



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



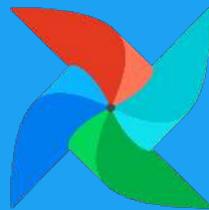
Airflow ML Stack



Google Cloud Platform



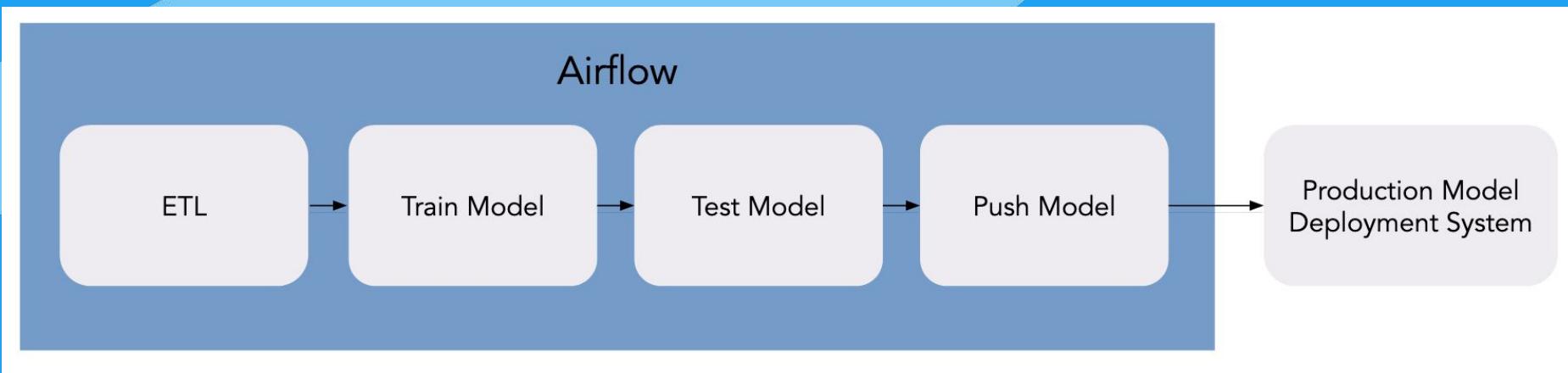
TensorFlow



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
example_bash_operator	0 0 * * *	airflow	0 0 0 0 0 0	2018-09-05 00:00:00	0 0 0 0 0 0	0 0 0 0 0 0
example_branch_dop_operator_v3	*/5 * * * *	airflow	0 0 0 0 0 0	2018-09-05 00:56:00	0 0 0 0 0 0	0 0 0 0 0 0
example_branch_operator	daily	airflow	0 0 0 0 0 0	2018-09-05 00:00:00	0 0 0 0 0 0	0 0 0 0 0 0
example_xcom	None	airflow	0 0 0 0 0 0	2018-09-05 00:00:00	0 0 0 0 0 0	0 0 0 0 0 0
latest_only	400000	Airflow	0 0 0 0 0 0	2018-09-07 16:00:00	0 0 0 0 0 0	0 0 0 0 0 0

Showing 1 to 5 of 5 entries

< > 1 < >

Show Paused DAGs

File Edit View Run Kernel Tabs Settings Help

first_notebook.ipynb

[1]: `print("Hola World!!")`

Hola World!!

