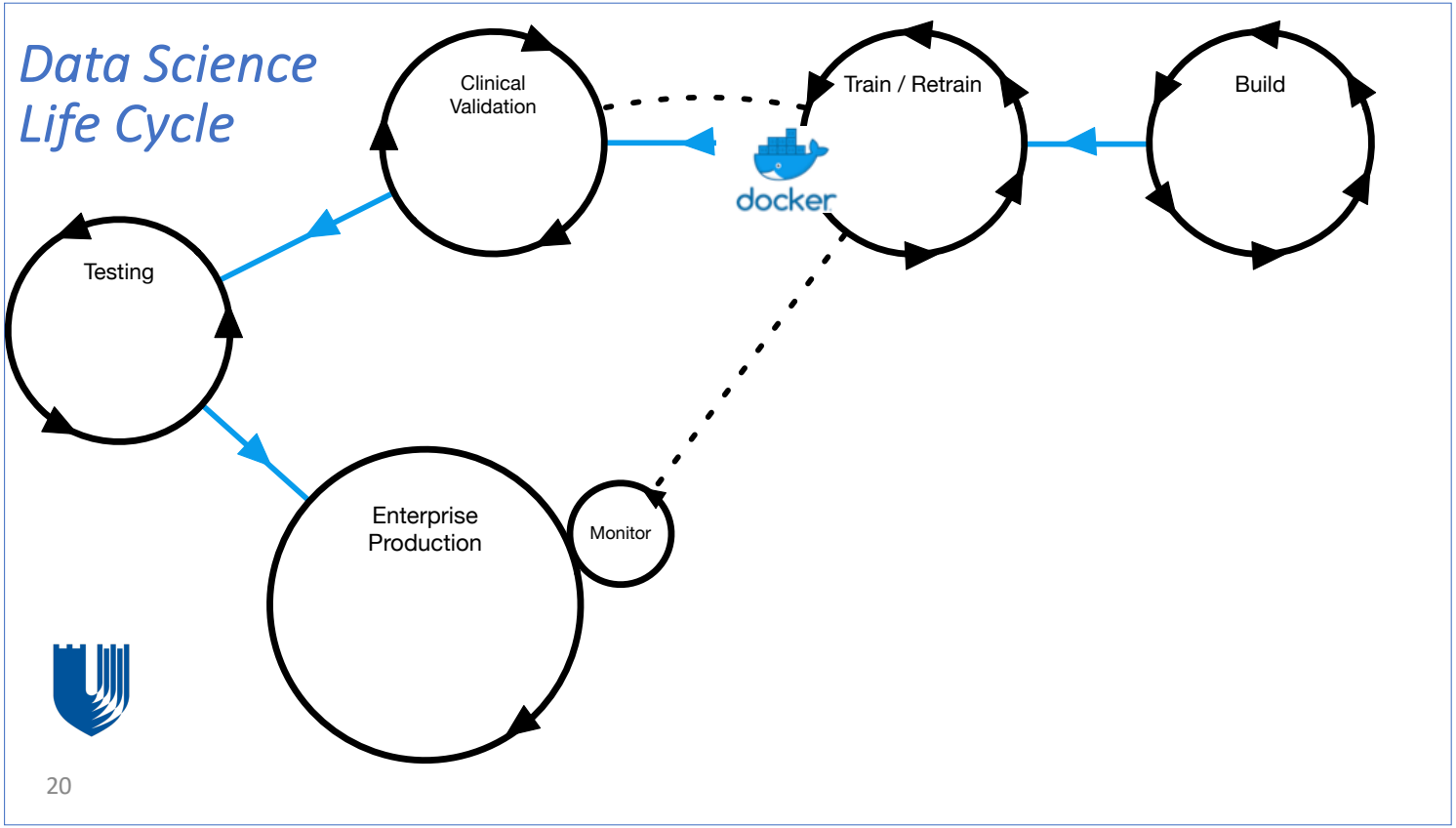


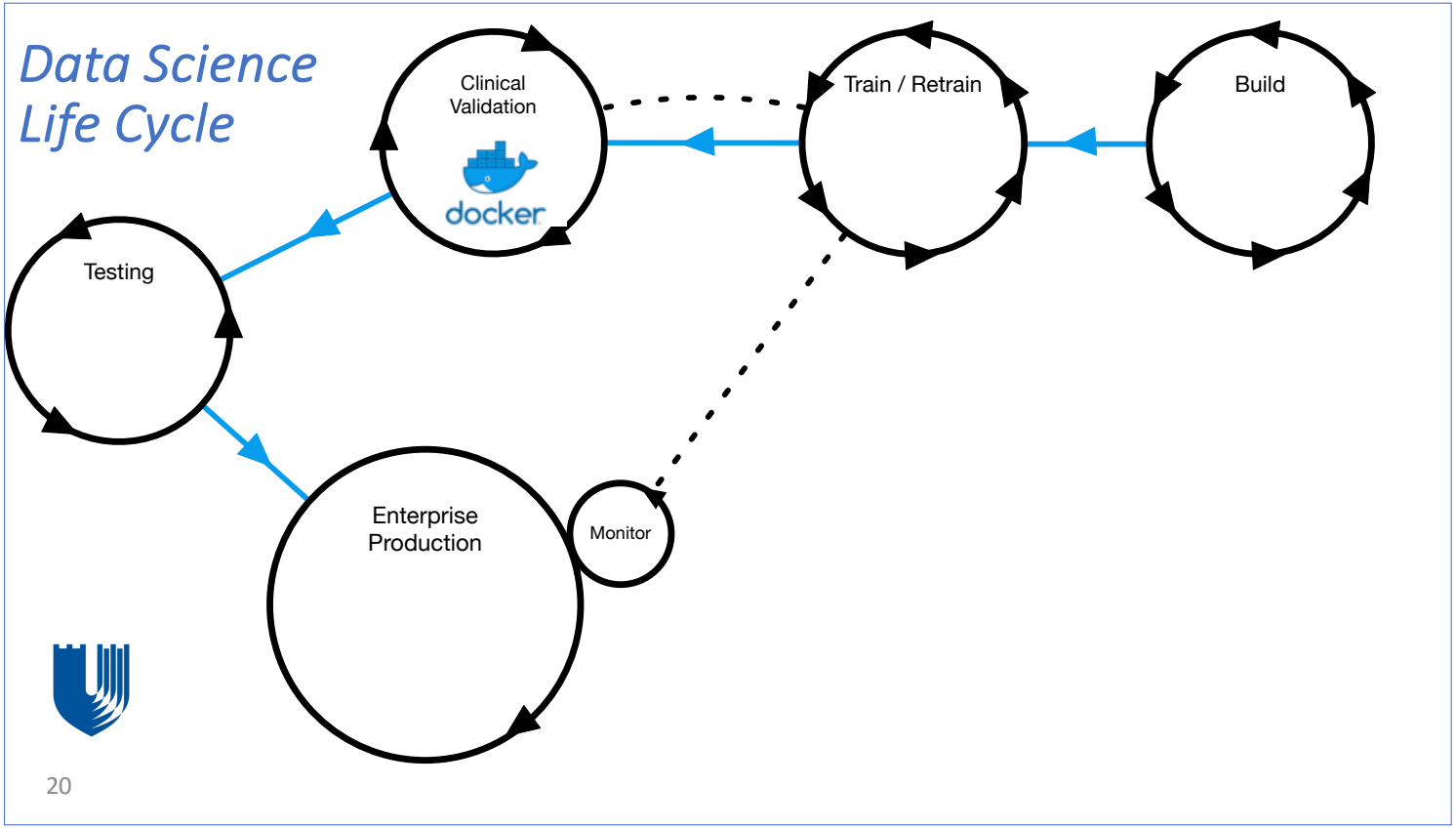
Anatomy of a “Data Pipeline”

