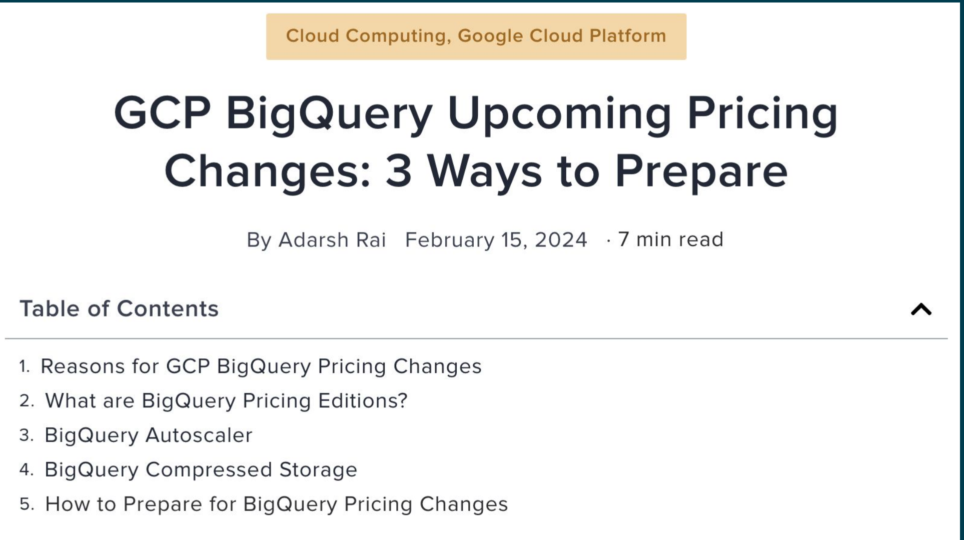
Using DBT with Spark (Experiment)

The grand unification of popular frameworks

/thoughtworks



Cloud Computing, Google Cloud Platform

GCP BigQuery Upcoming Pricing Changes: 3 Ways to Prepare

By Adarsh Rai February 15, 2024 · 7 min read

Table of Contents

> "Change is the only constant in life. Ones ability to adapt to those changes will determine your success in life."

Benjamin Franklin

Everything has costs

Time saved is money

There's **nothing wrong** using a Cloud Native Solution like BigQuery. **Cloud Native solutions abstracts** a lot of potential headaches and give us agility and **focuses on what matters**.

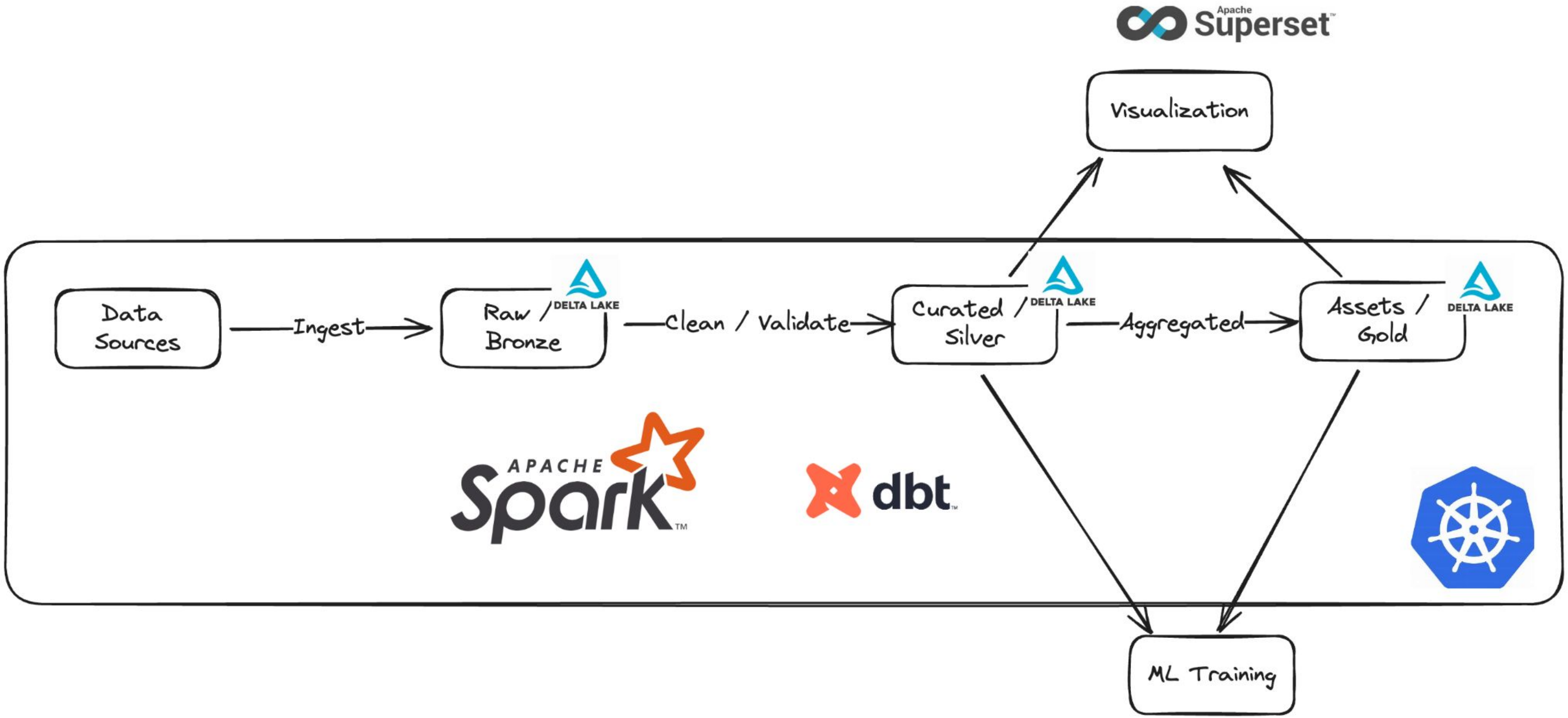
Building the **right things fast** by sacrificing on cost, is far better than **building wrong things right slowly**

