



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

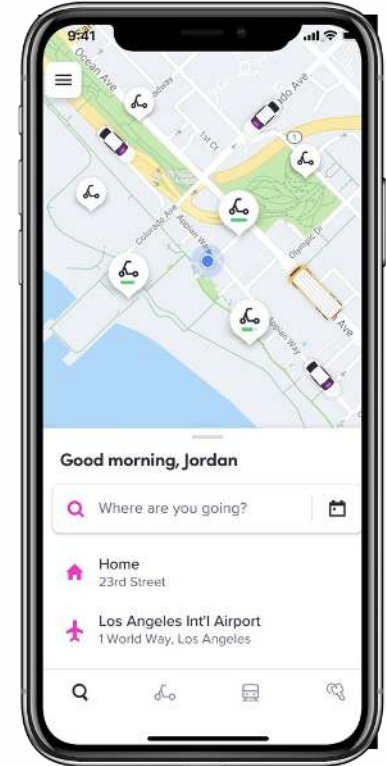
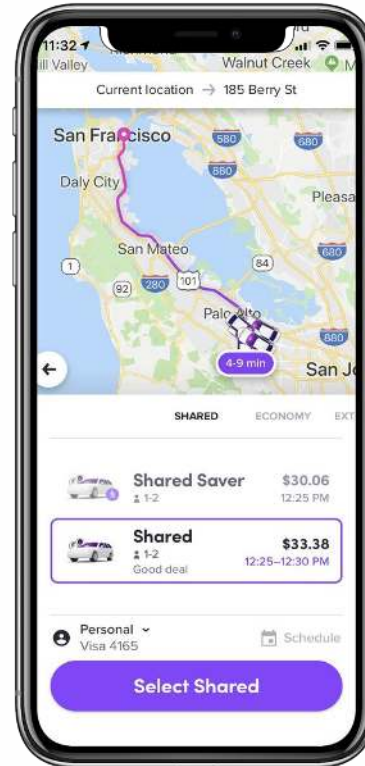
Agenda

- Data Platform @ Lyft
- Airflow Customization @ Lyft
- Current Focus For Airflow @ Lyft
- Summary