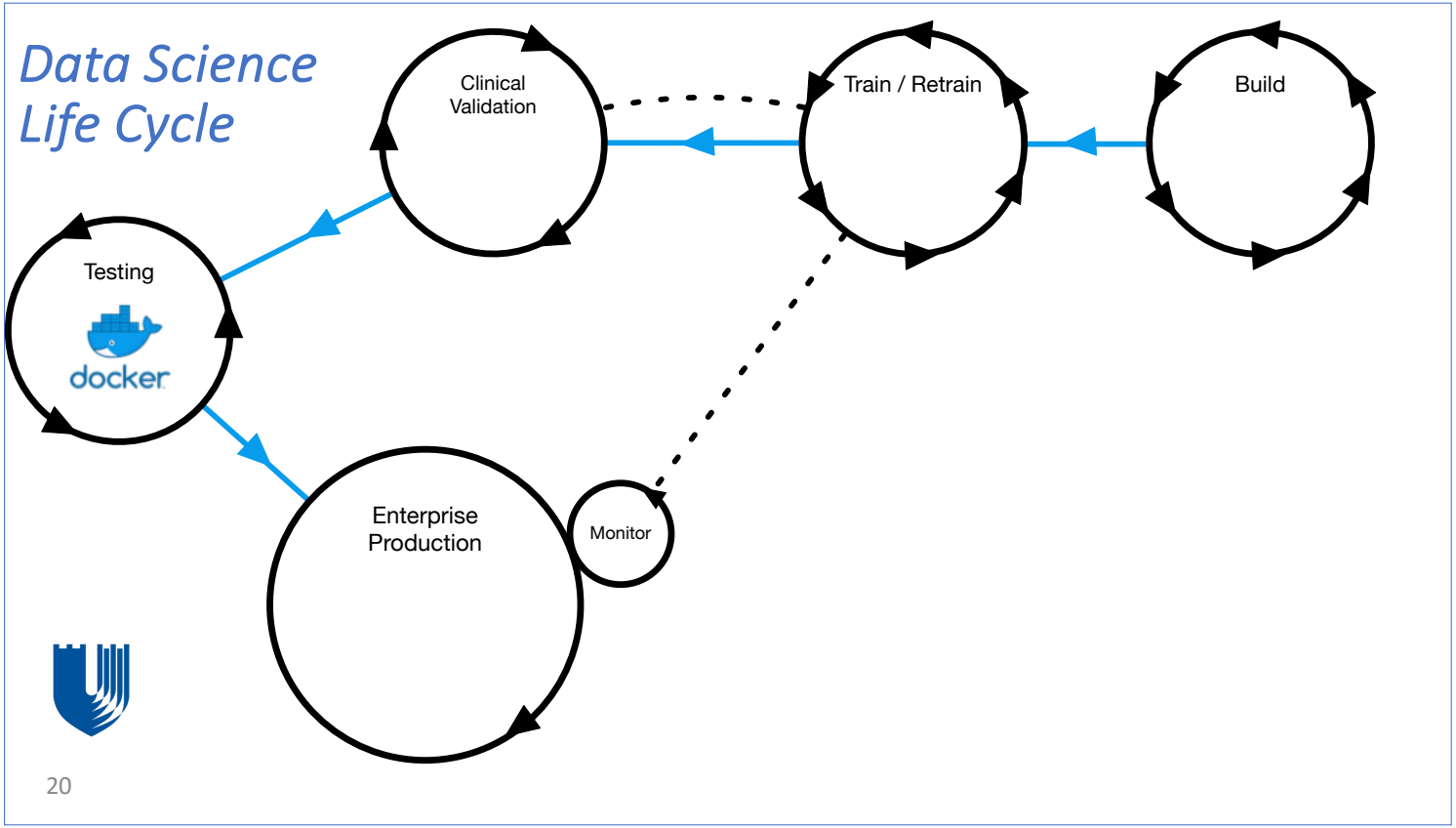
Metric definition can be customized right up to the last moment before delivery.

