Data version control with Delta Lake

Repeatable Analytics with data in motion

/thoughtworks

About us

/thoughtworks

We are a global technology consultancy that integrates strategy, design and engineering to drive digital innovation.

10,000+
Employees

27+
Years

48
Offices

17
Countries

Australia / Brazil / Chile / China
Ecuador / Finland / Germany / India
Italy / Netherlands / North America
Romania / Singapore / Spain
Thailand / United Kingdom

What we do in Data

Note: These offerings overlap!

Data Platforms

CD4ML

Data Mesh

Data Strategy and Governance

Why do we need Data Engineering practices and tools?

/thoughtworks

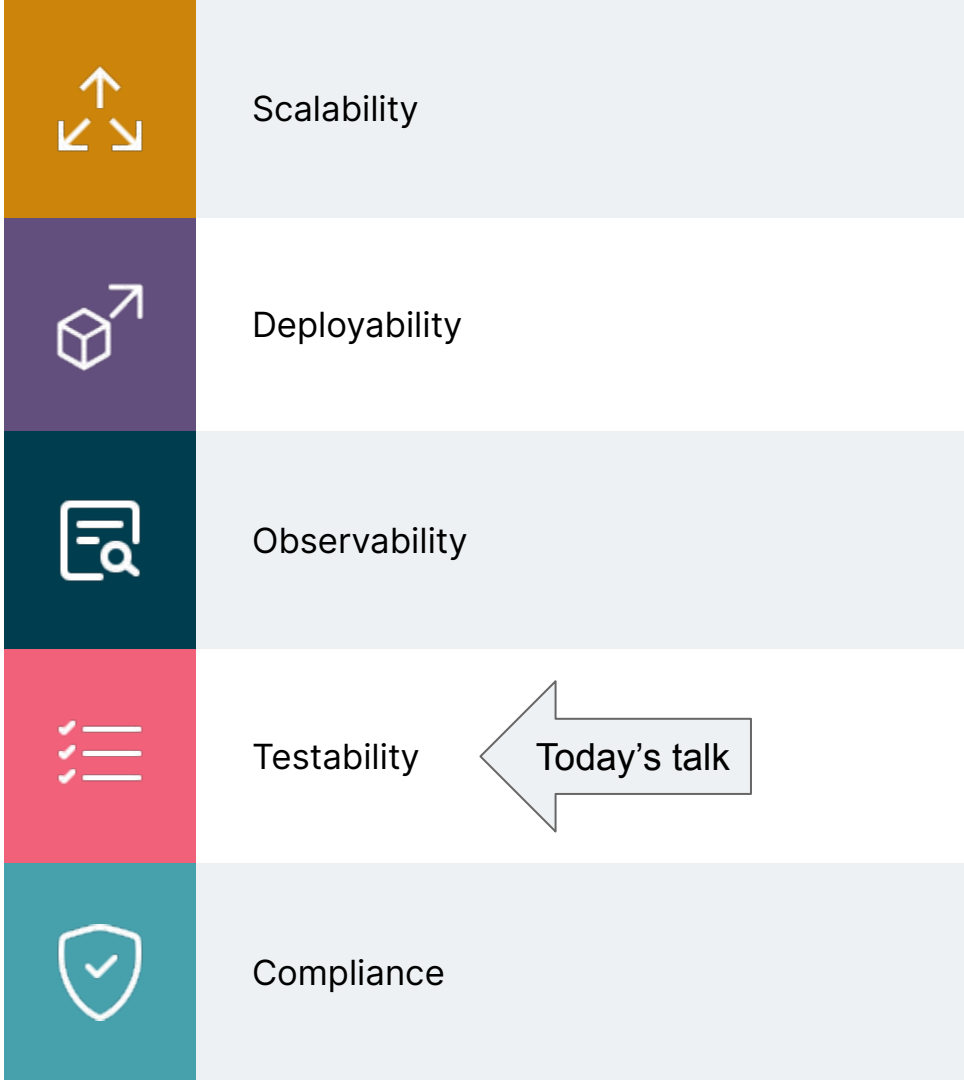
