SINGULARITY
TECH DAY

#STechDay2020

A Data Journey: from the Data Warehouse to the Data Lakehouse

SINGULARITY
TECH DAY

#STechDay2020

ORGANIZATION

plain concepts

SPONSORS

Microsoft

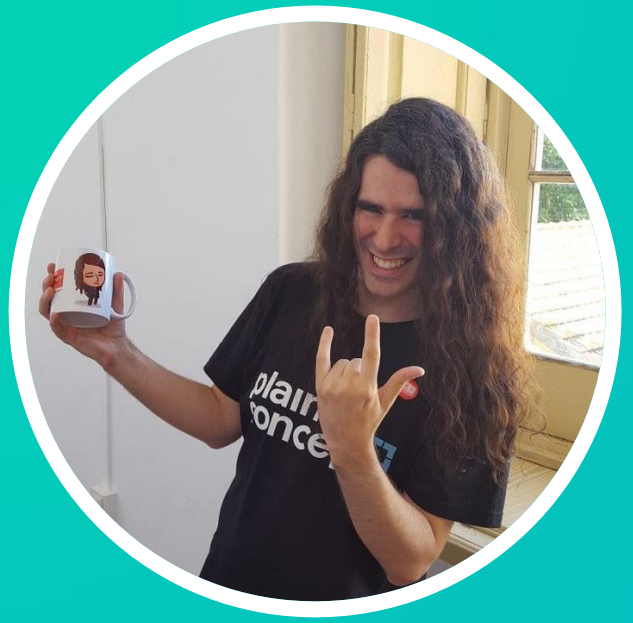
SUPPORT

World Central Kitchen

THANK YOU!

SINGULARITY
TECH DAY

#STechDay2020



Pablo A. Doval

Principal Data Architect – Plain Concepts UK

"I work with code and data, but don't tell my mom; she thinks I'm a piano player in a whorehouse."

@PabloDoval

palvarez@plainconcepts.com

Innovation through Data and AI…

Smart Maintenance	Video Content classifier	Real Time Churn Analysis	Knowledge Extraction	Lead Cases Selection
Anomaly Detection	Audience Segmentation	Automatic Case Classifier	Document Classification	Semantic Search
Behavioural Analysis	Smart Pricing Models	Case Outcome Forecasting	Clause Outlier Detection	Settlement Forecast

… requieres

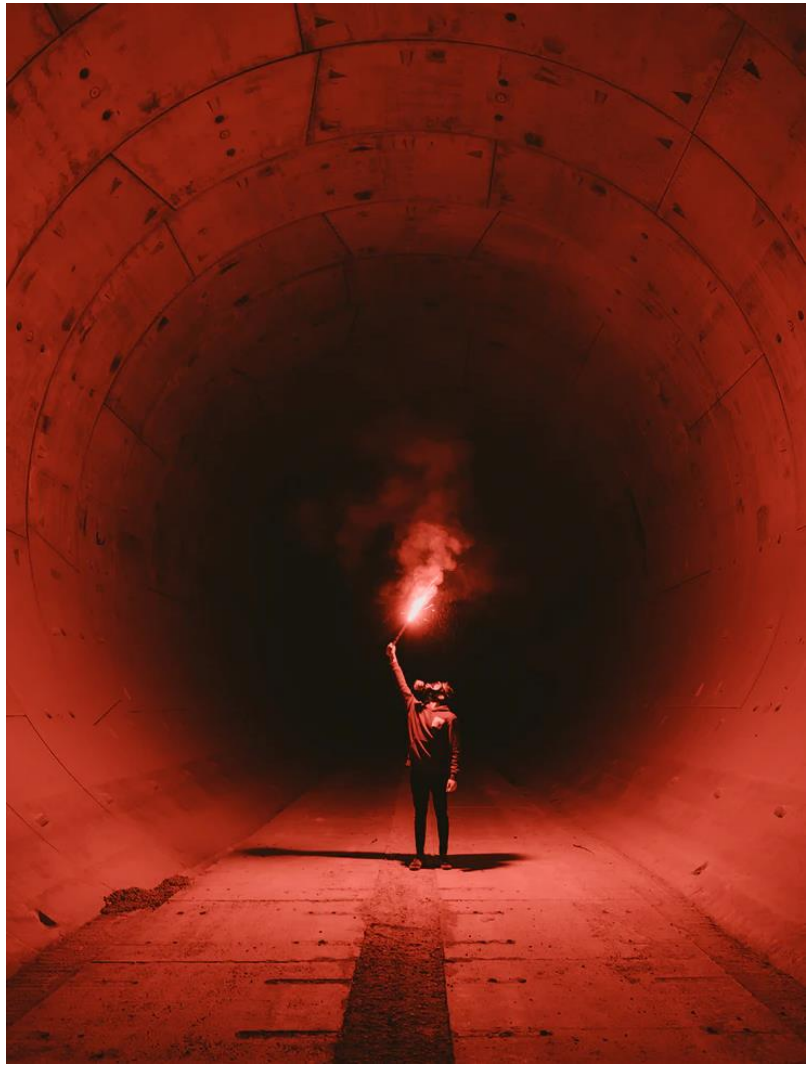
data first.

Where are we?

