



# 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

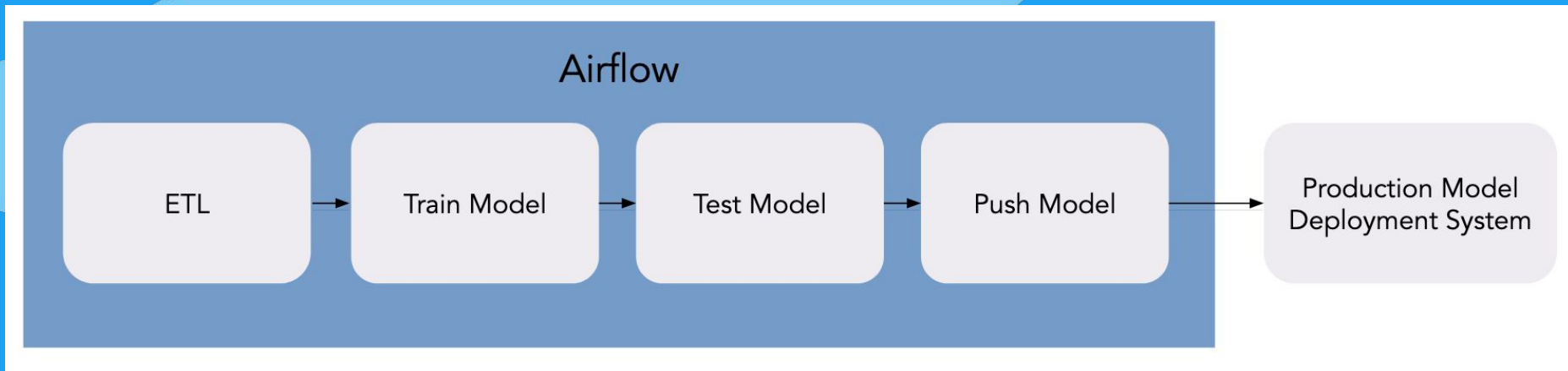


TensorFlow



kubernetes

# Typical ML Airflow Pipeline





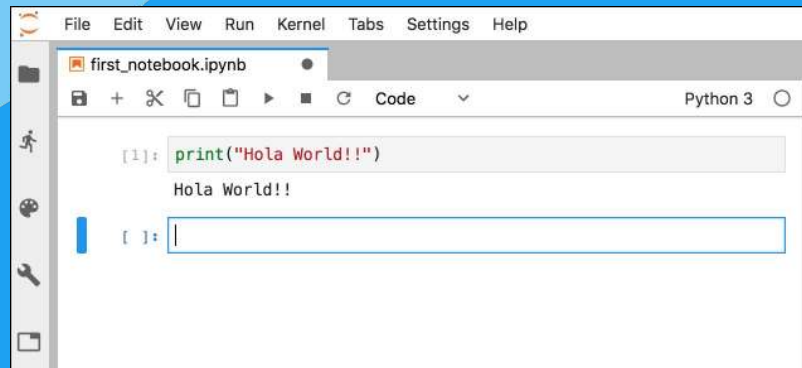
# Airflow Pain Points



# Development Speed



VS



Apache Airflow DAGs view showing a table of DAGs and their recent task status.

DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
example_bash_operator	*/5 * * * *	airflow		2018-09-06 00:00		
example_branch_dag_operator_v3	*/5 * * * *	airflow		2018-09-05 00:00		
example_branch_operator	*/5 * * * *	airflow		2018-09-05 00:00		
example_xcom	*/5 * * * *	airflow		2018-09-05 00:00		
latest_only	*/5 * * * *	Airflow		2018-09-07 16:00		

