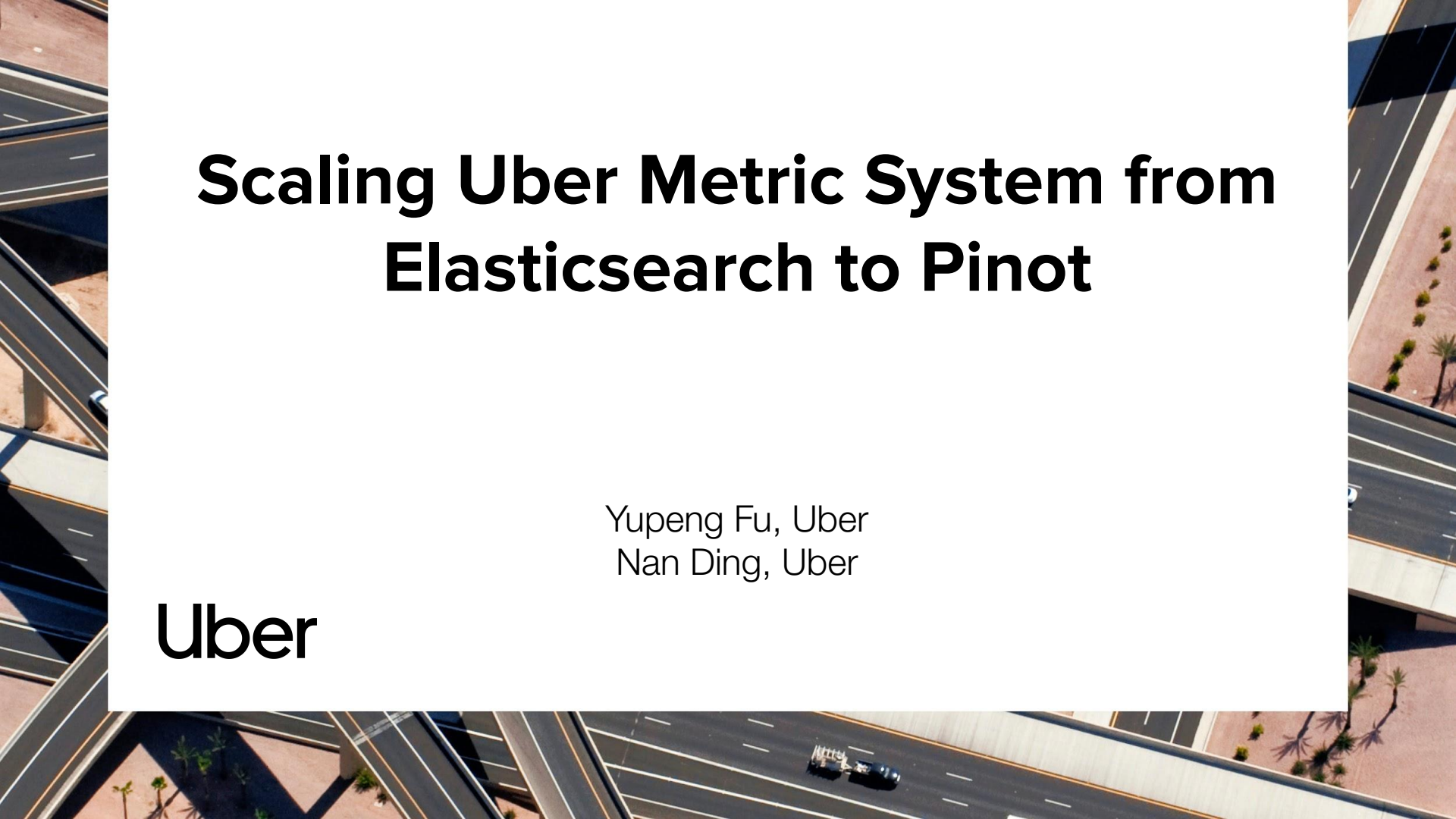


Scaling Uber Metric System from Elasticsearch to Pinot

Yupeng Fu, Uber
Nan Ding, Uber

Uber

Uber

About Us

Yupeng Fu (yupeng9@github)

- Principal Engineer @ Uber.Inc
- Real-time Data Platform
- Committer: Apache Pinot

Nan Ding

- Staff Engineer @ Uber.Inc
- Mobility & Platforms

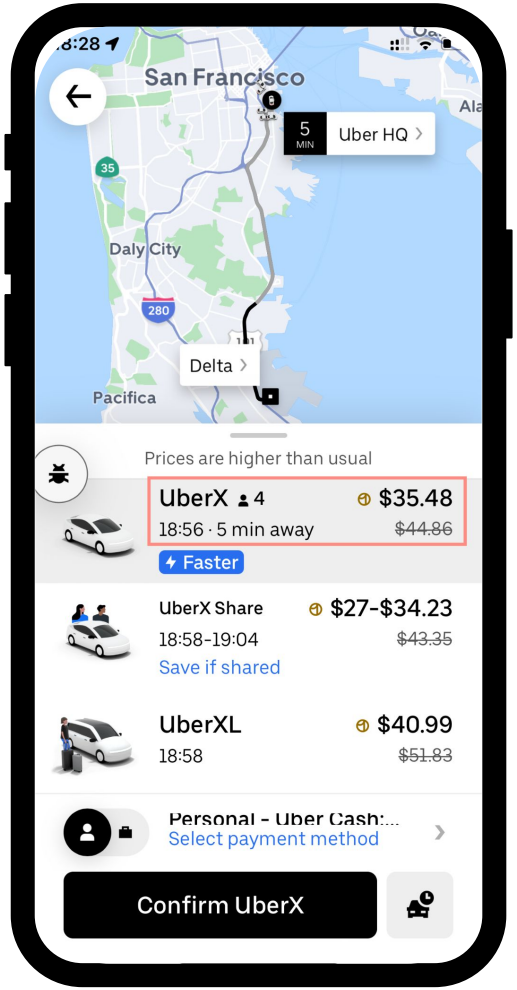
Uber

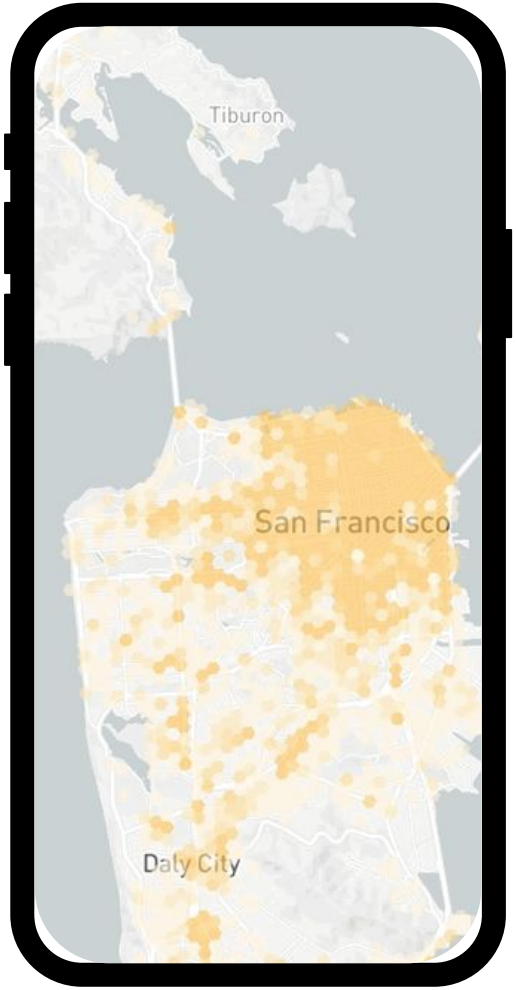
Agenda

- uMetric overview
- Challenges using Elasticsearch
- uMetric over Pinot - Challenge and Architecture
- Pinot at Uber
  - Why Pinot
  - Scale@Uber
  - Features contributed
- Future work