[1]: |

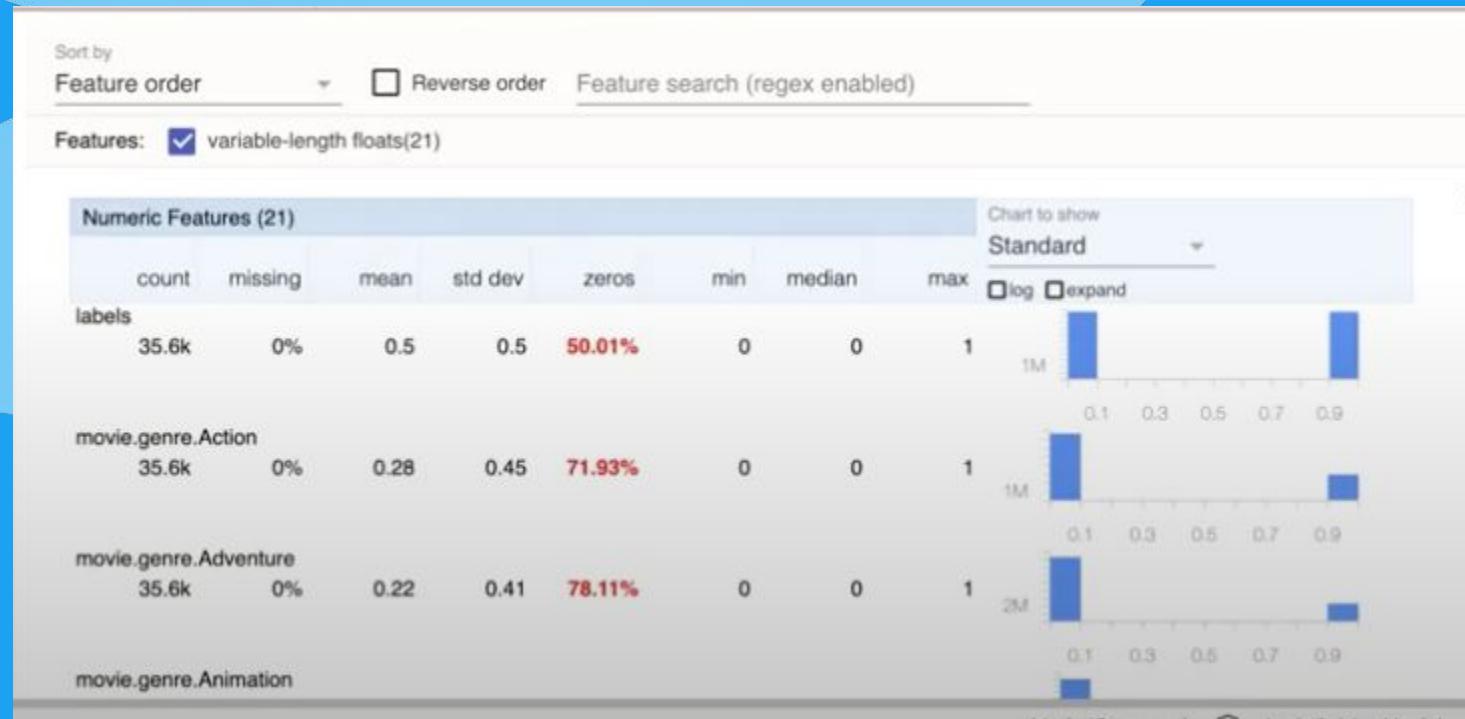


Clunky Interfaces

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

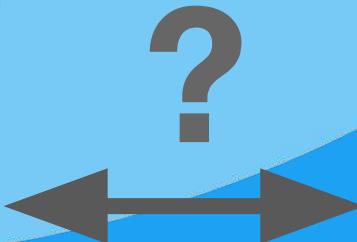
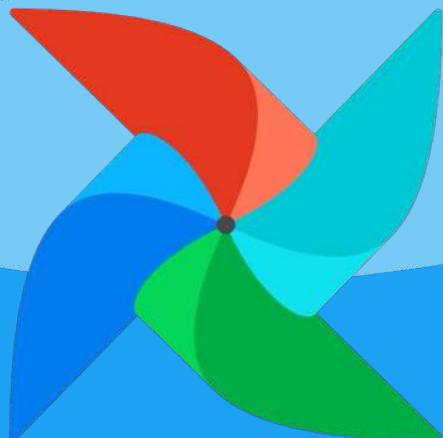