Showing 1 to 5 of 5 entries



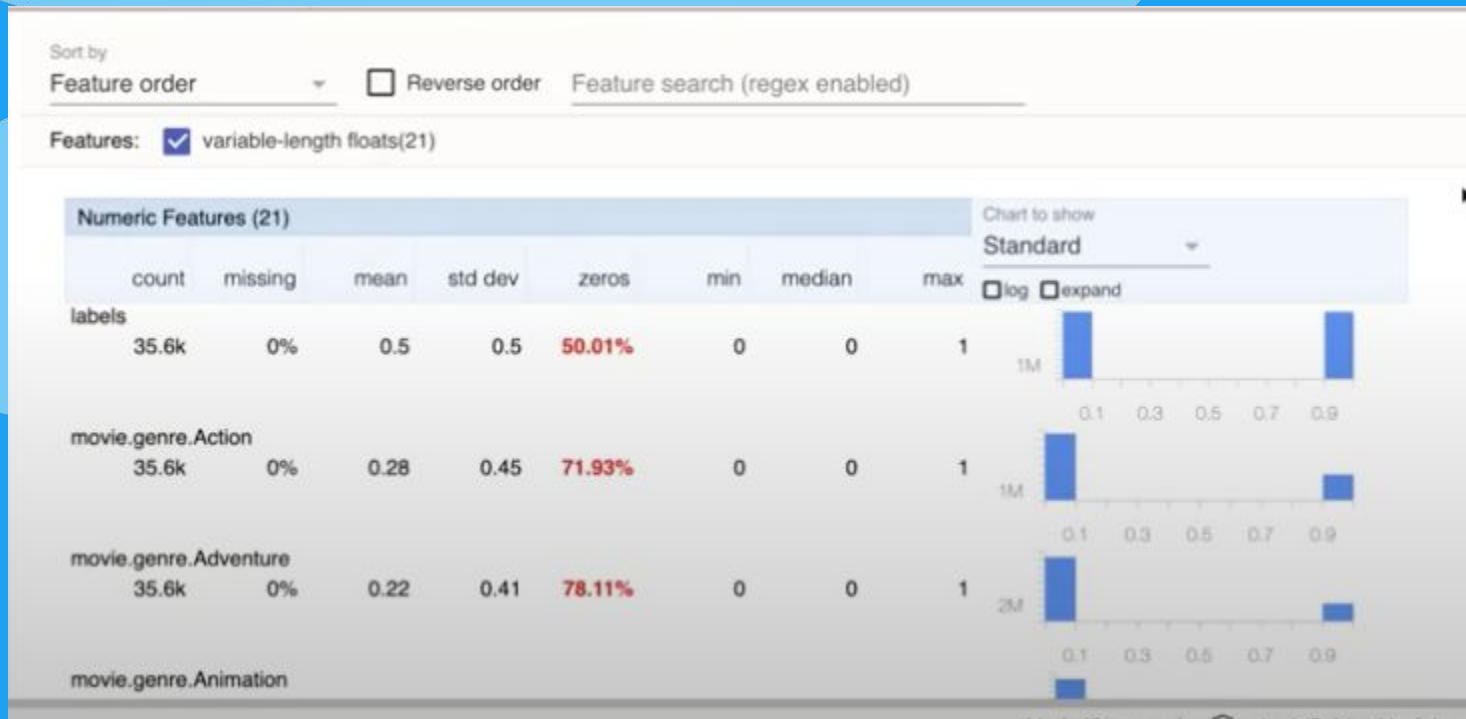


# Clunky Interfaces

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

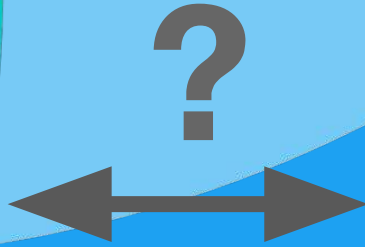
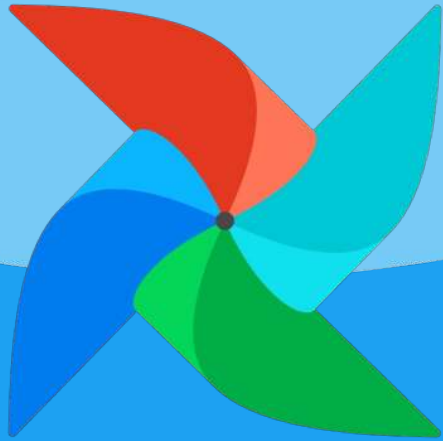


