



Airflow for ML at Twitter

@Dan Davydov

June 7th, 2020



Agenda

1. How Twitter uses Airflow for ML
2. Airflow pain points (focus on ML)
3. Future of Airflow @ Twitter



Airflow @ Twitter



~400 DAG Files

~30 Customer Teams

Mostly ML Use-Cases

Airflow ML Stack



Google Cloud Platform

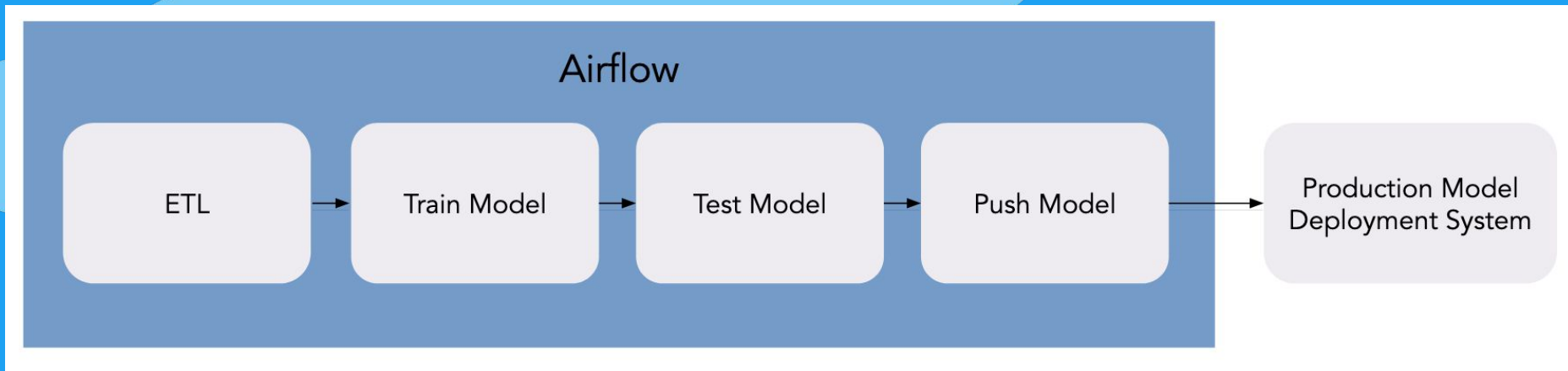


TensorFlow



kubernetes

Typical ML Airflow Pipeline





Airflow Pain Points



Development Speed



VS

Airflow DAGs Data Profiling Browse Admin Docs About 2018-09-07 22:14:10 UTC

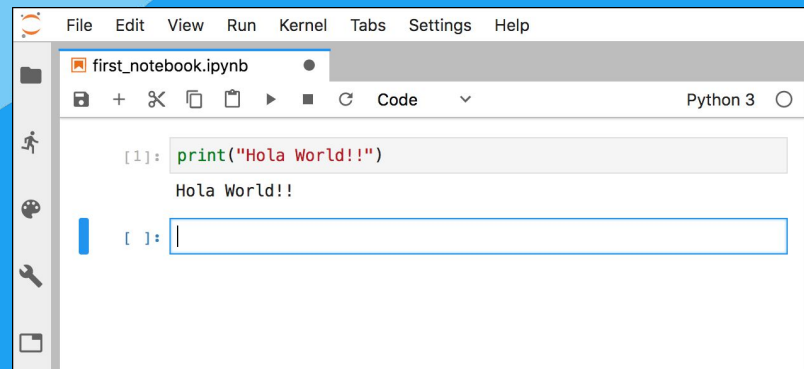
DAGs

Search:

	DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
On	example_bash_operator	<code>@@:*</code>	airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2018-09-06 00:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	🔍 📄 🔗 🔧 🔖 🔖 🔖 🔖 🔖 🔖
On	example_branch_dag_operator_v3	<code>@@:*</code>	airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2018-09-05 00:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	🔍 📄 🔗 🔧 🔖 🔖 🔖 🔖 🔖 🔖
On	example_branch_operator	<code>@@:*</code>	airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2018-09-06 00:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	🔍 📄 🔗 🔧 🔖 🔖 🔖 🔖 🔖 🔖
On	example_xcom	<code>@@:*</code>	airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2018-09-05 00:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	🔍 📄 🔗 🔧 🔖 🔖 🔖 🔖 🔖 🔖
On	latest_only	<code>@@:*</code>	Airflow	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2018-09-07 16:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	🔍 📄 🔗 🔧 🔖 🔖 🔖 🔖 🔖 🔖

Showing 1 to 5 of 5 entries

Show Paused DAGs



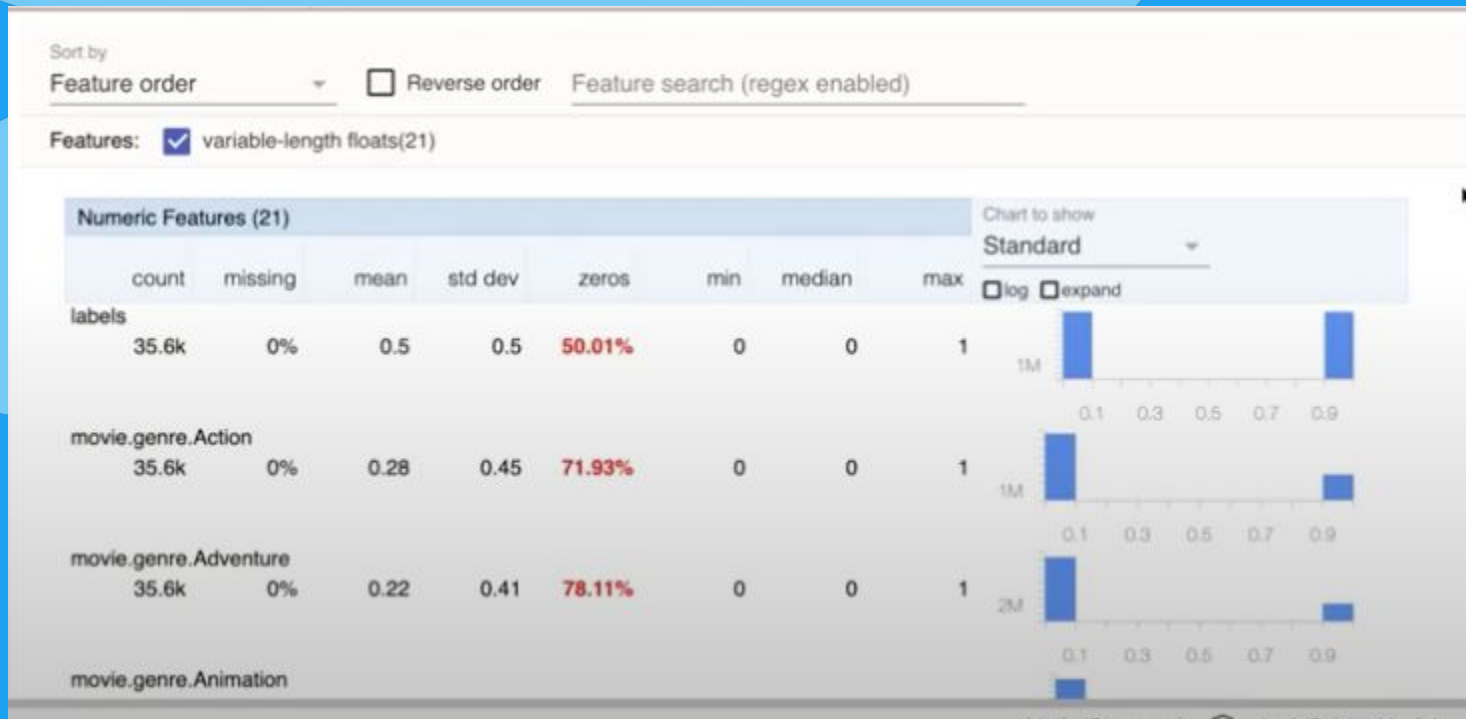


Clunky Interfaces

```
def push(**kwargs):  
    # pushes an XCom without a specific target  
    kwargs['ti'].xcom_push(key='value from pusher 1', value=value_1)
```

