Airflow @ Lyft

July 2020
Tao Feng | @feng-tao | Engineer, Lyft Data Platform

Blog: go.lyft.com/airflowblog

Who

- Engineer at Lyft Data Platform and Tools
- Apache Airflow PMC and Committer
- Working on different data products (Airflow, Amundsen, etc)
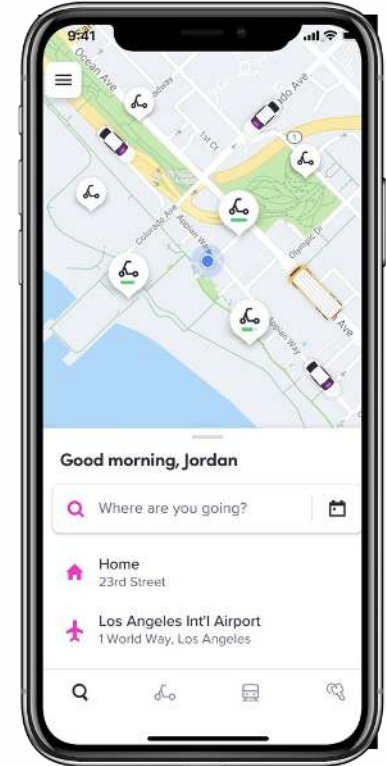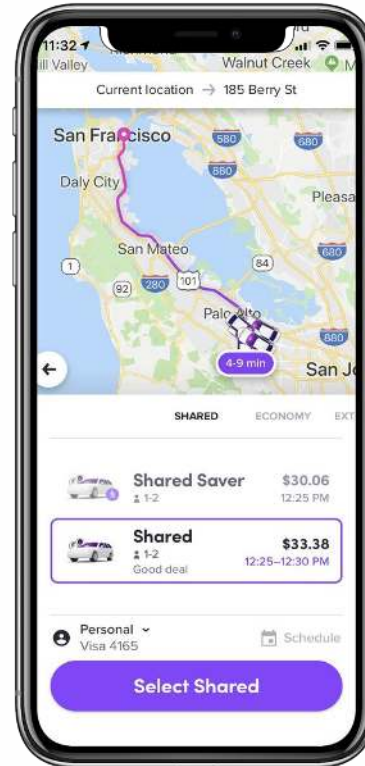- Previously at Linkedin, Oracle

Agenda

- Data Platform @ Lyft

- Airflow Customization @ Lyft
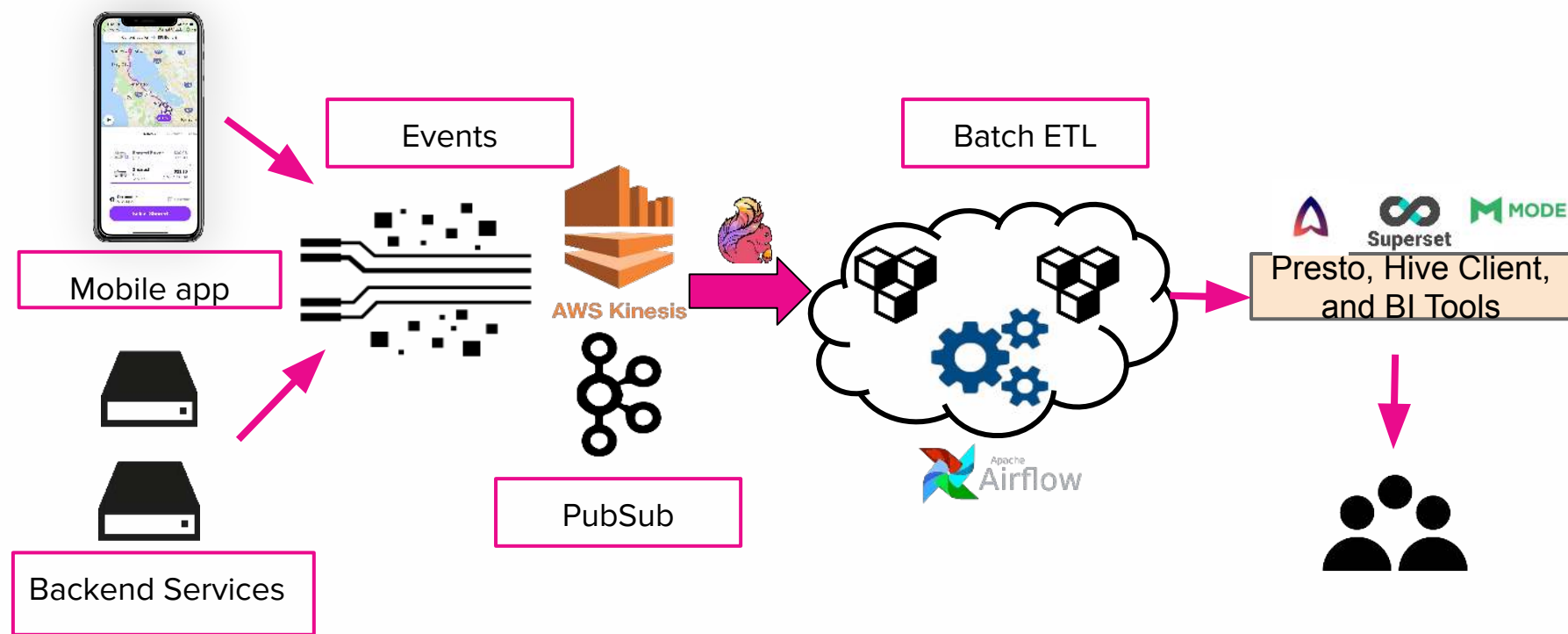
- Current Focus For Airflow @ Lyft