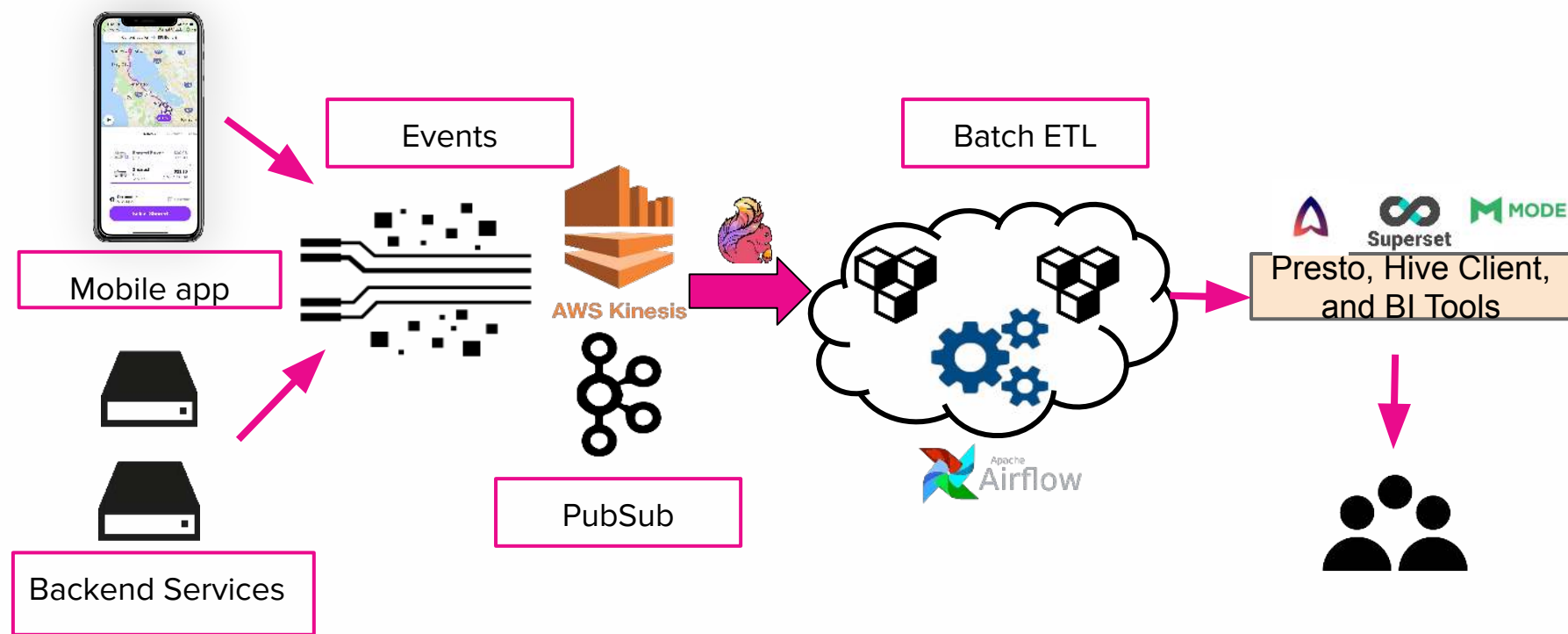
Data Platform @ Lyft

About Lyft