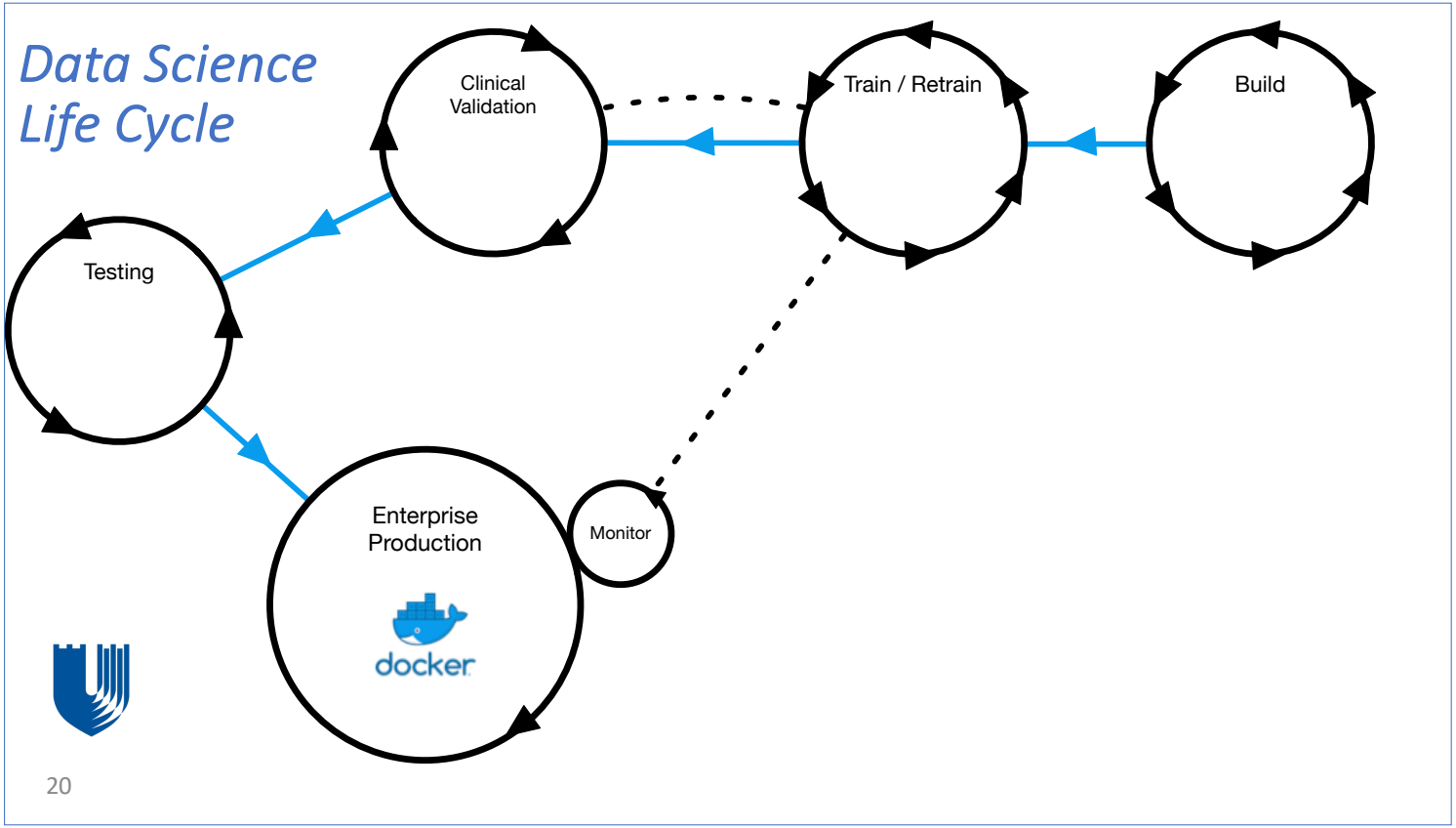


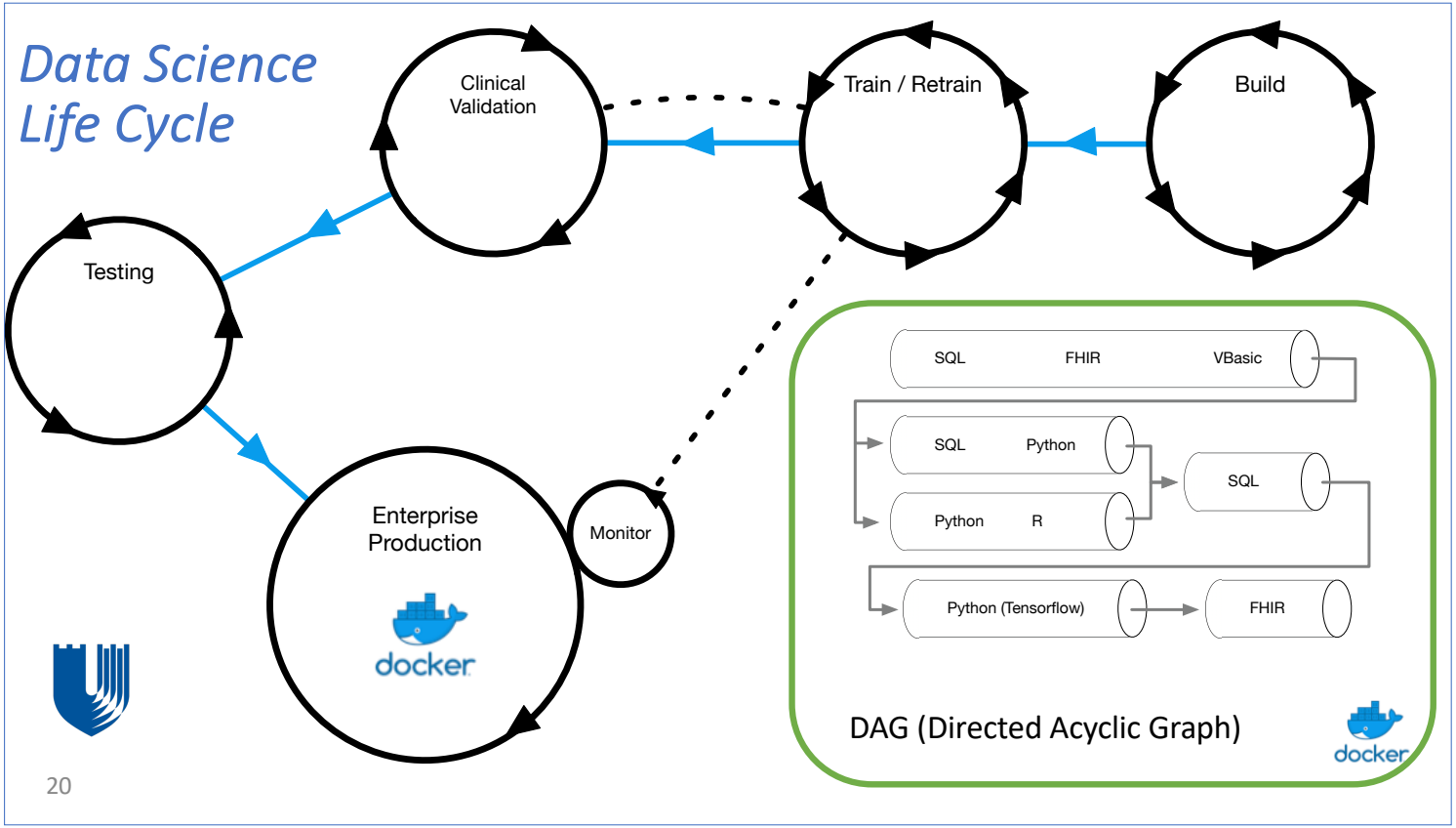

Data Science
Life Cycle
Clinical
Validation
Train / Retrain
Build
docker
Testing
Enterprise
Production
Monitor

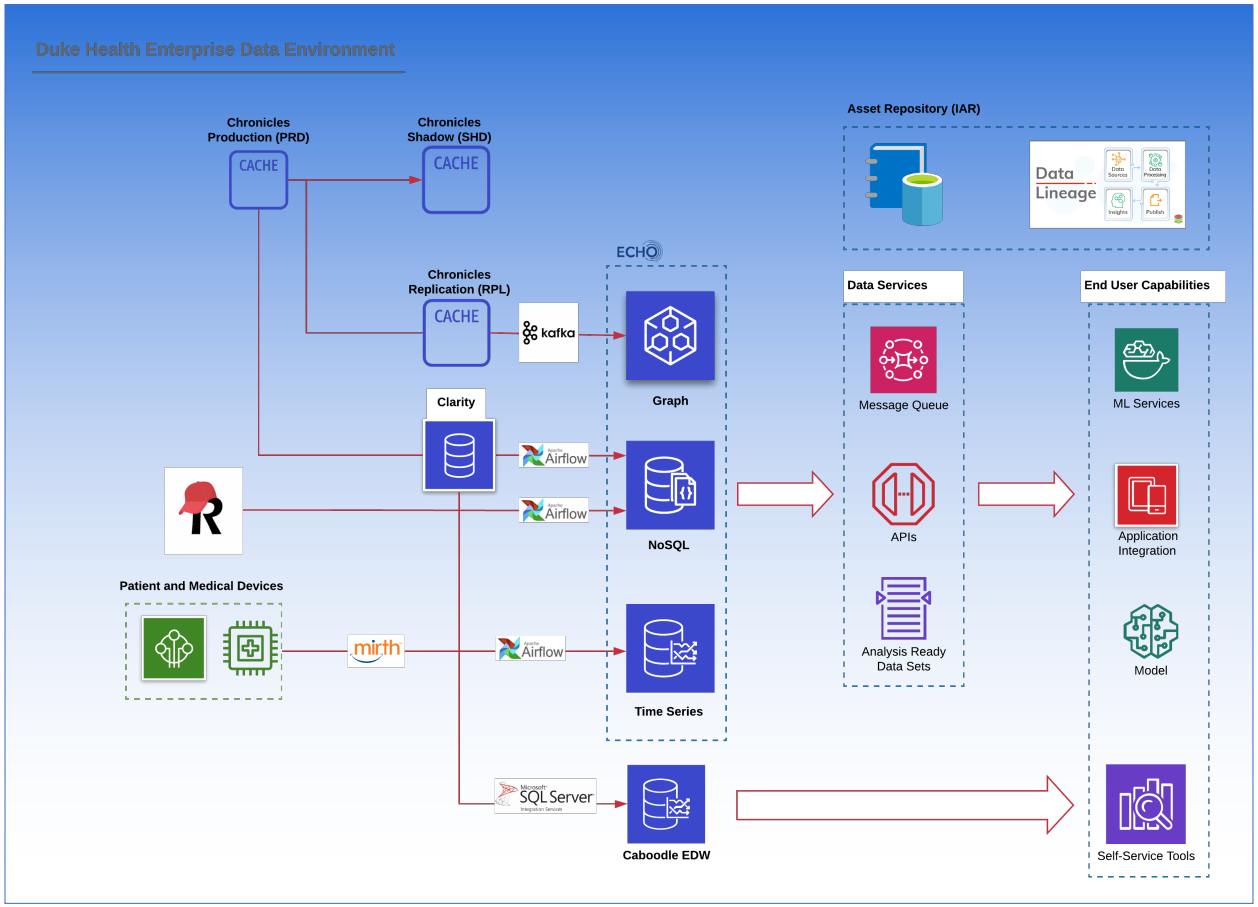

Data Science
Life Cycle
Clinical
Validation
Train / Retrain
Build
docker
Testing
Enterprise
Production
Monitor

