The Problem

What we hear from IT …

"ETL is causing our EDW project to fail"

"I wish we could do predictive analytics with our data"

"We failed at our last EDW project"

"We have lots of data today, but we don't know how to do analytics on it"

"We'd like to integrate social media but don't know how"

What we hear from business users …

“The data warehouse data is nightly, I need real-time data”

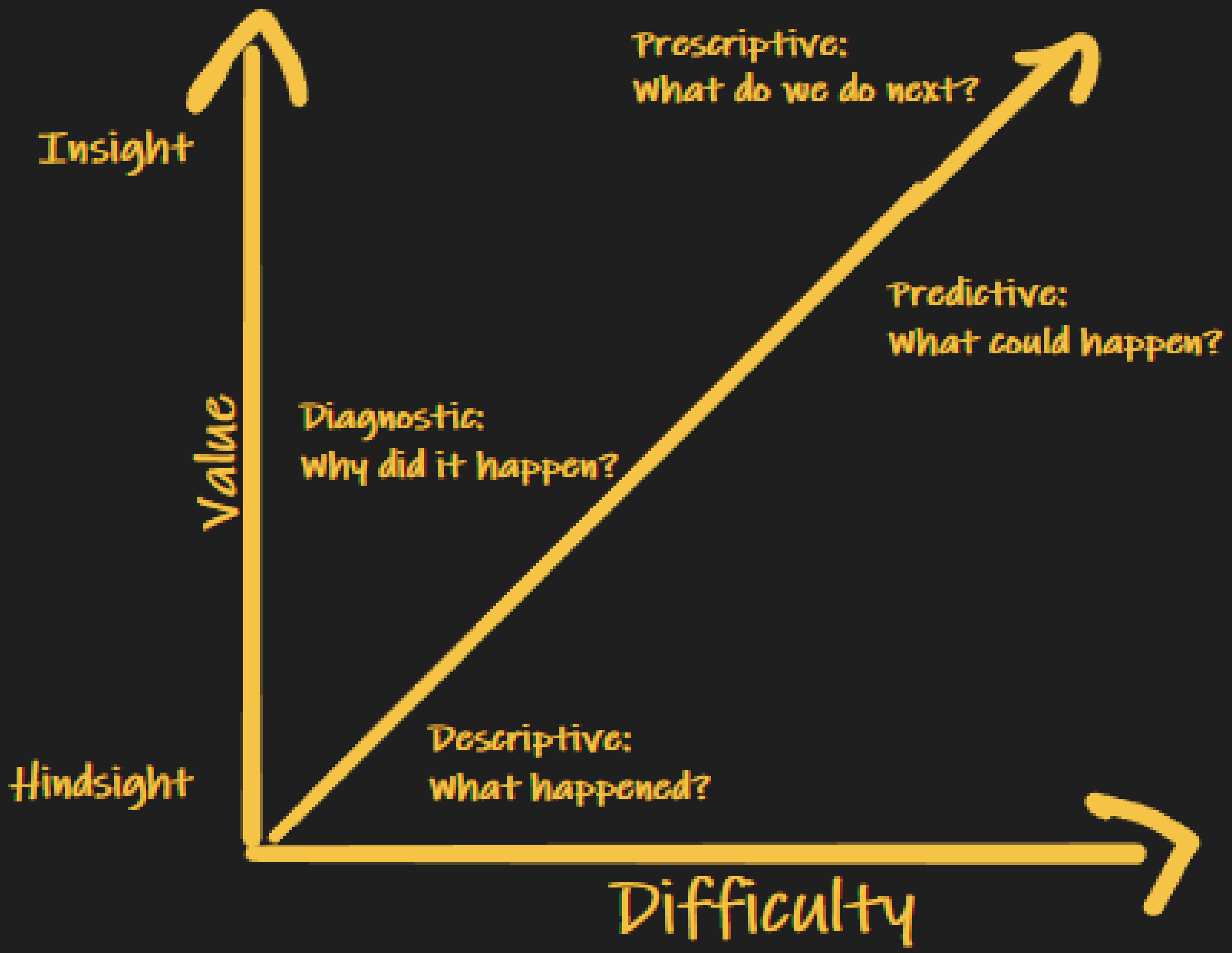
“I want to control my own data”

“Even with the data warehouse and reports, I do analysis in spreadmarts”

“I just want to get my job done”

“The data warehouse doesn’t answer my questions”

Analytics Maturity Models



Prescriptive:
What do we do next?

Predictive:
What could happen?

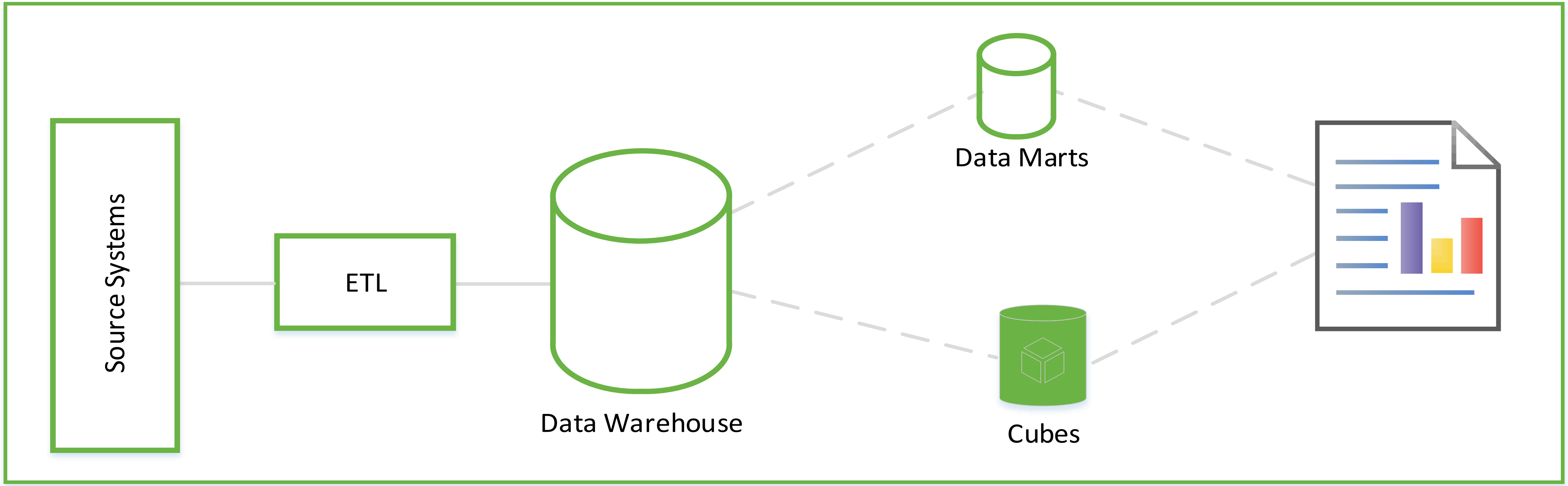
Diagnostic:
Why did it happen?

Descriptive:
What happened?