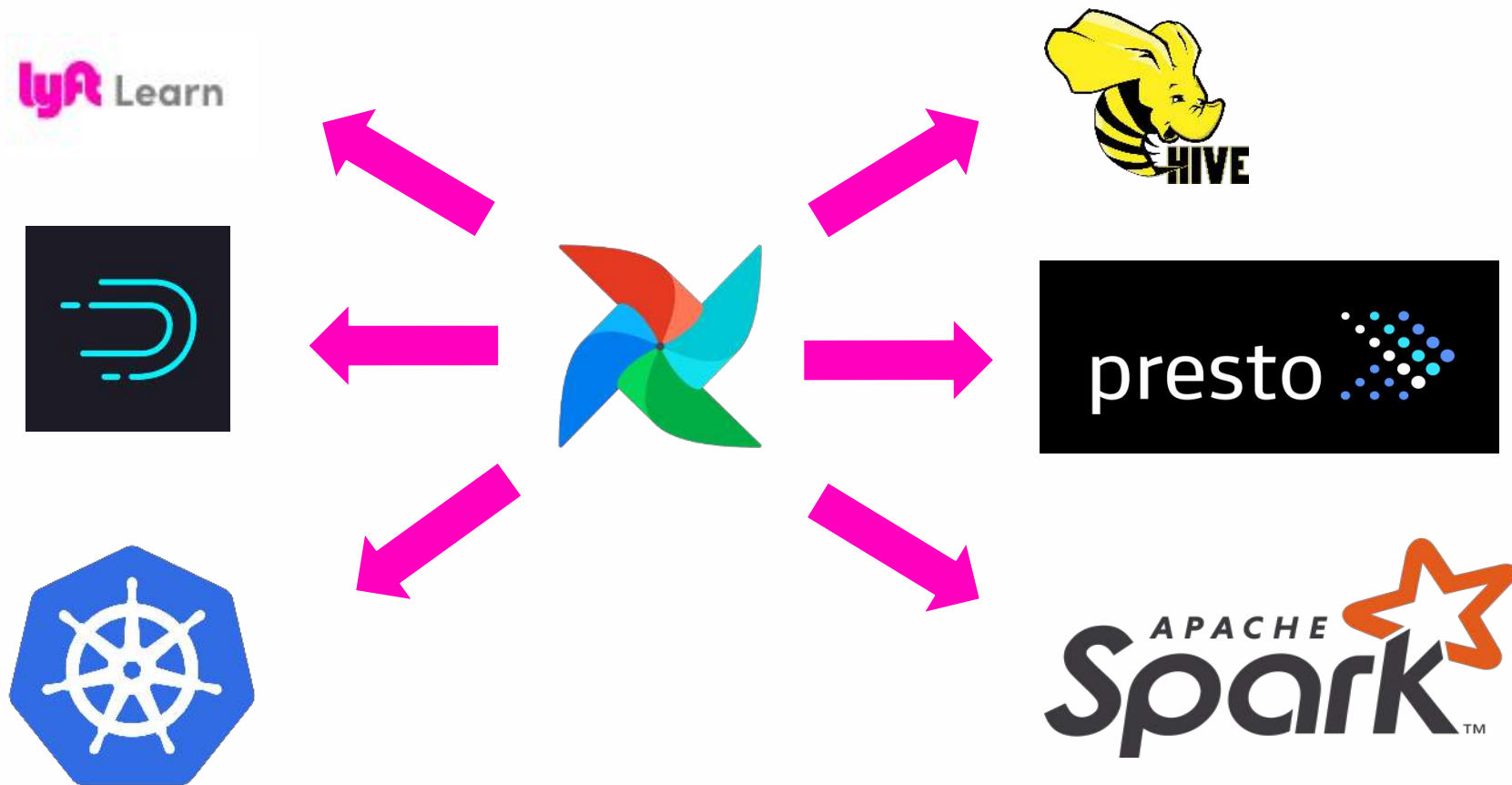
MISSION: Improve people's life with the world's best transportation