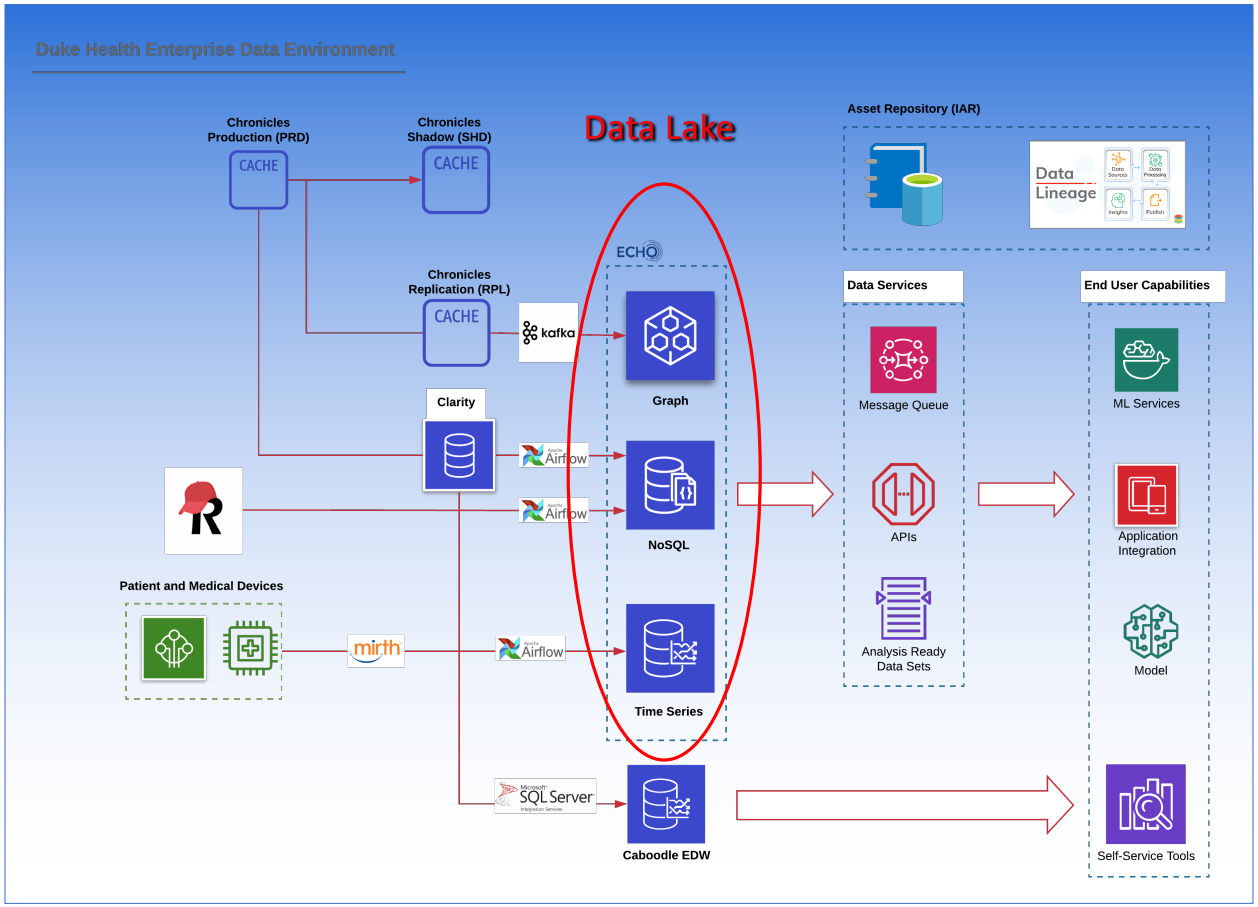

Data Science
Life Cycle
Clinical
Validation
docker
Train / Retrain
Build
Testing
Enterprise
Production
Monitor

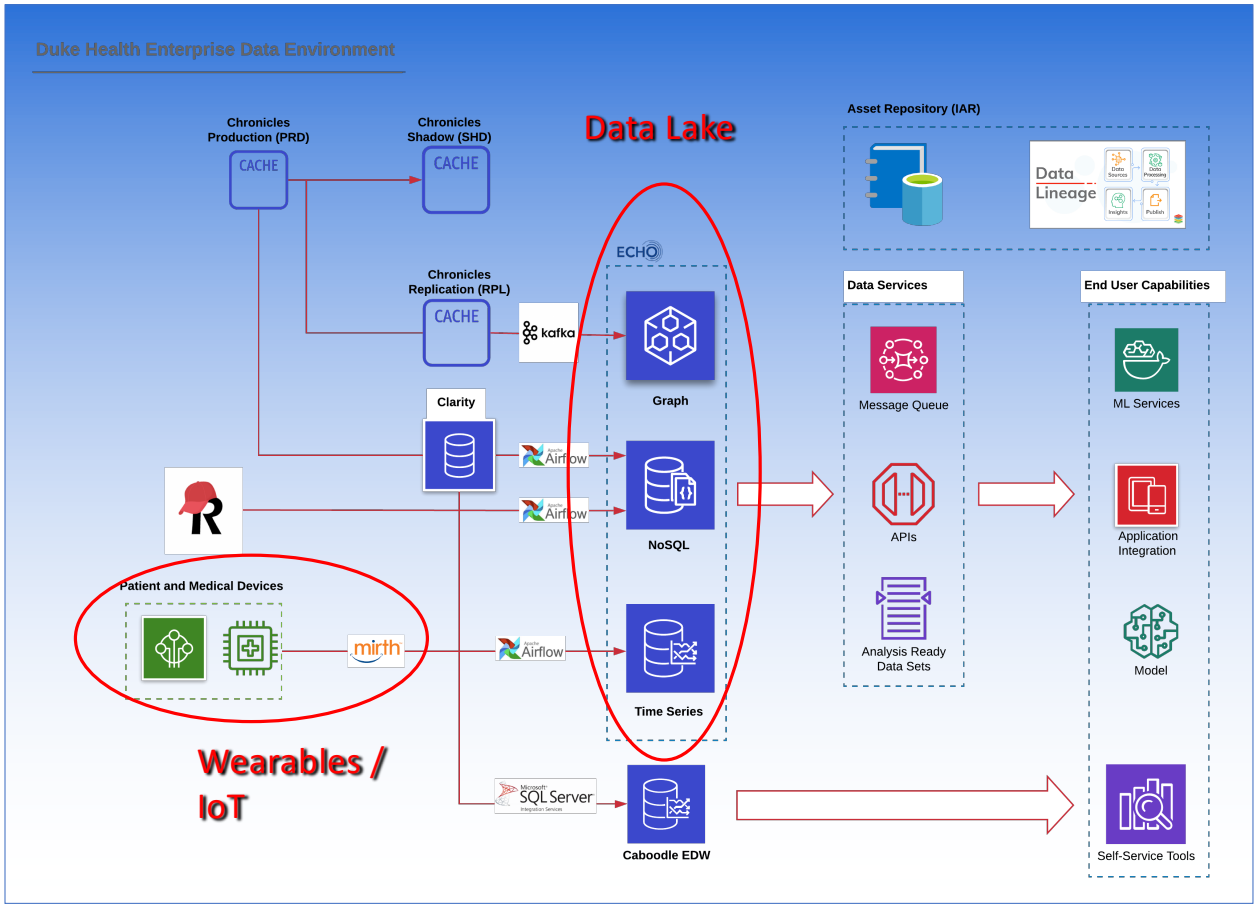

Data Science
Life Cycle
Clinical
Validation
Train / Retrain
Build
Testing
docker
Enterprise
Production
Monitor