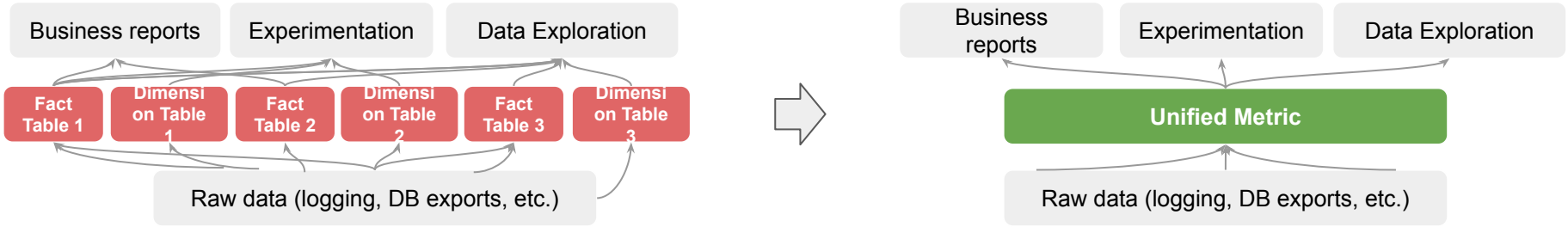
Data driven decision making

Capturing a snapshot of the market, and use the data in both real-time and long-retention history for decision-making.





Metric Unification & Standardization



Industry analogs:

Airbnb: Minerva

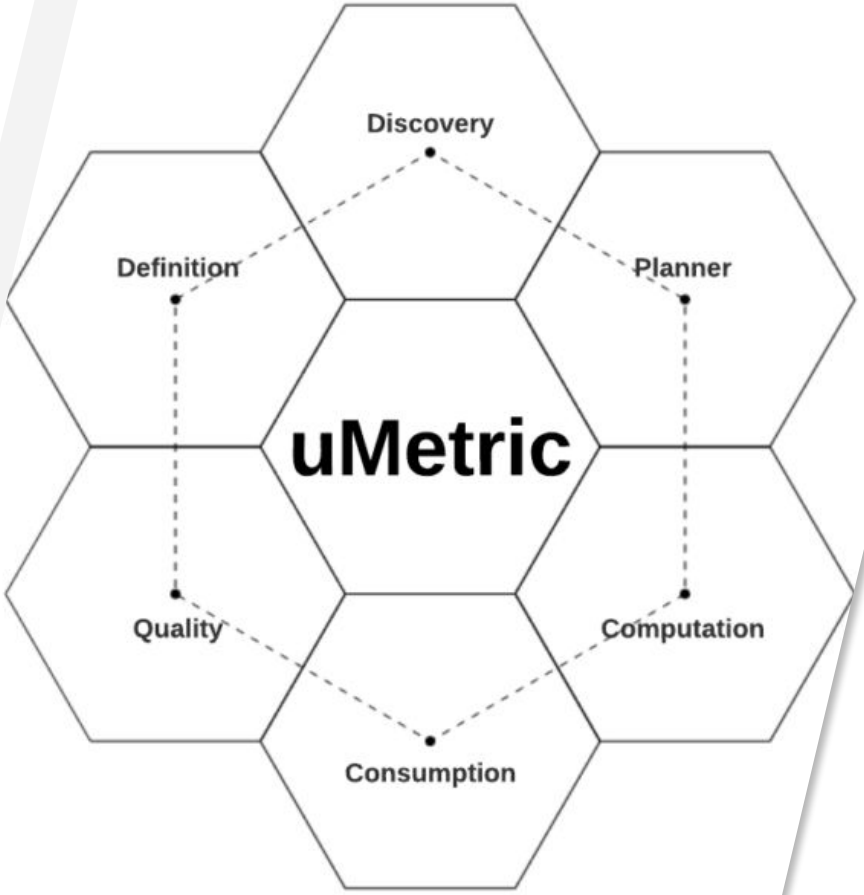
Linkedin: UMP

Transform.co

Liveramp: Metriql

uMetric

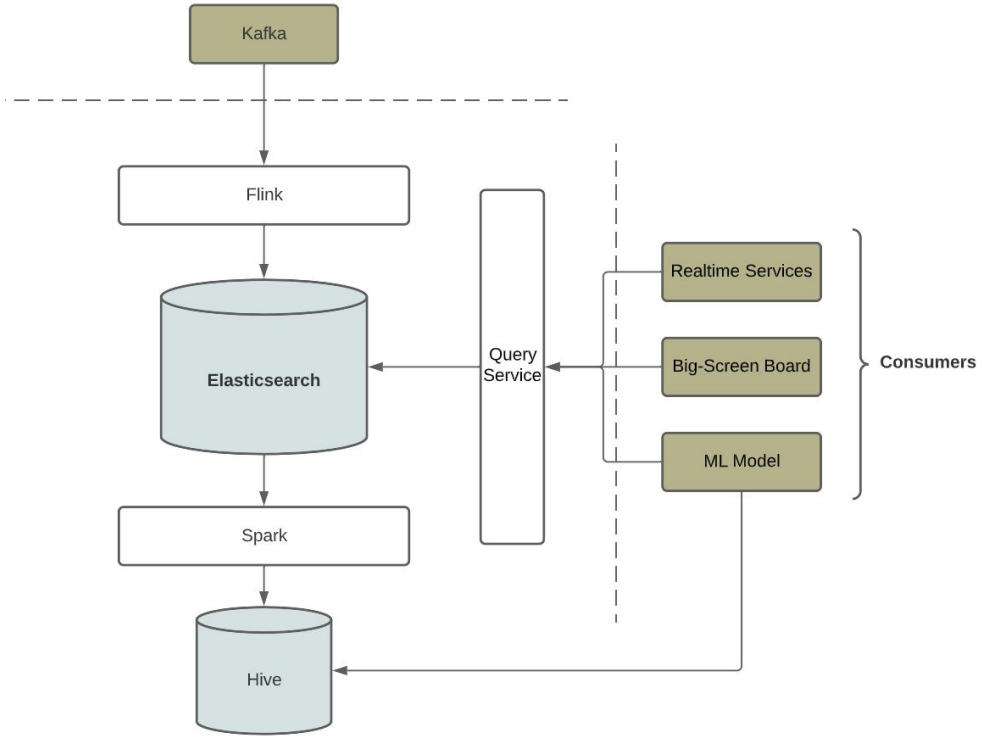
uMetric *is a unified metric platform at Uber that manages the full life cycle of a metric: definition, discovery, computation, verification, and serving*



Elasticsearch-Oriented Realtime Architecture

Since 2014

- Operable by small team
- Support idempotency insert (upsert)
- Support distributed aggregation
- Linearly scalable
- A matured system