Past Solutions

Legacy Thinking

The Philosophy: Model data » Transform data » Load data » *Understand* data

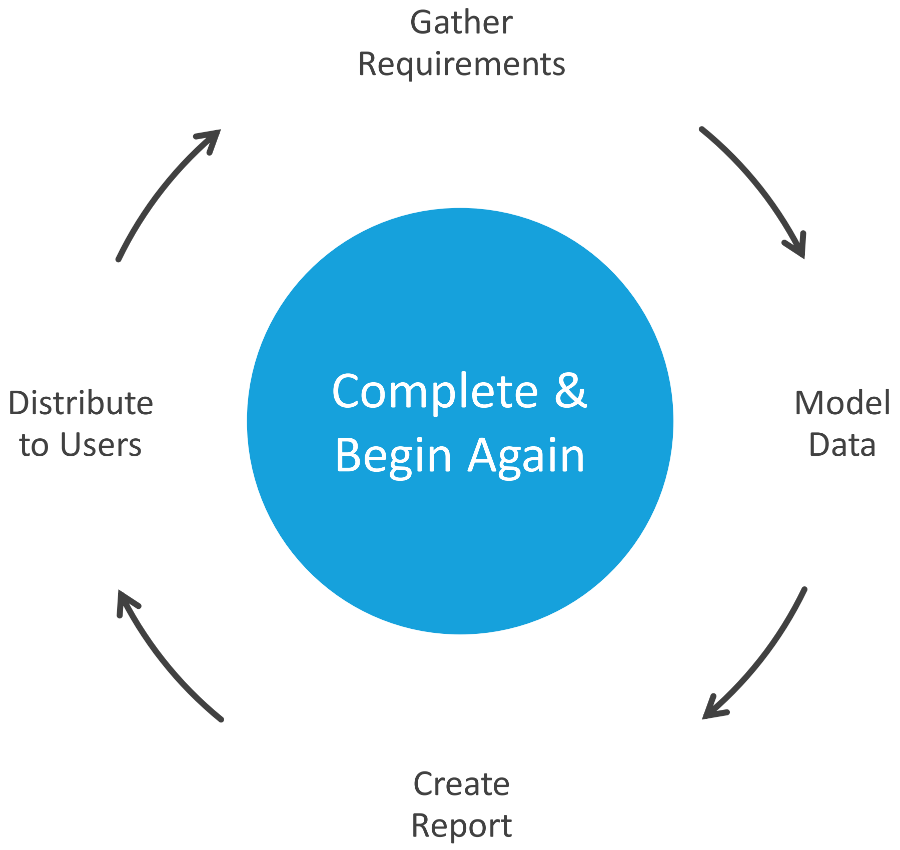


Data Projects have a high fail rate

Too much time is spent in:

- Requirements gathering
- Data modeling
- ETL

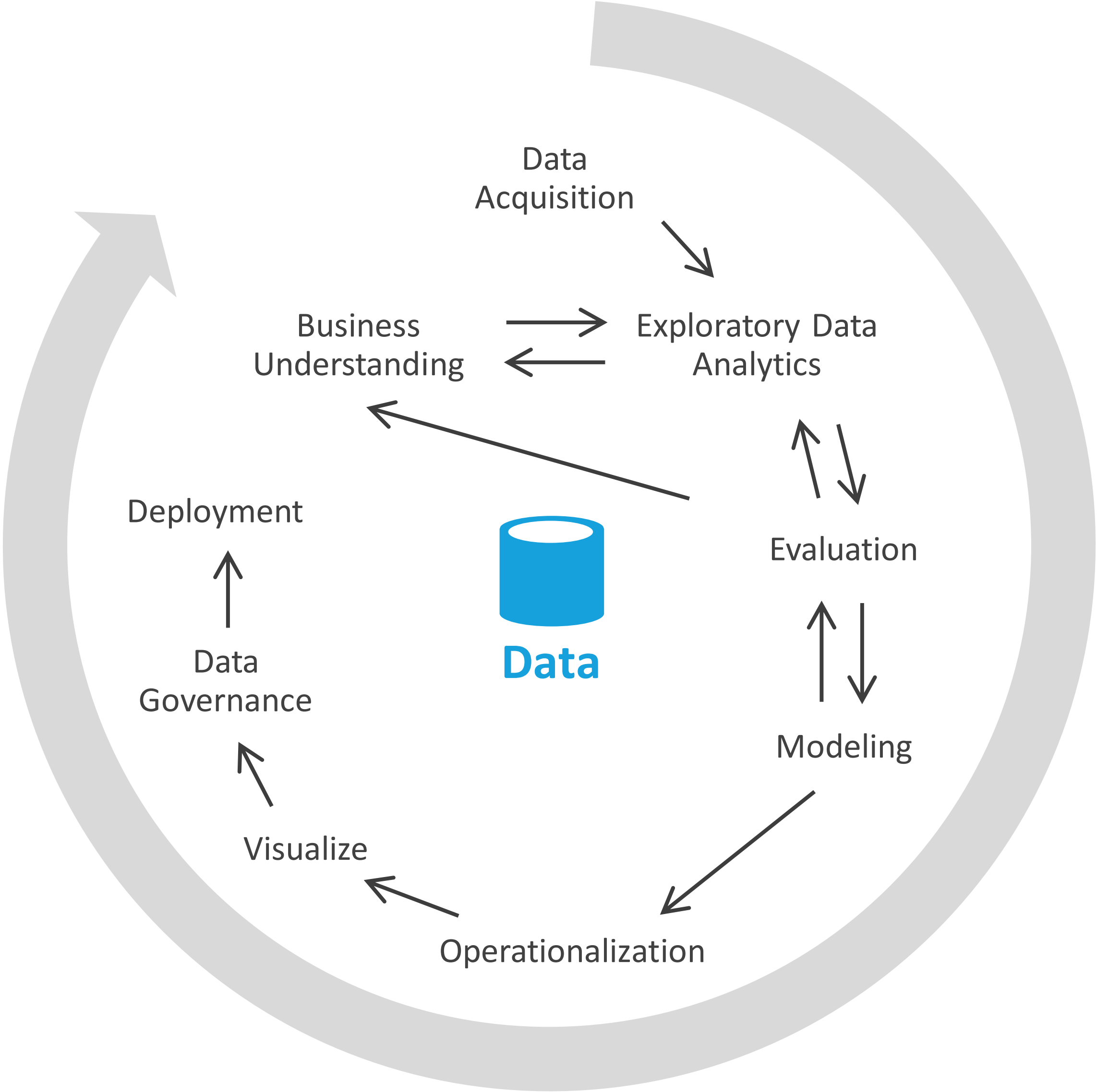
Users only see the fruits of the endeavor after the reports are created



Our Modern Approach

Data Sandboxing

- A robust and well-proven methodology.
- Data science-like.
- Iterative.
- Stresses up-front understanding of data.
- Modeling is done later in the process (schema-on-read).
- ETL might not be needed