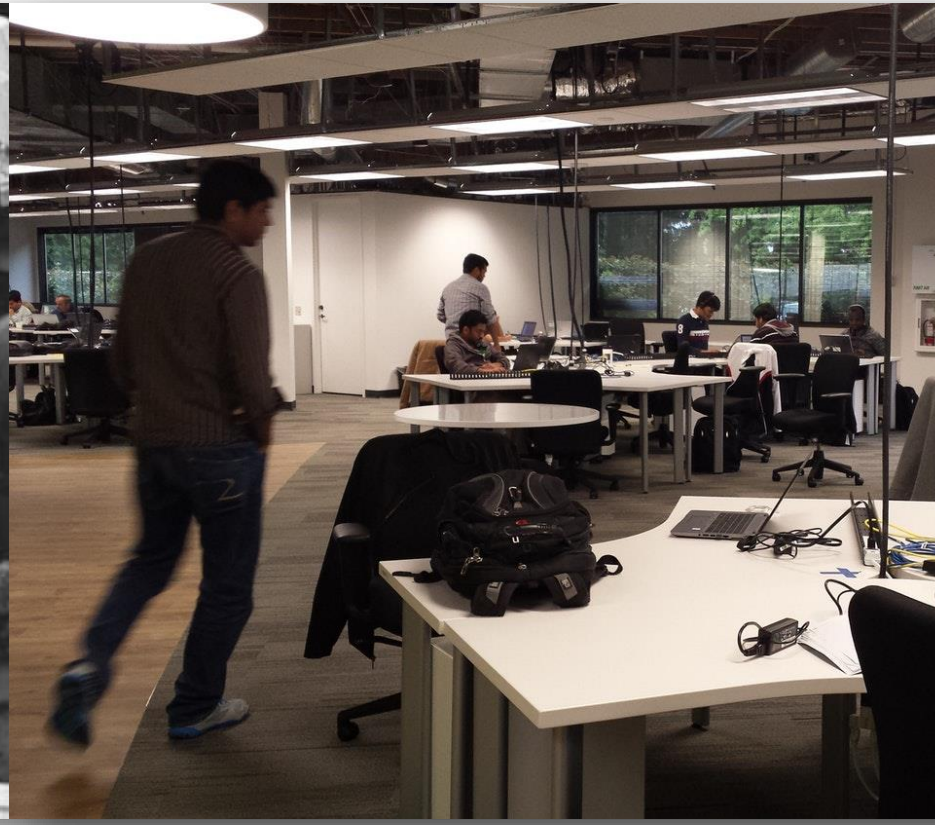
Unstructured, chaotic **data estate**

Lack of skills and productivity

Solutions not **Enterprise-ready**



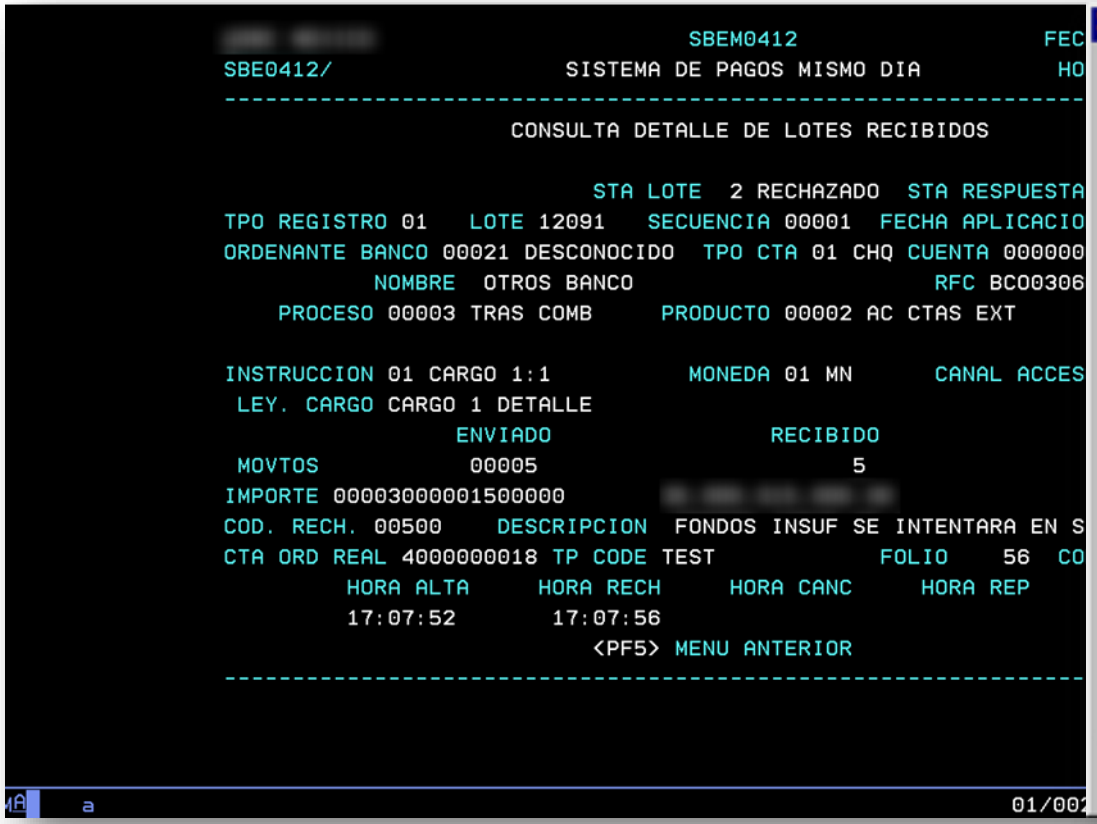
We have evolved and transformed…

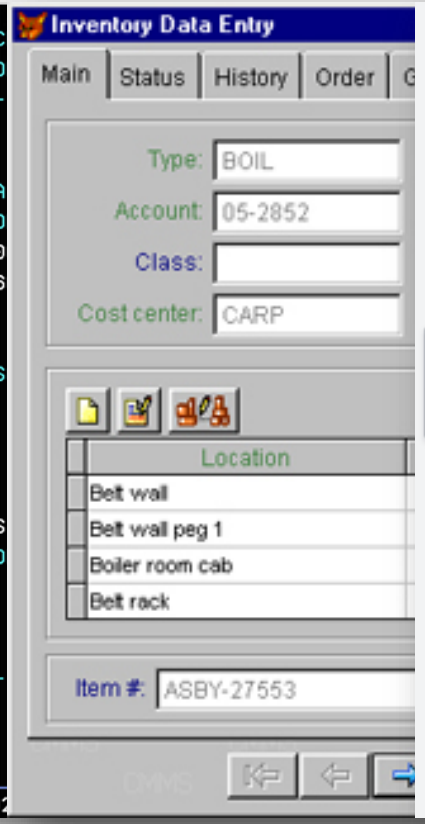


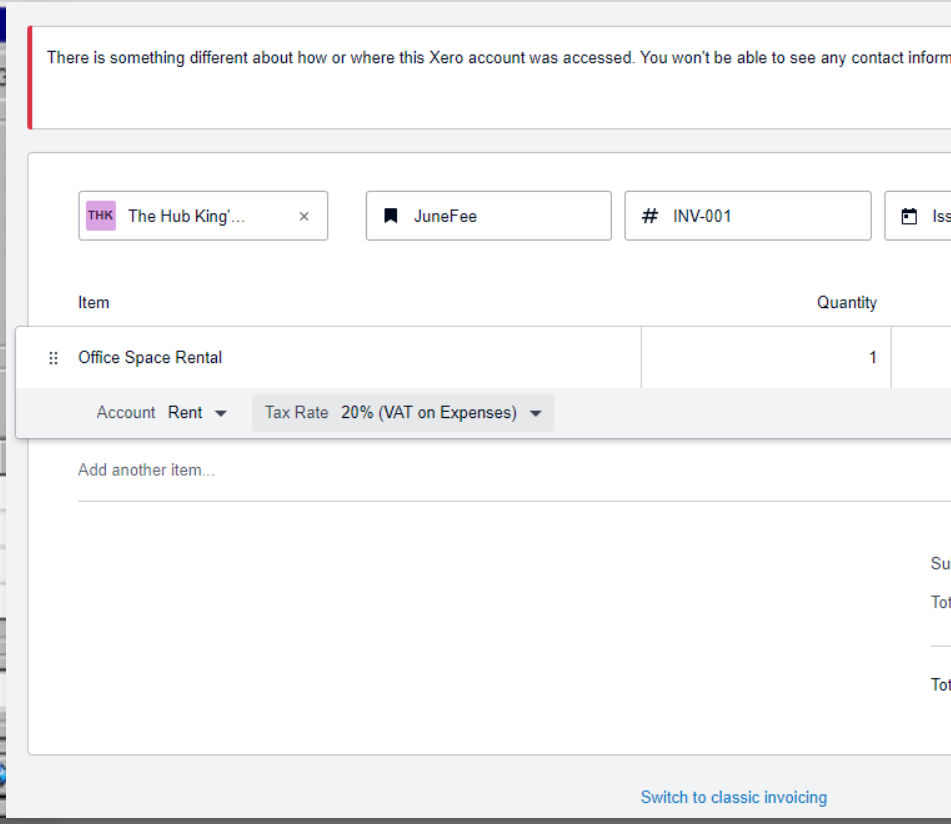
SINGULARITY
TECH DAY

#STechDay2019

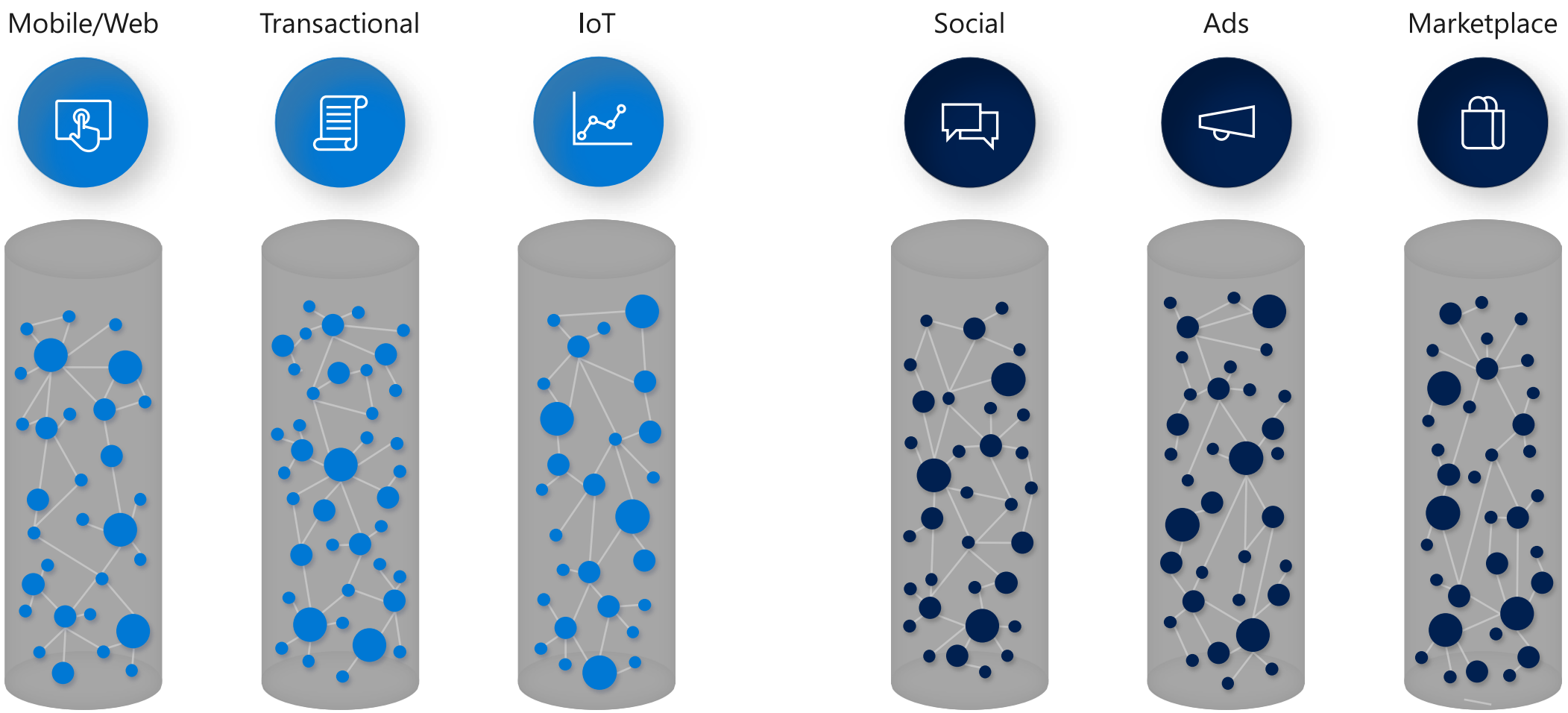
… but what about our data systems?



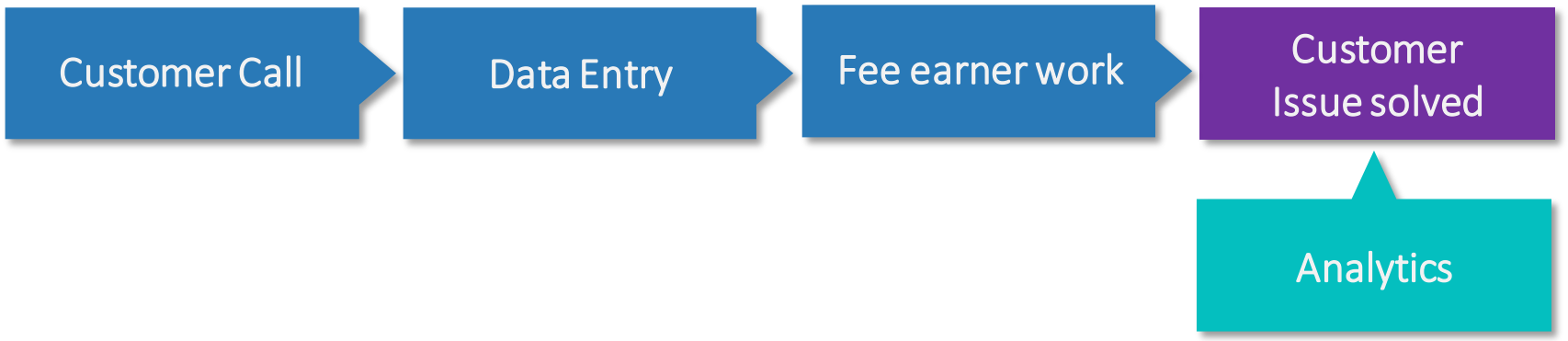




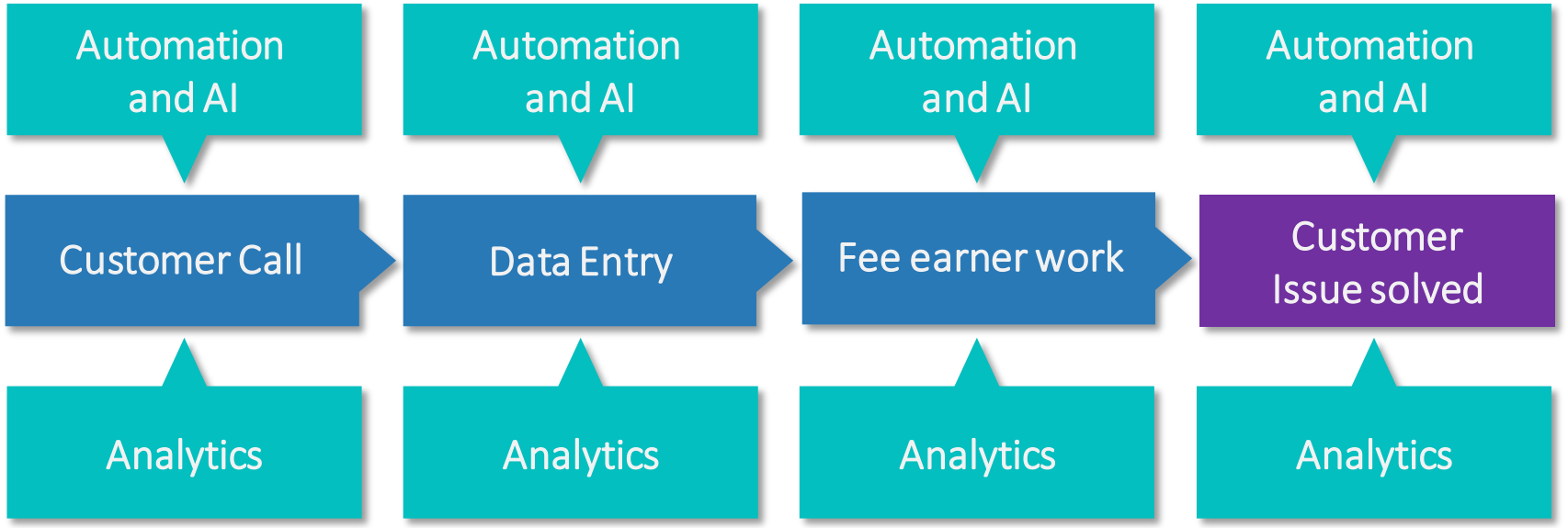
We have lots of data, but it is in silos



