**Apache Falcon**

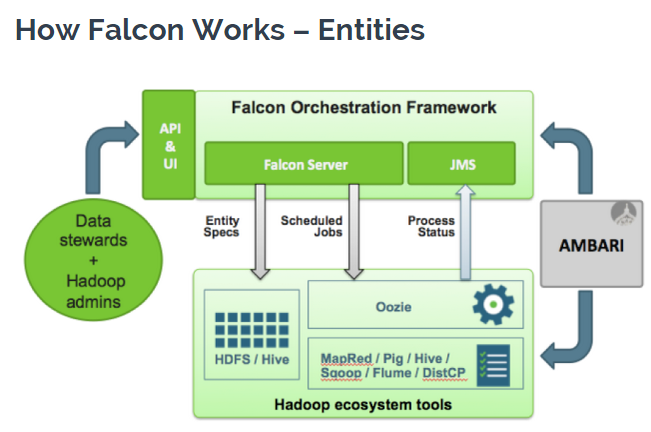
A framework for managing data life cycle in Hadoop Clusters.

Falcon is an XML based entity driven system that is made up of –

Clusters

Feeds

Processes

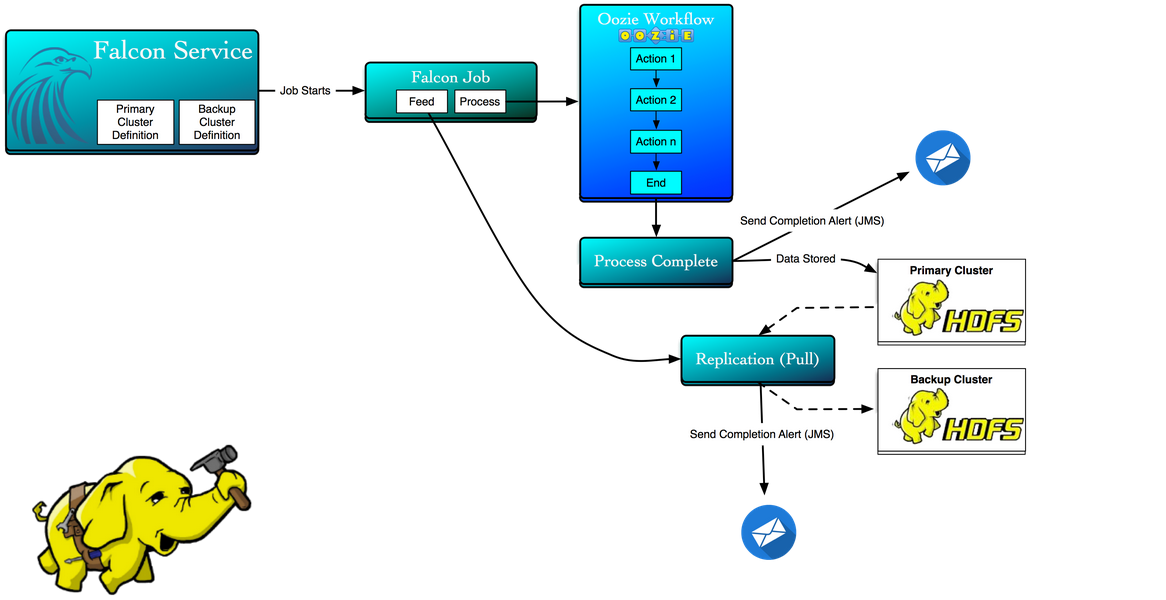


**A Cluster entity** defines where data, tools, and processes live on your Hadoop cluster.

**A Feed entity** defines where data lives on your cluster (in HDFS).

**A Process Entity**defines what action or “process” will be taking place in a pipeline.

Falcon’s a pipeline that you define in the form of entities (basically the pieces of the workflow) that define a workflow, its frequency, and a number of policies like lateness, retention, and replication.



**Defining and Submitting a Cluster entity –**

* Define cluster xml
* Create staging and working directories for Falcon to use (with specific ownership and permissions)
* Submitted a cluster to Falcon
* Checked existence of cluster in Falcon store
* Submitted a backup cluster to Falcon

<?xml version="1.0"?>

<cluster name="my-primary-cluster" description="my first cluster entity" colo="Primary Hadoop Cluster: CompanyName" xmlns="uri:falcon:cluster:0.1">

<interfaces>

<interface type="readonly" endpoint="hftp://name-node.com:50070" version="2.5.0" />

<interface type="write" endpoint="hdfs://name-node.com:54310" version="2.5.0" />

<interface type="execute" endpoint="job-tracker:54311" version="2.5.0" />

<interface type="workflow" endpoint="http://oozie.com:11000/oozie/" version="4.0.1" />

<interface type="messaging" endpoint="tcp://jms-server.com:61616?daemon=true" version="5.1.6" />

