

YAHOO!

Oozie: Workflow Engine for Hadoop



楊詠成/Yung-Cheng (Gibson) Yang
ycyang@yahoo-inc.com

2010/10/02

Outline

1. What is oozie
2. Do you need oozie
3. Understand oozie (**workflow/coordinator**)
4. How to use oozie
5. Use case sharing
6. Q & A

What Is Oozie ?

- Originally designed at Yahoo!
- Apache incubator project since 2011
- A web service that launches your jobs based on:
 - Time dependency
 - Data dependency
- Ability to rerun from last point of failure
- Monitoring

Do You Need Oozie ?

Q1: Having multiple jobs with dependency ?

Q2: Need to run jobs regularly ?

Q3: Need to check data availability ?

Q4: Need monitoring and operational support ?

If any one of your answer is YES,
then you should consider Oozie!

Understand Oozie - Workflow

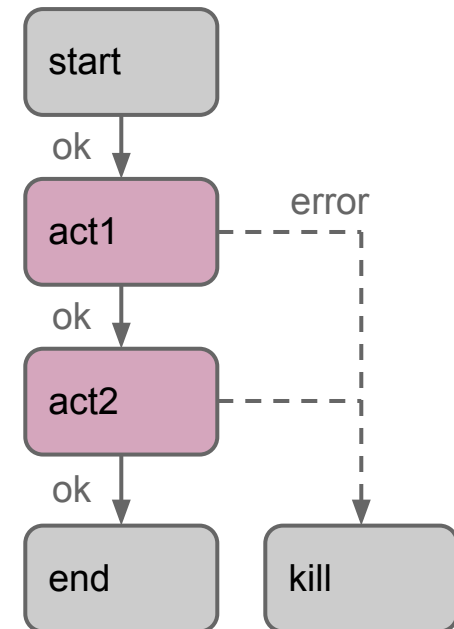
- Directed Acyclical Graph (DAG) of actions
- Supported Actions:
 - Map-Reduce action
 - Pig action
 - Java action
 - FS (HDFS) action
 - Email action
 - Shell action
 - Hive action
 - Sqoop action
 - Sub-workflow action
 - Writing a custom action

Understand Oozie - Workflow

- Basic flow control:

<ok to="..."/>, <error to="..."/>

```
<workflow-app name="demo1" xmlns="uri:oozie:workflow:0.3">
  <start to="act1"/>
  <action name="act1">
    ...omit...
    <ok to="act2"/>
    <error to="kill"/>
  </action>
  <action name="act2">
    ...omit...
    <ok to="end"/>
    <error to="kill"/>
  </action>
  <kill name="kill">
    <message>custom error message</message>
  </kill>
  <end name="end"/>
</workflow-app>
```



- More flow control nodes:

<fork ...>, <join ...>, <decision ...>

Understand Oozie - Coordinator

```
<coordinator-app name="demo2" frequency="60"  
  start="2012-10-02T00:00Z" end="2012-10-04T24:00Z" ...>  
  <action>  
    <workflow>  
      <app-path>hdfs://xxxx:9000/myworkflow/demo1</app-path>  
    </workflow>  
  </action>  
</coordinator-app>
```

- This coordinator works like a crontab
- runs the workflow every 60 mins

Understand Oozie - Coordinator

```
<coordinator-app name="demo3" frequency="${coord:days(1)}"
  start="2012-10-02T00:00Z" end="2012-10-04T24:00Z" ...>
  <datasets>
    <dataset name="logs" frequency="${coord:hours(1)}"
      initial-instance="2012-01-01T00:00Z" ...>
      <uri-template>
        hdfs://xxxx:9000/logs/${YEAR}/${MONTH}/${DAY}/${HOUR}
      </uri-template>
    </dataset>
  </datasets>
  <input-events>
    <data-in name="last_24hr_logs" dataset="logs">
      <start-instance>${coord:current(-23)}</start-instance>
      <end-instance>${coord:current(0)}</end-instance>
    </data-in>
  </input-events>
  <action>
    <workflow>
      <app-path>hdfs://xxxx:9000/myworkflow/demo1</app-path>
    </workflow>
  </action>
</coordinator-app>
```

this coordinator will be materialized:

@2012-10-02T00:00Z

@2012-10-03T00:00Z

@2012-10-04T00:00Z

@2012-10-02T00:00Z will wait for logs:

hdfs://xxxx:9000/logs/2012/10/01/01

hdfs://xxxx:9000/logs/2012/10/01/02

...

hdfs://xxxx:9000/logs/2012/10/01/23

hdfs://xxxx:9000/logs/2012/10/02/00

How To Use Oozie

1. **Deploy** your workflow on HDFS, this includes:

- oozie job definitions (workflow.xml)
- your codes: MR/pig/streaming/java etc.
- libraries (.so & .jar)

2. **Submit** your job

```
$ oozie job -run -config job.properties
```

```
Workflow ID: 0123-123456-oozie-wrkf-W
```


3. **Check** job status

```
$ oozie job -info 0123-123456-oozie-wrkf-W
```

```
$ oozie job -log 0123-123456-oozie-wrkf-W
```

(submit coordinator using the same way)

How To Use Oozie - Web Console

 [Apache Documentation](#) [Yahoo Documentation](#)

Oozie Web Console (v1) [/oozie/]

Workflow Jobs | **Coordinator Jobs** | Bundle Jobs | System Info | Instrumentation

All Jobs | Active Jobs | Done Jobs | Custom Filter ▾

user=newsddb;status=RUNNING

Custom Filter

Help

Job Id	User	Group	frequency	unit	Started	Next Materialization
1 0280329-120415013716421-oozie-	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
2 0277855-120415013716421-oozie-	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
3 0277851-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
4 0277850-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
5 0277849-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
6 0277848-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
7 0277847-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
8 0277846-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
9 0277845-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
10 0277844-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
11 0277843-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
12 0277842-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
13 0277841-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
14 0277840-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
15 0277839-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
16 0277838-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
17 0277837-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
18 0277835-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00
19 0277832-120415013716421-oozie-wr...	newsddb	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00


find my running jobs:
user=<user_name>;status=RUNNING

Page 1 of 4

1 - 50 of 152



How To Use Oozie - Web Console

 [Apache Documentation](#) [Yahoo Documentation](#)

Oozie Web Console (v1) [/oozie/]

Workflow Jobs **Coordinator Jobs** Bundle Jobs System Info Instrumentation

All Jobs Active Jobs Done Jobs Custom Filter ▼

Job Id	Group	frequency	unit	Started	Next Materialization
1 0277855-120415013716421-oozie-	users	1	DAY	Wed, 01 Aug 2012 00:00:00 G...	Sun, 30 Sep 2012 00:00:00 G...

id=0277855-120415013716421-oozie-wrkf-C
Custom Filter
Help

find a job by it's id:

id=01234-123456-oozie-wrkf-C

How To Use Oozie - Web Console

The screenshot displays the Oozie Web Console interface. It features three main panels:

- Coord Job Info:** Shows details for job 'newsdd_metric_web_search_NewZealand/coordJobId: 0277801-120415013716421-oozie-wrkf-C'. Fields include Job Id, Name, Status (RUNNING), User (newsddbe), Group (users), Frequency (1), Unit (DAY), Start Time (Wed, 01 Aug 2012 00:00:00 G), and Next Matd (Sun, 30 Sep 2012 00:00:00 G).
- Job Info:** Shows details for job 'newsdd_wf_metric_web/JobId: 0047351-120810053921497-oozie-wrkf-W'. Fields include Job Id, Name, App Path (hdfs://oxiumber-nn1.blue.ygrid.yah), Run (0), Status (SUCCEEDED), User (newsddbe), Group (users), Create Time, Nominal Time, Start Time, Last Modified, and End Time.
- Action Info:** Shows details for action 'web_param/JobId: 0047351-120810053921497-oozie-wrkf-W'. Fields include Name, Type (java), Transition (web_pig), Start Time, End Time, Status (OK), Error Code, Error Message, External ID (job_201208062139_437324), External Status (SUCCEEDED), Console URL (http://oxiumber-jt1.blue.ygrid.yahoo.com:50030/jobdetails.jsp?jobid=...), and Tracker URI (oxiumber-jt1.blue.ygrid.yahoo.com:50300).

At the bottom, there are two 'Actions' tables. The first table lists actions for the first job, with the second row highlighted. The second table lists actions for the second job, with the first row highlighted. A search icon in the top right of the Action Info panel is circled in red.