Reactive System Example

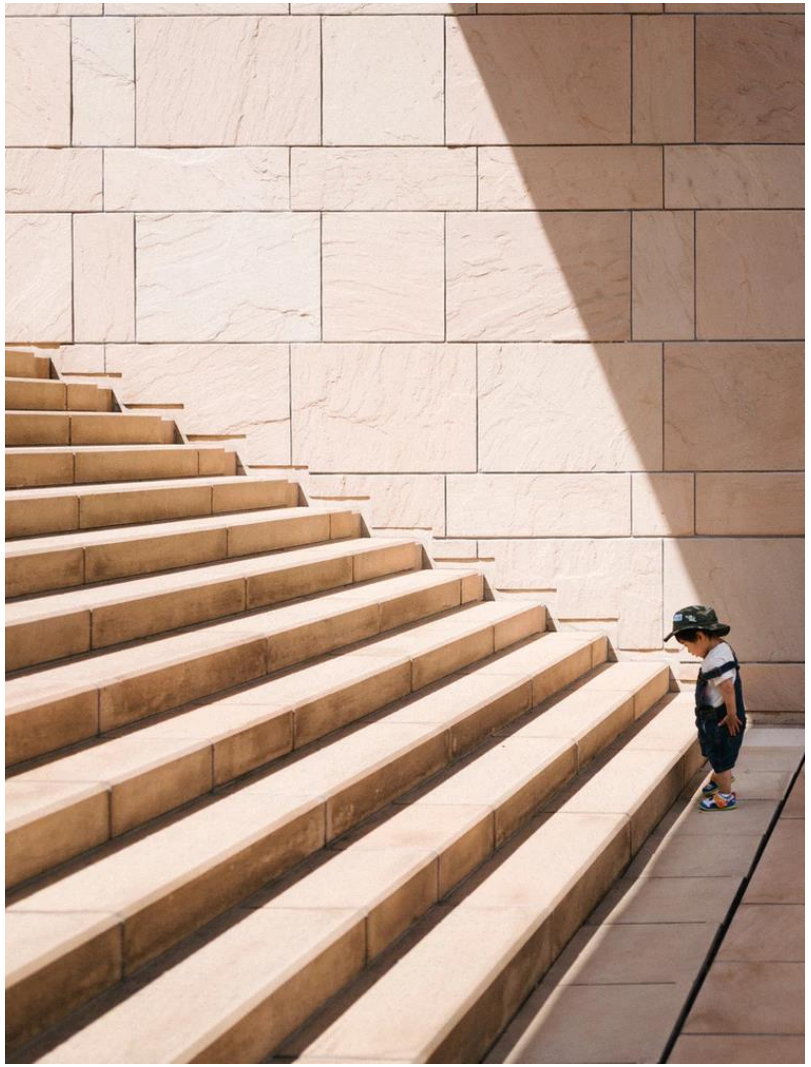


Proactive System Example



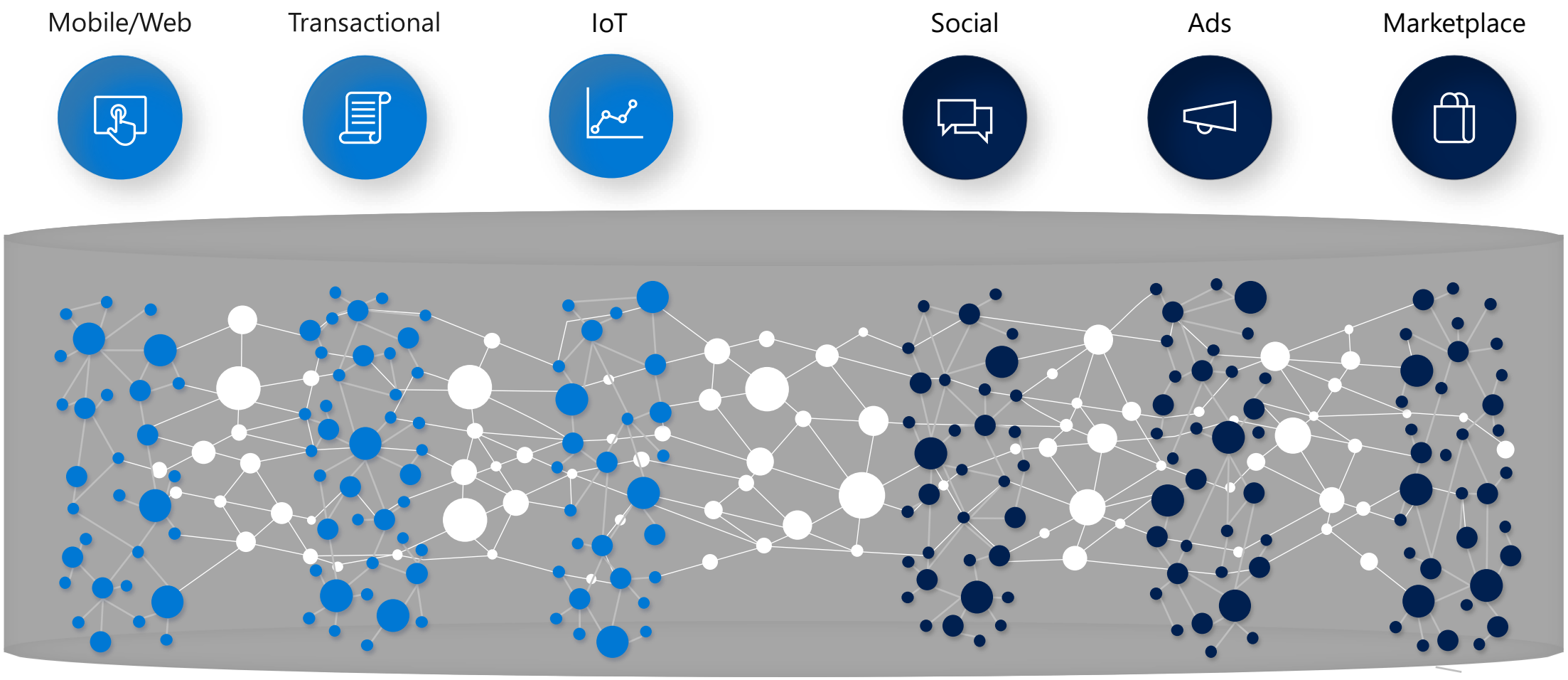
Some challenges ahead…

- Multi-modal support (tabular, video, audio…)
- Support advanced analytics through the data lifetime
- Promote rapid prototyping, and rapid productionalization
- Data dispersal due to proprietary systems and formats
- Lack of common semantic model

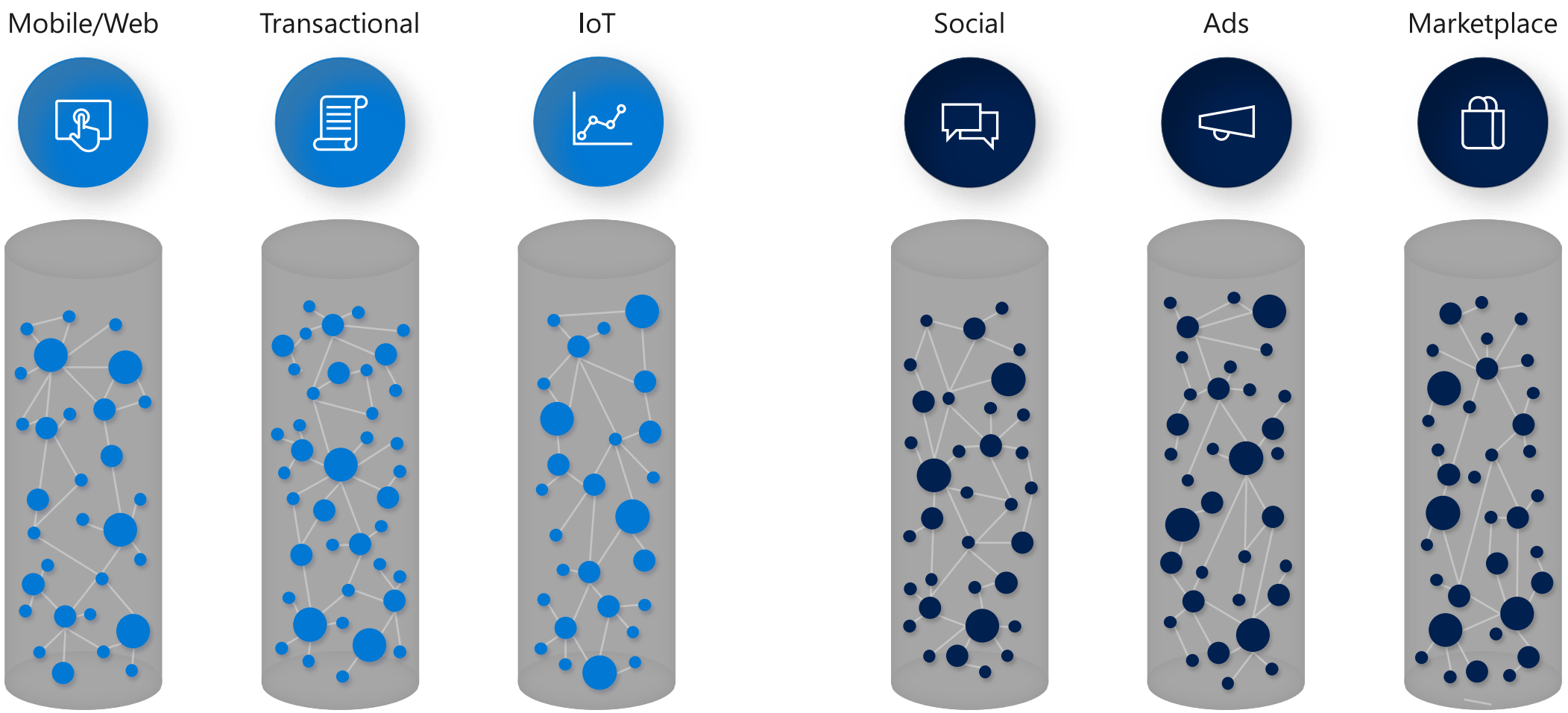


A bit of data history…

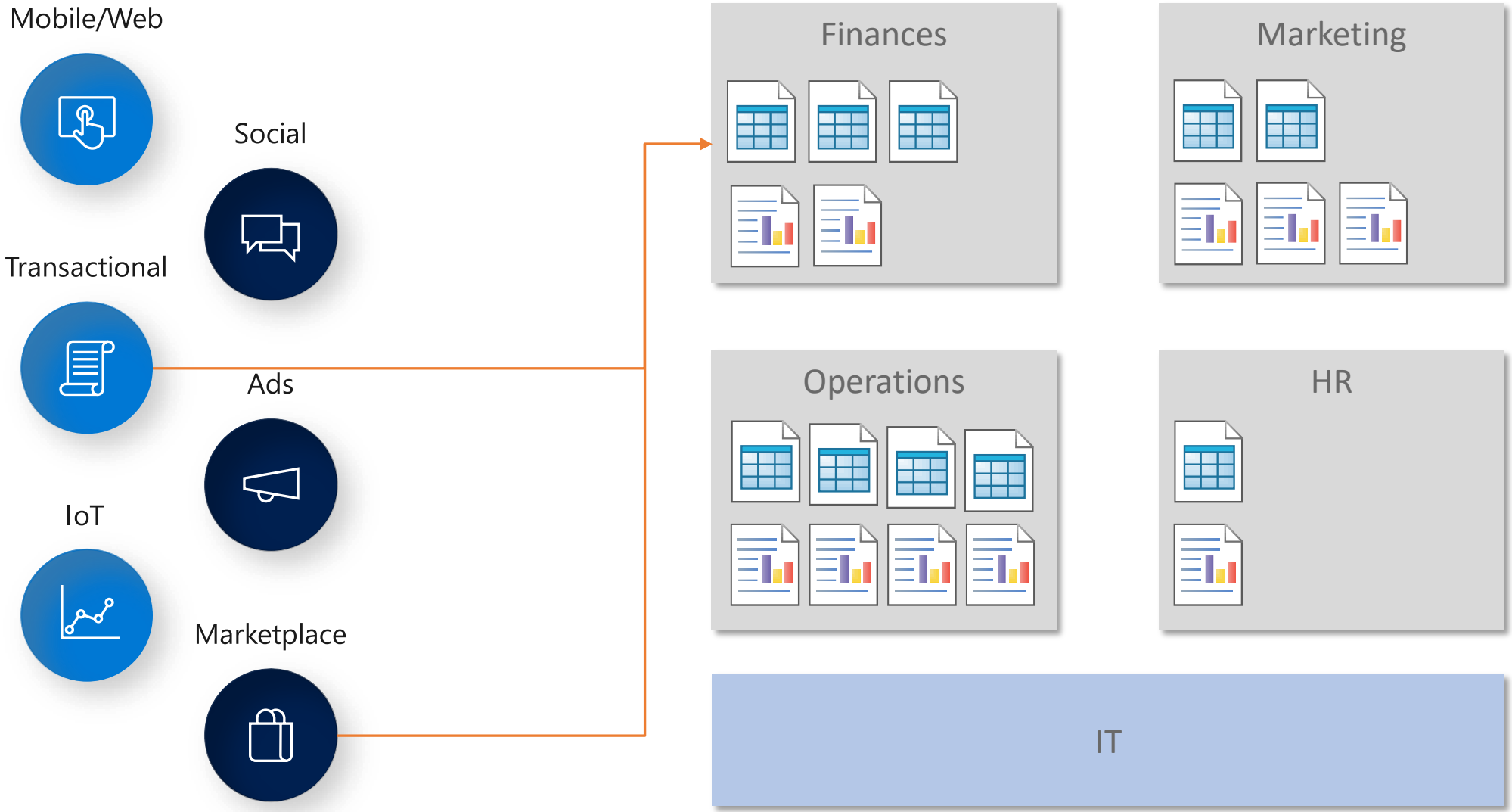
We need semantic models



Once upon a time…



Once upon a time…



Lessons Learned

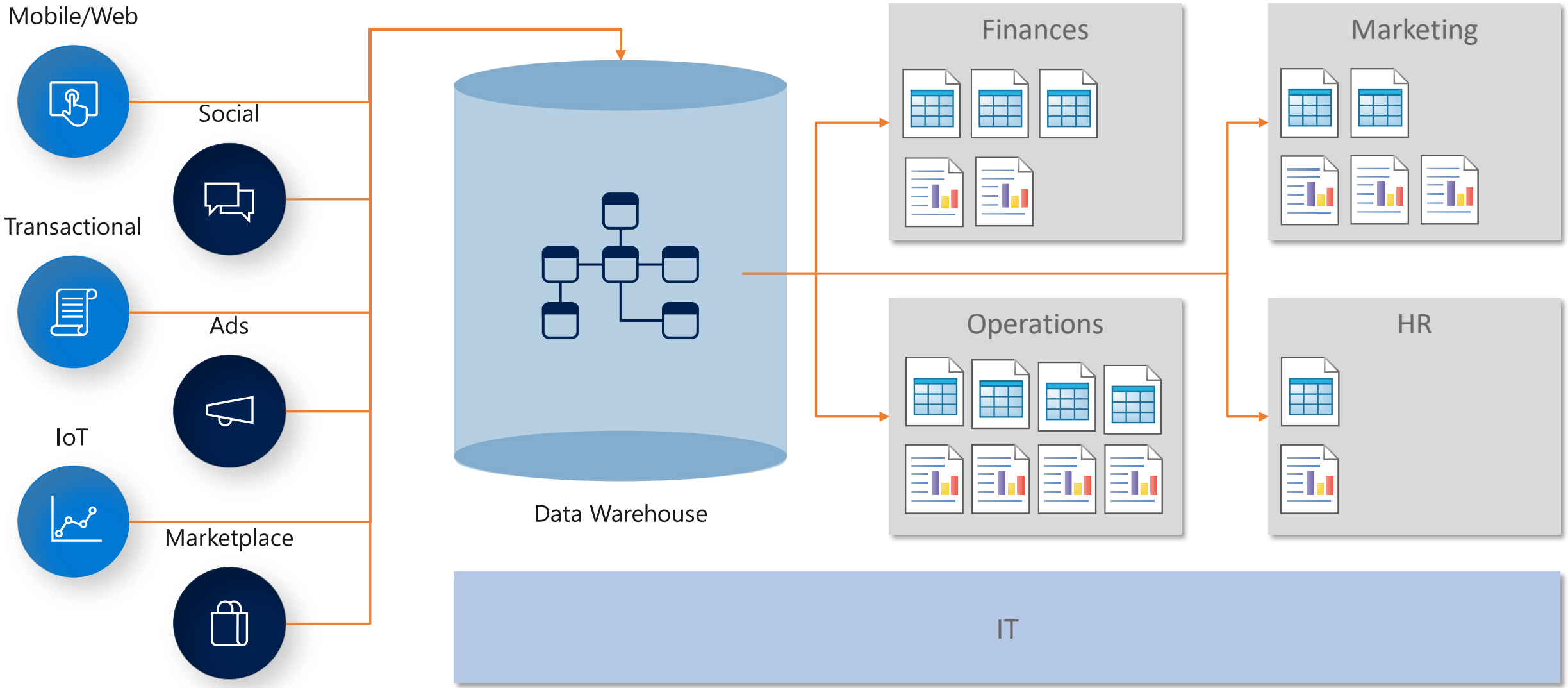
End users / departments require certain **technical skills**