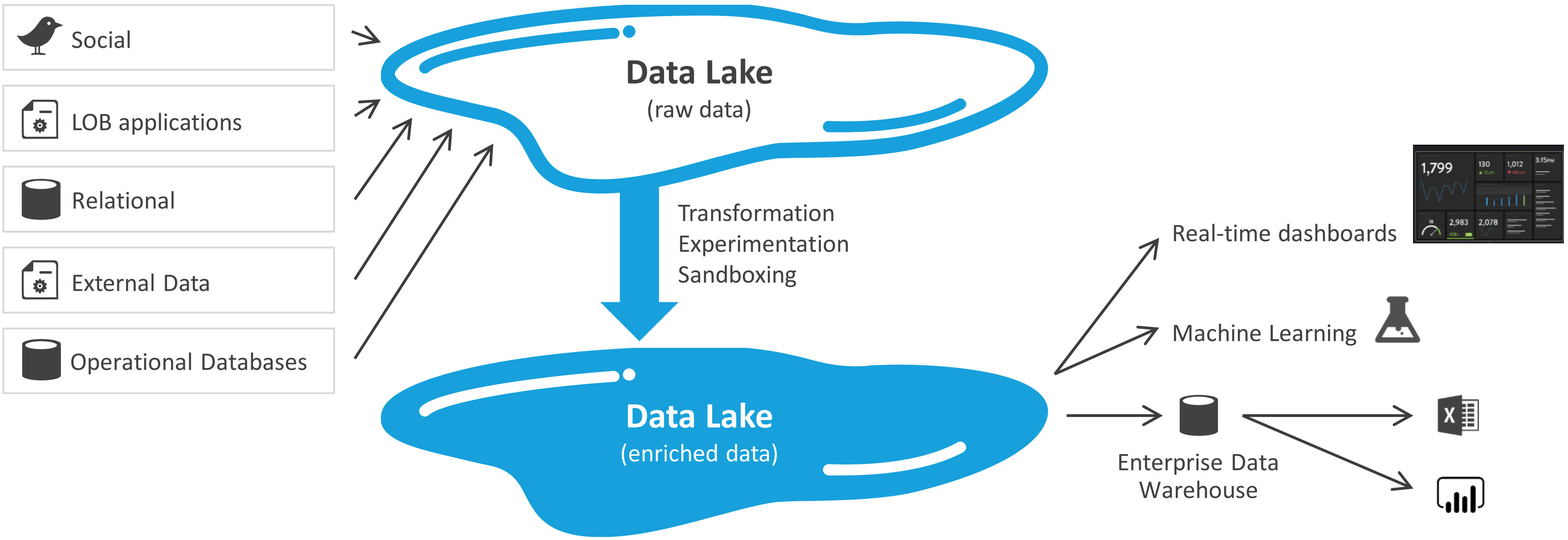


Ingest all Data
Extract and Load, NO Transform

Store all data
In native format

Do analysis
Using almost any tool

Operationalize
Create schemas and pipelines



Social

LOB applications

Relational

External Data

Operational Databases

Data Lake
(raw data)

Transformation
Experimentation
Sandboxing

Data Lake
(enriched data)

Real-time dashboards

Machine Learning

Enterprise Data
Warehouse

Data Catalog

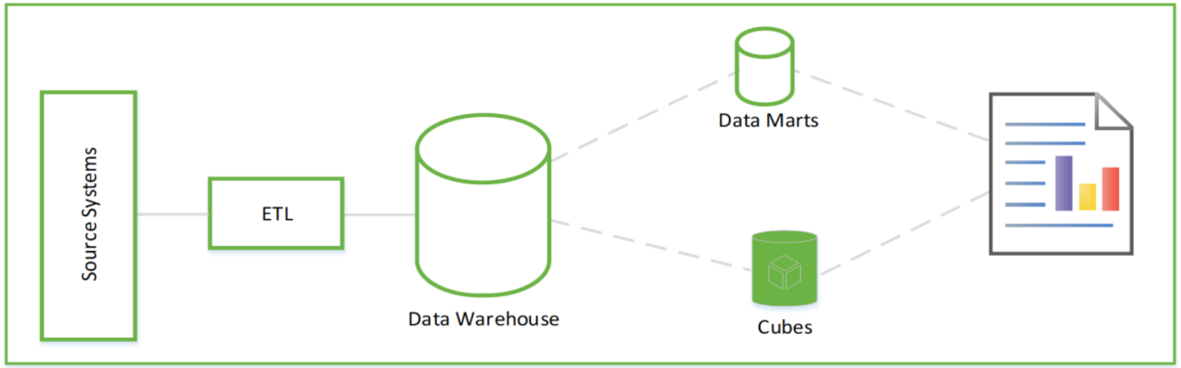
Self-Service Enabler

A Data Lake solves 80% of analytical needs

It is not meant to provide operational reporting

- Data ingestion is more real-time, enabling prediction
- The Data Lake, as a source of all data, is built to efficiently feed a data warehouse.
- Fetches all data, no longer have to go back to source systems for minor changes

Real-World Example – Customer 360



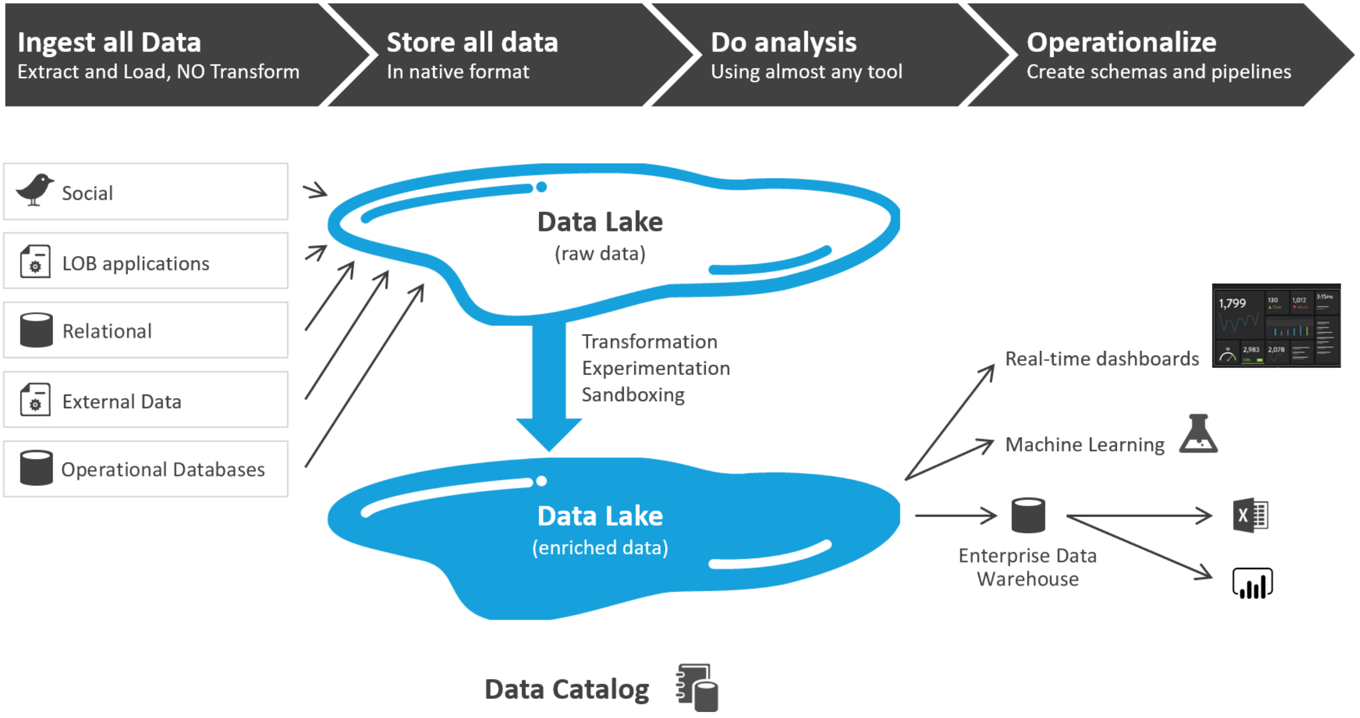
Source Systems

ETL

Data Warehouse

Data Marts

Cubes



Ingest all Data
Extract and Load, NO Transform

Store all data
In native format

Do analysis
Using almost any tool

Operationalize
Create schemas and pipelines

Social

LOB applications

Relational

External Data

Operational Databases

Data Lake
(raw data)