26%

Poor data quality impacts 26% of their companies' revenue

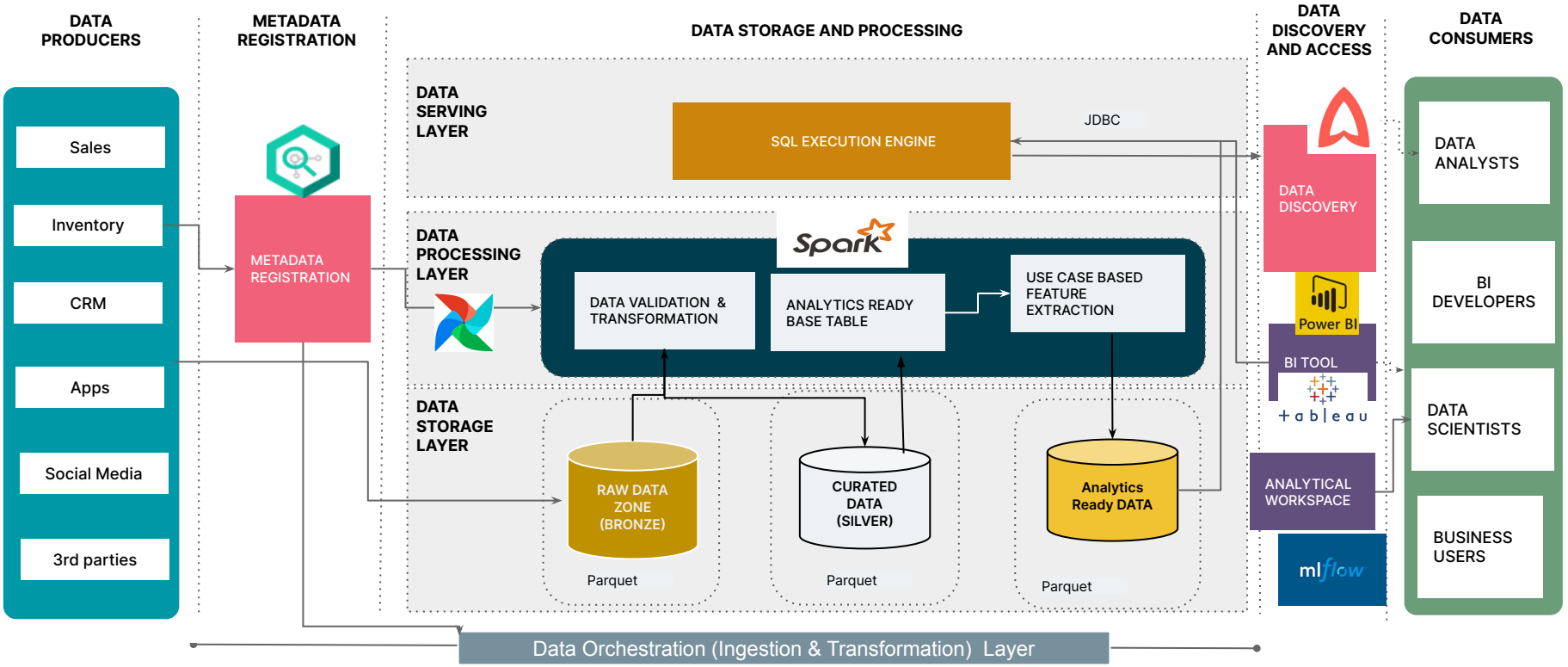
Common CFRs for Data Platform

Not only do we need to make it work, it has to work right as well.

Not only does it need to work right today, it should also work right in the future as well.



Reference Data Platform Architecture



"Instead of focusing on the code, companies should focus on developing systematic engineering practices for improving data in ways that are reliable, efficient, and systematic. In other words, companies need to move from a model-centric approach to a data-centric approach."

Andrew Ng

How can we make our workload repeatable?