Keeping things platform agnostic

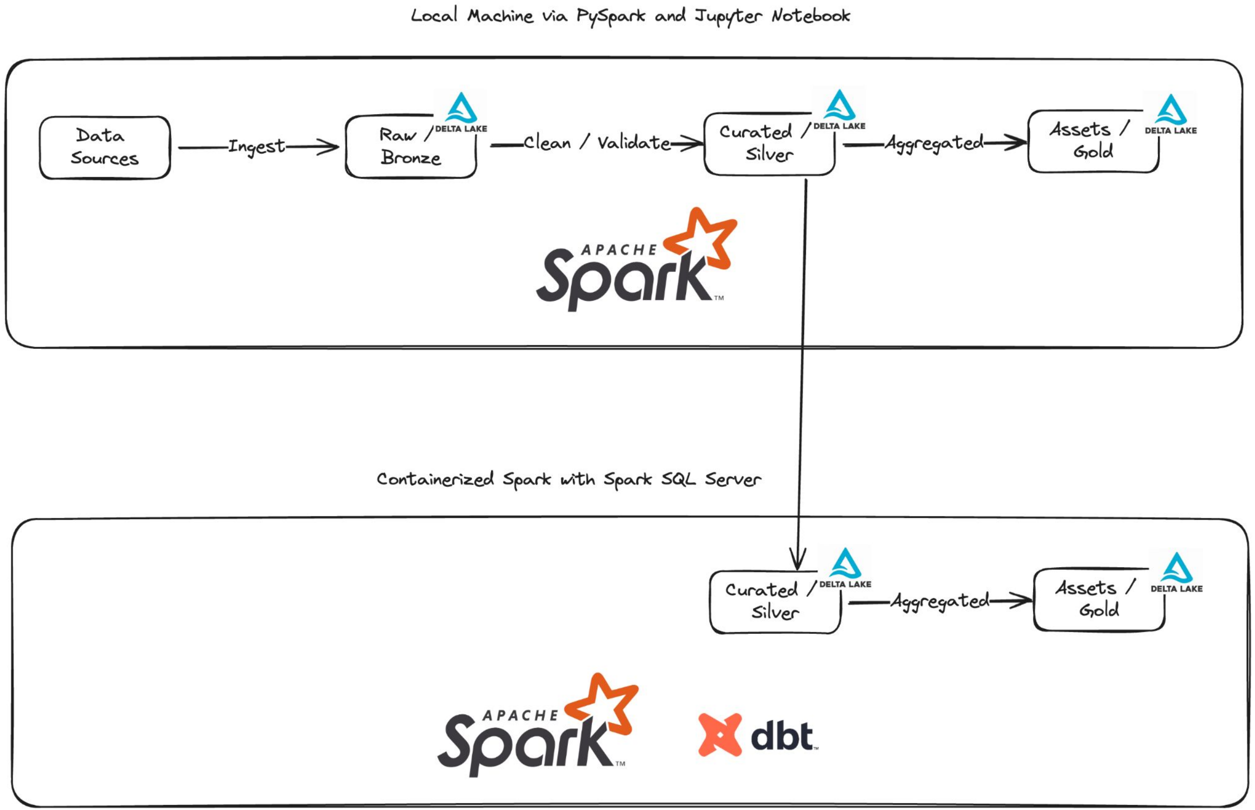
One way we can have complete control over data pipelines

