What is an Operator?

“How to run your tasks”

Definition

While DAGs describe how to run a workflow, Operators determines what actually gets done.

An operator describes a single task in a workflow. Operators are usually (but not always) atomic, meaning they can stand on their own and don't need to share resources with any other operators. The DAG will make sure that operators run in the correct certain order; other than those dependencies, operators generally run independently. In fact they may run on two completely different machines (for scalability)

Airflow Documentation

Key Points

- Definition of a single task
- Should be idempotent
  - Meaning your operator should produce the same result regardless of how many times it is run.
- Retry automatically
  - In case of a failure
- A Task is created by instantiating an Operator class.
- An Operator defines the nature of this Task and how should it be executed.
- When an Operator is instantiated, this task becomes a node in your DAG.

Airflow Provide Many Operators

- BashOperator
  - Executes a bash command
- PythonOperator
  - Calls an arbitrary Python function
- EmailOperator
  - Sends an email
- MySqlOperator, SqliteOperator, PostgreOperator…
  - Executes a SQL command

Types of Operators

- All Operators inherit from BaseOperator.
- There are actually 3 types of operators:
  - Action operators that perform an action (BashOperator, PythonOperator, EmailOperator ...)
  - Transfer operators that move data from one system to another (PrestoToMysqlOperator, SftpOperator ...)
  - Sensor operators waiting for data to arrive at a defined location.

Transfer Operators

- Operators that move data from one system to another.
- Data will be pulled out from the source, staged on the machine where the executor is running, and then transferred to the target system.
- Don’t use these operators if you are dealing with a large amount of data.

Sensor Operators

- Sensor operators inherit of BaseSensorOperator (BaseOperator being the superclass of BaseSensorOperator)
- They are useful for monitoring external processes like waiting for files to be uploaded in HDFS or a partition appearing in Hive.
- They are basically long running task.
- The Sensor Operator has a poke method called repeatedly until it returns True (it is the method used for monitoring the external process)

Coding Time!

Let’s create some amazing tasks using the operators we’ve just learned !