Getting Data Right at Uber Scale

Uber has revolutionized the transportation industry by connecting millions of bikes, riders, drivers, and restaurants. Behind all this transformation is a complex data stack. In this blog post, adapted from [this presentation](https://m.facebook.com/Engineering/videos/data-quality-at-uber-how-to-get-data-right-at-uber-scale/1960560874116690/) at Meta’s Data Observability Learning Summit, we look at some of the challenges Uber has faced in maintaining data quality at Uber scale, and the solutions they have implemented to overcome them.

# History of Uber’s Data Infrastructure

Uber's data infrastructure has evolved significantly over the years. In the early days, the company had a monolithic data pipeline that was responsible for collecting, storing, and processing all data. This data pipeline consisted of:

* A sharded MySQL database as the “online database”
* A CDC pipeline powered by Hive that took data from the online database and pushed it to the data lake (24-hour process). Once in the data lake, these were considered “raw tables.”
* A data warehouse team that would take the raw tables and build dimension tables and fact tables off of them
* Utilities and tools on top of the data warehouse for the use of data scientists.

At the height of this platform, there were 300000+ unowned datasets.

# The Need to Build a Data Observability Platform

With this huge number of pipelines and datasets, several issues arose. In particular,

* **Data duplication:** because no one knew what data existed, everyone just went ahead and created their own version of the data.
* **Lack of understanding of data lineage and freshness**: no one knew when data was landing in tables.
* **Data quality:** no one knew whether the data they were looking at (e.g. in dashboards) was of high quality.

In the years of Uber’s hypergrowth (2015-2016), effort was concentrated mostly on scaling the data infra itself, rather than investing in the data product. Eventually, though, Uber realized that it needed to take a more fundamental approach.

# Uber’s Principles for Data

Uber applied the following principles to arrive at a better data culture:

* **Data as code**: Data is treated as code and is managed in a similar way to software. The artifacts are reviewed, and any schema change that is done in production goes through the review process. Producers of the data as well as consumers of the data are tagged during the review process. This approach makes it easier to track changes, version data, and collaborate with others.
* **Data is owned**: This principle mainly focuses on data ownership. The data must be owned by the business or functional teams that use it. The teams must clearly define the intent of the data product and artifact, own it, and provide guarantees around the data. This approach ensures that teams are responsible for the quality of the data they use and they are motivated to improve it.
* **Data quality is known**: Data quality is continuously monitored and measured. The SLA targets are used as part of the assertions. All datasets are categorized with tiering levels, which are defined as criteria to set default SLA values. This approach enables teams to identify and fix data quality issues quickly and easily.

With the implementation of these principles, Uber moved from a platform of self-serving tools to a more regulated, owned, and responsible data platform.

# Data Observability at Uber in 2021

Fast forward to the present day: Uber has built out a data observability platform with the following components:

## Tiering

At Uber, not all data is equally important. The company implemented a tiering concept for its data assets (tables, pipelines, ML models, and dashboards), with Tier 1 being an extremely important dataset and Tier 5 being an individually owned dataset, generated in staging environments, that doesn’t have any guarantees. At Uber, for example, Kafka as a service is tier 0.

After all datasets were tiered, the company ended up identifying 2500 Tier 1 and Tier 2 tables (out of 130k+) that were extremely important. This approach enables Uber to focus its efforts on ensuring the quality of the most important data, while still providing visibility into all data.

## Databook, Uber’s data catalog

Uber has an in-house catalog, [Databook](https://www.uber.com/blog/databook/), that serves as a user interface on top of dataset metadata like:

* Quality signals
* Tiers
* Who owns these particular data assets
* Products that are enabled by the data

Databook makes data exploration and discovery much easier for Uber’s engineers, data scientists, and operations teams.

## Lineage

Databook also provides information about lineage or the relationships between different datasets. This helps engineers understand how data flows through the pipeline from source to destination.

## Data Quality System

To ensure data quality, Uber has implemented a Data Quality System.

Once a certain dataset is labeled as Tier 1 or Tier 2, it is automatically onboarded into a certain set of data quality checks and foundational guarantees including that the data is:

* Documented
* Owned
* Connected to PagerDuty on-call

These guarantees essentially mean that the data asset is treated like a service.

Additionally, all Tier 1/Tier 2 data assets are monitored on a set of metrics including:

* **Freshness**: measures how recent the data in a dataset is. It can be determined by comparing the timestamp of the data to the current time, or by comparing it to a known source of truth.
* **Completeness**: measures how much of the expected data is present in a dataset. It can be determined by comparing the number of rows or columns to a known expected value. This metric also ensures that all data that is present in upstream is also present in downstream.
* **xDC consistency**: measures the consistency of data across different data centers. It can be determined by comparing data in different data centers for the same key or by using a hashing function to compare data across data centers.
* **Duplicates**: measures the number of duplicate records in a dataset. It can be determined by comparing primary keys or by using a hashing function to compare records.

Additionally, Uber allows users to set up custom checks on top of these standard metrics. This gives users the ability to define specific checks that are relevant to their use case and to monitor the data in a way that is most meaningful to them.

### Conclusion

In conclusion, Uber's approach to data reliability and quality is built on the principles of data as code, data ownership, and data quality. To put these principles into practice, the company has set up processes and tooling that provide visibility into all data and metadata, with a particular emphasis on key pipelines and datasets.