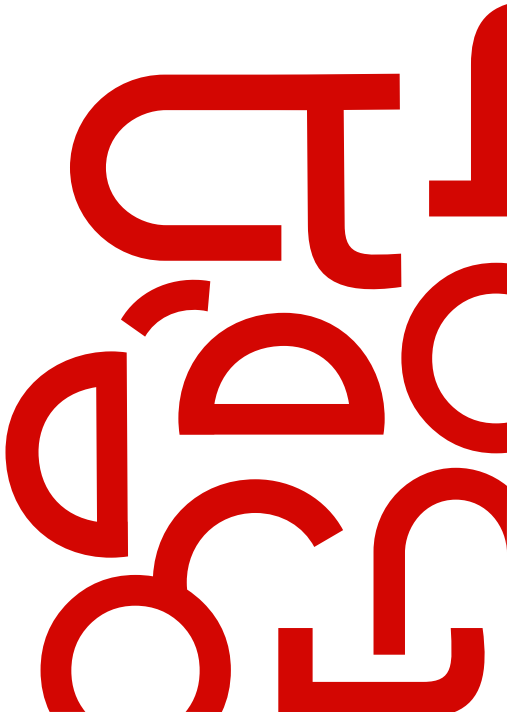
Building Tecton's Feature Engineering Platform on DuckDB

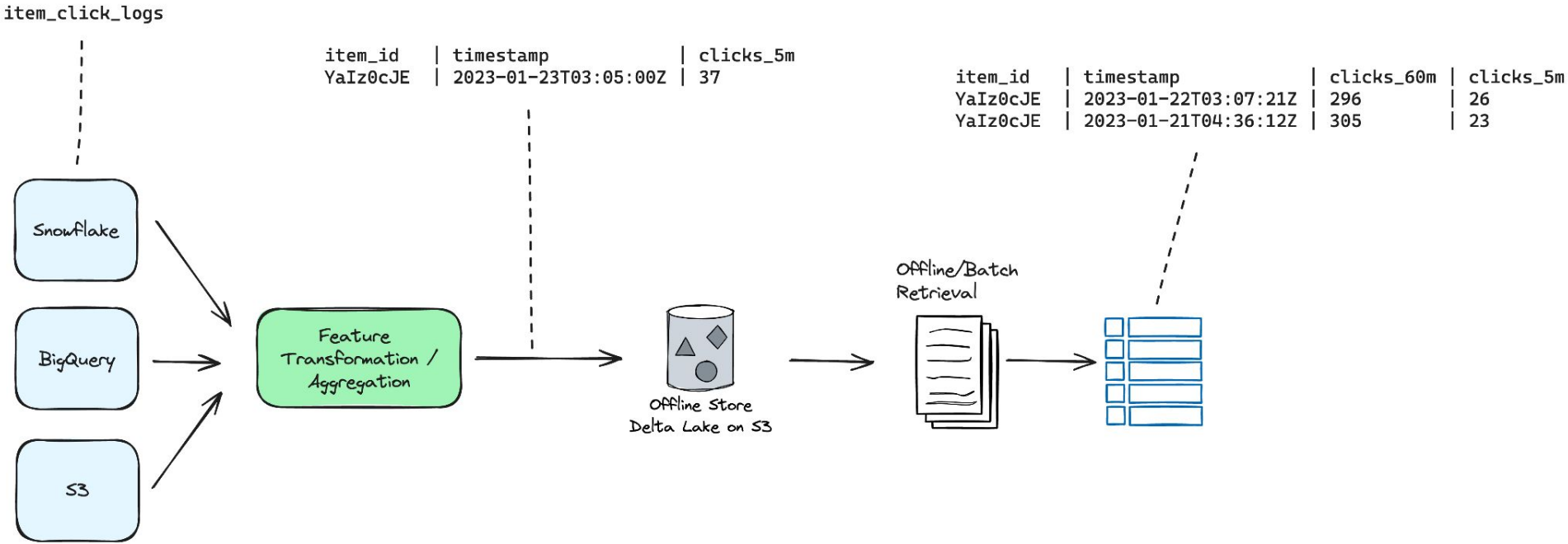
Mike Eastham

tecton



Tecton

- Series C, ~5 years old, ~50 engineers
- We make a Feature Engineering platform which allows customers to define Features for ML models as transformations on top of various data sources. We orchestrate pipelines to compute, store, and serve these values to models.


item_click_logs
item_id | timestamp | clicks_5m
YaIz0cJE | 2023-01-23T03:05:00Z | 37
item_id | timestamp | clicks_60m | clicks_5m
YaIz0cJE | 2023-01-22T03:07:21Z | 296 | 26
YaIz0cJE | 2023-01-21T04:36:12Z | 305 | 23
Snowflake
BigQuery
S3
Feature Transformation / Aggregation
Offline Store
Delta Lake on S3
Offline/Batch Retrieval

Problems

Spark is Confusing

Customers respond with confusion when exposed to Spark configuration or debugging

Deployment Complexity

We require a managed Spark vendor (Databricks, EMR, Dataproc) for deployment

Heavyweight

Spark clusters take 10+ minutes to start up, have thousands of configuration parameters. Overkill for many of our customers with datasets in the 10s of GiBs range or less