Transformation
Experimentation
Sandboxing

Data Lake
(enriched data)

Real-time dashboards

Machine Learning

Enterprise Data
Warehouse

Data Catalog

Data Lake Design–Folder Structures

A Data Lake Is Just a Folder Structure with Smart Organization and significant processing power

Authentication/Governance

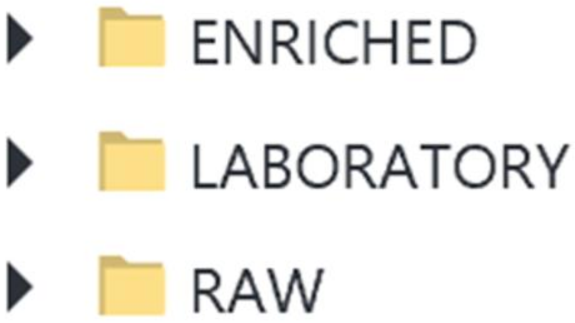
- Accounts/Folders/Files

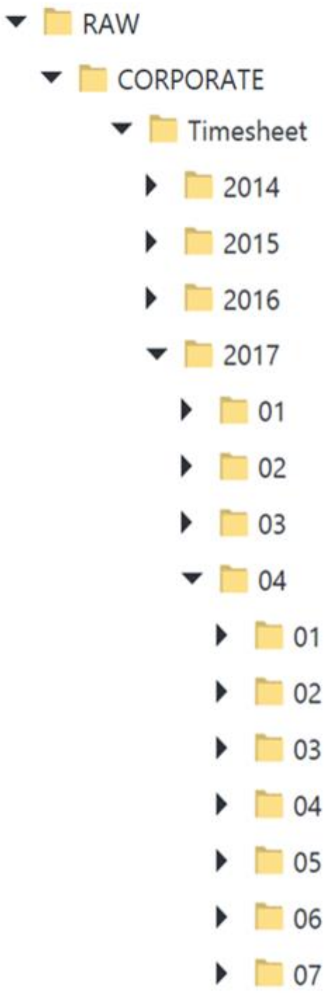
Obvious, Self-documenting Paths

- dev/raw/{datasource}/{object}/YYYYMMDD/
- dev/reject/{datasource}/{object}/YYYYMMDD/
- prod/snapshot/{datasource}/{object}/YYYYMMDD/
- laboratory/jsmith

