The Road to Data Correctness

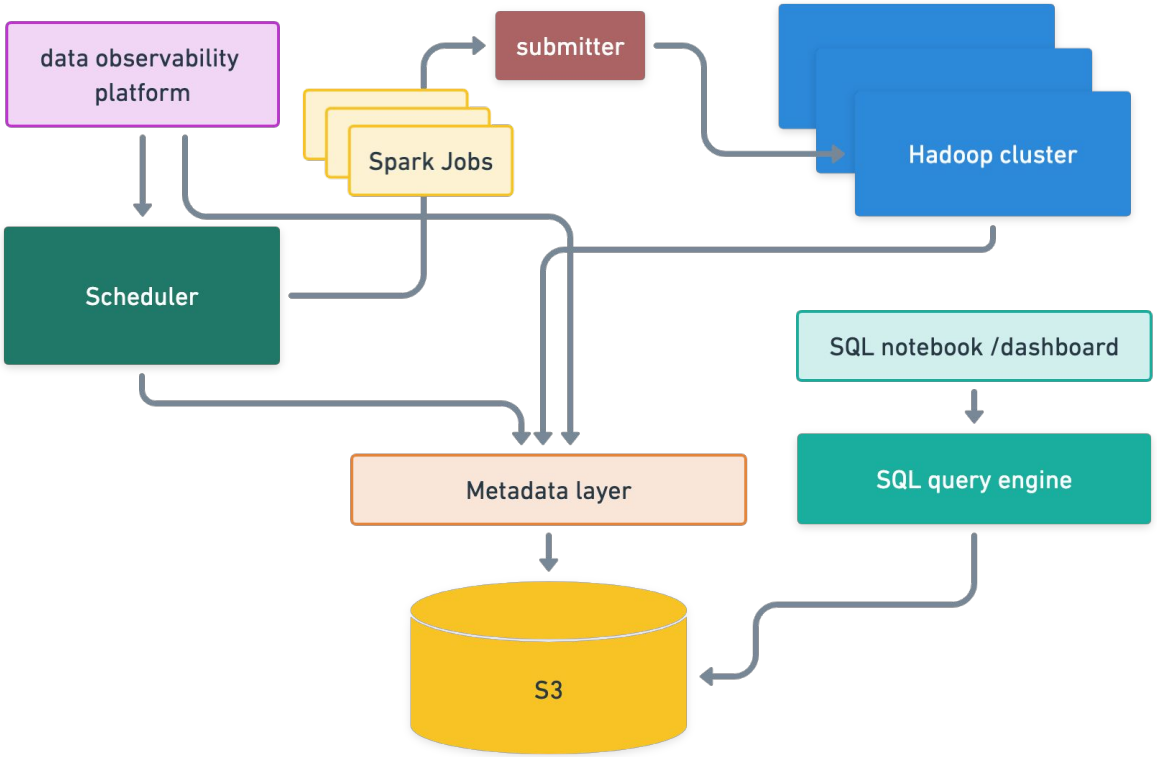
Emma Tang @emmaytang

Why exceptional correctness?

How Data was used

- **Treasury** - How much money do we have in our banks?
- **Reconciliation** - Which user transactions are linked with which transactions at the bank?
- **Billing** - How much do we bill each user this month?
- and more…

Data Platform (Simplified)



How to be correct

Data quality checks

- Full suite of post-job checks that users can choose for their jobs with decorators, executes automatically
- Examples
  - `@is_monotonically_increasing`
  - `@unique_keys(column_name)`
  - `@primary_key_in(table_name, column_name)`

Fallback behavior

Job failure decorators

- Use yesterday’s output in place of today’s output
- Rerun with yesterday’s input data
- Block pipeline and page owning team

S3 eventual consistency

- Custom scheduler based and command line tooling for checking consistency before allow downstream jobs to continue.
- Write special metadata to signal consistency.

MOAR METADATA!

Typesafe Spark

- Custom encoders for all data types, e.g. special currency float types
- Much AST fun!

Data observability platform

- Shape of the data (columns, types)
- Size of the data (num rows, bytes)
- Historical runs
- Estimate of completion of entire pipeline / critical downstream
- Alerting

UI based tool to specify checks and fallback behavior, cost attribution

Recompute the universe

Trade-offs

Latency

Cost

Silver lining

- Moving to more efficient data storage and table formats (Apache Iceberg) for reduced data lift and compute
- Allows for cool things like data locality

Conclusion

Build a robust system designed for exceptional data correctness

Allow your users to sleep well at night & profit!

Thank you

Emma Tang  @emmaytang