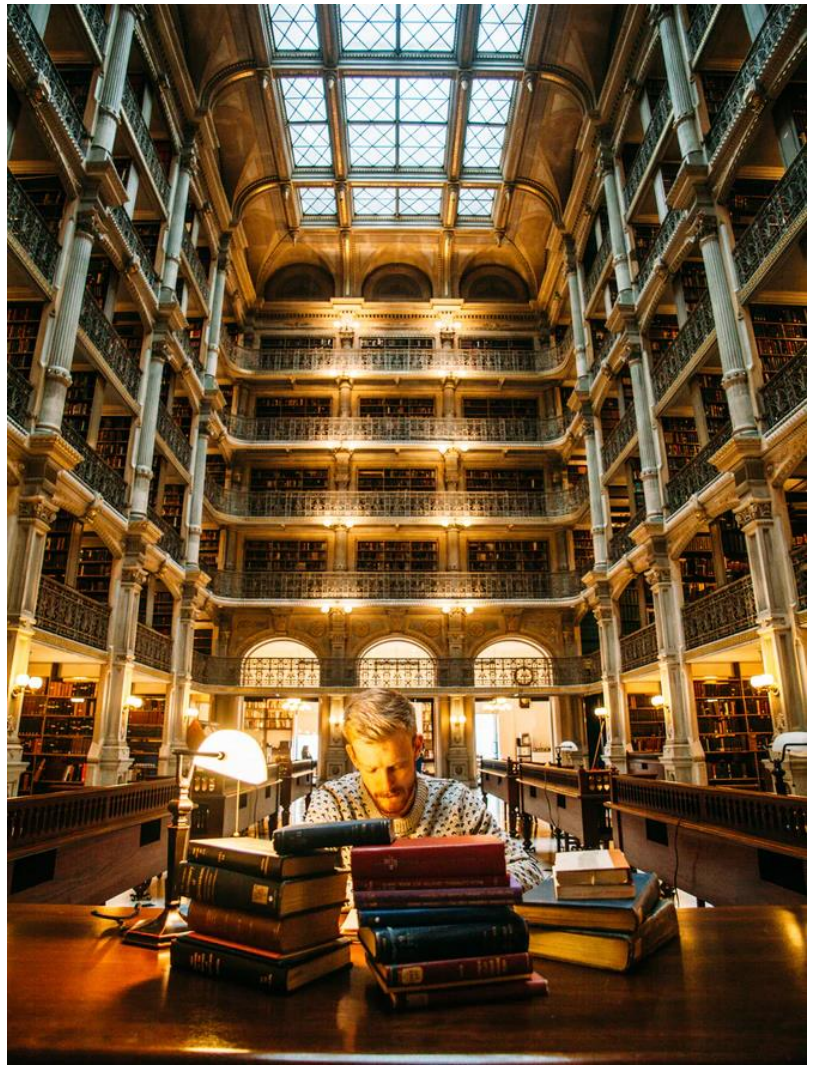
Ensure **veracity and data quality** is practically impossible

Data lineage management is also nearly impossible

Balance between **cost** and **data governance**

The mighty Data Warehouse…





Lessons Learned

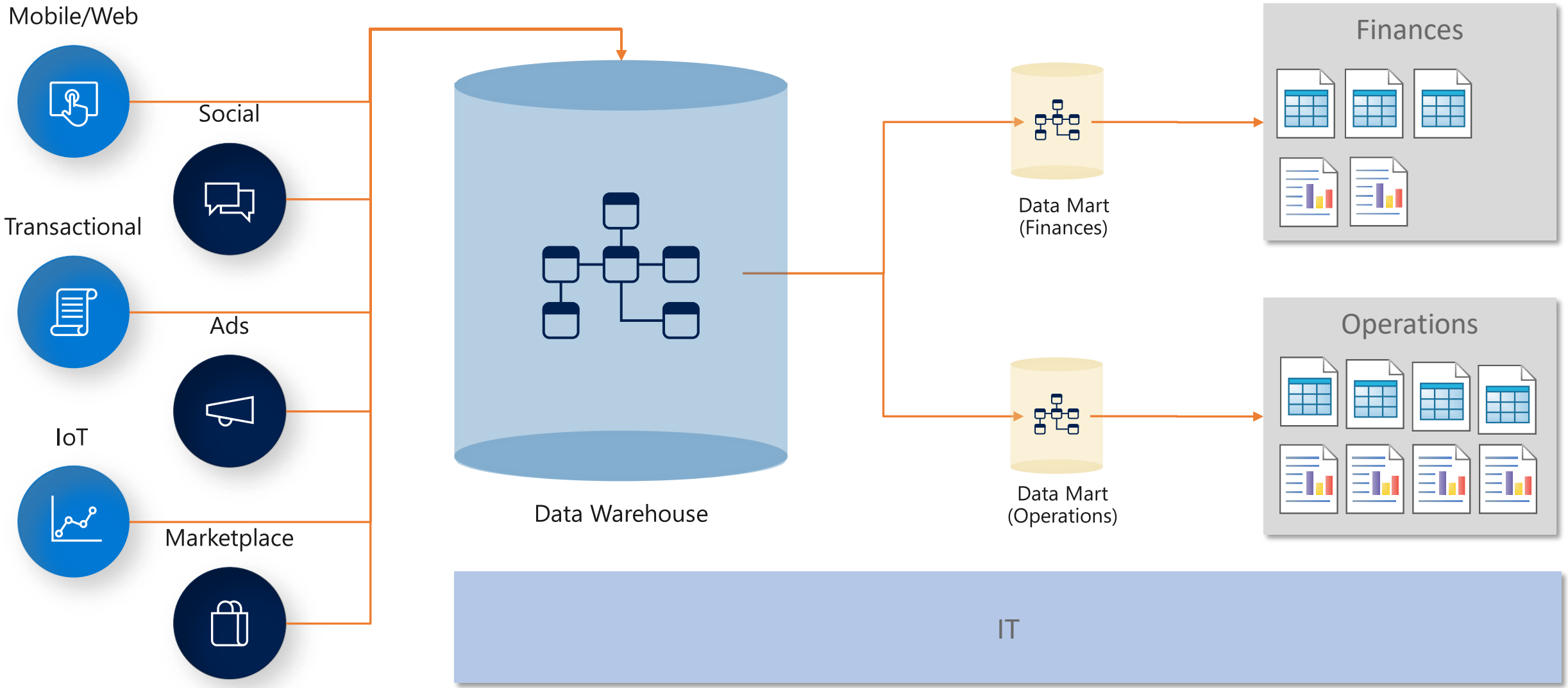
High cost in trying to create a **single version on truth**

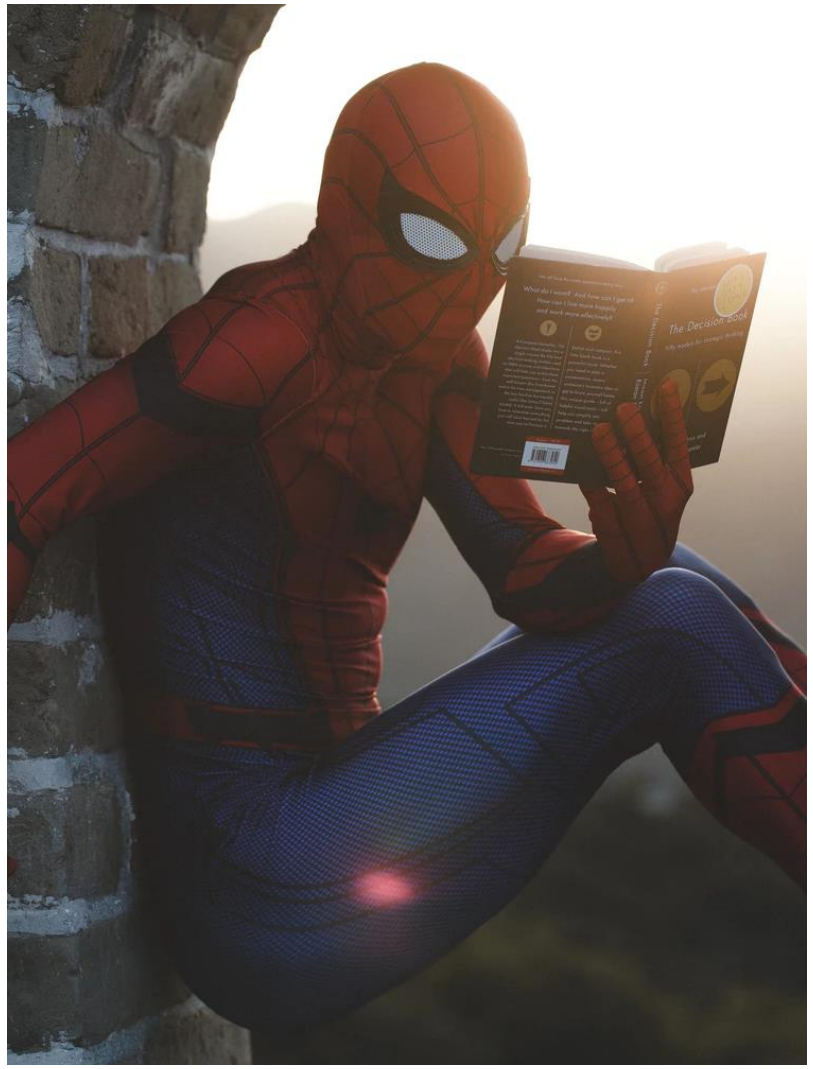
Difficult to change without impacting the rest of the org

Lack of multi-modal support (Video, Audio…) and **ML**

Limited Support for **Streaming**

The Data Mart approach…





Lessons Learned

Manage **master data i**s a very complex challenge

Requires **extra capacity and skills** on IT team

Lack of multi-modal support (Video, Audio…) and **ML**

Limited Support for **Streaming**

The New World

Real Time Data

Highly Volatile Data Structures

Hybrid and Multi-vendor Ecosystems

AI/ML Capabilities

SINGULARITY
TECH DAY

#STechDay2019

Data Lakes

The promise of the Data Lake

Social

Transaccional

Ads

IoT

Marketplace

1 Collect Everything

Storage

Data Lake Storage (Optimized)