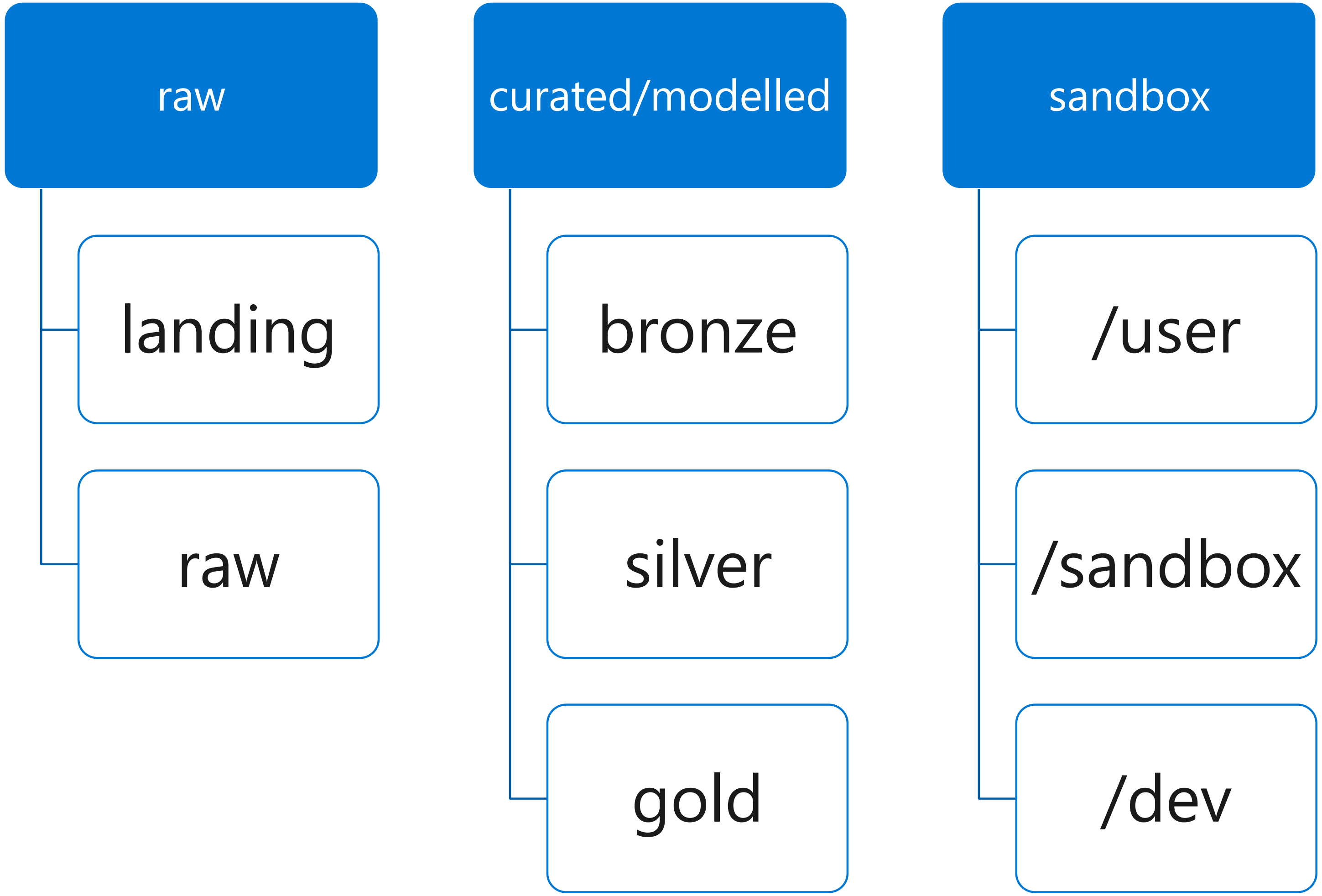
Time partitioning schemes are important

- AVRO/TPFS format
- .../YYYY/MM/DD/HH/MM



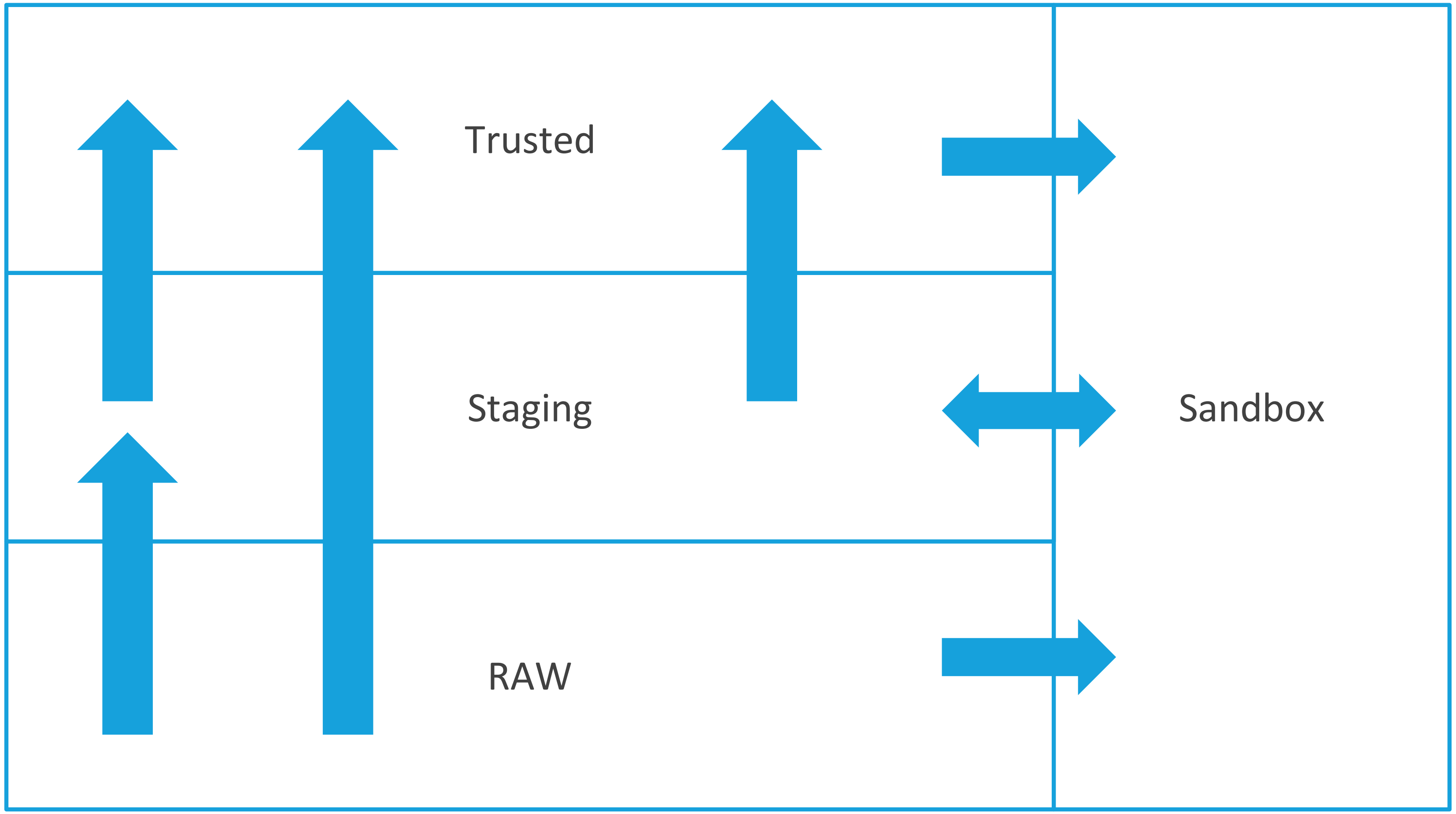


Physical Structure



- raw
  - landing
  - raw
- curated/modelled
  - bronze
  - silver
  - gold
- sandbox
  - /user
  - /sandbox
  - /dev

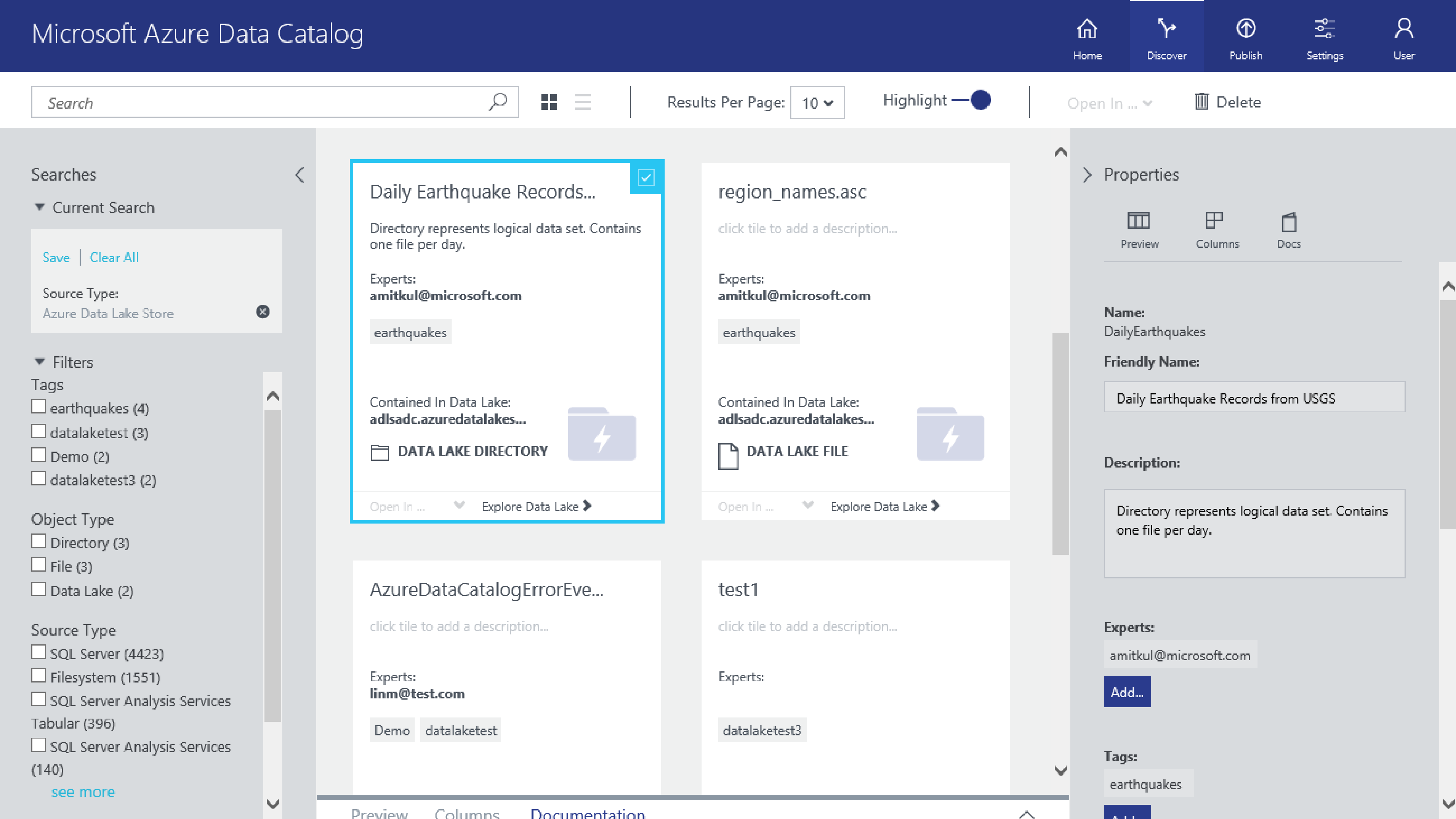
Analytics workflow



1. Development / Data Ingestion
2. Production
3. Data Scientists / Business Analysts/ Developers
4. Development
5. Production