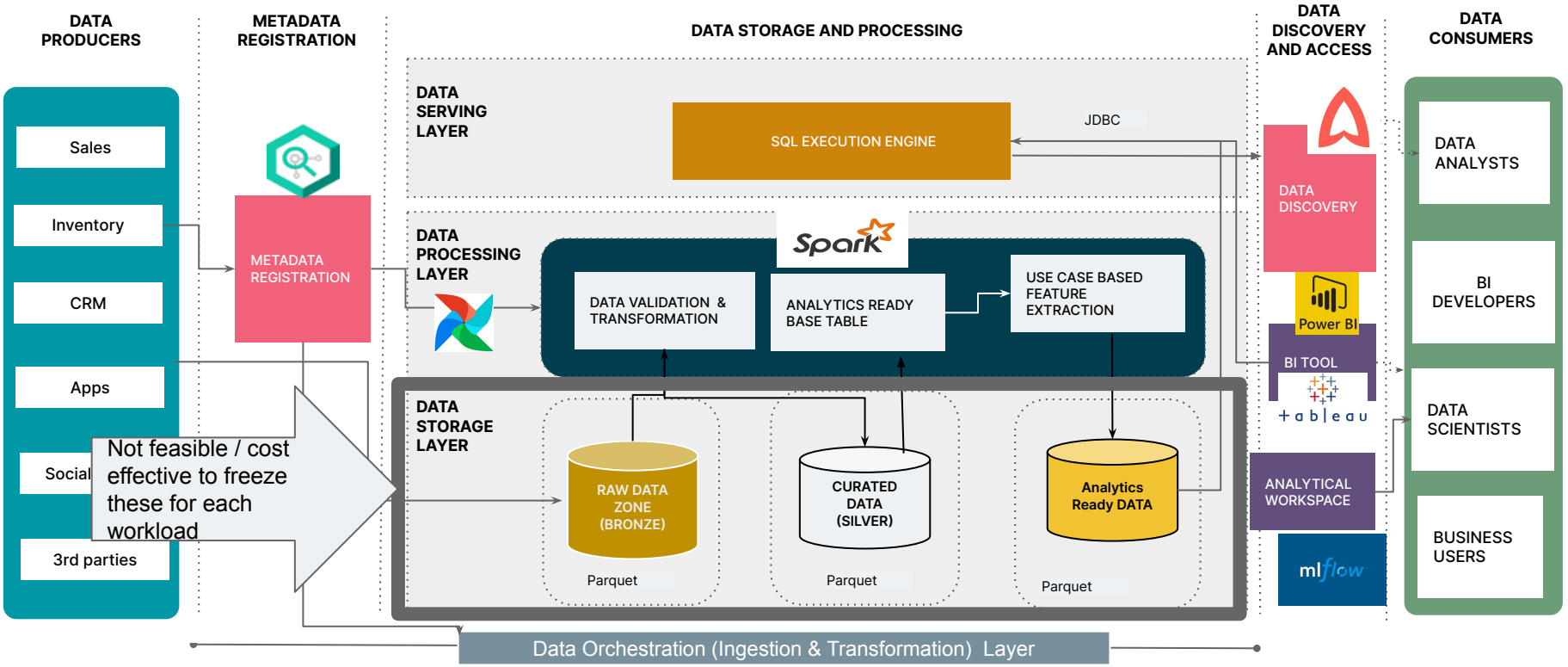
Because some workload can be **adversely affected** from some **variation of data**.

Having the ability to use **versioned data** on the **same infrastructure** can prove to be invaluable.