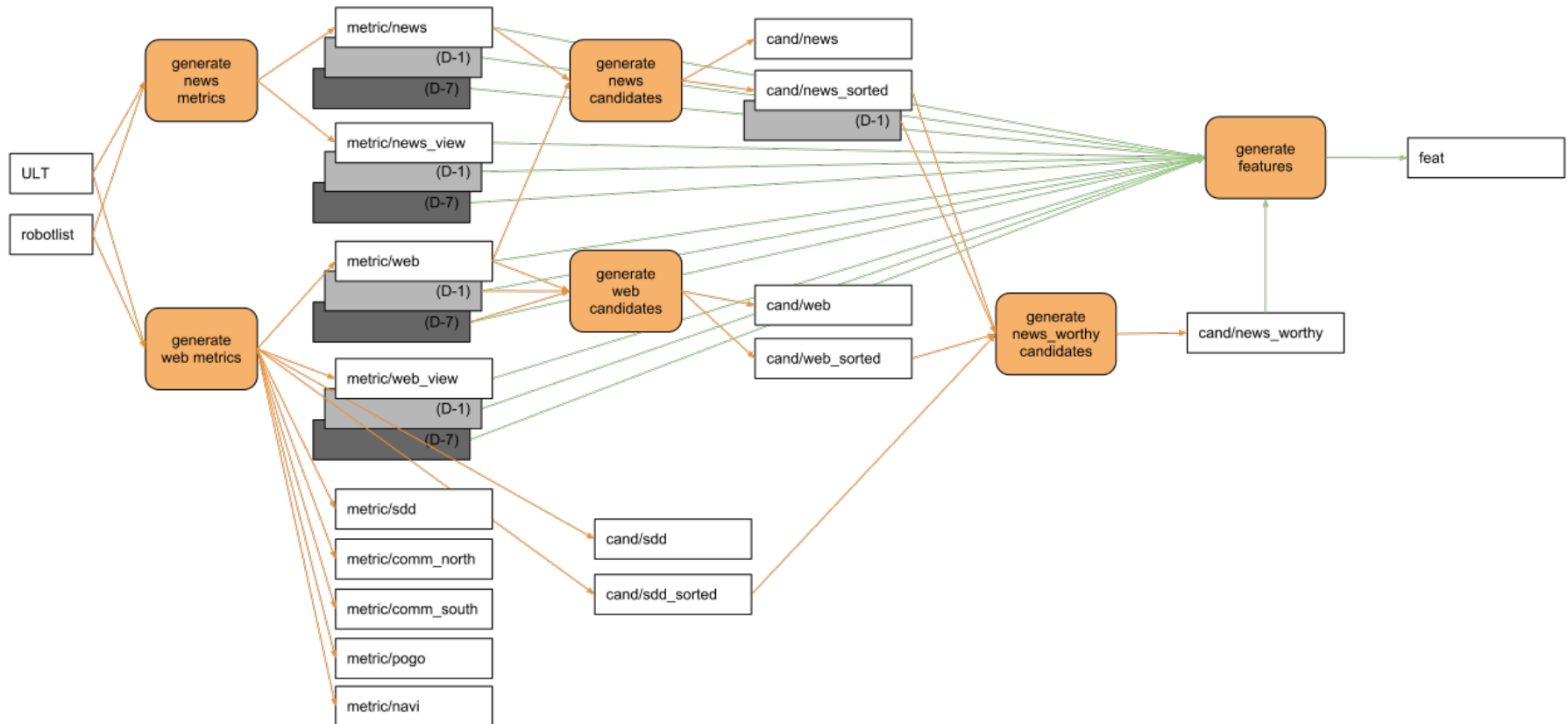
Action Id	Status
1 0277801-120415013716421-oozie-wrkf-C@25	SUCCEE
2 0277801-120415013716421-oozie-wrkf-C@26	SUCCEE
3 0277801-120415013716421-oozie-wrkf-C@27	SUCCEE
4 0277801-120415013716421-oozie-wrkf-C@28	SUCCEE
5 0277801-120415013716421-oozie-wrkf-C@29	SUCCEE
6 0277801-120415013716421-oozie-wrkf-C@30	SUCCEE
7 0277801-120415013716421-oozie-wrkf-C@31	SUCCEE
8 0277801-120415013716421-oozie-wrkf-C@32	SUCCEE
9 0277801-120415013716421-oozie-wrkf-C@33	SUCCEE
10 0277801-120415013716421-oozie-wrkf-C@34	SUCCEE

Action Id	Name	Type	Status	Transition	StartTime	EndTime
1 0047351-120810053921497-oozie-wrkf-W@...	web_param	java	OK	web_pig	Mon, 27 Aug 2012 08:43:37 G...	Mon, 27 Aug 2012 08:44:00 G...
2 0047351-120810053921497-oozie-wrkf-W@...	web_pig	pig	OK	end	Mon, 27 Aug 2012 08:44:02 G...	Mon, 27 Aug 2012 08:44:02 G...



Use Case Sharing

- Data mining for query terms that have news intent
- Complex data dependency



Use Case Sharing

- Was using crontab + python scripts
- After porting to oozie:
 - Reduce code size (4906 -> 1708 lines)
 - More smooth processing (1 week delay -> 3 days)
 - More stable

Q & A

<http://incubator.apache.org/oozie/>