Data Lake vs. Data Warehouse

Data Lake	Data Warehouse
Complementary to the EDW	Can be sourced from the Data Lake
Load first, understand later	Understand first, load later
Schema-on-read	Schema-on-write
System of Insight	System of Record
Detailed Data	Refined Data
Supports varied data formats	Structured data
Adapts to changing requirements	Difficult to change structure
Optimized for Cost	Optimized for Performance



Source Type:
Azure Data Lake Store

Tags
□ earthquakes (4)
□ datalaketest (3)
□ Demo (2)
□ datalaketest3 (2)

Object Type
□ Directory (3)
□ File (3)
□ Data Lake (2)

Source Type
□ SQL Server (4423)
□ Filesystem (1551)
□ SQL Server Analysis Services Tabular (396)
□ SQL Server Analysis Services (140)
see more

☑ Daily Earthquake Records...

Directory represents logical data set. Contains one file per day.

Experts:
amitkul@microsoft.com

earthquakes

Contained In Data Lake:
adlsadc.azuredatalakes...

DATA LAKE DIRECTORY

region_names.asc

click tile to add a description...

Experts:
amitkul@microsoft.com

earthquakes

Contained In Data Lake:
adlsadc.azuredatalakes...

DATA LAKE FILE

AzureDataCatalogErrorEve...

click tile to add a description...

Experts:
linm@test.com

Demo datalaketest

test1

click tile to add a description...

Experts:

datalaketest3

Name:
DailyEarthquakes

Friendly Name:
Daily Earthquake Records from USGS

Description:
Directory represents logical data set. Contains one file per day.

Experts:
amitkul@microsoft.com

Tags:
earthquakes

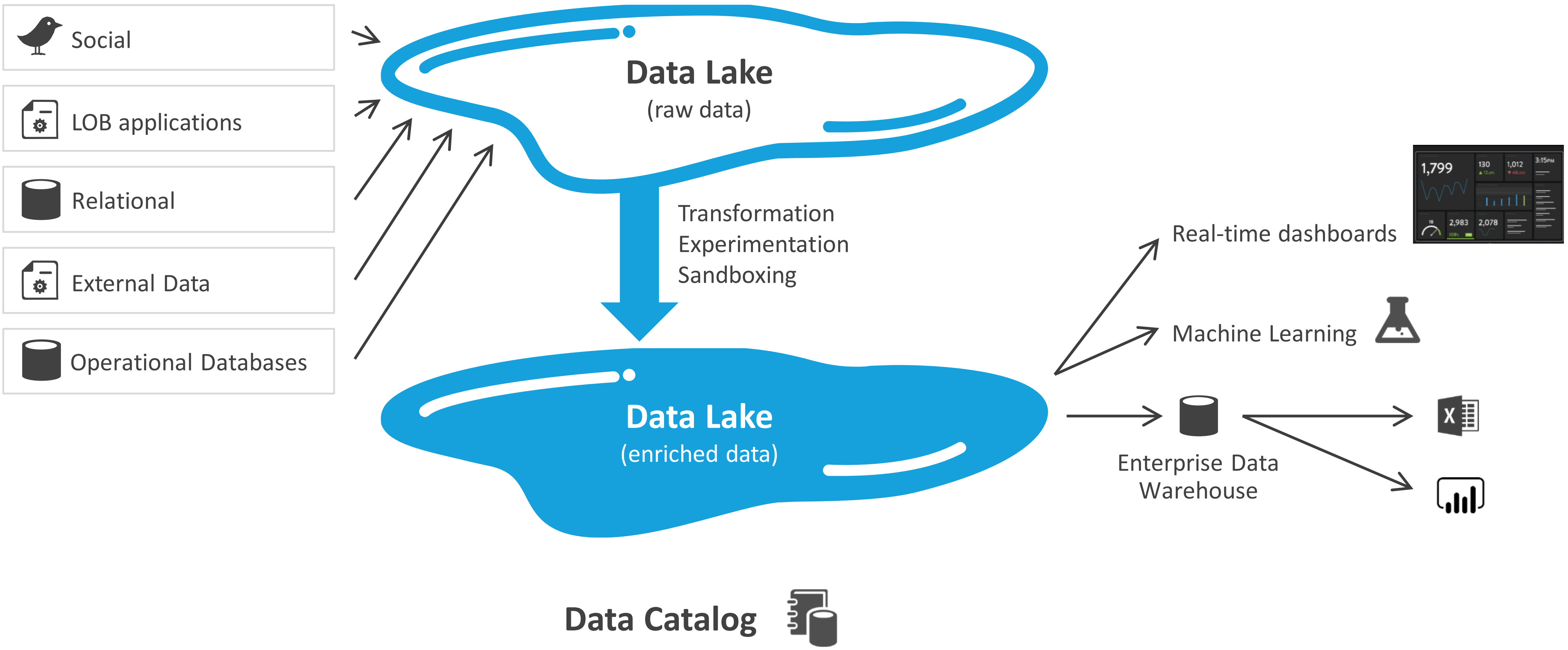
We have a winning idea, help us operationalize it

Ingest all Data
Extract and Load, NO Transform

Store all data
In native format

Do analysis
Using almost any tool

Operationalize
Create schemas and pipelines



Social

LOB applications

Relational

External Data

Operational Databases

Data Lake
(raw data)