Scale

1.5PB

Dataset Size

4.5T

Documents

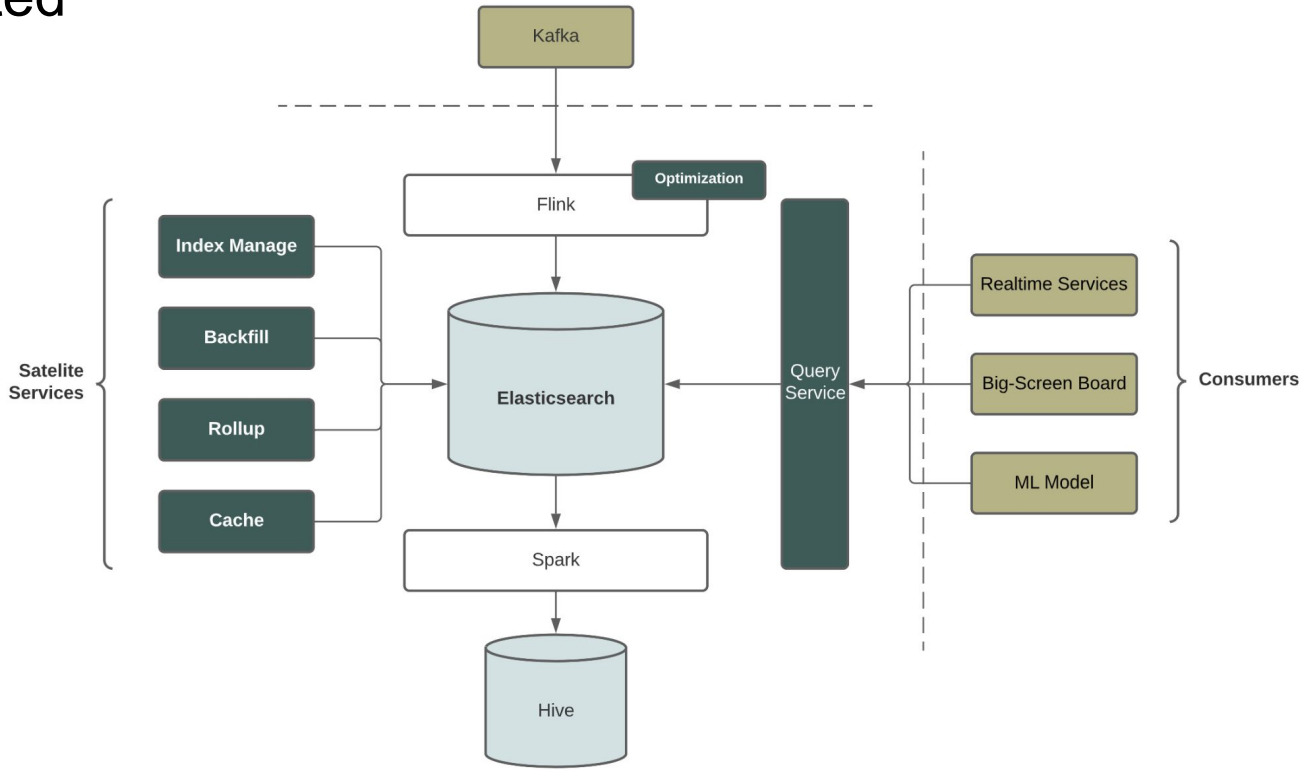
1.3M/s

Write per Sec

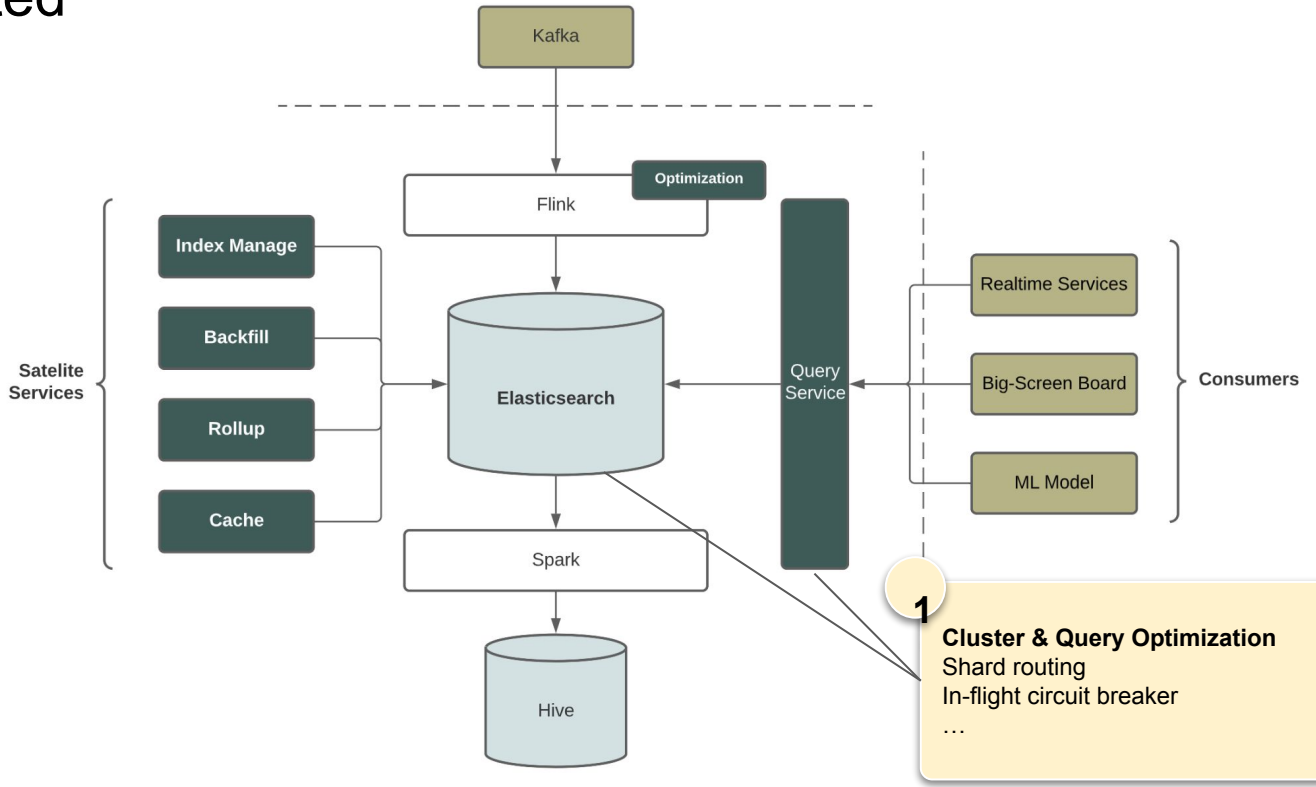
700M/s

Doc Scanned per Sec

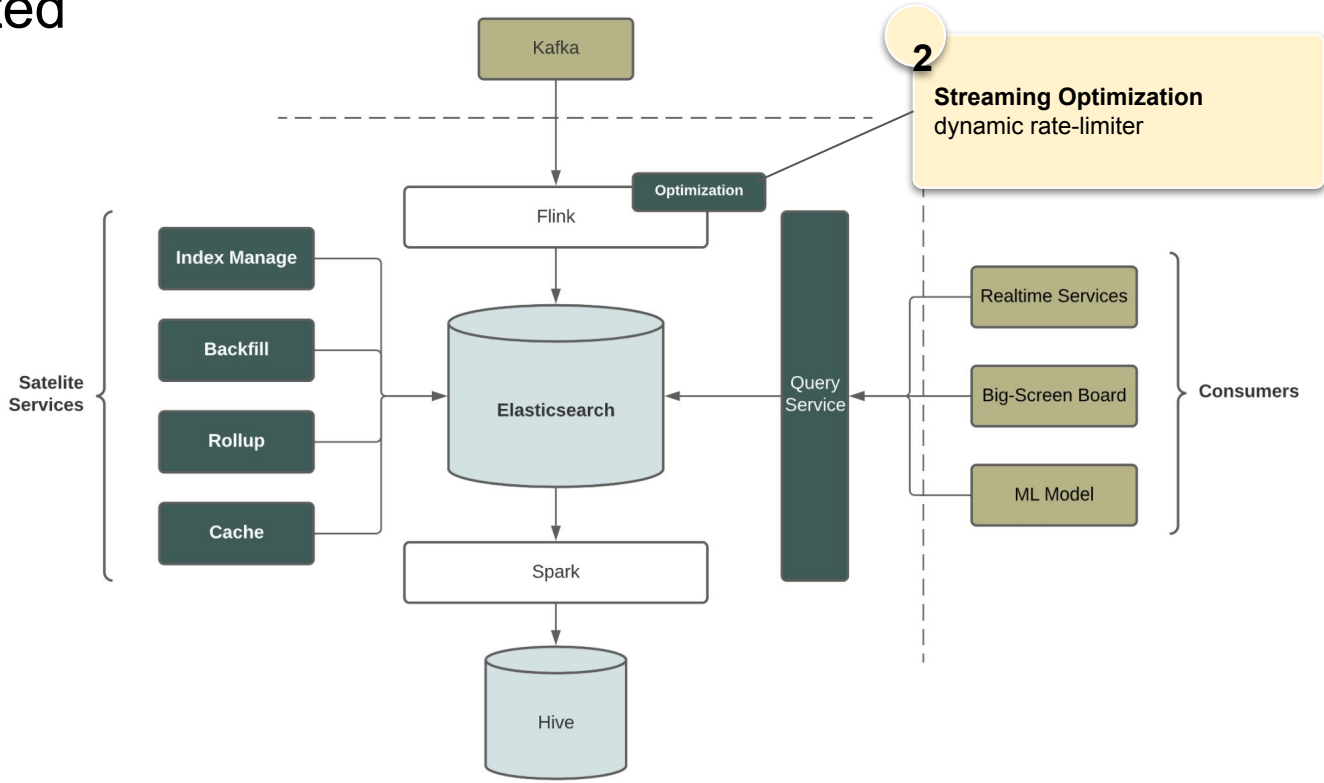
Elasticsearch-Oriented Arch Optimization