Moving Data

Versioned

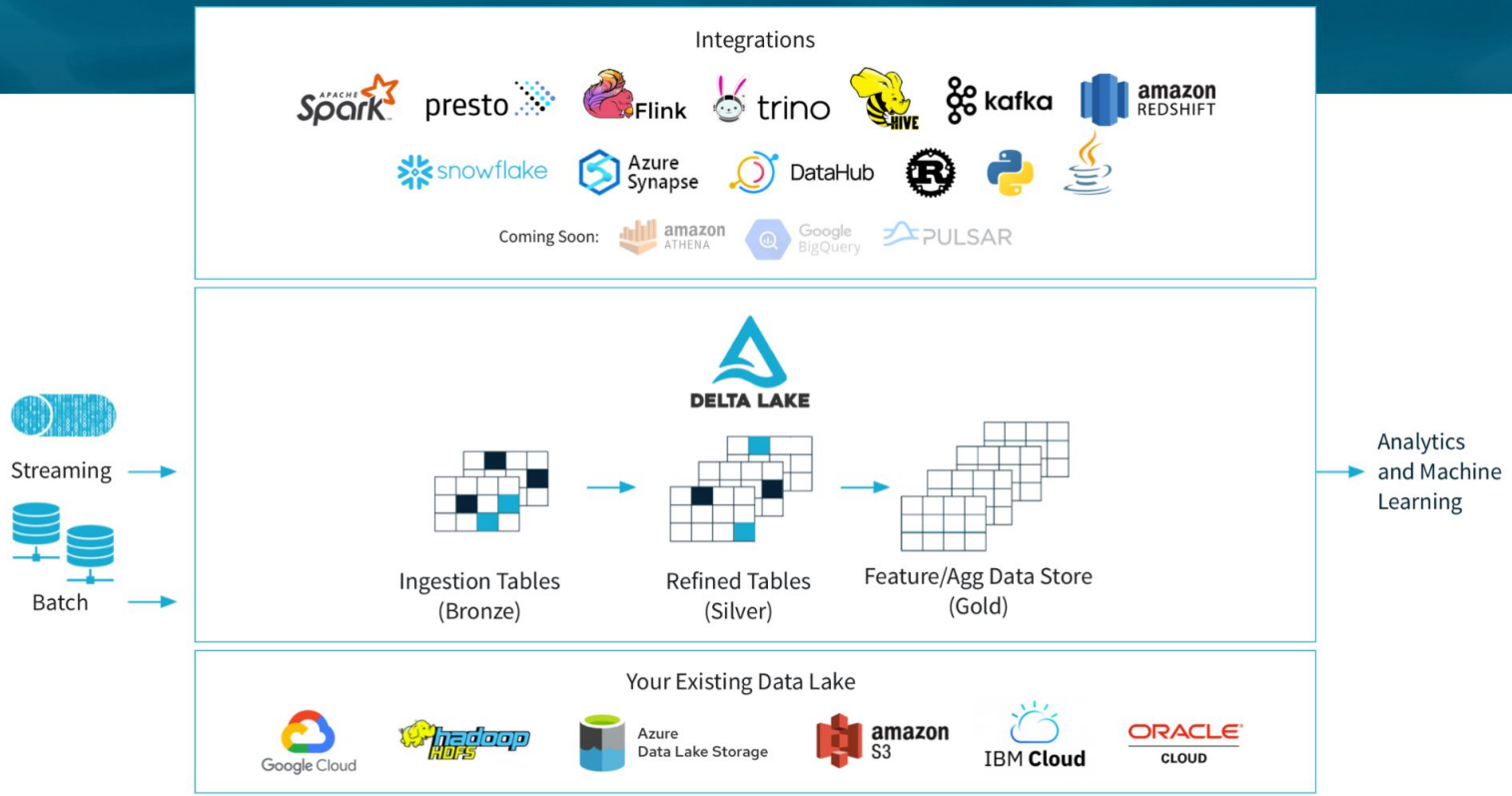
Reference Data Platform Architecture



Delta Lake (Grilling data in motion)

/thoughtworks

DELTA LAKE



databricks

Commercializes by

Key Features

ACID Transactions

Protect your data with serializability, the strongest level of isolation

Scalable Metadata

Handle petabyte-scale tables with billions of partitions and files with ease

Time Travel

Access/revert to earlier versions of data for audits, rollbacks, or reproduce

Open Source

Community driven, open standards, open protocol, open discussions

Unified Batch/Streaming

Exactly once semantics ingestion to backfill to interactive queries

Schema Evolution / Enforcement

Prevent bad data from causing data corruption

Audit History

Delta Lake log all change details providing a fill audit trail

DML Operations

SQL, Scala/Java and Python APIs to merge, update and delete datasets

Walkthrough of today's simulation

Ingesting Data

ingest.py

1. Download trading data from Binance
2. Unzip the file
3. Insert the file into Delta lake with PySpark
4. Remove the downloaded file
5. Go back to #1 with different date

This is purely for real time data ingest demo, there are much better ways to do this!!!