Requirements

Familiar Devex

- Pandas
- SQL

Local First

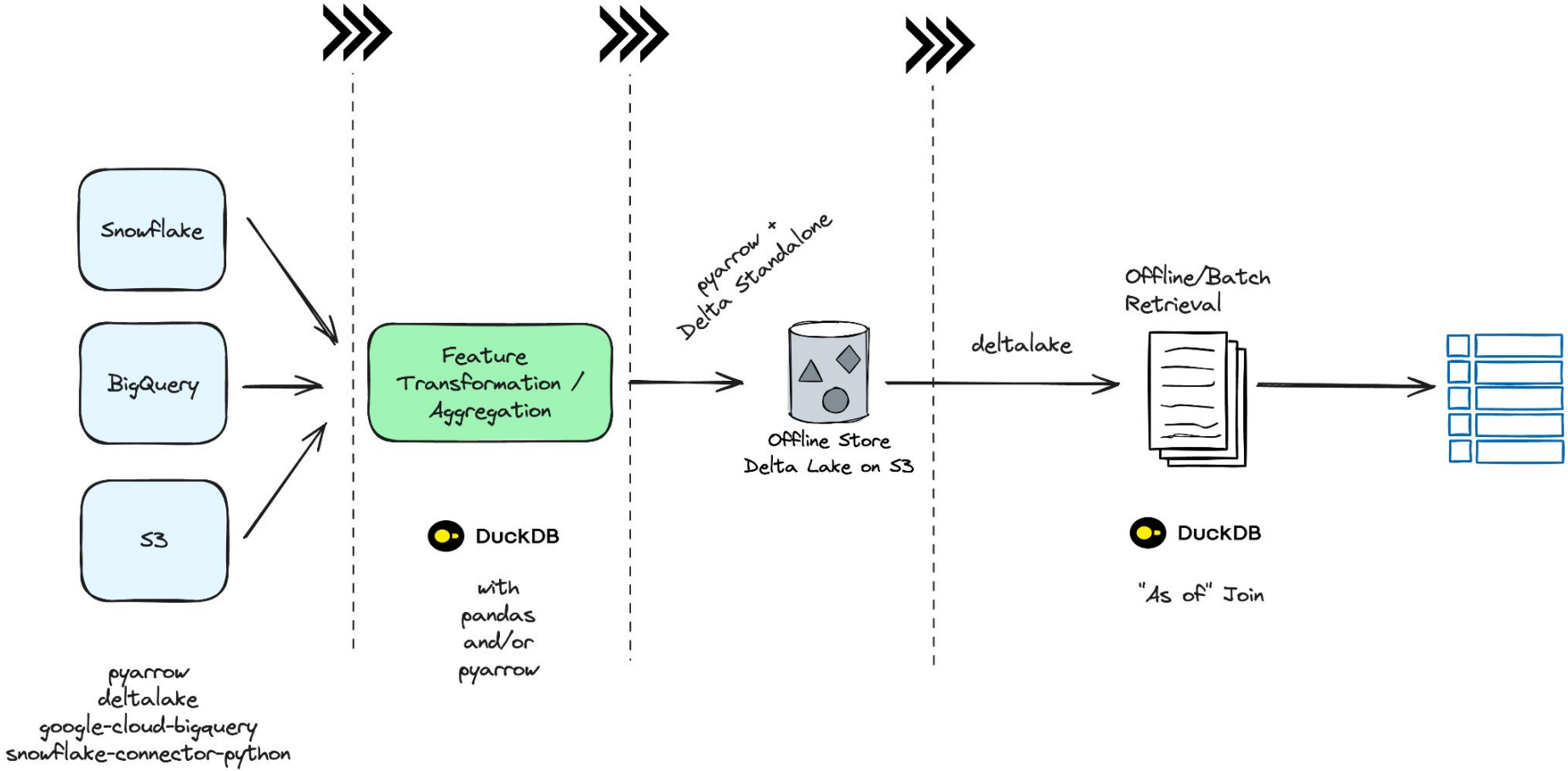
- Fast to start
- Standalone
- Performant

Easy to Deploy

No *mandatory* third party vendors like we have with Spark

Integrations

- Snowflake
- BigQuery
- RedShift
- Postgres
- ...


Snowflake
BigQuery
S3
pyarrow
deltalake
google-cloud-bigquery
snowflake-connector-python
Feature Transformation / Aggregation
DuckDB
with pandas and/or pyarrow
pyarrow + Delta Standalone
Offline Store
Delta Lake on S3
deltalake
Offline/Batch Retrieval
DuckDB
"As of" Join

Why DuckDB

Arrow

Seamless integration allows us to flexibly integrate new data sources, transformation modes, etc.

Performance

Fast & Lightweight

Can process non-trivial datasets with modest amounts of RAM due to streaming, out-of-core

Simplicity

Eliminates the complexities of distributed query engines, Spark/Hadoop configuration, etc., which are unnecessary for the datasets we're typically dealing with.

Results (so far)

- Currently in Private Preview
- Many jobs now complete faster than a Spark cluster can start up
- Scales up to datasets of 10s of GiBs (both input and output) without problems
- Demos and Deployments have been substantially faster

Challenges

- Extension development / distribution
- Managing AWS Credentials
- Delta Lake integration
- pandas / arrow version wrangling

That's It!

- Find me later if you have questions!
- We're hiring in SF, NYC, and remote!