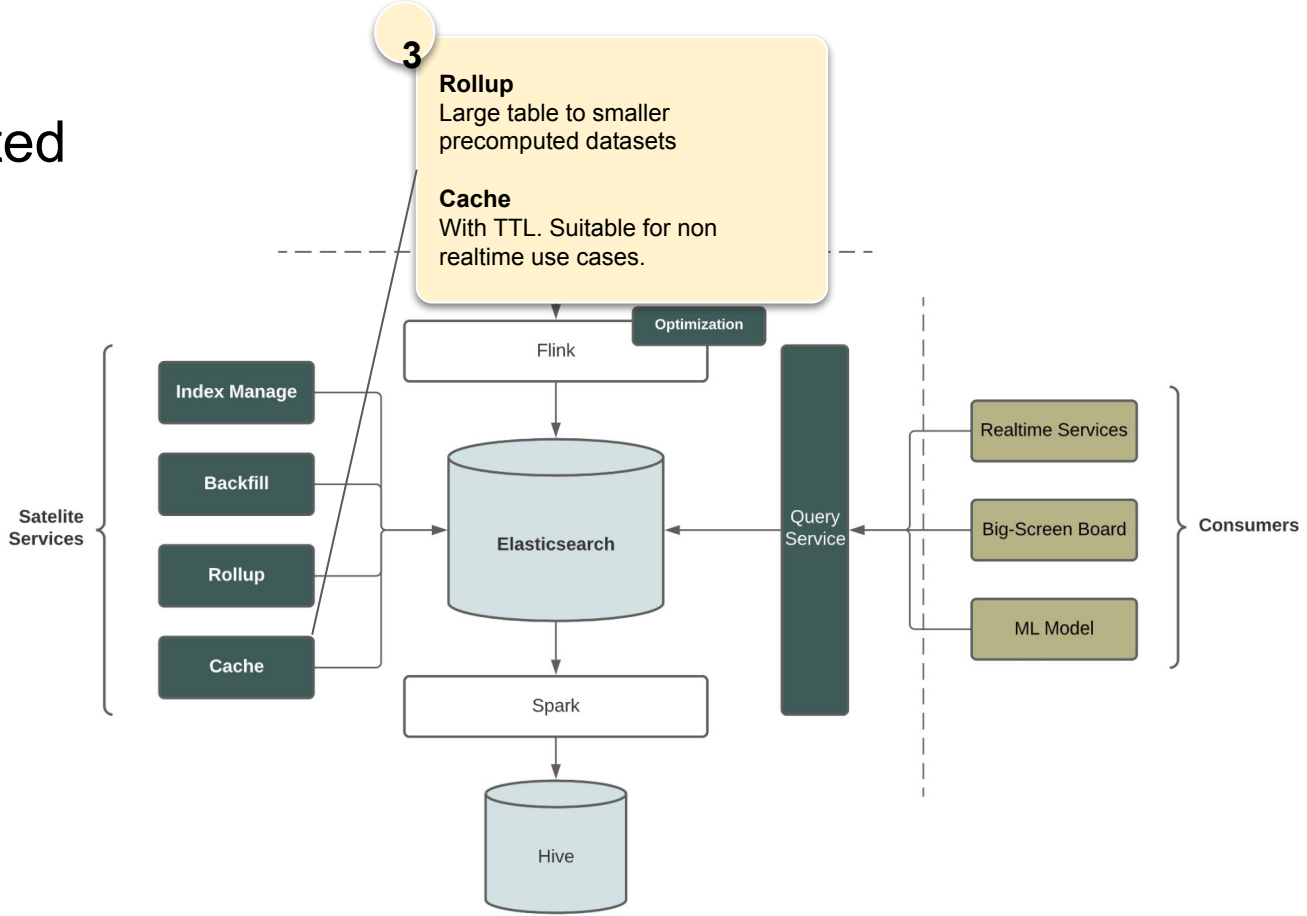
Elasticsearch-Oriented Arch Optimization



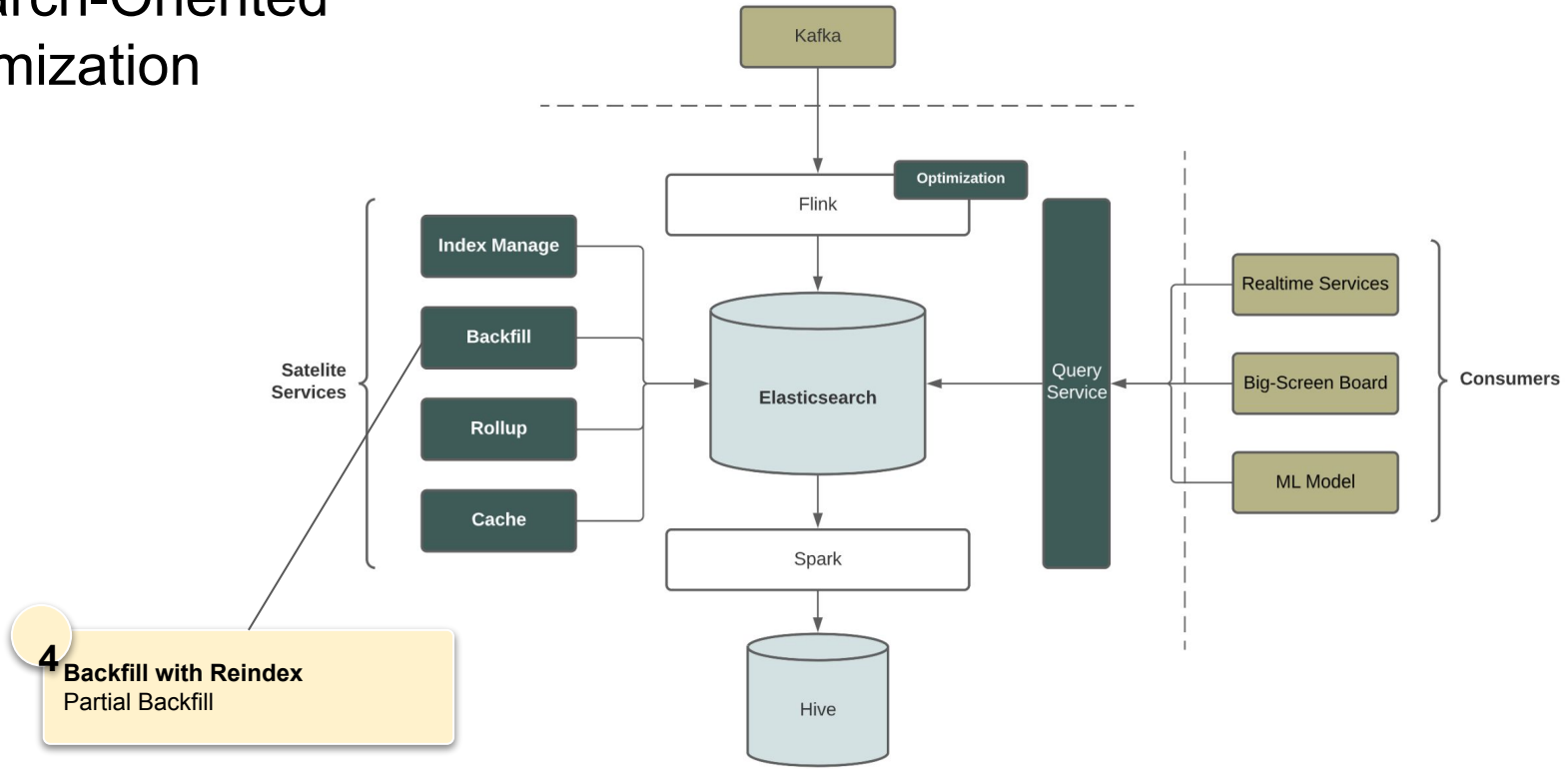
Elasticsearch-Oriented Arch Optimization