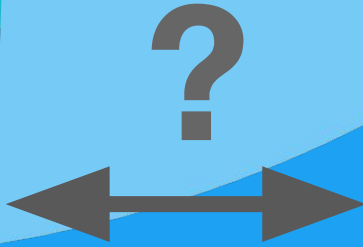


Lack of First-Class ML Tooling Integration





Lack of OSS ML Operators

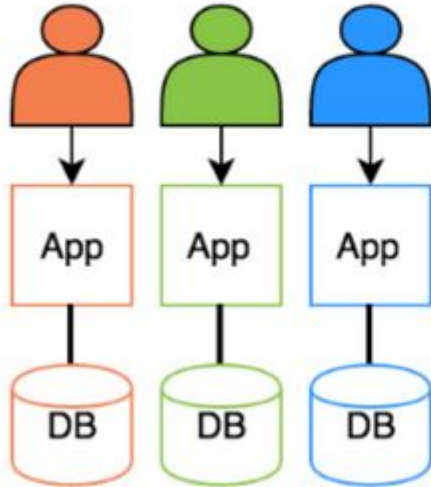


TensorFlow



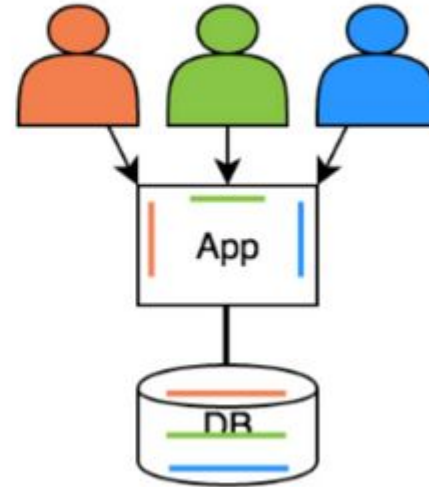
No Multi-tenancy

Single-Tenant



Vs

Multi-Tenant

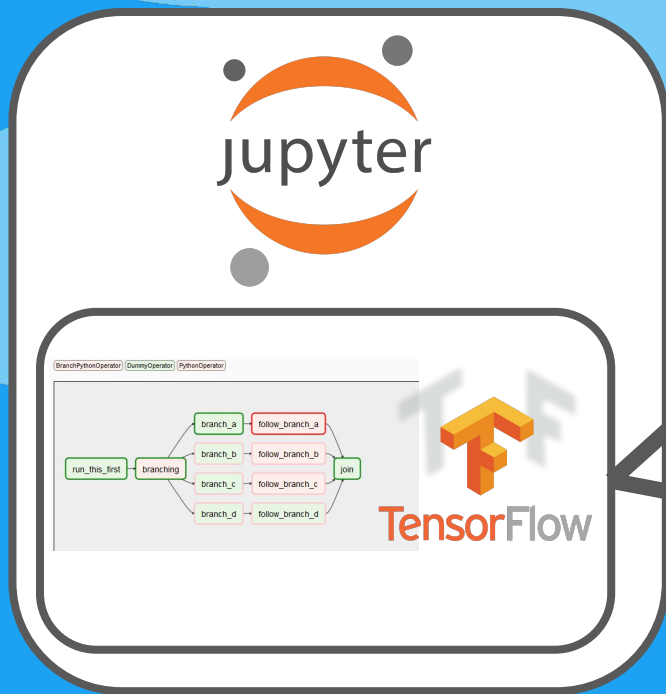




Future of Airflow for ML @ Twitter



Airflow as Job Dispatch



Production Pipelines





Thanks For Watching!