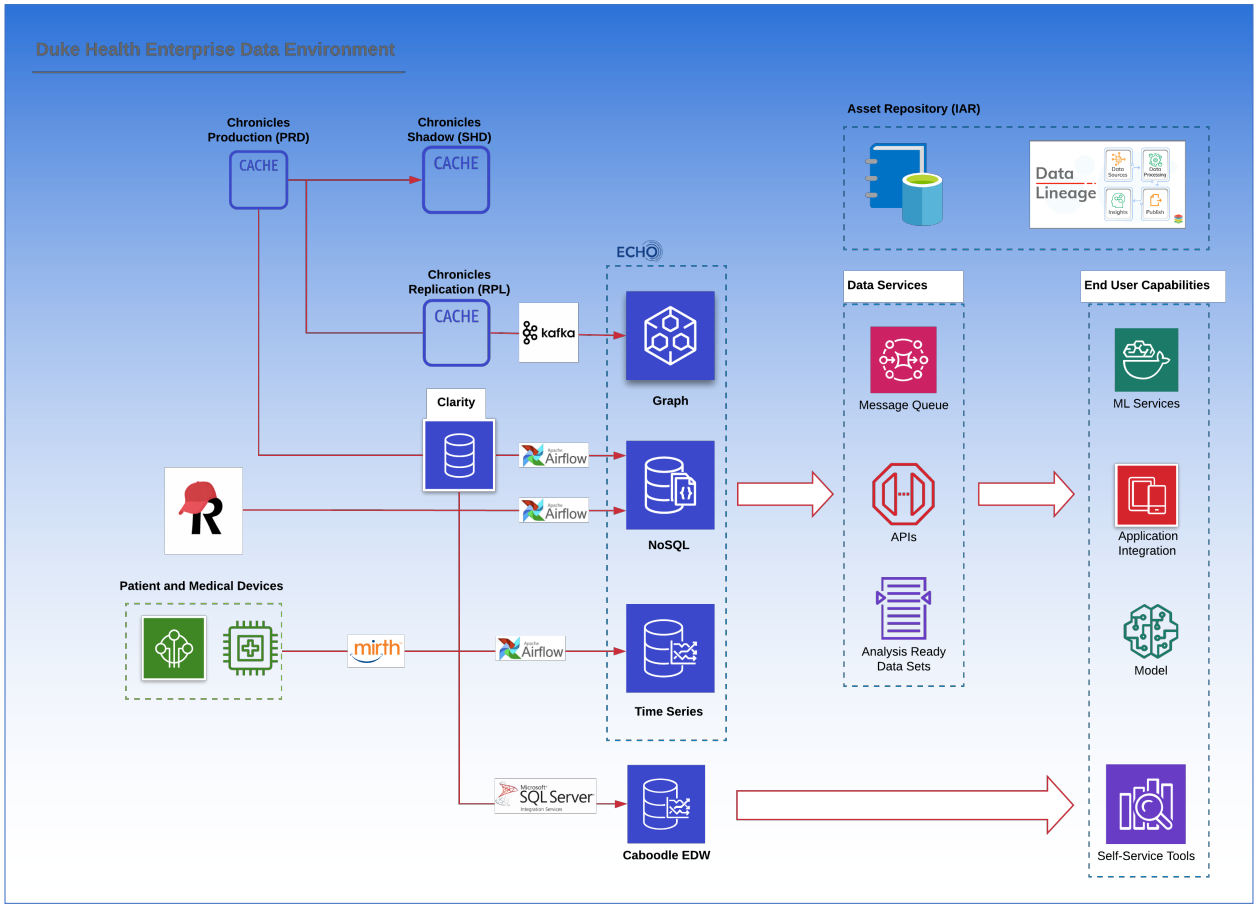

Data Science
Life Cycle
Clinical
Validation
Train / Retrain
Build
Testing
Enterprise
Production
Monitor
docker

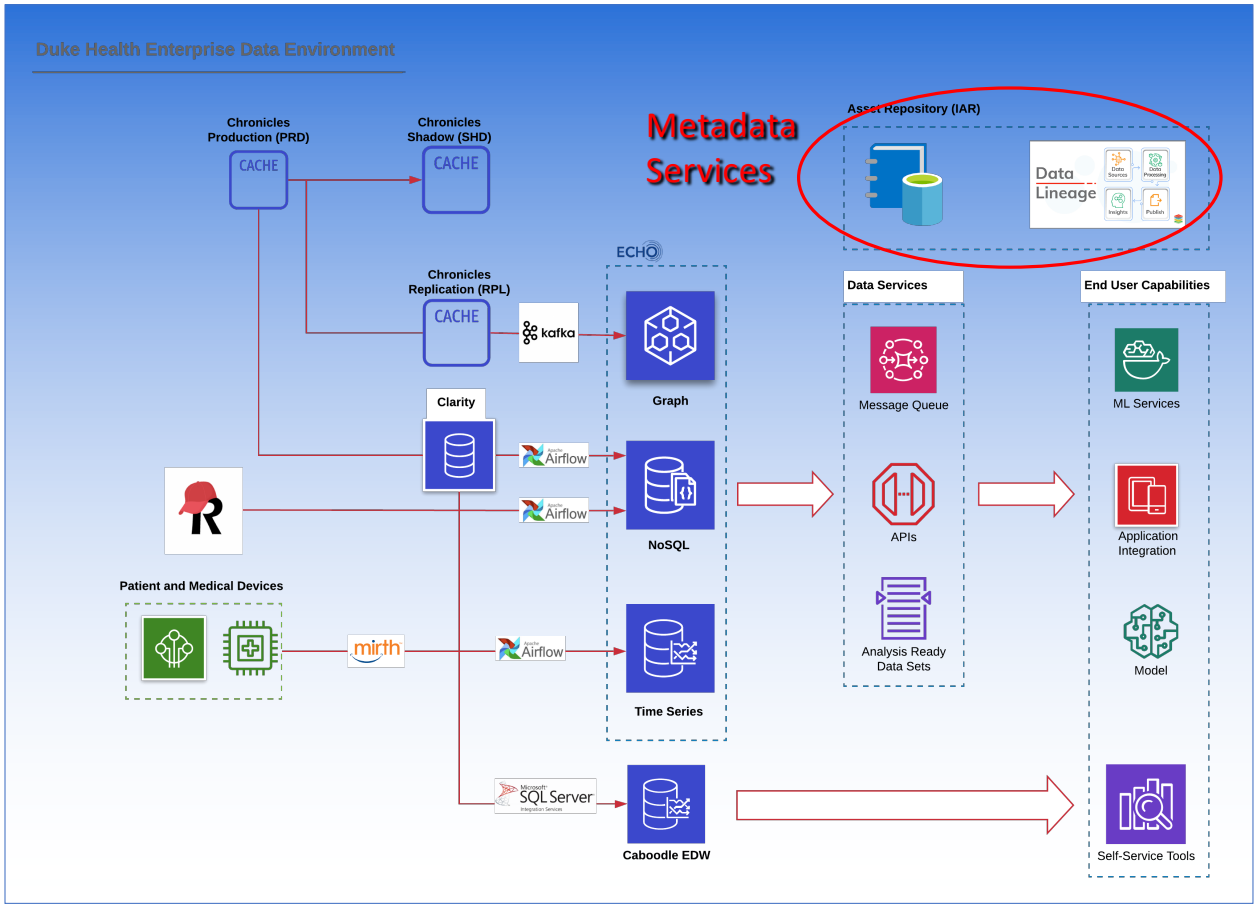

Data Science Life Cycle
Clinical Validation
Train / Retrain
Build
Testing
Enterprise Production
Monitor
docker
SQL
FHIR
VBasic
SQL
Python
SQL
Python
R
Python (Tensorflow)
FHIR
DAG (Directed Acyclic Graph)
docker