Elasticsearch-Oriented Arch Optimization



Elasticsearch-Oriented Arch Optimization



Challenges on Elasticsearch Stack

Reliability

- Cluster constantly in high load
- Frequent SLA breach
- Frequent minor data-loss
- Lack effective mitigation
- Issue cannot be root-caused

Scalability

- Rapidly growing use cases
- Linearly scale cluster / data node has low ROI, and cause derivative issues

Engineering Cost

- Build optimization systems
- High oncall load

Migration toward Pinot

Feature Gaps

- Upsert
- Backfill
- Spark connector
- Nested col support

Safe Migration

- Perf / DQ / Reliability comparison between Pinot and ES
- Drain traffic by table / metric / user / retention / etc.

Performance Tuning

- Multi-cluster by tier
- Dedicated machine spec per cluster
- Fine-tuning per query, with indexes and algorithm optimization

After Migration

Reliability

- Significantly improved overall reliability
- Fine-grained table-level failover

Scalability

- Unblocked new user onboarding
- Scale table up on demand

Engineering Cost

- Deprecated 6 customized system
- Reduced oncall load

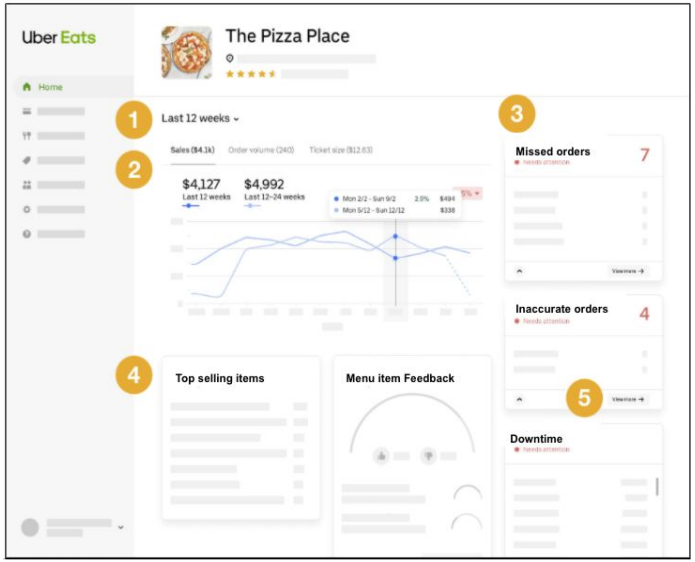
Uber

Pinot @ Uber

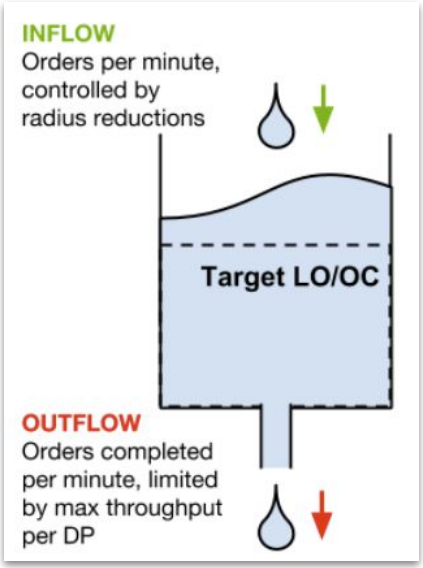
Real-time Analytics at Uber

1. **Real-time** and **actionable** insights
2. **Time-sensitive** decisions
3. User **engagement** growth

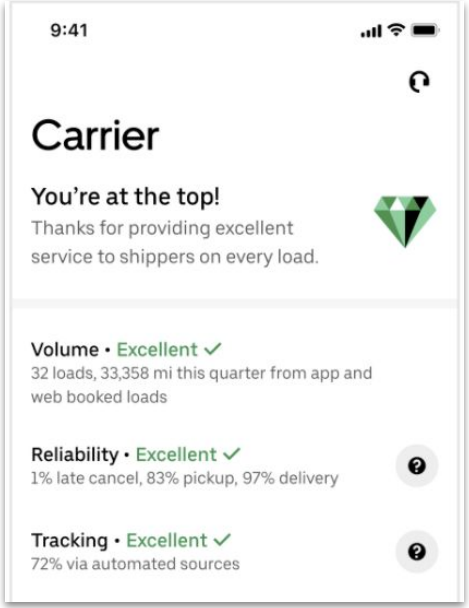
Fast Access to Fresh Data at Scale



Restaurant Performance View



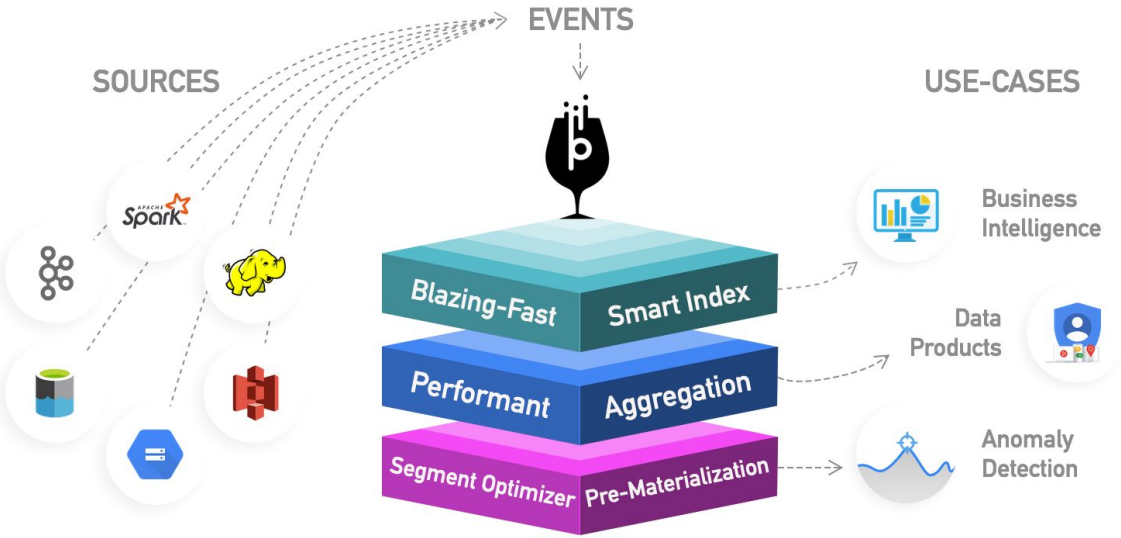
Demand/Supply Management



Freight Carrier Scorecard

Apache Pinot: Fast, Distributed OLAP

- Started by Linkedin for Metrics System
- Highly available
- Horizontally scalable
- Low latency/High throughput
- Immutable data