2 Store it all on the Data Lake

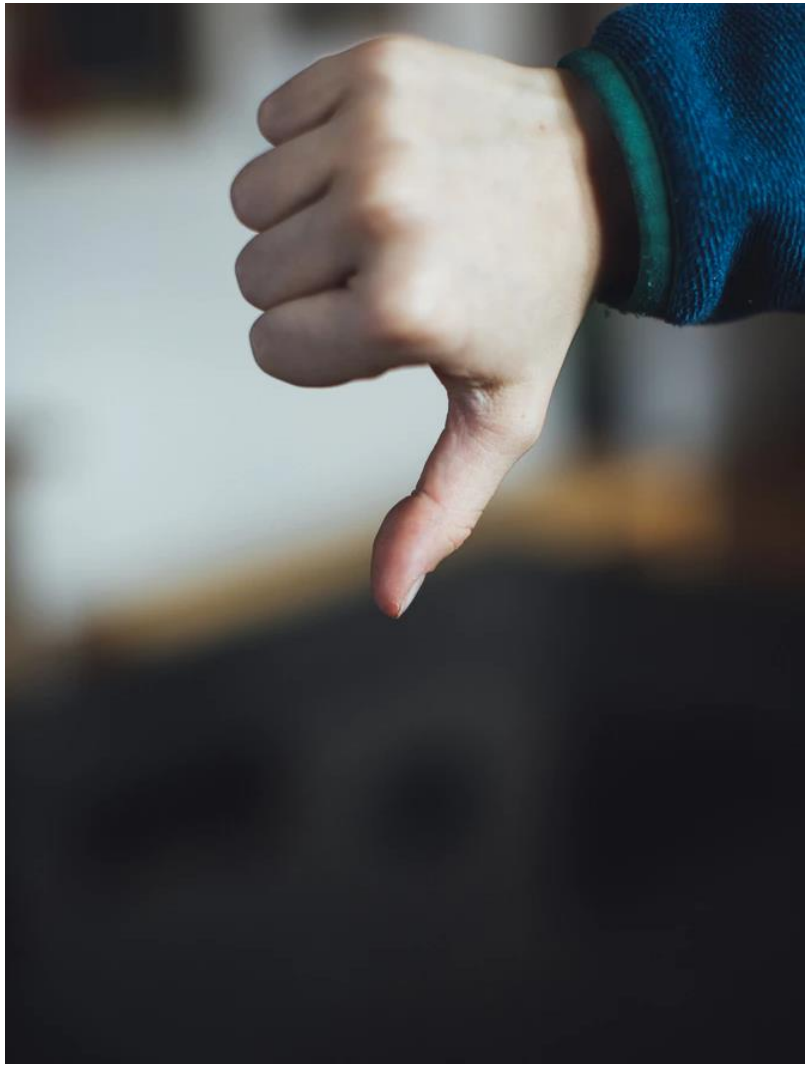
3 Develop business use cases

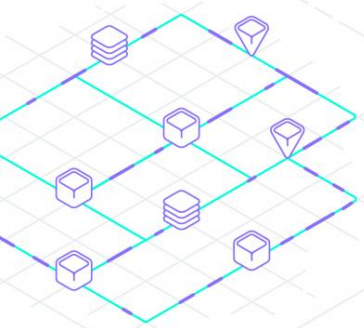
What a Data Lake is *not*

Just storage / Azure Data Lake Storage ☺

Just a Hadoop/Spark/HPC cluster

An more modern kind of Data Warehouse





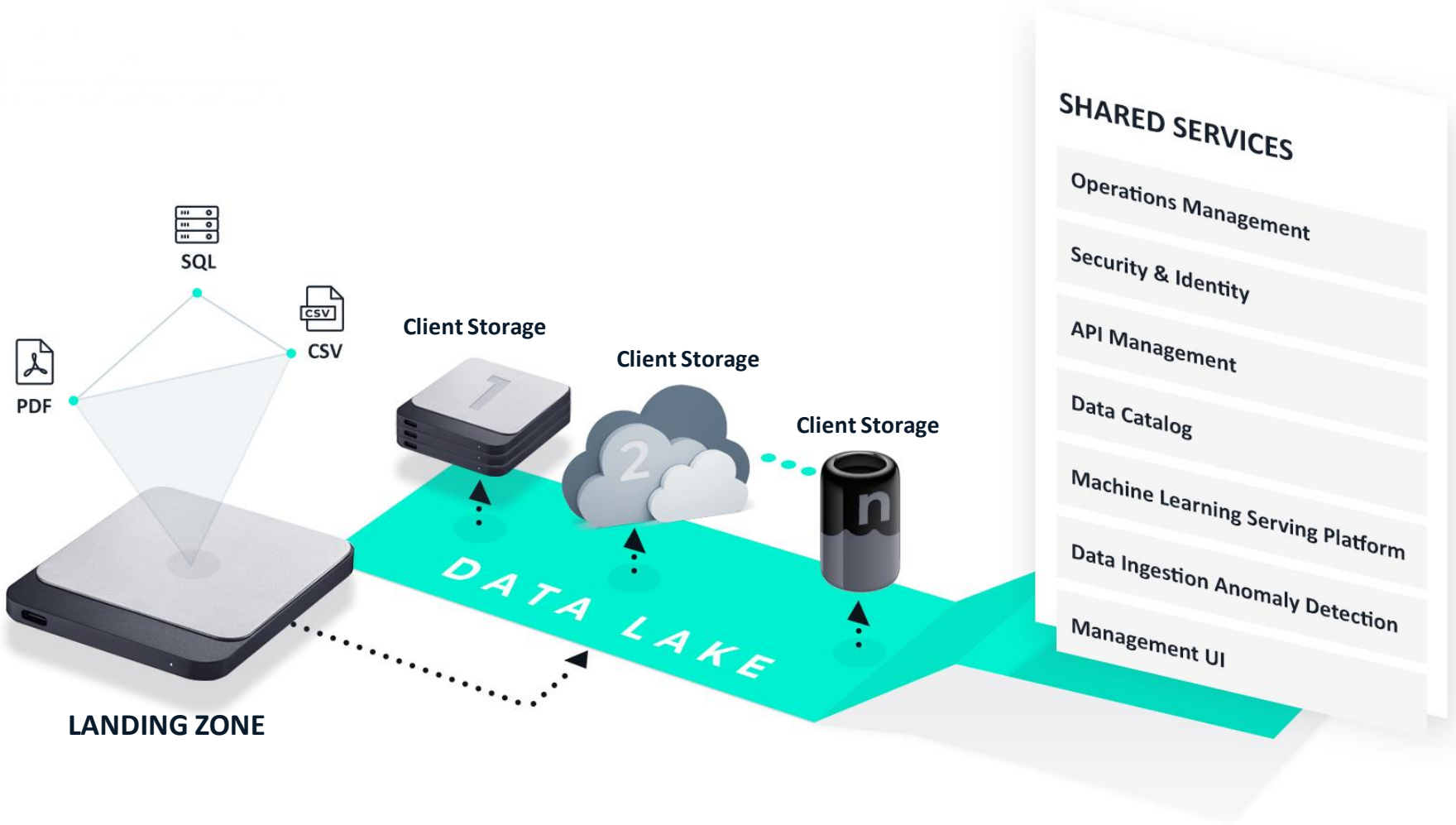
sidra
Data Platform

Disclaimer!

The next slides will be heavily based on our approach to data lakes, as implemented by Sidra Data Platform. Far from a sales pitch, I will use them to explain our approach to some technical challenges; what has worked and what needed to be improved over time.