Transformation
Experimentation
Sandboxing

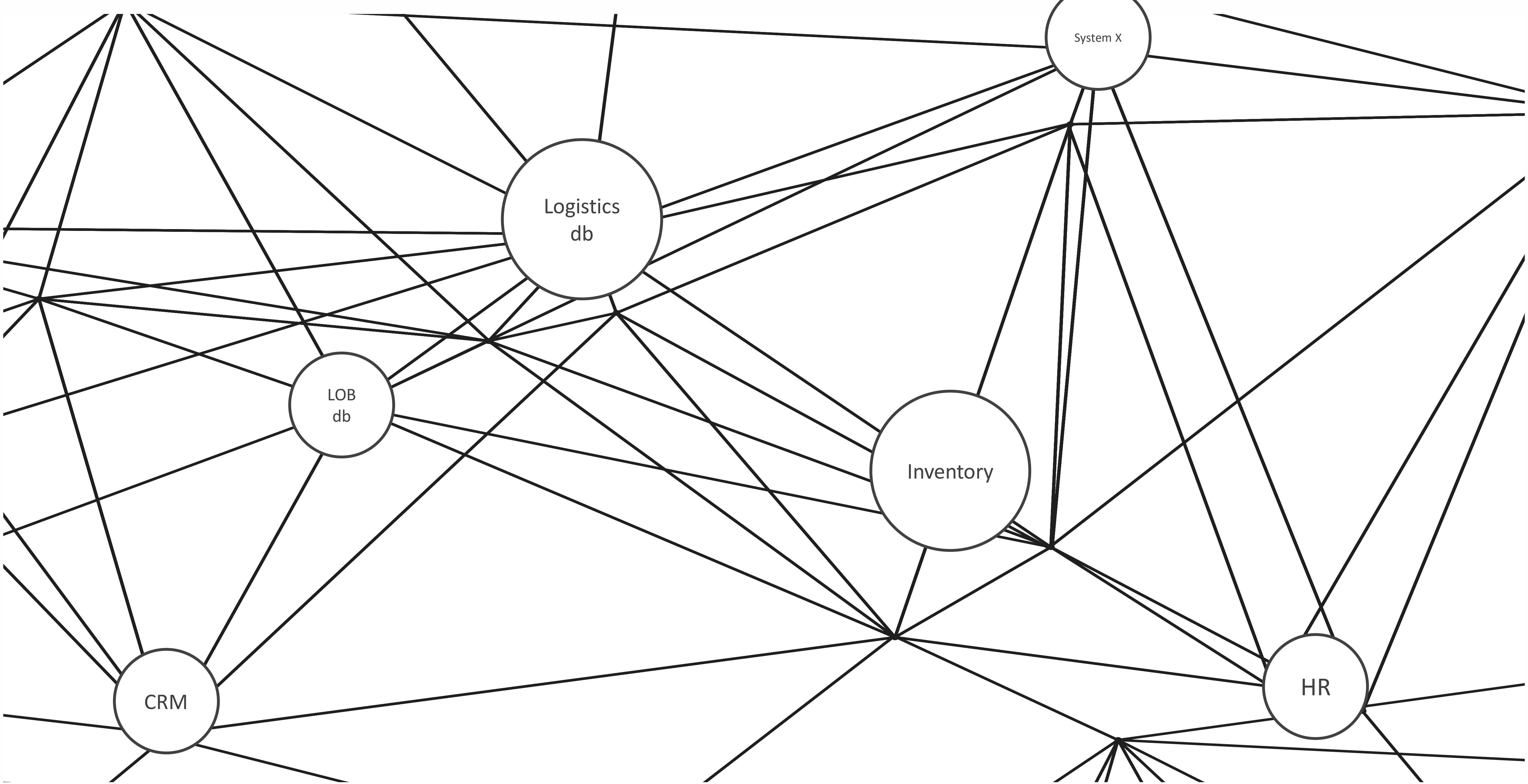
Data Lake
(enriched data)