1M+

Events/sec

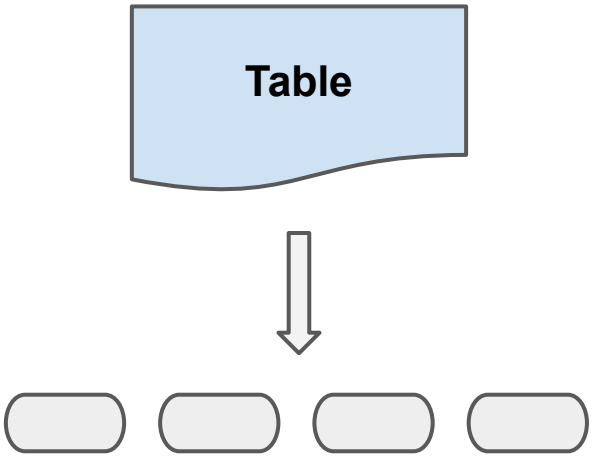
170k+

Peak QPS

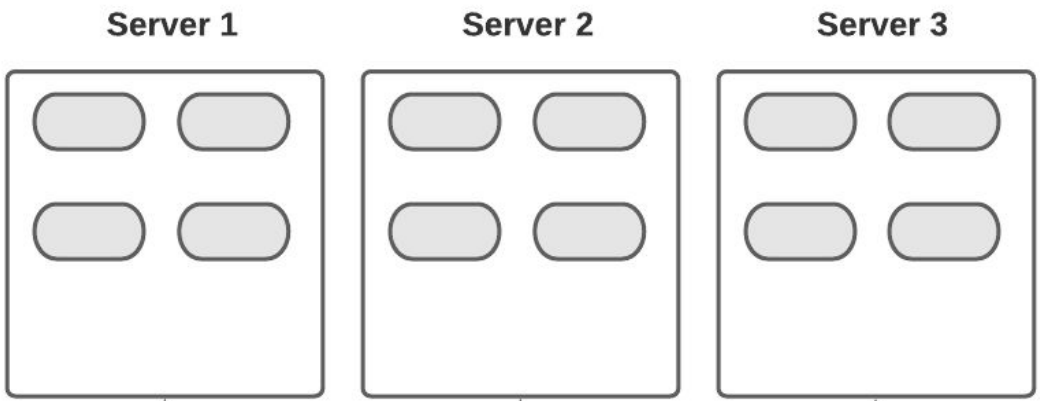
ms

Query Latency

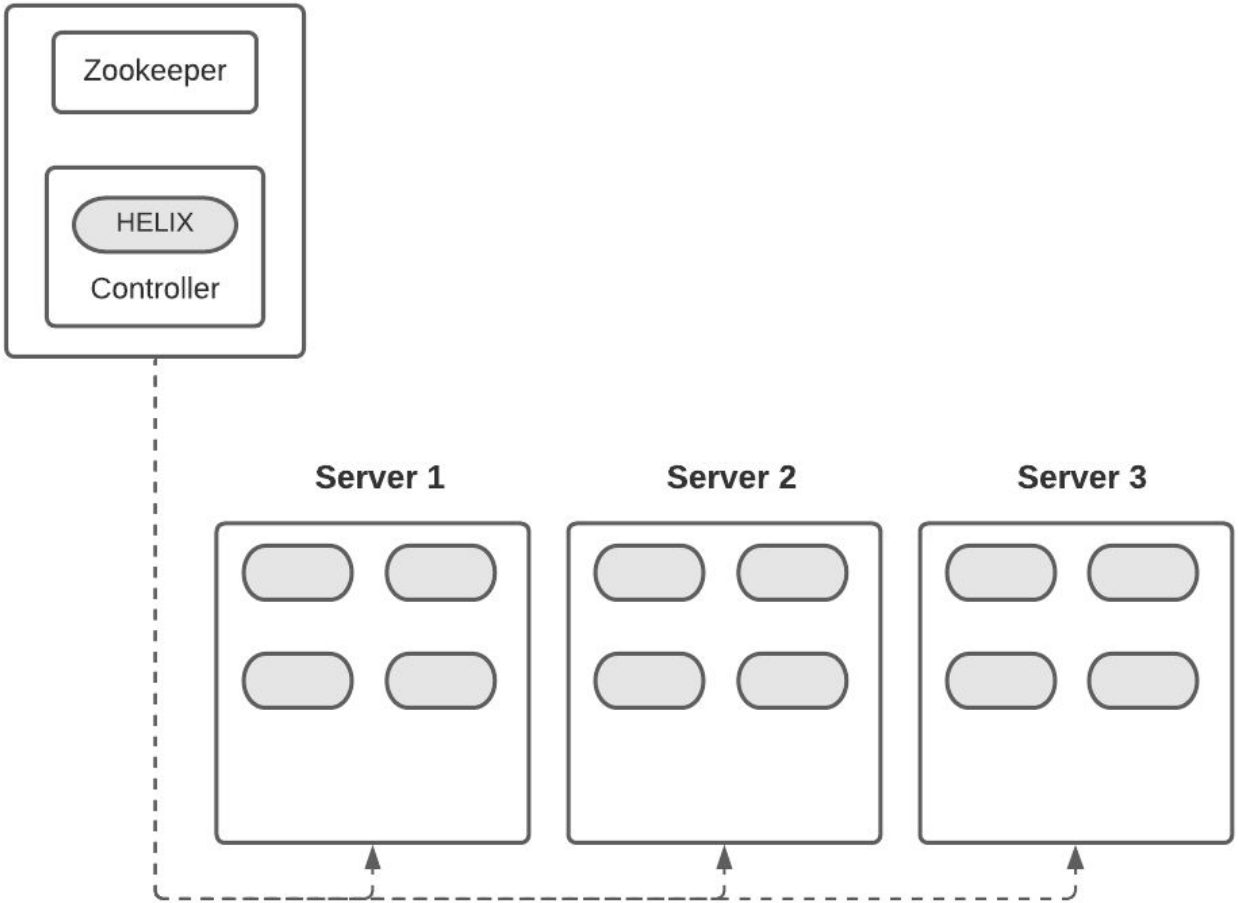
High Level Architecture



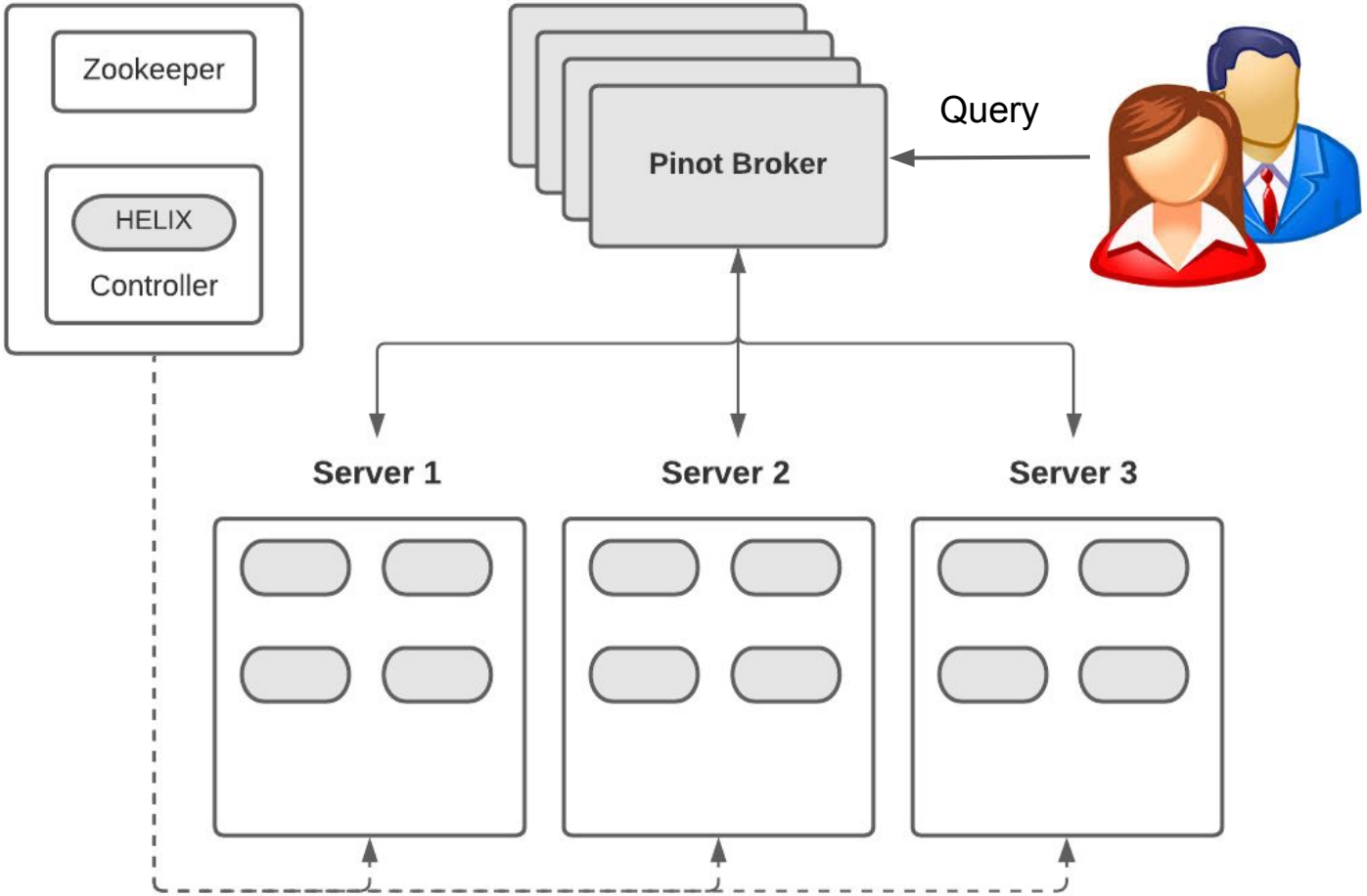
High Level Architecture



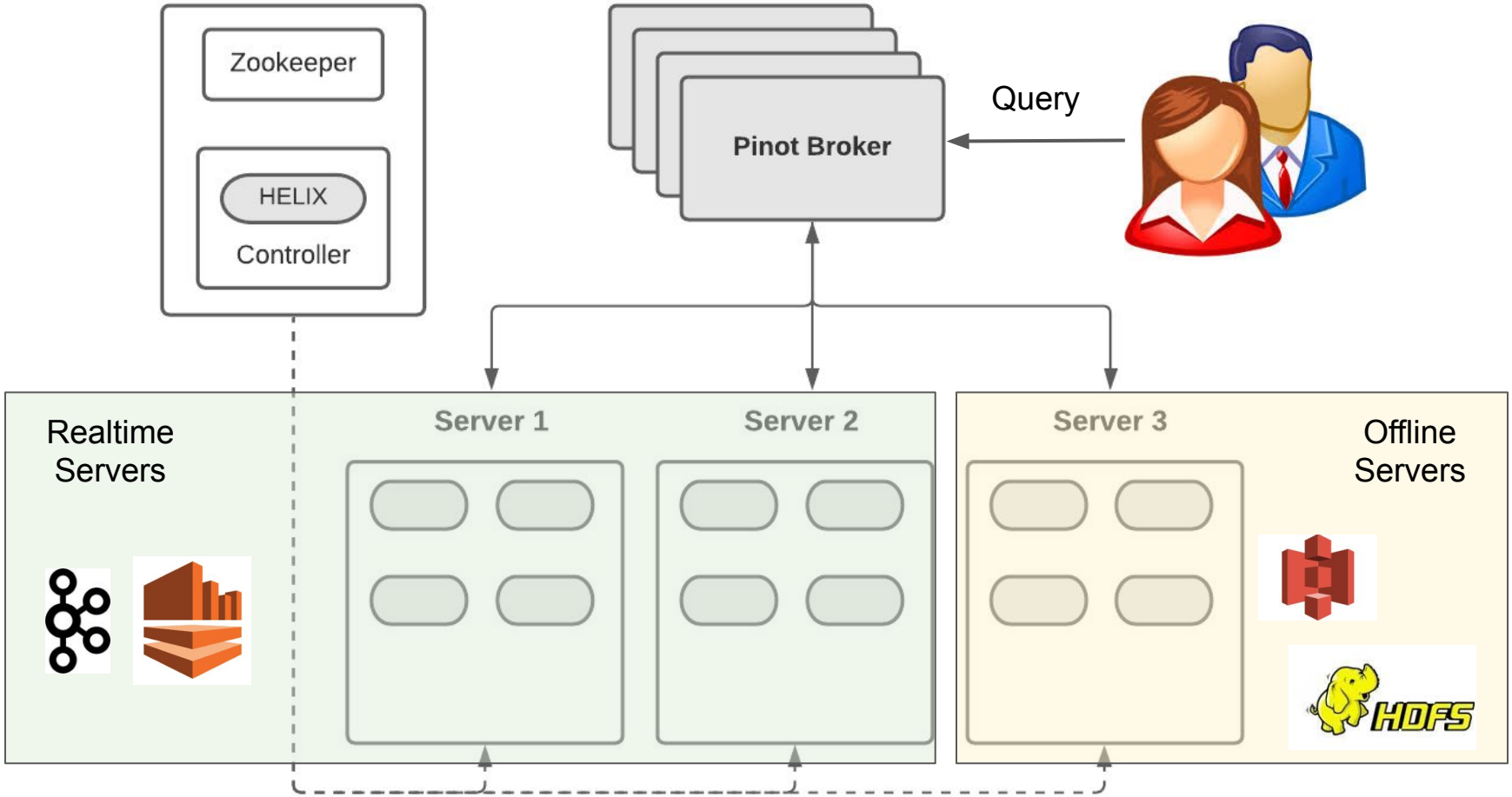
High Level Architecture



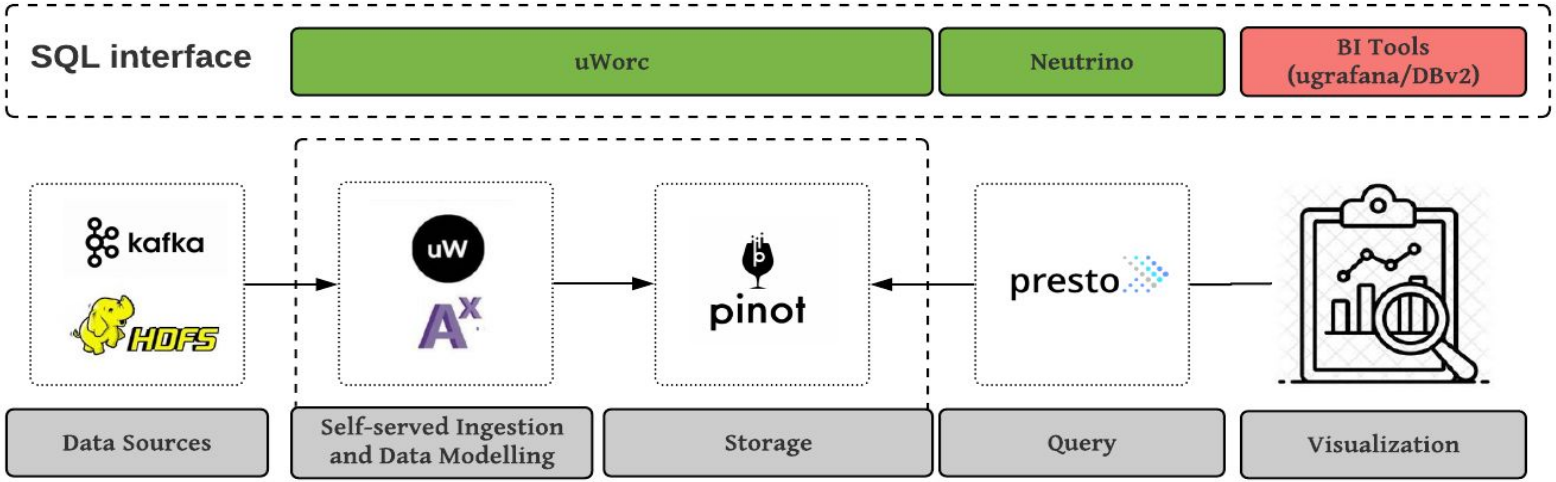
High Level Architecture



High Level Architecture



EVA: Building a World-class RTA platform for Uber



Tier 0 platform, 99.99% uptime

Built on top of **Apache Pinot**

Seconds data freshness

Self service Onboarding (via uWorc)