# Lack of First-Class ML Tooling Integration





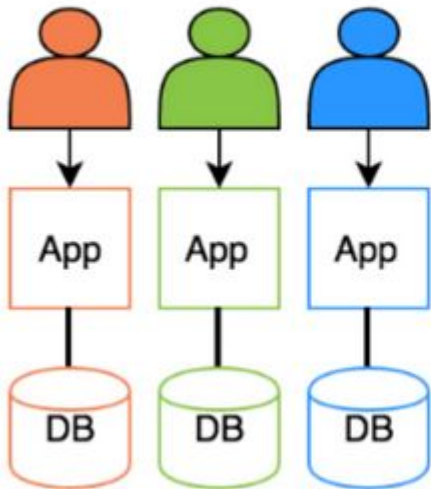
# Lack of OSS ML Operators



TensorFlow

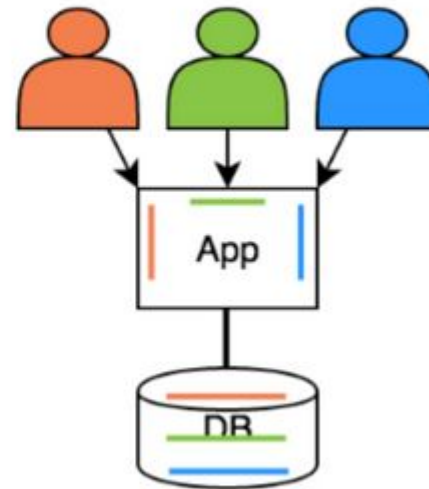
# No Multi-tenancy

Single-Tenant



Vs

Multi-Tenant

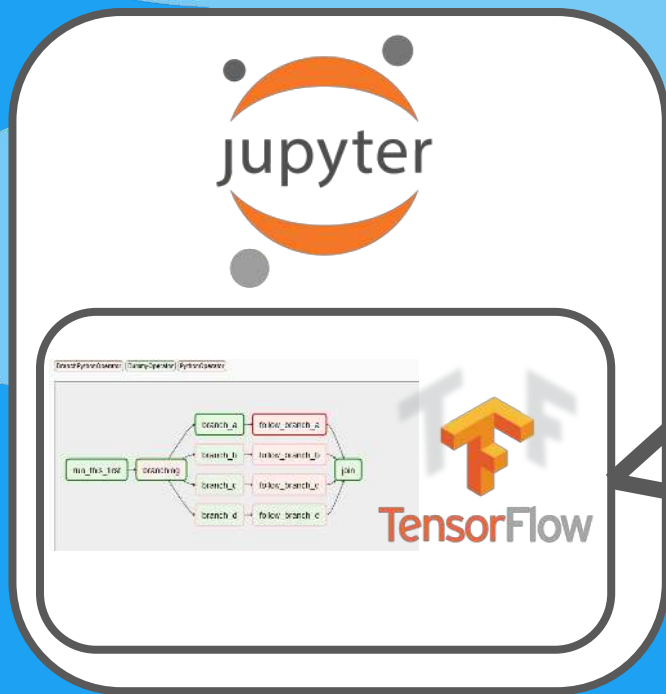




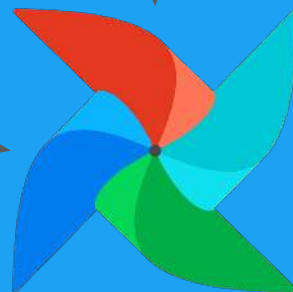
# Future of Airflow for ML @ Twitter



# Airflow as Job Dispatch



Production Pipelines





Thanks For Watching!