- Summary

Data Platform @ Lyft

About Lyft
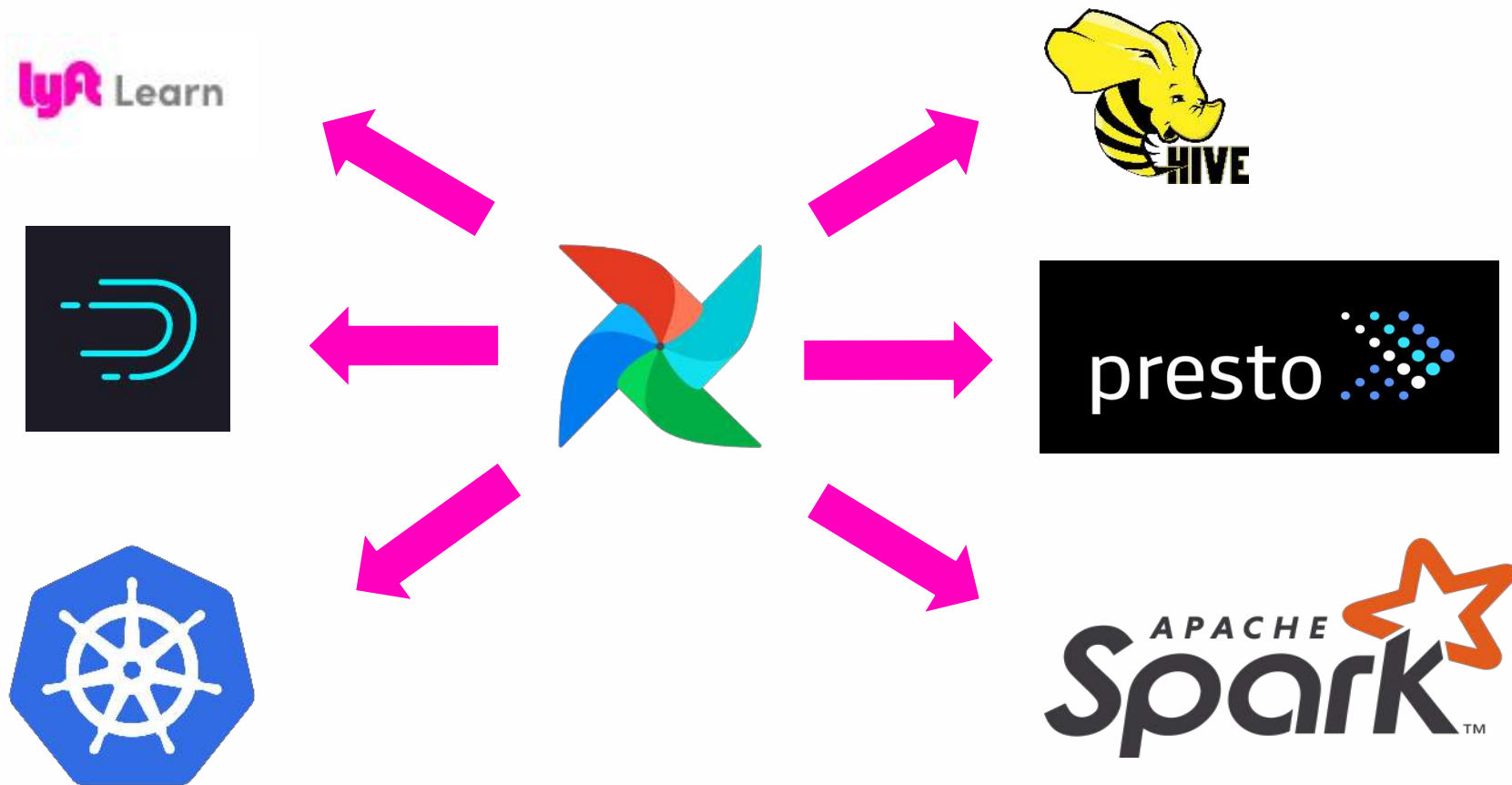
MISSION: Improve people's life with the world's best transportation