SQL API (via Presto / Neutrino)

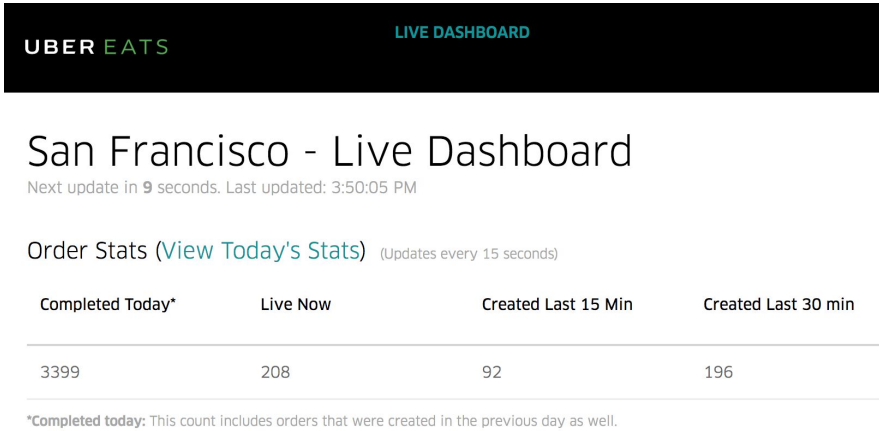
< 100ms @P99 query latency

https://www.uber.com/blog/operating-apache-pinot/

Upserts

- First real-time OLAP to support upsert
- Challenge: data stored as immutable segments
- Data partitioning to share nothing
- Released in Pinot 0.6.0
- Partial upsert

```
SELECT current_status,
       count(*)
FROM   uberEatsOrders
WHERE  regionid = 1366
       AND MinutesSinceEpoch
       BETWEEN 25432140 AND 25433580
GROUP  BY current_status
TOP    10000
```

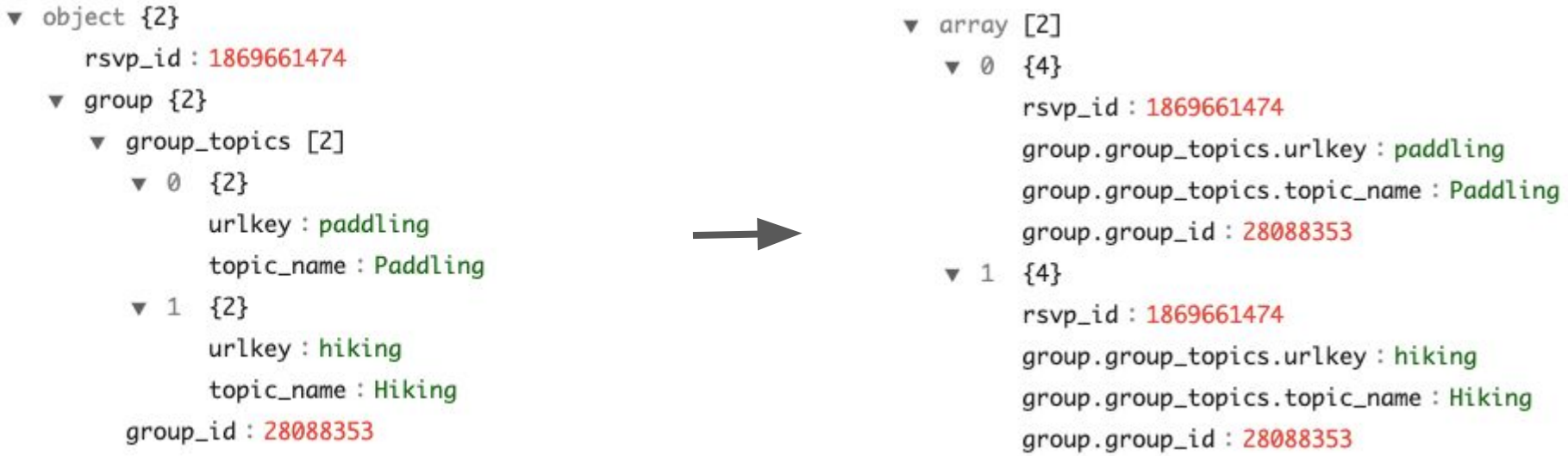


https://www.youtube.com/watch?v=CnSnLKQLuXc&t=1s