sidra
Data Platform

#STechDay2019



sidra
Data Platform

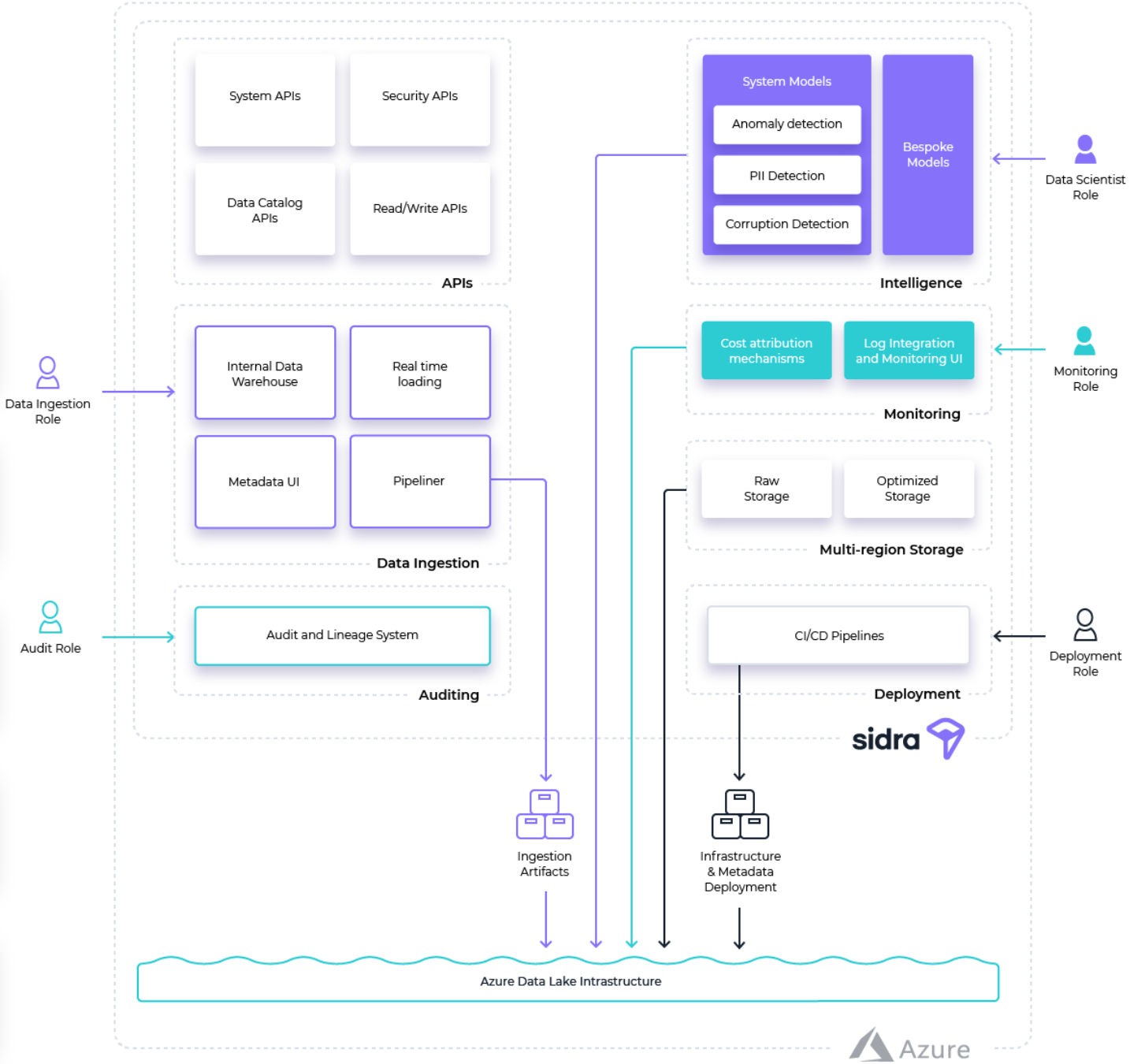
DATA INTAKE

USE CASES / CLIENT APPLICATIONS

SECURITY AND COMPLIANCE

DATA GOVERNANCE

AI & ML ENABLEMENT

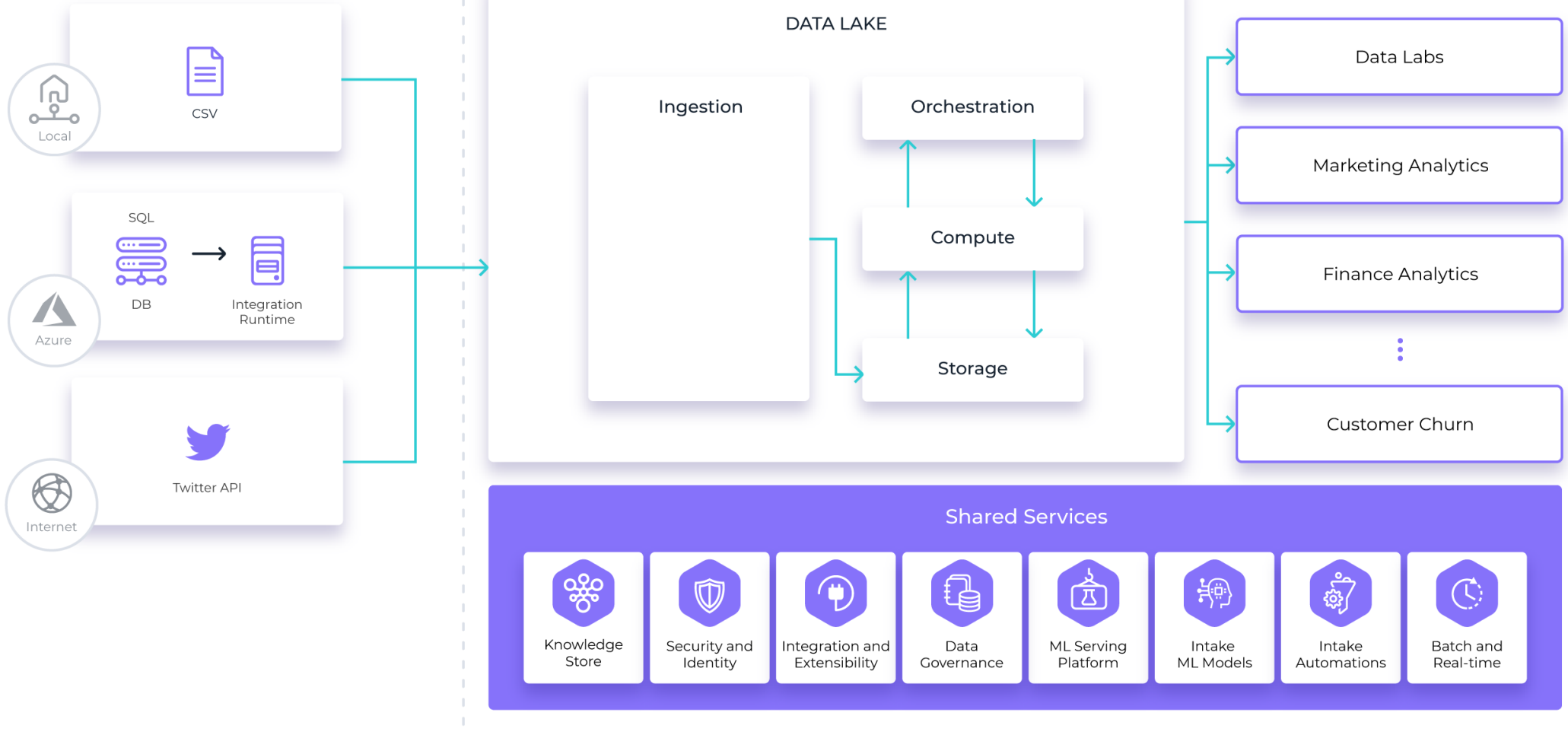


sidra
Data Platform

#STechDay2019

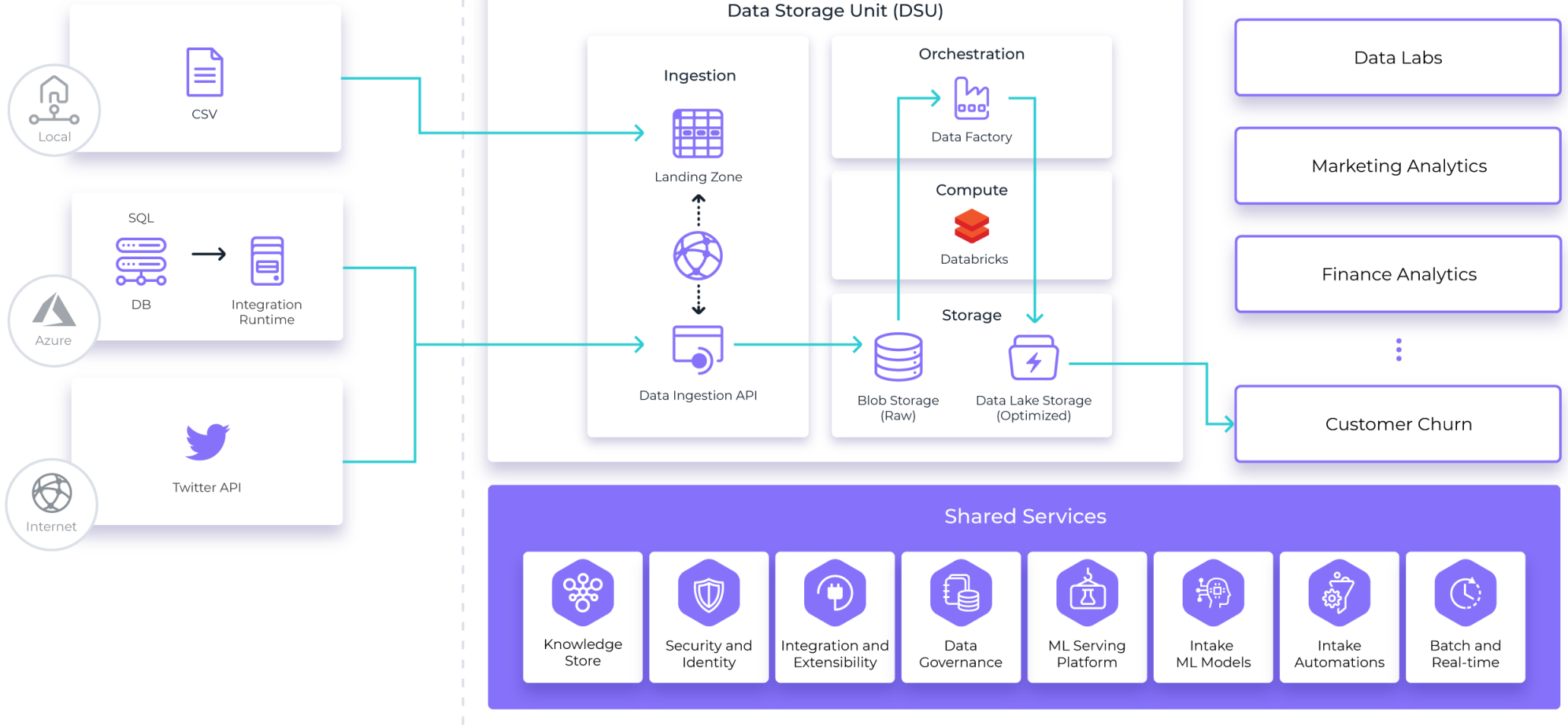
sidra
Data Platform

#STechDay2019



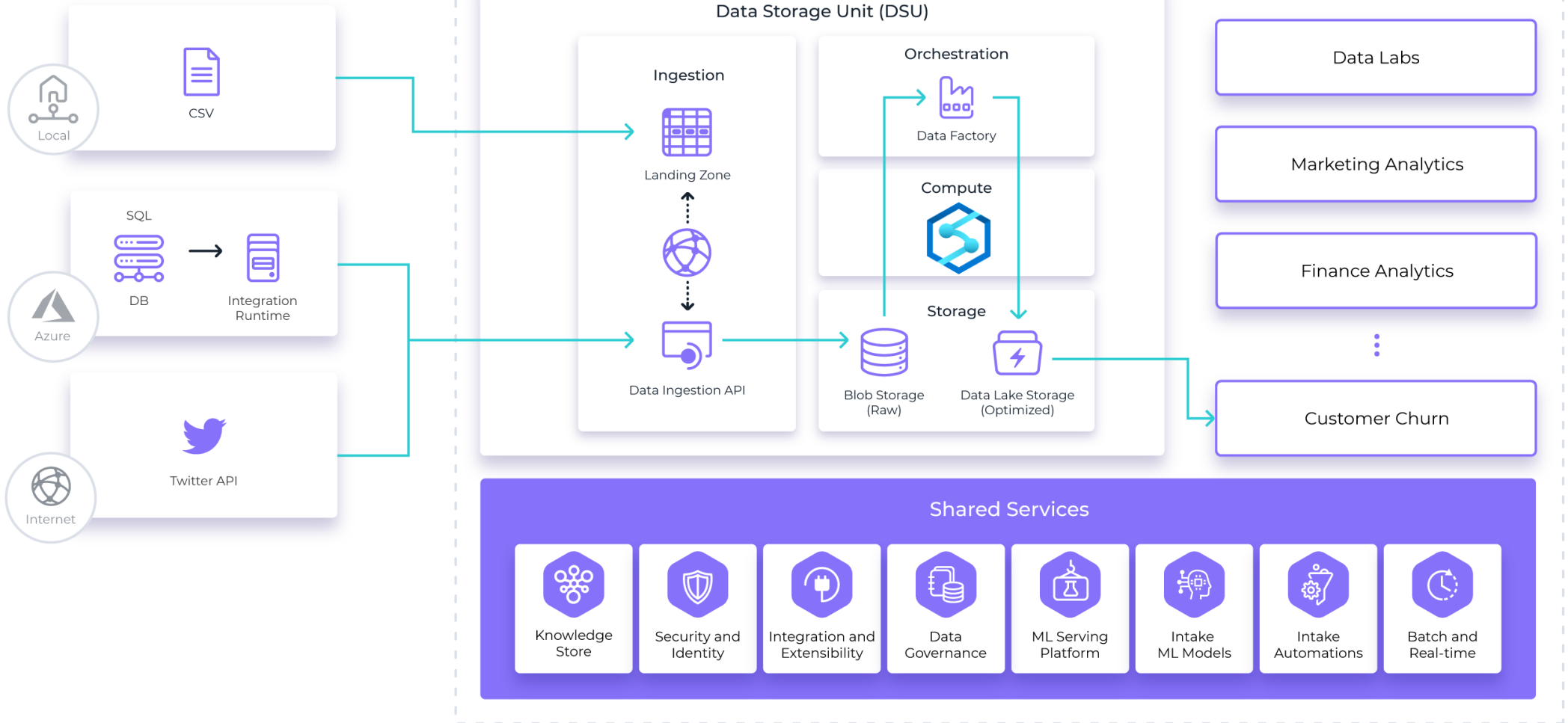
sidra
Data Platform

#STechDay2019



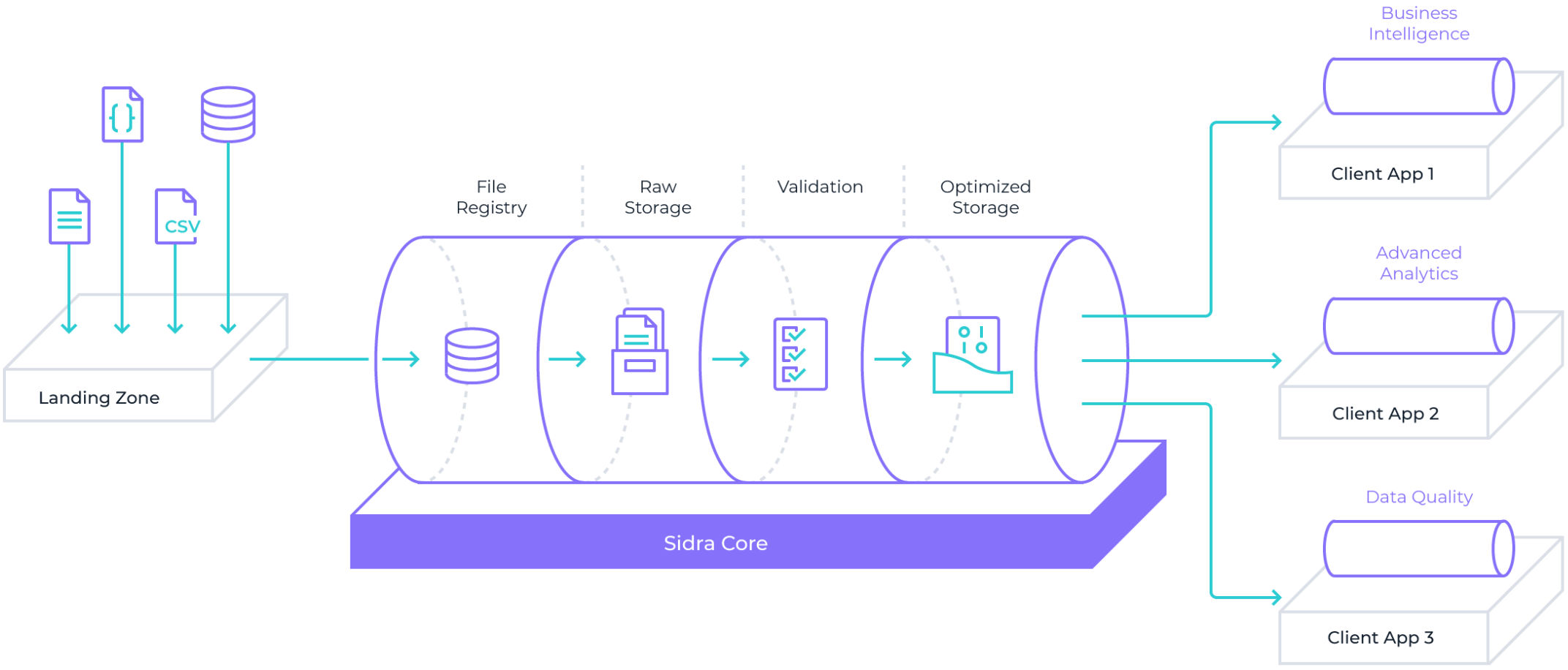
sidra
Data Platform

#STechDay2019



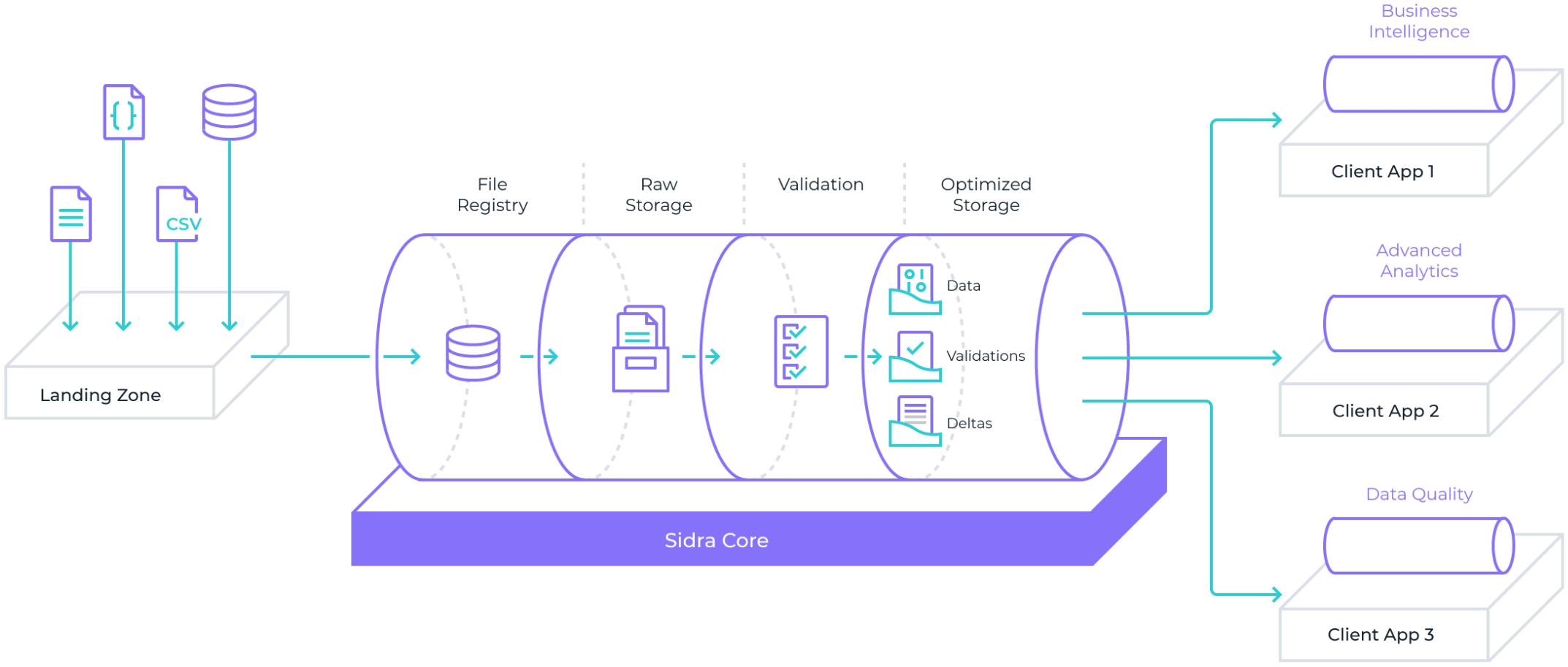
sidra
Data Platform

#STechDay2019



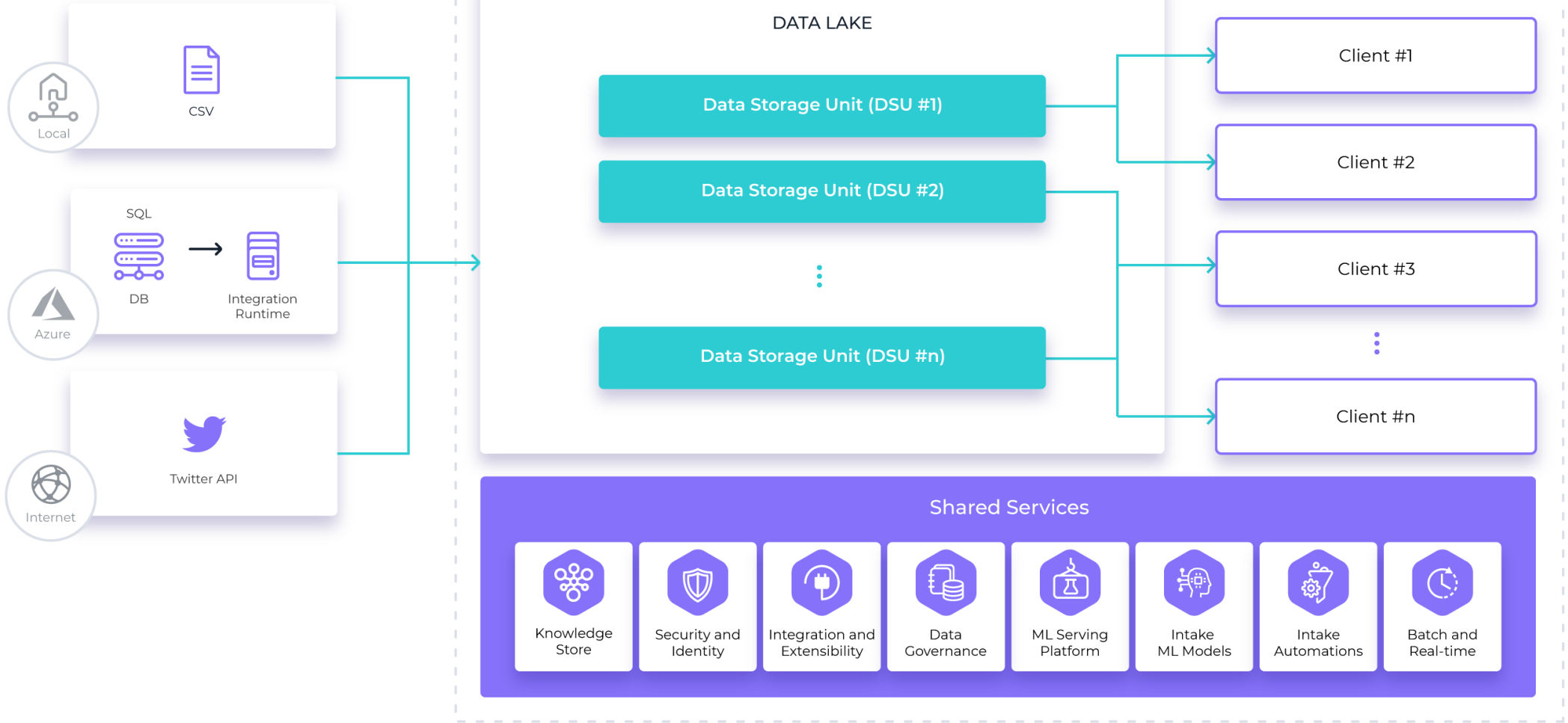
sidra
Data Platform

#STechDay2019



sidra
Data Platform