Lyft's data analytics platform architecture



Airflow main use cases @ Lyft



Airflow usage @ Lyft

Total Active DAGs

1.21k

Total Active DAGs

2. Task Run Yesterday

72k

Tasks Run Yesterday

Total DAG Run Yesterday

4.44k

DAG Run Yesterday

- Two Clusters
- Celery Executors

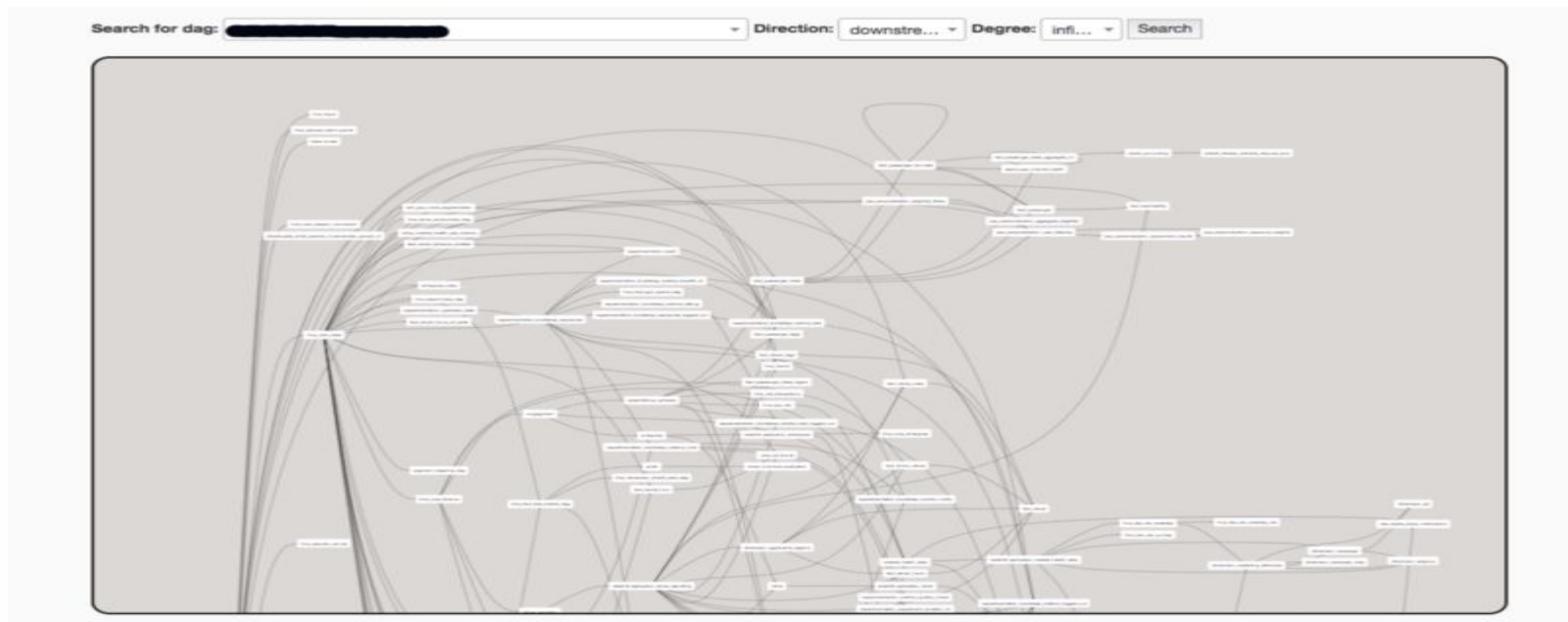
Airflow Customization @ Lyft

Airflow customization @ Lyft

- UI auditing

24914052	12-11T04:27:58.789532	None	ui_action			('message': 'turn off dag - fact_user_messages')
----------	-----------------------	------	-----------	--	--	--

- DAG dependency graph



Airflow customization @ Lyft

- Extra link for task instance UI panel

mozart_sonata_task ▼ on 2020-05-12T07:20:00

Task Instance Details Rendered Task Instances View Log

Clear Past Future Upstream Downstream Recursive

Mark Success Past Future Upstream Downstream

Hive_Logs Validate_SQL Dr_Elephant_Report **Hive_Job_Analysis**

Close

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

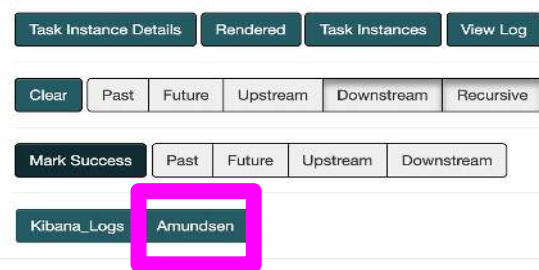
Hive Job Analysis ☆

Edit dashboard ▼

S3 Data Read	HDFS read	Files read	Files Created	Input Records	Total Tasks
51.6B	0	89.2B	250k	3.27B	20k
S3 Data Written	HDFS written	Files Written	Dynamic Partition Files Created	Skipped Input	Failed Tasks
4.15B	0	158B	0	19.9M	409

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.



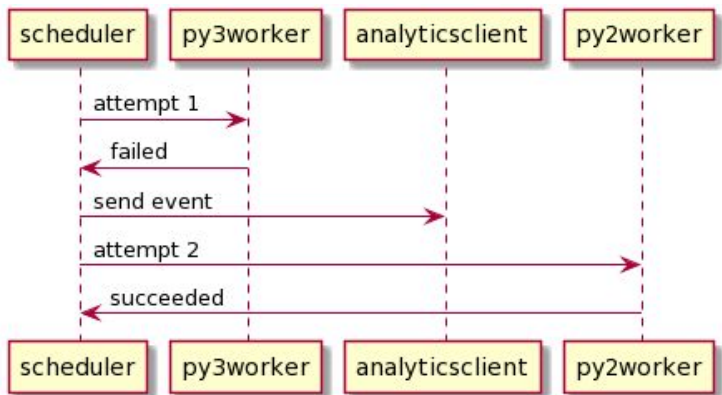
A screenshot of the Amundsen interface. The top navigation bar is dark purple with the Lyft logo and 'AMUNDSEN' text. A search bar is present. On the right, there are links for 'Announcements', 'Browse', 'FAQ', and a user profile icon. The main content area shows a dataset page for 'default.event_amundsenfrontend_user_action' with a star icon. Below the dataset name, it says 'Datasets • Hive'. A tooltip is visible over the 'Airflow' icon, showing the path 'event_hive/hive.default.event_amundsenfrontend_user_action'. The 'Airflow' icon is highlighted by a pink rectangle. To the right of the 'Airflow' icon are 'Lineage' and 'github' icons, and 'Preview' and 'Explore' buttons. The bottom section shows the 'Description' of the dataset: 'User action event from Amundsen frontend'. The 'Request Description' is also visible. On the right, there are tabs for 'Columns (23)' and 'Dashboards (0)'. The 'Columns (23)' tab is active, showing a column named 'ds' with the description 'Partition key in the format: YYYY-MM-DD' and a data type of 'string'.

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