Complex-type (Array, Map) support in Pinot

Uber

- Released in Pinot 0.8.0

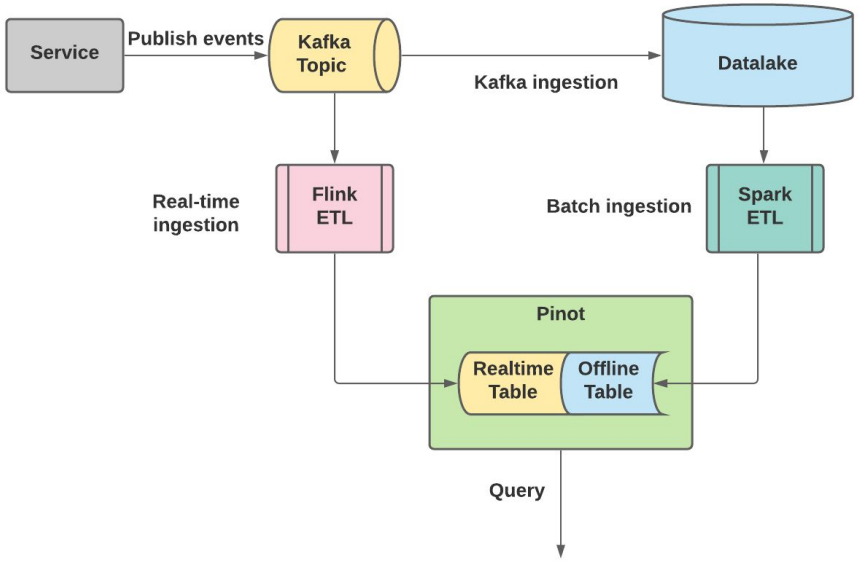


https://docs.pinot.apache.org/basics/data-import/complex-type

Uber

Spark/Flink connectors with Pinot

- Spark Connector
  - Pinot -> Hive dispersal
  - Perf improvement via GRPC
- Flink Connector
  - Streaming/batch unification for Pinot ingestion
  - Backfill large Pinot Upsert tables



Next

Uber

- Leverage StarTree index to replace custom rollup pipelines
- Better cluster isolation and tiering for ease of deployment
- Upsert table compaction / TTL

Q&A

How upsert works