#STechDay2019



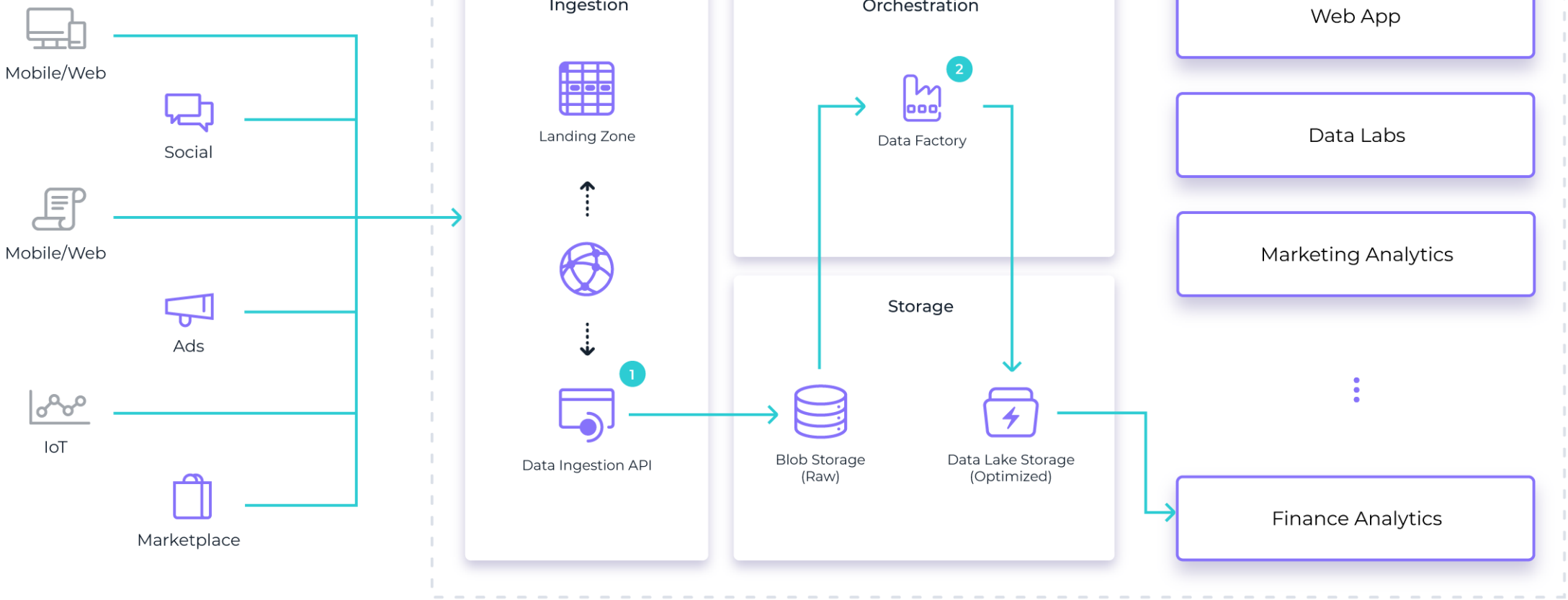
SINGULARITY
TECH DAY

#STechDay2019

Enabling Business Cases via Client Apps

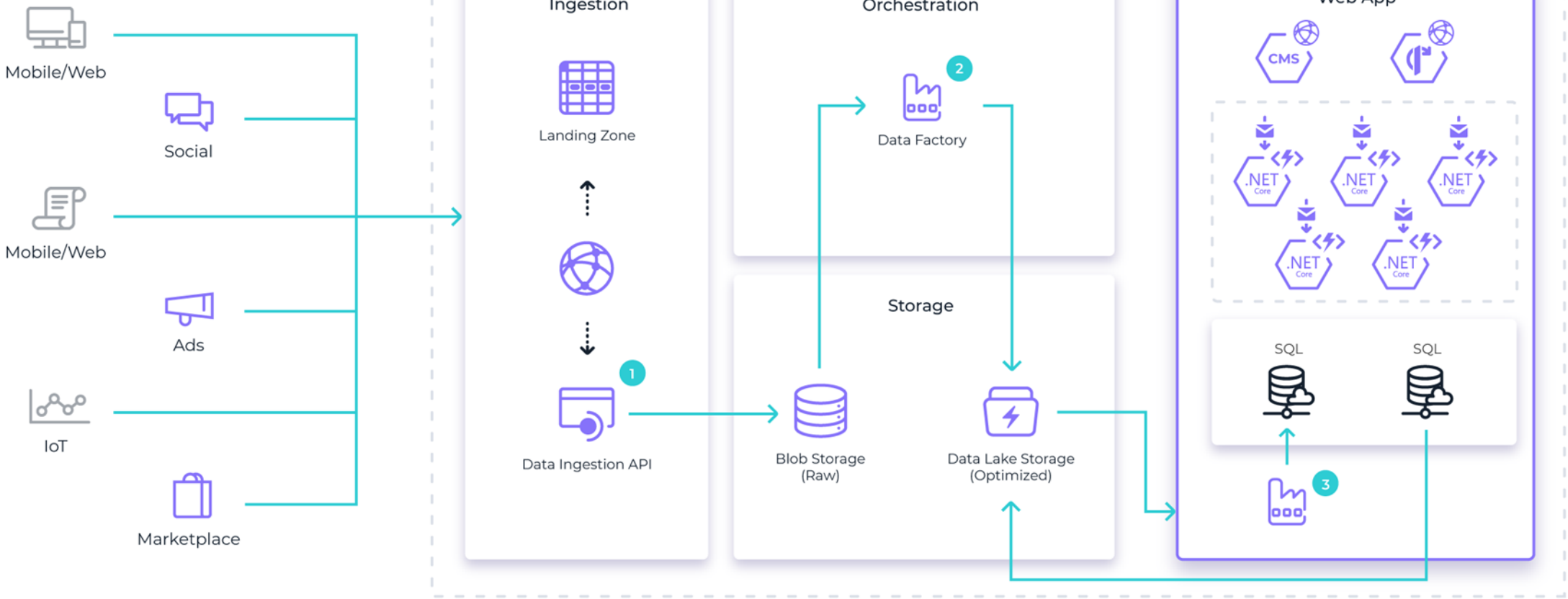
sidra
Data Platform

#STechDay2019



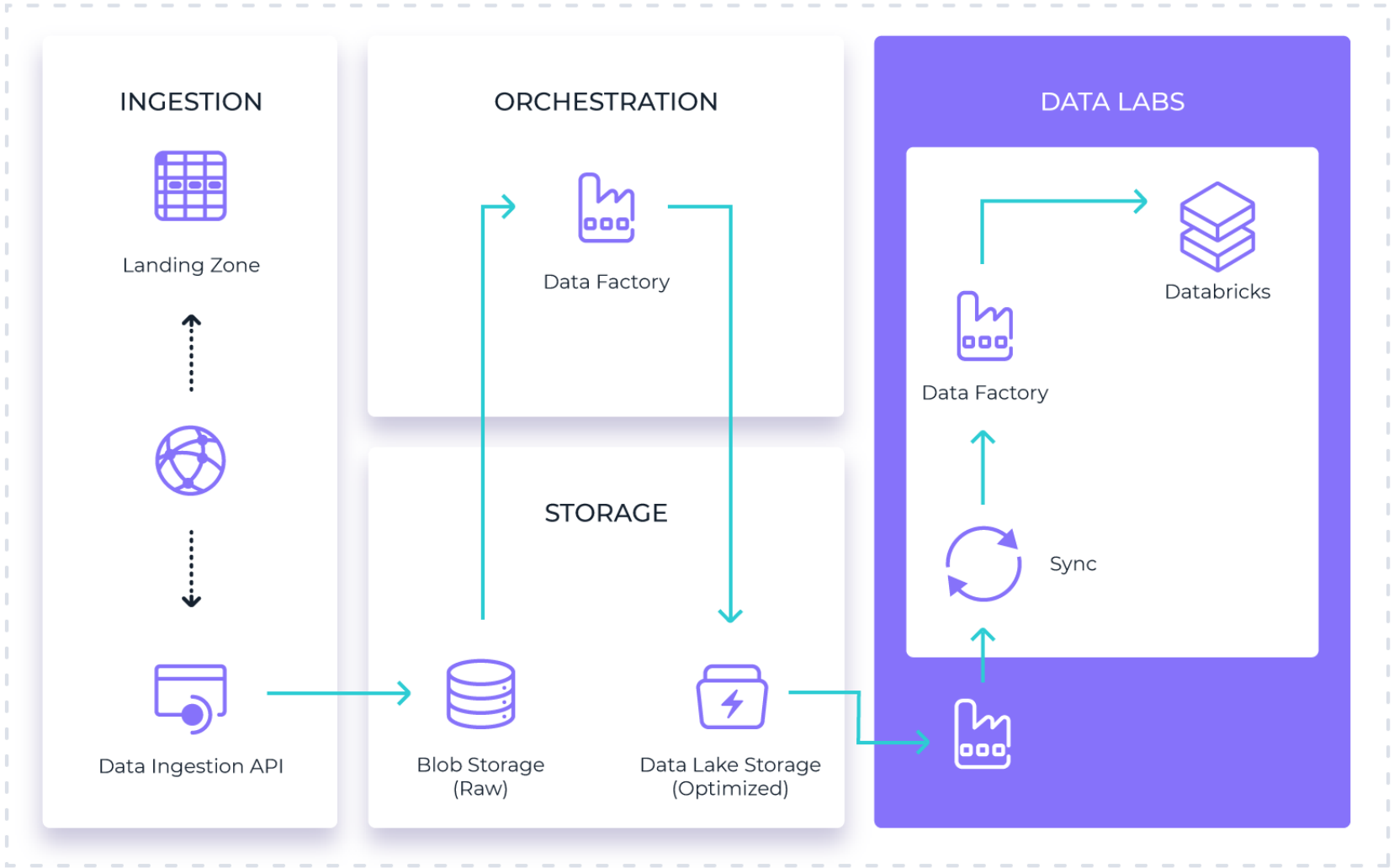
sidra
Data Platform

#STechDay2019



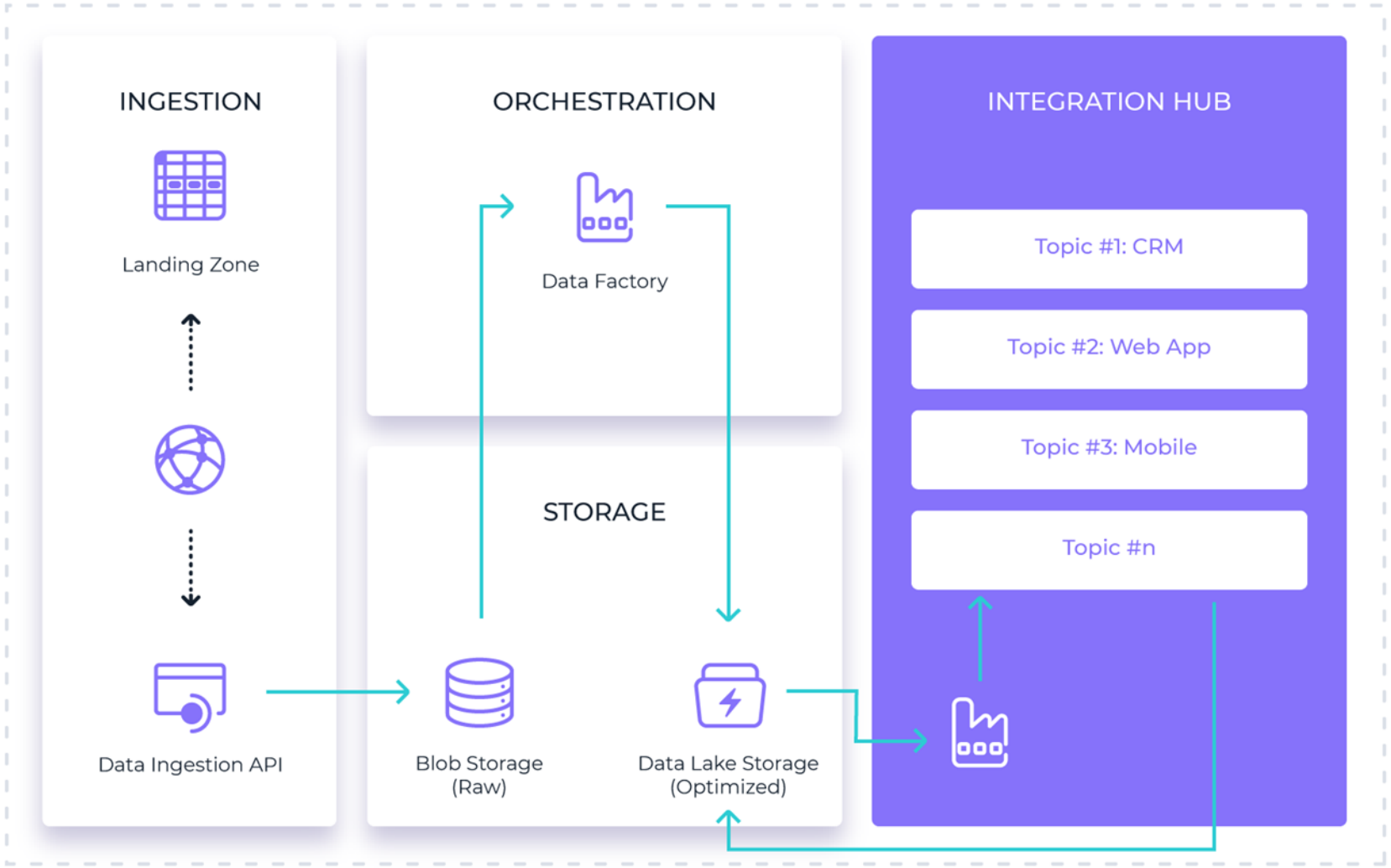
sidra
Data Platform

#STechDay2019



sidra
Data Platform

#STechDay2019



SINGULARITY
TECH DAY

#STechDay2019

The road ahead...

Challenges with Data Lakes

- Deployment Complexity
- Real-Time (lambda architectures)
- Reprocessing (Validations, etc…)
- Append, Update and Merge (Right to be forgotten, etc…)
- Keeping Historical Versions of Data

Enter the Delta Engine