Real-time dashboards

Machine Learning

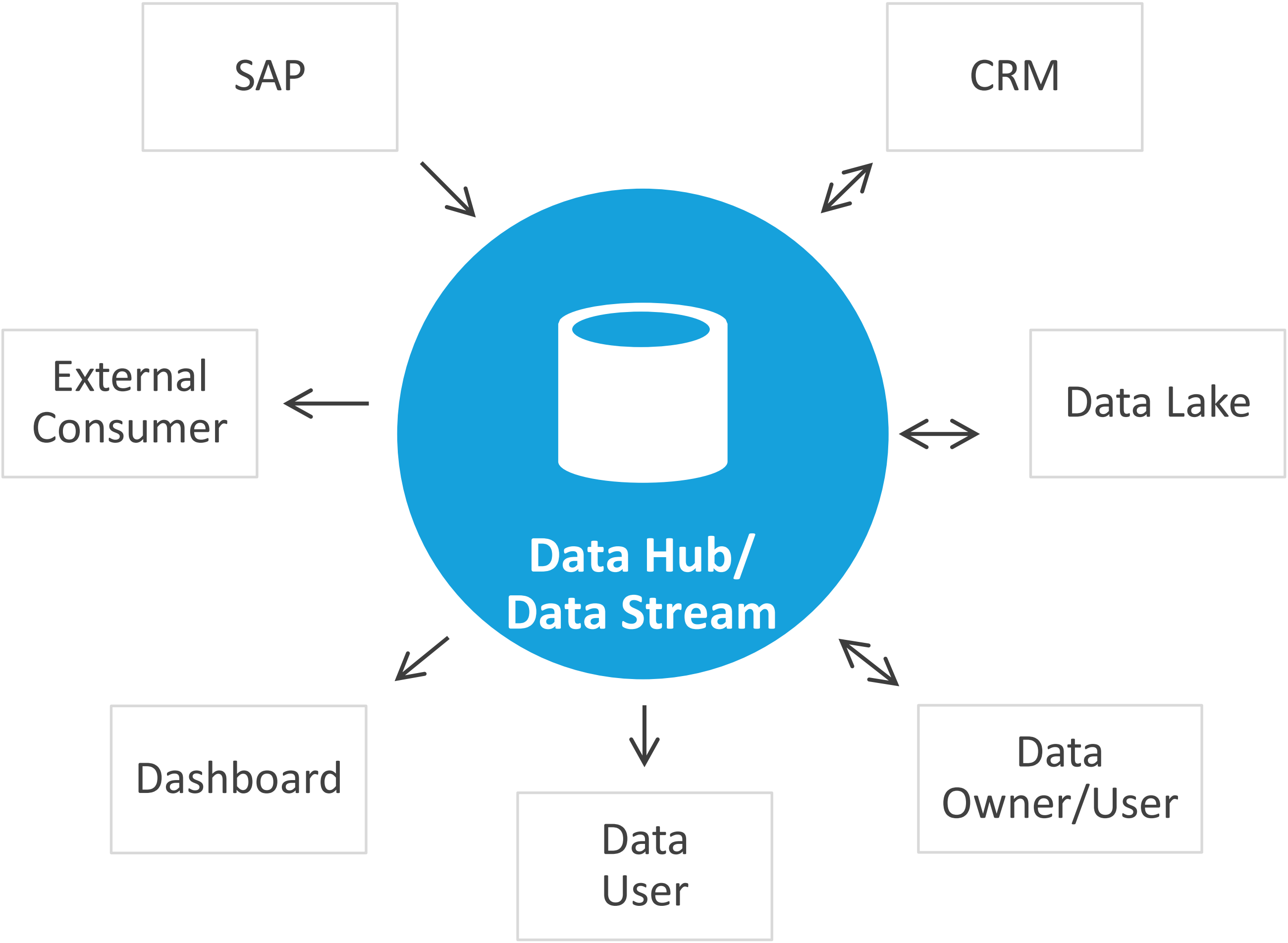
Enterprise Data
Warehouse

Data Catalog

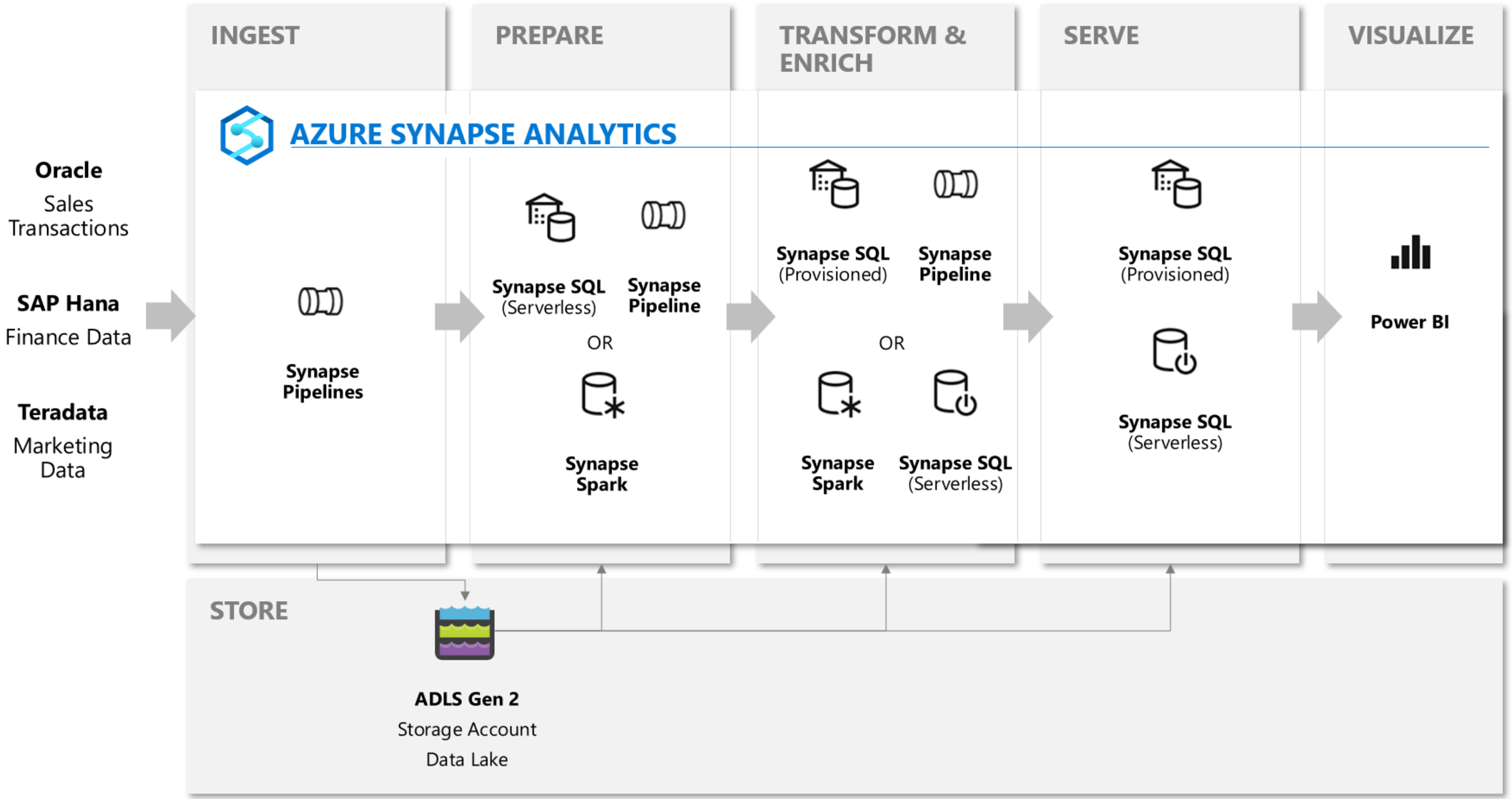

System X
Logistics db
LOB db
Inventory
CRM
HR

Data Ingestion Factory

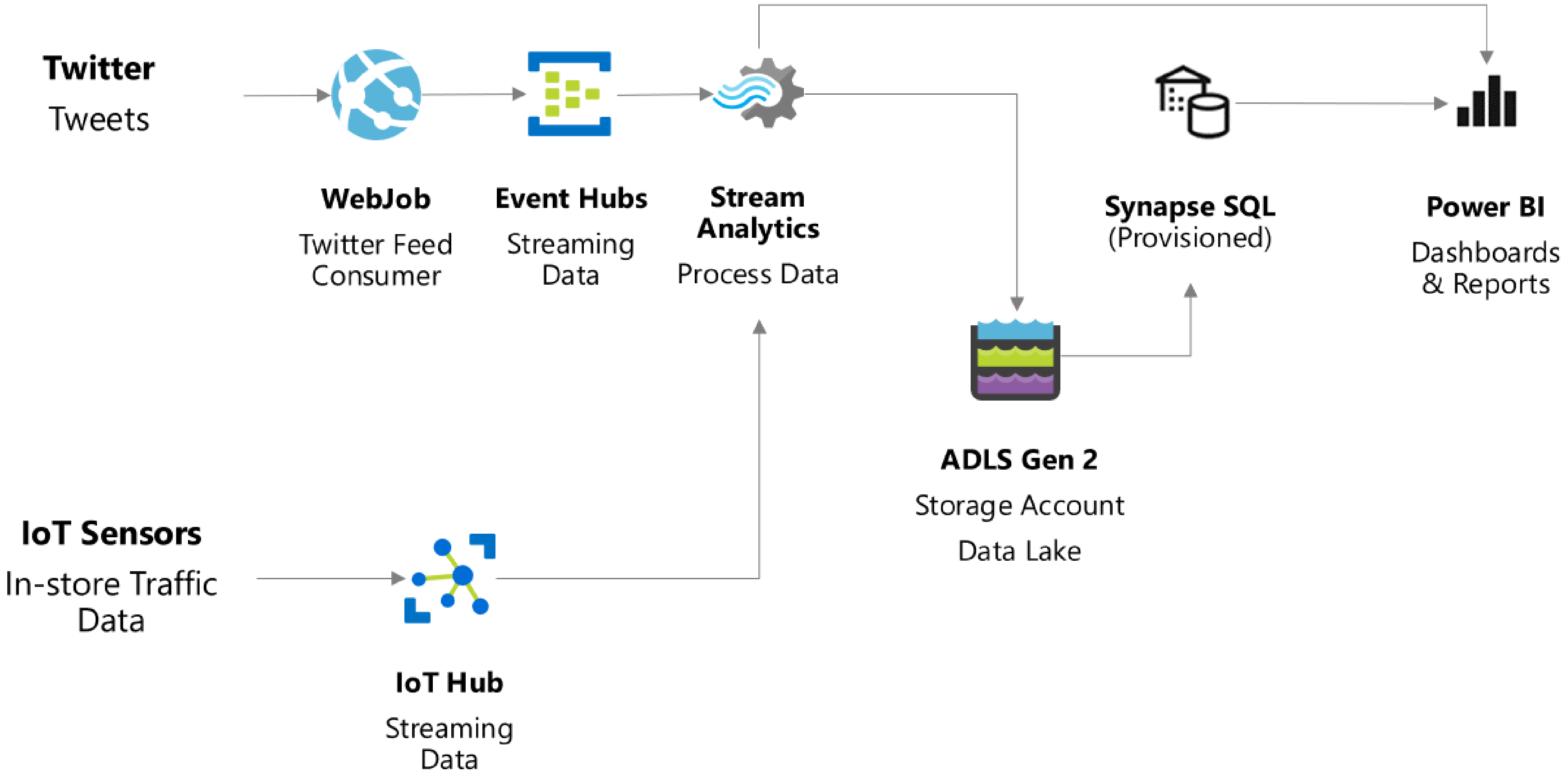
(Kappa Architecture)



Modern Data Warehouse



Handling Streaming Data


Twitter
Tweets
WebJob
Twitter Feed Consumer
Event Hubs
Streaming Data
Stream Analytics
Process Data
Synapse SQL
(Provisioned)
Power BI
Dashboards & Reports
ADLS Gen 2
Storage Account
Data Lake
IoT Sensors
In-store Traffic Data
IoT Hub
Streaming Data