Querying moving data

Find the average value of price on the dataset

```
import pandas as pd

pandaDf = deltaDf.limit(100).toPandas()
pandaDf.head()
```

	tradeId	price	qty	quoteQty	time	isBuyerMaker	isBestMatch	tradingPair	year	isoDate
0	706248993	58976.08	0.001694	99.905480	1615766400007	False	True	BTCUSDT	2021	2021-03-15
1	706248994	58976.08	0.041042	2420.496275	1615766400007	False	True	BTCUSDT	2021	2021-03-15
2	706248995	58976.07	0.001694	99.905463	1615766400049	False	True	BTCUSDT	2021	2021-03-15
3	706248996	58971.26	0.026764	1578.306803	1615766400064	True	True	BTCUSDT	2021	2021-03-15
4	706248997	58976.07	0.033417	1970.803331	1615766400075	False	True	BTCUSDT	2021	2021-03-15

```
from pyspark.sql.functions import avg

deltaDf.select(avg(deltaDf.price)).show()
```

```
+------------------+
|        avg(price)|
+------------------+
|53774.985631620526|
+------------------+
```

Straight to the demo!

What we've learned today

- How to query moving data deterministically
- How to restore the table back in time
- How Delta Lake works behind the scene
- Next steps (Can be a bit tedious, consider using Databricks)
  - Expose Delta Table with Hive for JDBC connection
  - Table constraints, setup audit logs
  - Streaming use cases

Q & A

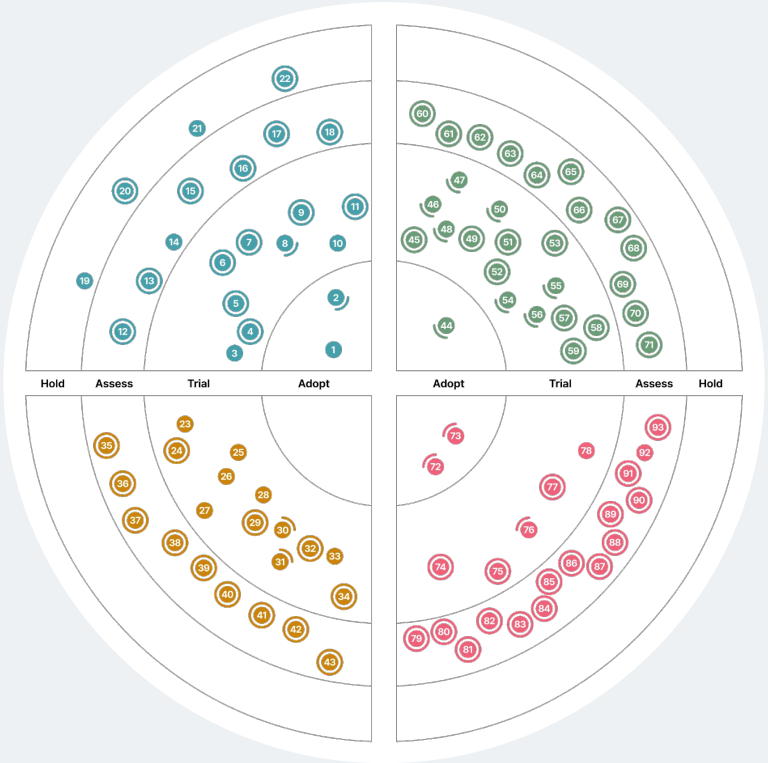
Feedbacks welcome ->

Demo repo: github.com/pee-tw/delta-lake

/thoughtworks

Technology Radar volume 27 is soon to be released

thoughtworks.com/radar



Special preview webinar

October 19th | 12.30 pm

We are hiring

Would you like to be part of a team of passionate technologists? Use technology to drive change and impact the world we live in? Be part of a team that prides itself in engineering excellence and solving complex problems?

If this sounds like you, join our fast growing team at Thoughtworks Thailand!

thoughtworks.com/en-th/careers

Delivering Extraordinary Impact together

/thoughtworks

Sign Up. See More.

Access Thoughtworks careers

If you're curious about what it's like to work at Thoughtworks, sign up for Access Newsletter.

Once a month, we'll send an email with relevant job opportunities, invites to career and tech events near you, and your first look at fresh content like Tech Radar and books from some of our thought leaders.

thoughtworks.com/access

Access Thoughtworks Careers

/thoughtworks