/thoughtworks

One way to be completely platform agnostic



Today's sharing is only about DBT with Spark



A short recap of what each tech are for

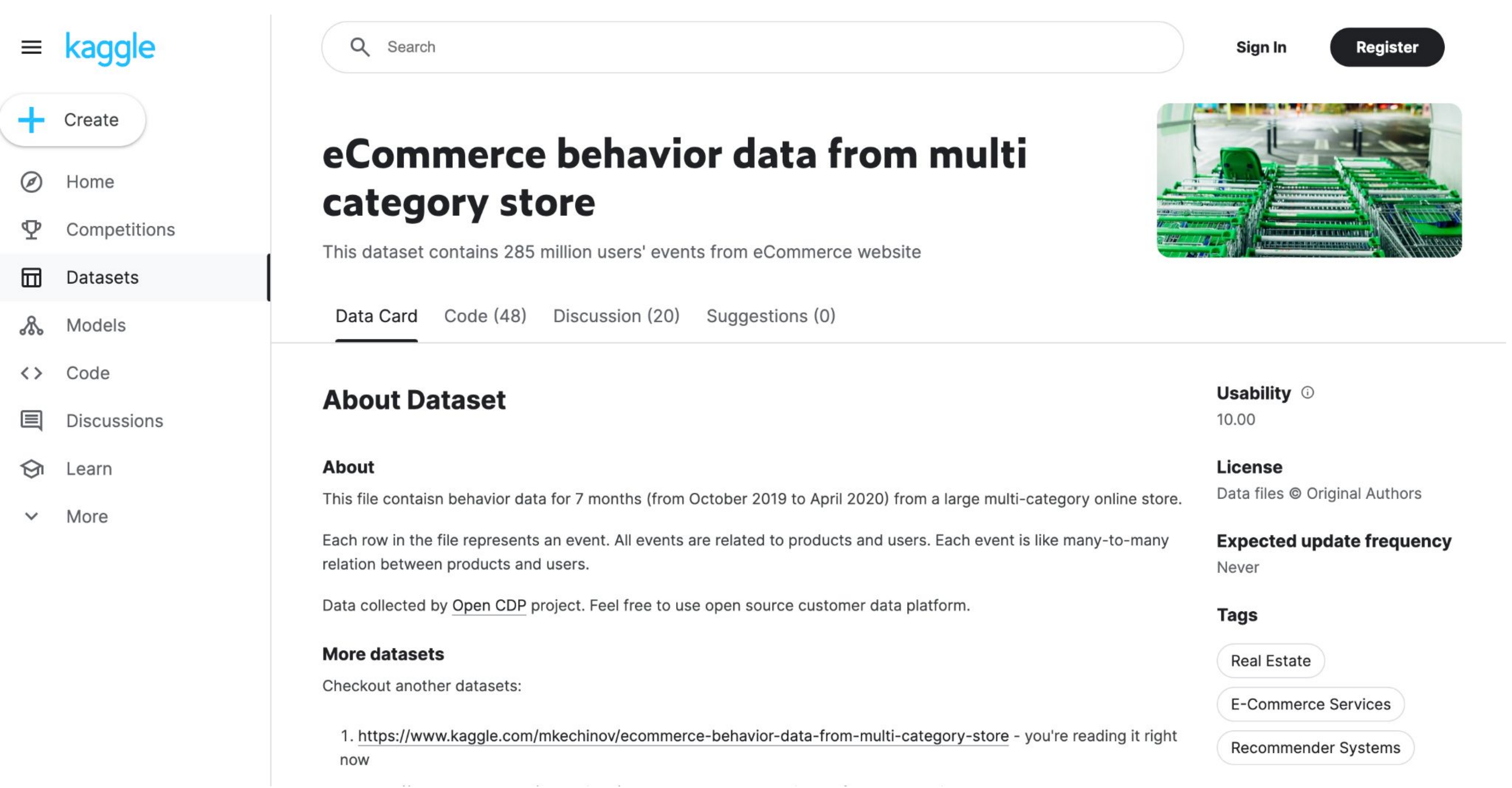
APACHE Spark™ — Distributed Data Processing Engine

DELTA LAKE — Columnar Data Storage Format

dbt™ — SQL templating engine

Scenario: Adhoc Analysis on eCommerce data

/thoughtworks



eCommerce behavior data from multi category store

This dataset contains 285 million users' events from eCommerce website

About Dataset

About

This file contaisn behavior data for 7 months (from October 2019 to April 2020) from a large multi-category online store.

Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users.

Data collected by Open CDP project. Feel free to use open source customer data platform.

More datasets

Checkout another datasets:

1. https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store - you're reading it right now

Usability

10.00

License

Data files © Original Authors

Expected update frequency

Never

Tags

Real Estate

E-Commerce Services

Recommender Systems

Naive Business Questions

As an experiment to demonstrate the capability of PySpark for Adhoc analysis

We will be jumping from PySpark to dbt to demonstrate compatibility

- Do people spend more on Dyson products?
- Which brand has the highest revenue?
- Which product has the most view?

Live demo:

https://github.com/pee-tw/dbt-spark

/thoughtworks

Recap of what we've learnt

- There's nothing wrong with Cloud native solutions, it's a tradeoff between freedom vs agility
- It's possible to **run** data pipeline workload **anywhere**
- We can **keep using existing dbt** transformation logic if we want

Q & A

Things not yet covered:

- Authentication / Authorization
- Iceberg / Hudi / Delta
- S3 storage
- Spark Connect (Remote Spark)

Feedbacks, please