Lyft's data analytics platform architecture

Events

Mobile app

Backend Services

AWS Kinesis

PubSub

Batch ETL

Apache Airflow

Presto, Hive Client, and BI Tools

Superset

MODE

Airflow main use cases @ Lyft

Airflow usage @ Lyft

Total Active DAGs

1.21k
Total Active DAGs

Total DAG Run Yesterday

4.44k
DAG Run Yesterday

2. Task Run Yesterday

72k
Tasks Run Yesterday

- Two Clusters
- Celery Executors

Airflow Customization @ Lyft

Airflow customization @ Lyft

- UI auditing

- DAG dependency graph
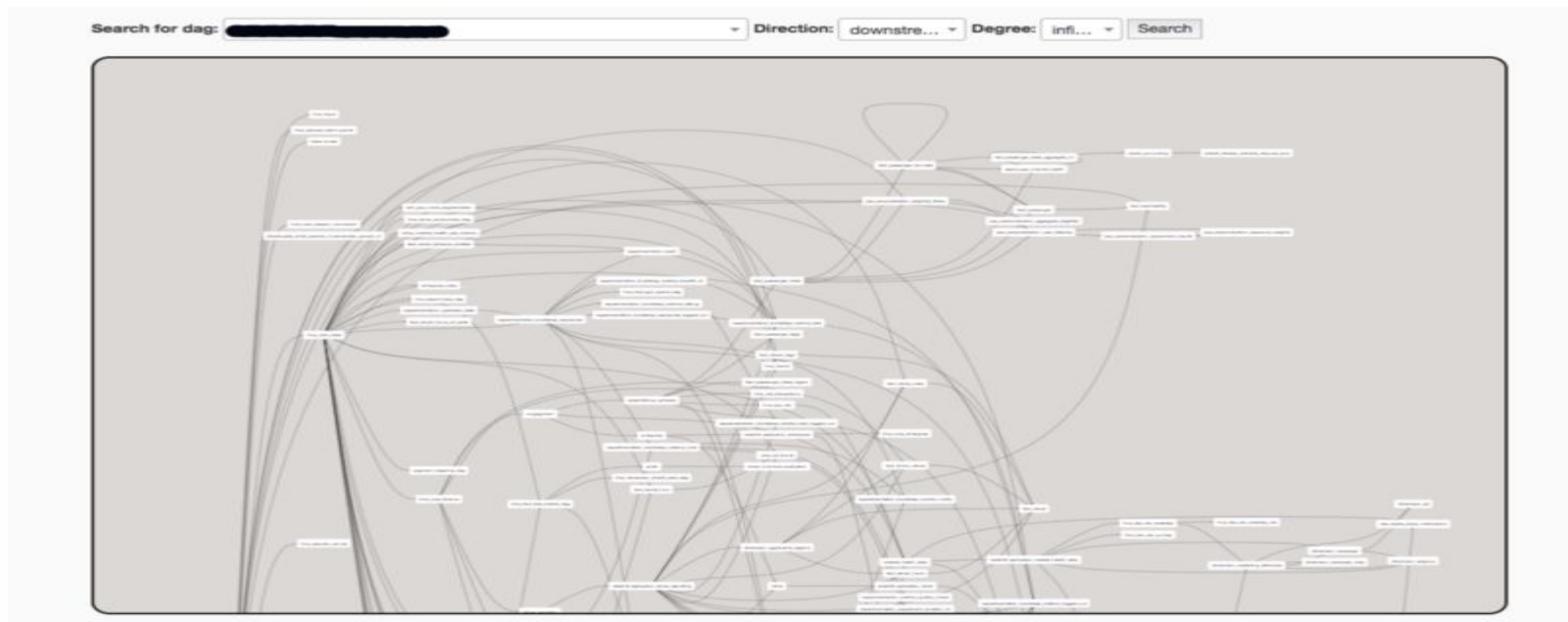
Airflow customization @ Lyft