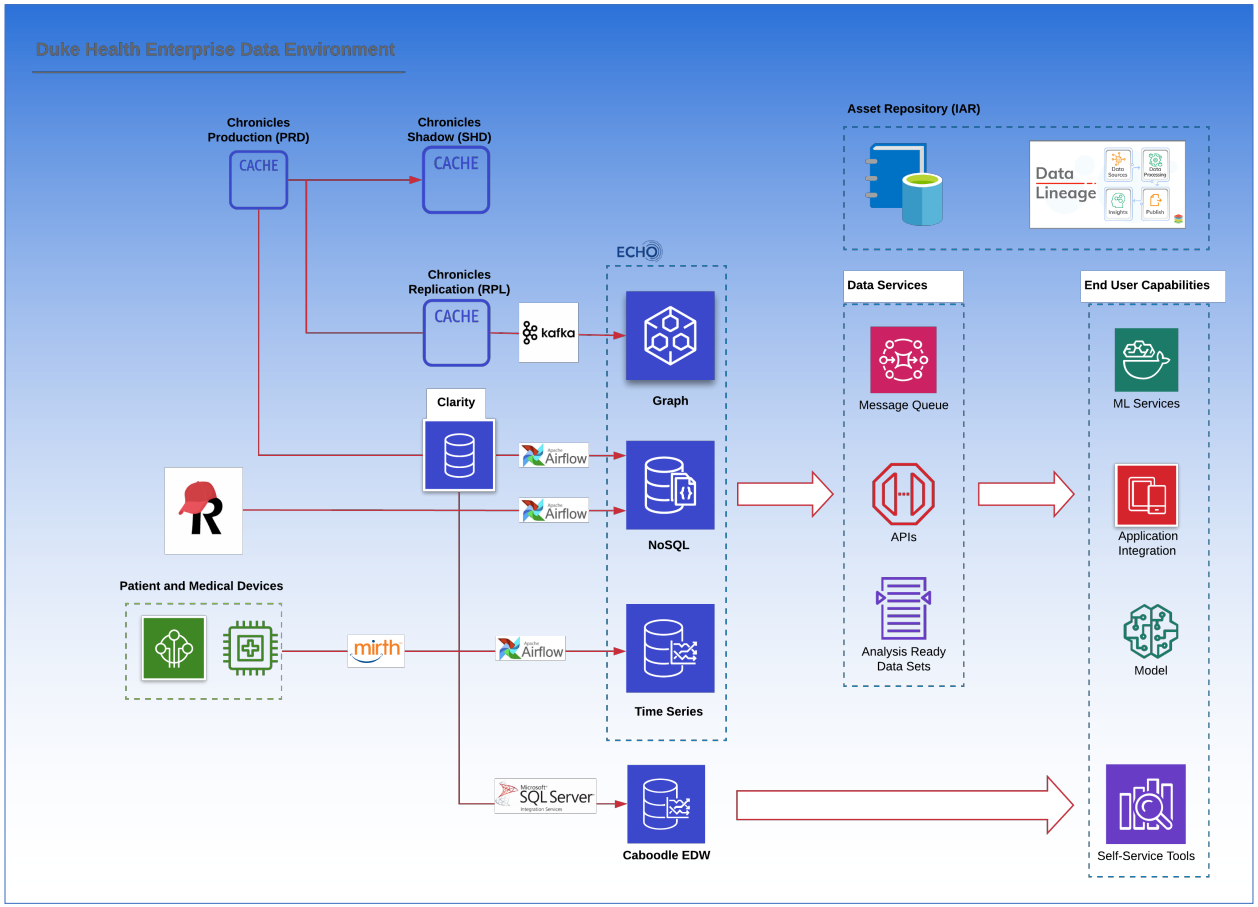

Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Chronicles Replication (RPL)
CACHE
kafka
Clarity
Airflow
Airflow
Patient and Medical Devices
mirth
Airflow
Microsoft SQL Server
ECHO
Graph
NoSQL
Time Series
Caboodle EDW
Asset Repository (IAR)
Data Lineage
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools

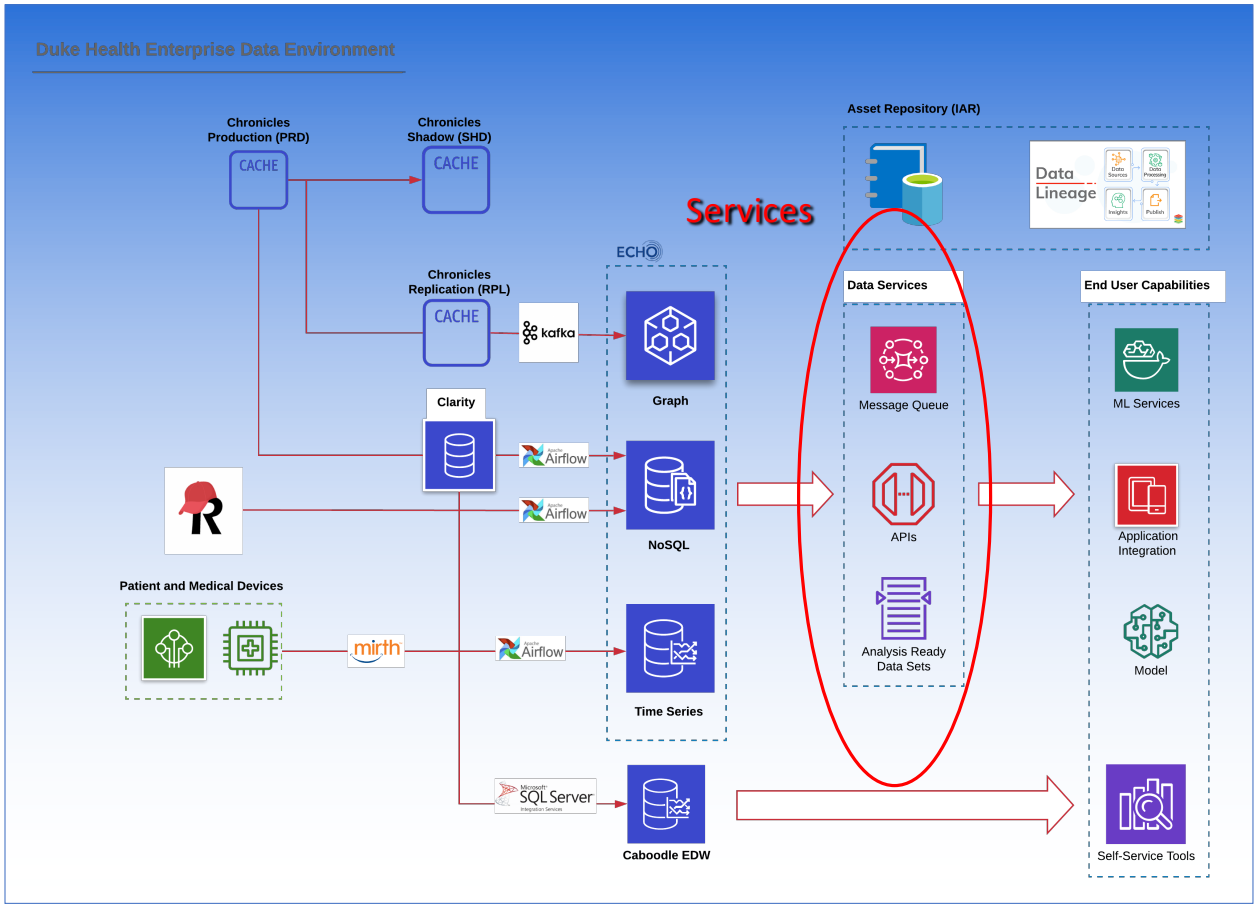

Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Data Lake
Asset Repository (IAR)
Data Lineage
ECHO
Chronicles Replication (RPL)
CACHE
kafka
Graph
Data Services
End User Capabilities
Message Queue
ML Services
Clarity
Airflow
Airflow
NoSQL
APIs
Application Integration
Patient and Medical Devices
mirth
Airflow
Time Series
Analysis Ready Data Sets
Model
Microsoft SQL Server
Caboodle EDW
Self-Service Tools

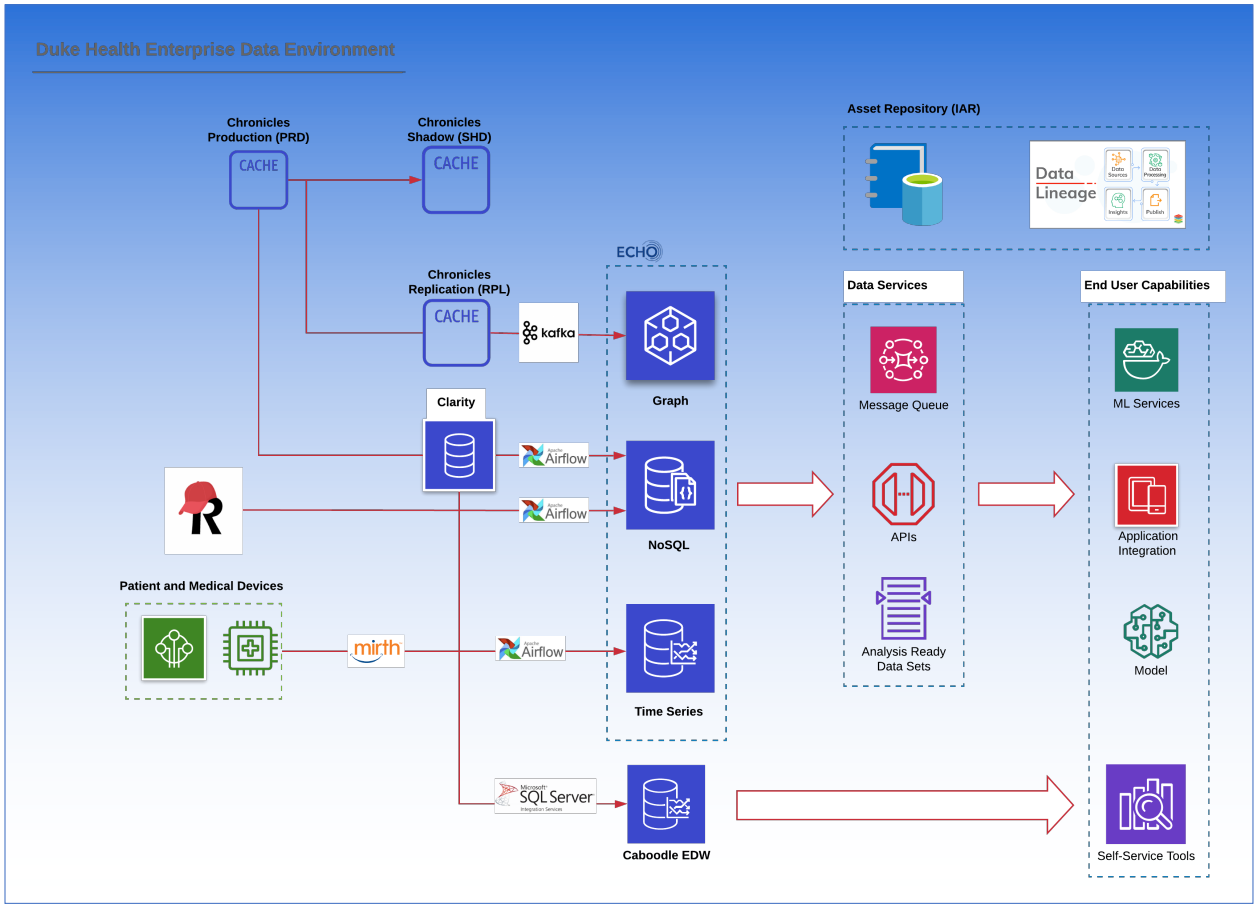

Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Data Lake
Asset Repository (IAR)
Data Lineage
ECHO
Chronicles Replication (RPL)
CACHE
kafka
Graph
Clarity
Airflow
NoSQL
Airflow
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Patient and Medical Devices
mirth
Airflow
Time Series
Wearables / IoT
SQL Server
Caboodle EDW
Self-Service Tools