</interfaces>

<locations>

<location name="staging" path="/apps/falcon/my-primary-cluster/staging"/>

  <location name="temp" path="/tmp"/>

 <location name="working" path="/apps/falcon/my-primary-cluster/working"/>

  </locations>

</cluster>

**Staging**: Falcon copies the artifacts of submitted/scheduled processes and feeds to distinct child folders under this location. (Permissions 777)

**Working**: Falcon copies the jars it requires for execution to this location to support processes and feeds. (Permissions 755)

**Submiting a Cluser to Falcon -**

falcon entity -type cluster -submit -file my-primary-cluster.xml

**checking –**

falcon entity -type cluster –list

o/p - (CLUSTER) my-primary-cluster

**Defining and Submitting a Feed entity –**

* We define a feed xml
* Submit feed to Falcon
* Checked existence of feed in Falcon store
* Schedule a feed in Falcon

<?xml version="1.0" encoding="UTF-8" standalone="yes"?>

<feed name="my-example-feed" description="my-example-feed" xmlns="uri:falcon:feed:0.1">

 <frequency>days(1)</frequency>

 <timezone>UTC</timezone>

 <clusters>

 <cluster name="my-primary-cluster" type="source">

 <validity start="2015-05-01T23:00Z" end="2099-12-31T23:00Z"/>

 <retention limit="months(9999)" action="delete"/>

 <locations>

 <location type="data" path="/path/to/my/data/"/>

 </locations>

 </cluster>

 <cluster name="my-backup-cluster" type="target">

 <validity start="2015-05-01T23:00Z" end="2099-12-31T23:00Z"/>

 <retention limit="months(9999)" action="delete"/>

 <locations>

 <location type="data" path="/path/to/my/data/"/>

 </locations>

 </cluster>

 </clusters>

 <locations>

 <location type="data" path="/path/to/my/data/"/>

 <location type="stats" path="/none"/>

 <location type="meta" path="/none"/>

 </locations>

 <ACL owner="someUnixUser" group="someUnixGroup" permission="0755"/>

 <schema location="/none" provider="none"/>

</feed>

The above feed

* replicates data at ‘/path/to/my/data/’ daily at 23:00 UTC from my-primary-cluster to my-backup-cluster as someUnixUser
* retains data in ‘/path/to/my/data/’ for 9999 months

**Submitting a Feed to Falcon -**

falcon entity -type feed -submit -file my-example-feed.xml

**checking –**

falcon entity -type feed –list

O/p - (CLUSTER) my-example-feed

**Scheduling a Feed to Falcon -**

falcon entity -type feed -schedule -name my-example-feed

**Defining and Submitting a Process entity –**

* We define a process xml
* We define an oozie workflow
* Submit the process to Falcon
* Check existence of process in Falcon store
* Schedule a process in Falcon

<?xml version="1.0" encoding="UTF-8" standalone="yes"?>

<process name="my-example-process" xmlns="uri:falcon:process:0.1">

  <clusters>

    <cluster name="my-primary-cluster">

      <validity start="2015-05-14T23:00Z" end="2099-03-10T23:00Z"/>

    </cluster>

  </clusters>

  <parallel>1</parallel>

  <order>FIFO</order>

  <frequency>days(1)</frequency>

  <timezone>UTC</timezone>

  <outputs>

    <output name="my-example-feed" feed="my-example-feed" instance="now(0,0)"/>

  </outputs>

<properties>

     <property name="workflowName" value="my-example-workflow" />

     <property name="sshCommand" value="/path/to/my/script/cool\_script.sh"/>

     <property name="inputFile" value="data.csv" />

     <property name="outputPath" value="/path/to/store/output/" />

   </properties>

  <workflow name="my-example-workflow" version="2.0.0" engine="oozie" path="/tmp/my-example-workflow/"/>

  <retry policy="periodic" delay="minutes(15)" attempts="3"/>

  <ACL owner="someUnixUser" group="someUnixGroup" permission="0755"/>

</process>

It represents a process that:

* executes a process daily at 23:00 UTC from my-primary-cluster starting on May 14, 2015 and will do so daily until 2099-03-10 (basically forever)
* execution calls an Oozie Workflow, stored in HDFS
* Oozie workflow has action(s) in it, which is truly where we tell our workflow what to do

**Definition of an Oozie Workflow –**

<workflow-app name="my-example-workflow" xmlns="uri:oozie:workflow:0.1">

 <start to="example-action"/>

 <action name="example-action">

  <ssh xmlns="uri:oozie:ssh-action:0.1">

      <host>${host}</host>

      <command>${sshCommand} </command>

      <args>${inputFile}</args>

      <args> ${outputPath}</args>

      <capture-output/>

   </ssh>

   <ok to="end"/>

   <error to="kill"/>

 </action>

 <kill name="kill">

   <message>Action failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>

 </kill>

 <end name="end"/>

</workflow-app>

**Submitting a Process to Falcon -**

falcon entity -type process -submit -file my-example-process.xml

**checking –**

falcon entity -type process –list

O/p - (PROCESS) my-example-process

**Scheduling a Process in Falcon -**

falcon entity -type process -schedule -name my-example-process