Lack of First-Class ML Tooling Integration





Lack of OSS ML Operators

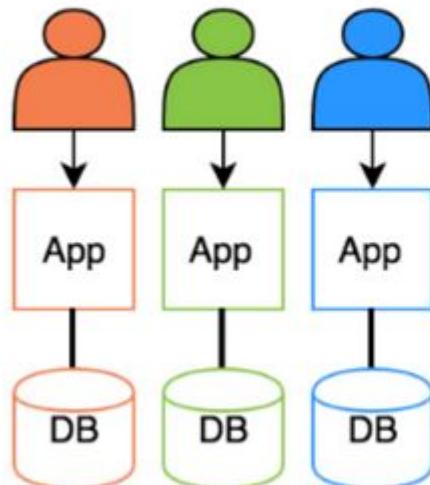


TensorFlow

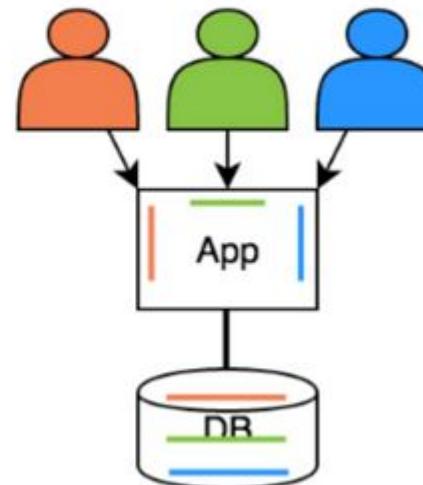


No Multi-tenancy

Single-Tenant



Multi-Tenant



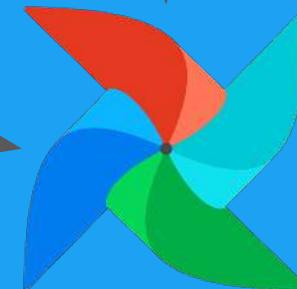
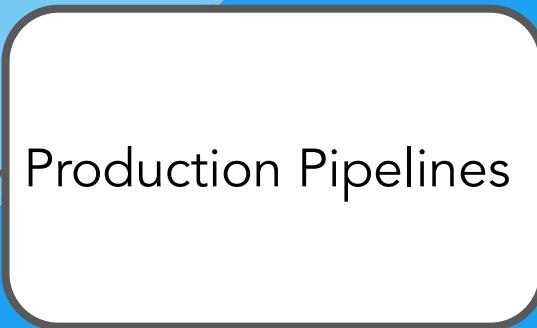
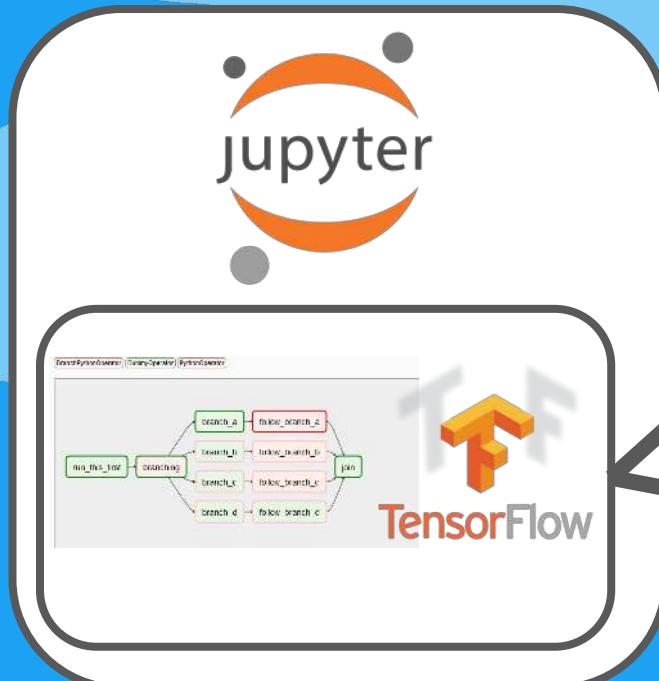
VS



Future of Airflow for ML @ Twitter



Airflow as Job Dispatch





Thanks For Watching!