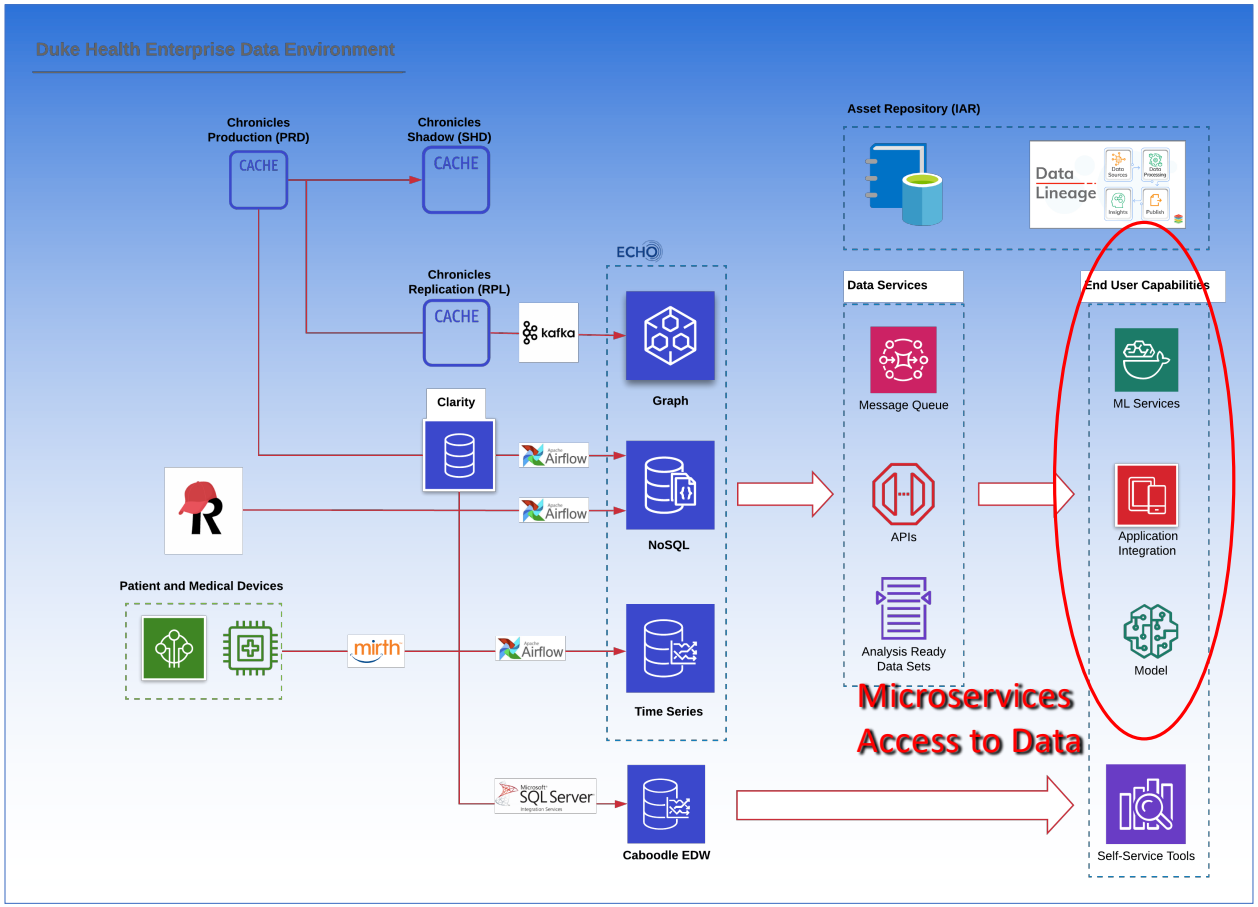

Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Chronicles Replication (RPL)
CACHE
kafka
Clarity
Airflow
Airflow
Patient and Medical Devices
mirth
Airflow
Microsoft SQL Server
ECHO
Graph
NoSQL
Time Series
Caboodle EDW
Asset Repository (IAR)
Data Lineage
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools


Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Metadata Services
Asset Repository (IAR)
Data Lineage
ECHO
Chronicles Replication (RPL)
CACHE
kafka
Graph
Clarity
Airflow
Airflow
NoSQL
Patient and Medical Devices
mirth
Airflow
Time Series
Microsoft SQL Server
Caboodle EDW
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools

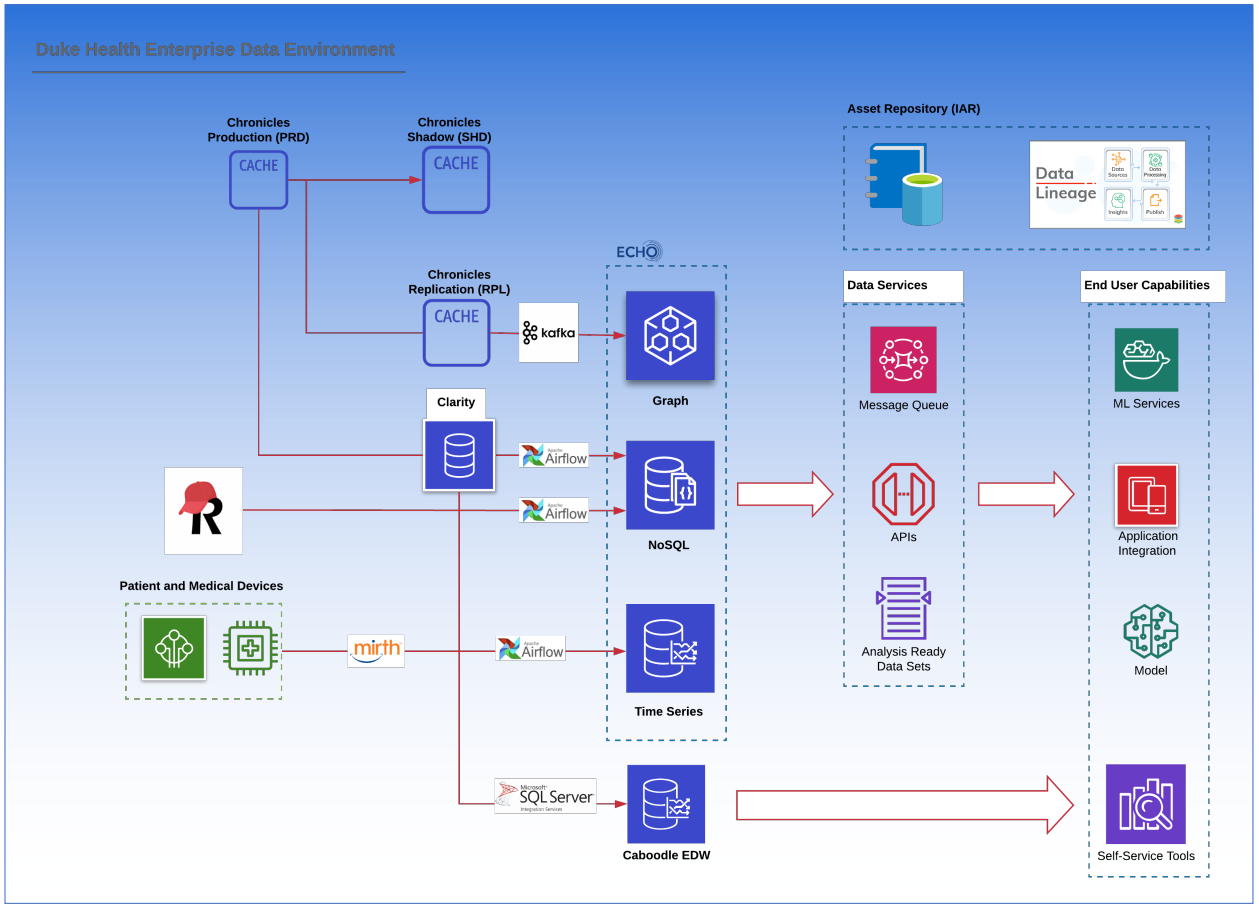

Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Chronicles Replication (RPL)
CACHE
kafka
Clarity
Airflow
Airflow
Patient and Medical Devices
mirth
Airflow
Microsoft SQL Server
ECHO
Graph
NoSQL
Time Series
Caboodle EDW
Asset Repository (IAR)
Data Lineage
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools


Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Chronicles Replication (RPL)
CACHE
kafka
Clarity
Airflow
Airflow
Patient and Medical Devices
mirth
Airflow
Microsoft SQL Server
Integration Services
ECHO
Graph
NoSQL
Time Series
Caboodle EDW
Services
Asset Repository (IAR)
Data Lineage
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools


Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Chronicles Replication (RPL)
CACHE
kafka
Clarity
Airflow
Airflow
Patient and Medical Devices
mirth
Airflow
Microsoft SQL Server
ECHO
Graph
NoSQL
Time Series
Caboodle EDW
Asset Repository (IAR)
Data Lineage
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools


Duke Health Enterprise Data Environment
Chronicles Production (PRD)
CACHE
Chronicles Shadow (SHD)
CACHE
Chronicles Replication (RPL)
CACHE
kafka
ECHO
Graph
Clarity
Airflow
Airflow
NoSQL
Patient and Medical Devices
mirth
Airflow
Time Series
Microsoft SQL Server
Caboodle EDW
Asset Repository (IAR)
Data Lineage
Data Services
Message Queue
APIs
Analysis Ready Data Sets
End User Capabilities
ML Services
Application Integration
Model
Self-Service Tools
Microservices Access to Data

Where *Will* the Warehouse Be?



Where *Will* the Warehouse Be?