Full ACID Transactions

Based on Parquet – Open Standard, Open Source

Runs on Spark

Enter the Delta Engine

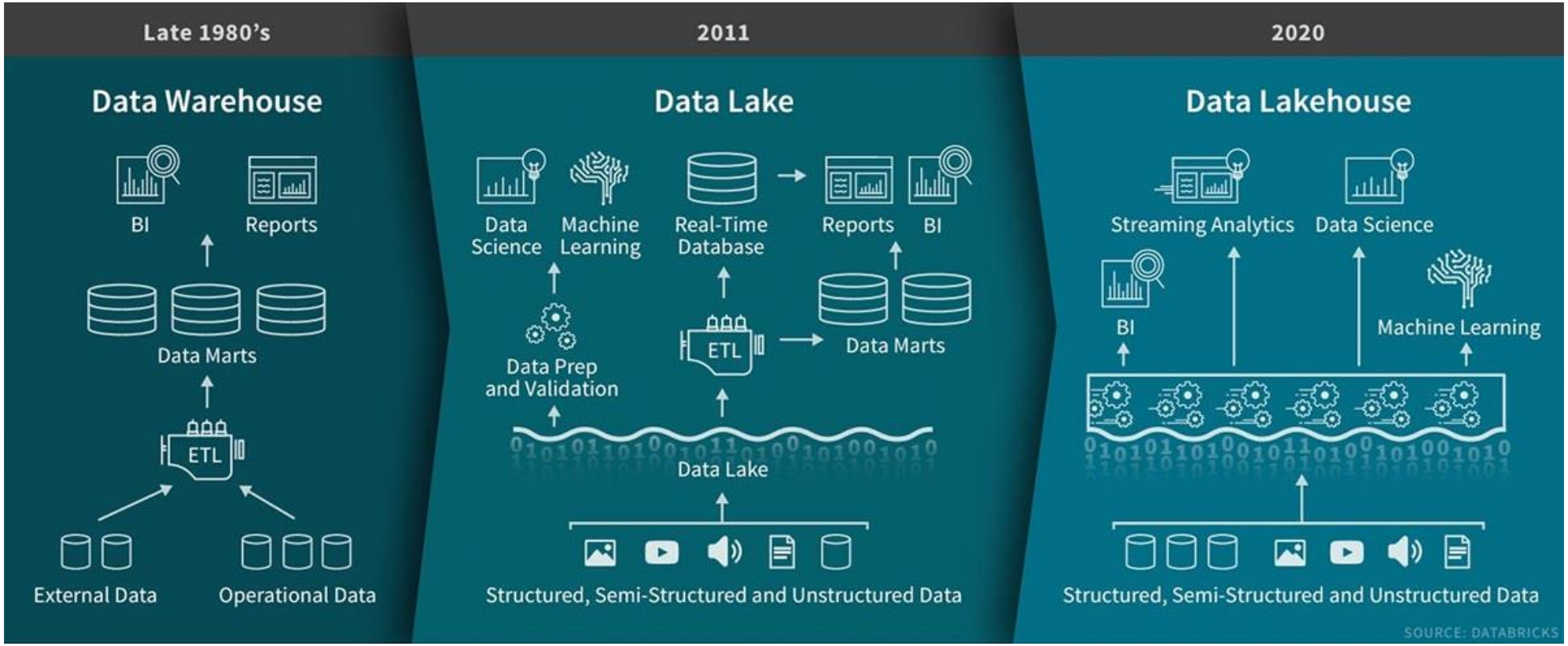
DELTA LAKE

Ingestion Tables
(Bronze)

Refined Tables
(Silver)

Business-level Aggregates
(Gold)

Delta Lake and the Data Lakehouse

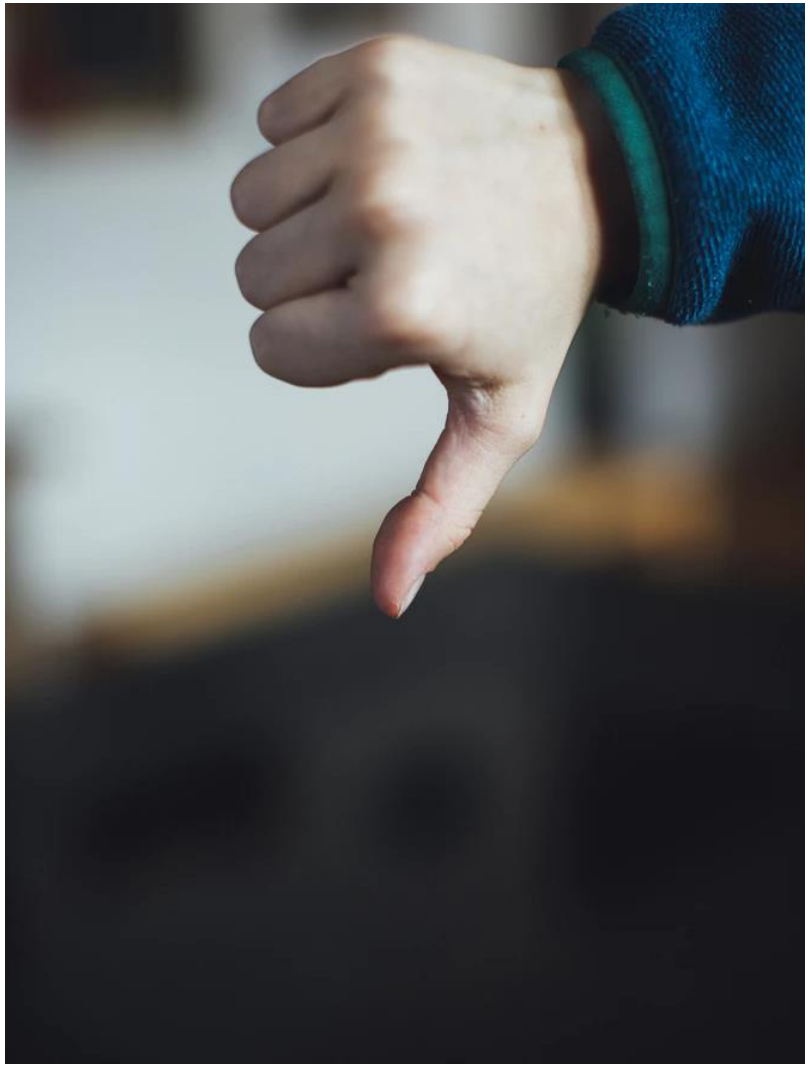


Source: What is a Lakehouse? - The Databricks Blog (30-Jan-2020)

Our lessons learnt…

What have we learnt?

- Integrated Data Quality processes
- Integrated Data Mastering processes
- 3-tier Approach using Transactional Tables



SINGULARITY
TECH DAY

#STechDay2020

THANKS AND...

SEE YOU SOON!

ORGANIZATION

plain concepts

SPONSORS

Microsoft

SUPPORT

World Central Kitchen