- Extra link for task instance UI panel

- Hive query log
- *Dr elephant report* for performance tuning
- Hive job analysis dashboard

Airflow customization @ Lyft

- [Amundsen](#) is an **open-sourced** data discovery portal.
- It is integrated with Airflow to show the *task and table* **lineage**.
- It is currently used by **18+** companies.

Current Focus For Airflow @ Lyft

ETL Expiration System

- Lots of ETLs are not well maintained with no clear ownership.
- Built an ETL Expiration system to:
    - Disabled DAGs with expired TTLs (DAG owner needs to renew the TTL every six months).
    - Disabled DAGs that produced unused datasets
    - Disabled DAGs that are failing for a long time

PY2 -> PY3
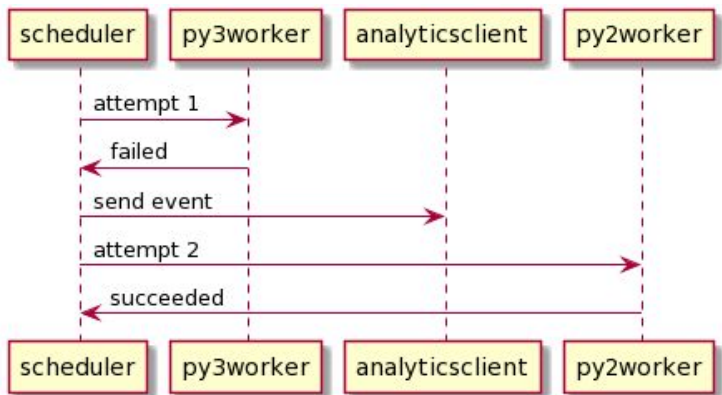
- Built a dashboard to understand PY3 issue.
  - Most issues are related to *string encoding* or *string and integer comparison*.
- DAG loading time is higher in py3 compared to py2
  - Cherry pick a few performance improvement patches from upstream

DAGs failing on py3

	dag_id	num_dag_runs	latest_dag_run	num_failures
1	event_hive	8	2020-06-23T02:30:00	3311
2	hive_core_rides	6	2020-06-23T00:00:00	34
3	fact_green_charges	3	2020-06-23T00:00:00	26
4	driving_supply_control_open_driving_hour	1	2020-06-12T17:00:00	2
5	paymongodep_beta_hive_pax_growth_subscriptions_dag	1	2020-06-06T00:00:00	6

Airflow Upgrade

- Leverage new features:
    - DAG serialization
    - RBAC
    - Data Lineage
    - Performance Improvements

- Current status:
    - Built a new multi-tenant cluster to onboard new use cases.
    - Finishing PY3 upgrade for legacy DAGs.
    - Converting the existing legacy mono DAG repo as another tenant on the new cluster.

Summary

Summary

- Covers Lyft data platform in general
- Discusses about Airflow customization at Lyft
- Discusses about Airflow current work at Lyft

Acknowledgement

- Members who maintain Airflow at Lyft
    - Andrew Stahlman
    - Bhanu Renukuntla
    - Chao-han Tsai (committer)
    - Jinhyuk Chang
    - Junda Yang
    - Max Payton
    - Sherry Zhao
    - Shenghu Yang (EM)
    - Tao Feng (committer)
- Thanks Maxime for his guidance

Tao Feng | @feng-tao
Blog at go.lyft.com/airflowblog