Airflow & Zeppelin: Better together

Jeff Zhang (zjffdu)
Agenda

- What is Zeppelin
- Why Airflow + Zeppelin
- Demo
What is Zeppelin

Web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala, Python and more.
Supported Interpreters

- Spark (Scala/PySpark/SparkR/Sql)
- Flink (Scala/PyFlink/Sql)
- Jdbc (Mysql, Hive, Presto and etc.)
- Python
- R
- Cassandra
- Shell
- Markdown
- ....
Why Zeppelin (Alternative to Jupyter)

- Better Support of Big Data
- Built-in Visualization
- Easy Configuration
- Self-Contained
Agenda

- What is Zeppelin
- Why Airflow + Zeppelin
- Demo
Move Spark Job to Production

```python
sql_job = SparkSqlOperator(
    sql="SELECT * FROM bar",
    master="yarn",
    executor_cores=1,
    executor_memory="2g",
    task_id="sql_job"
)
```

```python
submit_job = SparkSubmitOperator(
    application="${SPARK_HOME}/examples/src/main/python/pi.py",
    executor_cores=2,
    num Executors=5,
    driver_memory="2g",
    task_id="submit_job"
)
```
Move Spark Job to Production

- Develop Spark Job
- Copy Code & Configuration to Airflow Operator
- Run Spark Job in Airflow
- Update your Spark Job

- The second step is error prone and inefficient
- Environment in development stage is different from that in production stage
Airflow Worker

- PythonOperator
- BashOperator
- SparkSubmitOperator
- SparkSqlOperator
- ...

- How to launch task
- Where to launch task
- Job Configuration
- Job Code
Airflow + Zeppelin

- Airflow Worker
- ZeppelinOperator
- Zeppelin

- How to launch task
- Where to launch task
- Job Configuration
- Job Code

- Python Interpreter
- Flink Interpreter
- Spark Interpreter
- ...

ZeppelinOperator and Zeppelin are connected through a flow diagram showing how to launch tasks with different interpreters.
Zeppelin Operator

```python
pyspark_task = ZeppelinOperator(
    task_id='pyspark_note',
    conn_id='zeppelin_default',
    note_id='2EWM84JXA',
    version='0.1' # not supported yet
)
```
 Agenda

● What is Zeppelin

● Why Airflow Zeppelin

● Demo
Future Work

- Git Integration
- K8s Support
- HA Support
- Integrate with other modern data stack (e.g. Great Expectations)
References

- http://zeppelin.apache.org/
- https://zeppelin.apache.org/docs/0.10.1/interpreter/spark.html
- https://zeppelin.apache.org/docs/0.10.1/interpreter/flink.html
- https://github.com/zjffdu/zeppelin_airflow (demo docker)
Thank You!