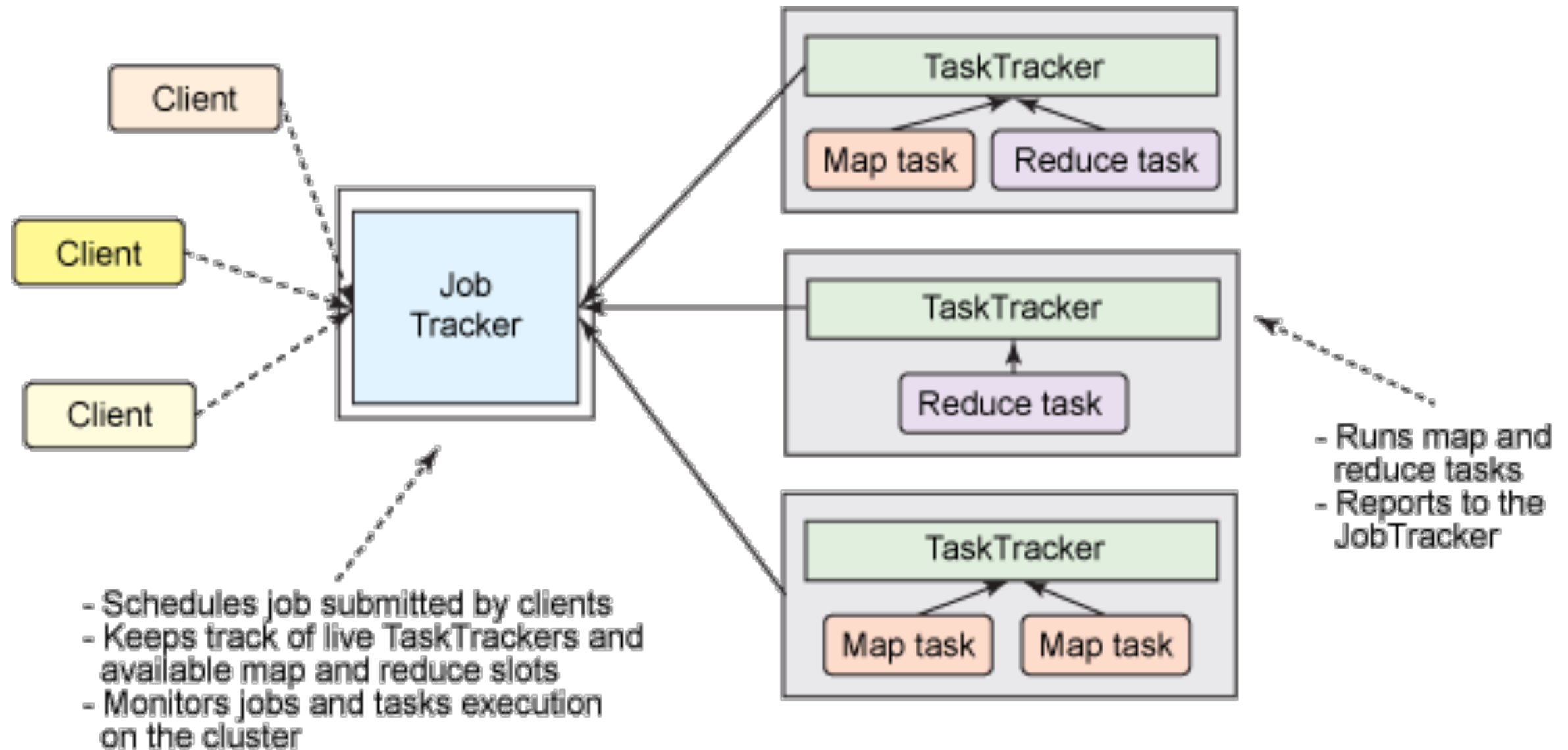


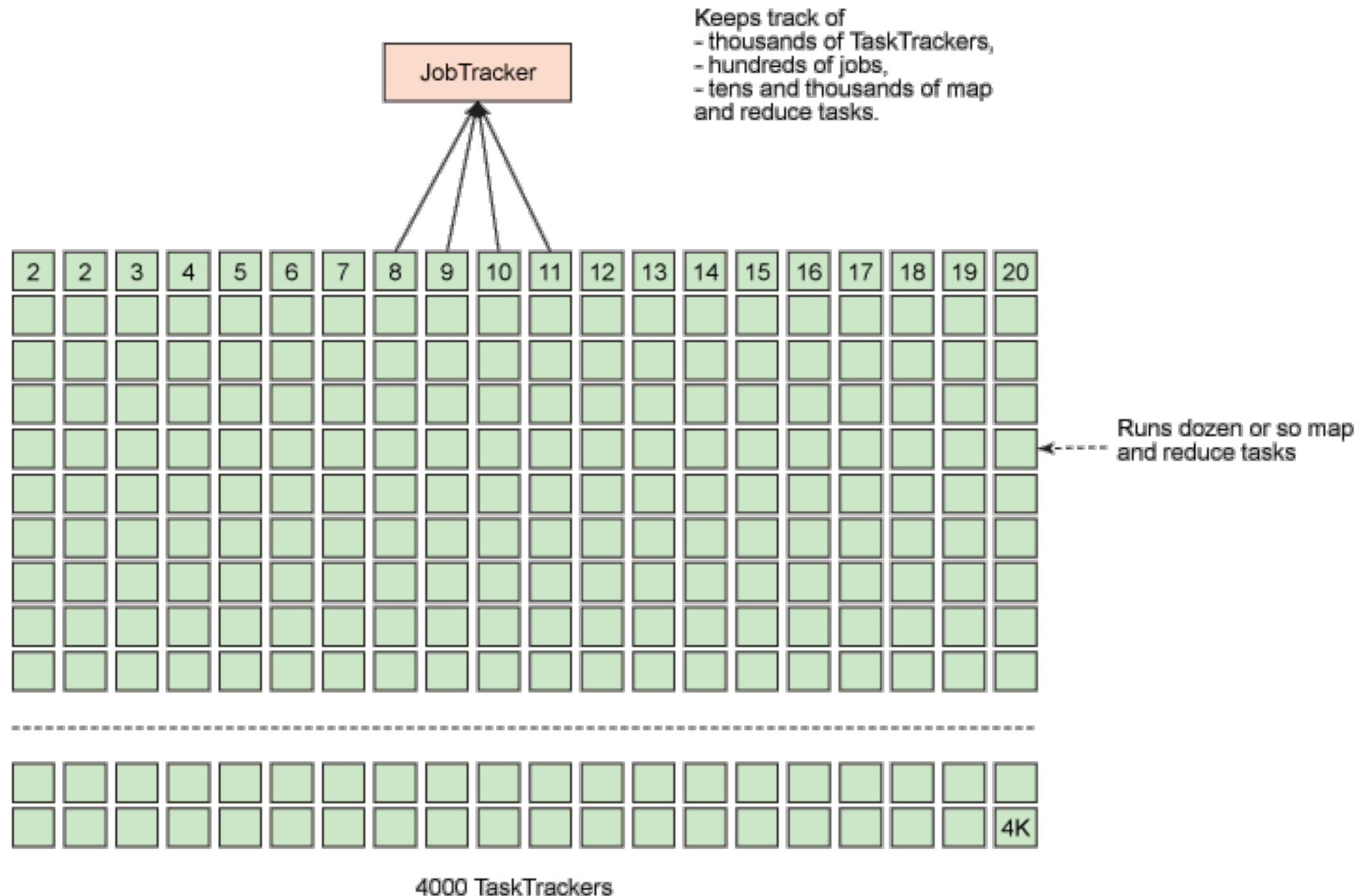
# YARN

Yet Another Resource Negotiator

# MR 1



# Job Tracker



# Job Tracker WebUI

**bda1node03 Hadoop Map/Reduce Administration**  
State: RUNNING  
Started: Mon Jan 23 18:28:30 PST 2012  
Version: 0.20.2-cdh3u2, 95a024e4005b2a94fe1c11f1ef9db4c672ba40cb  
Compiled: Thu Oct 10 21:51:41 PDT 2011 by root from Unknown  
Identifier: 201201231828

**Cluster Summary (Heap Size is 517.12 MB/910.25 MB)**

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Occupied Reduce Slots	Reserved Map Slots	Reserved Reduce Slots	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node
0	0	10	14	0	0	0	0	168	168	24.00

**Scheduling Information**

Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)   
Example: "user:smith 3200" will filter by "smith" only in the user field and "3200" in all fields

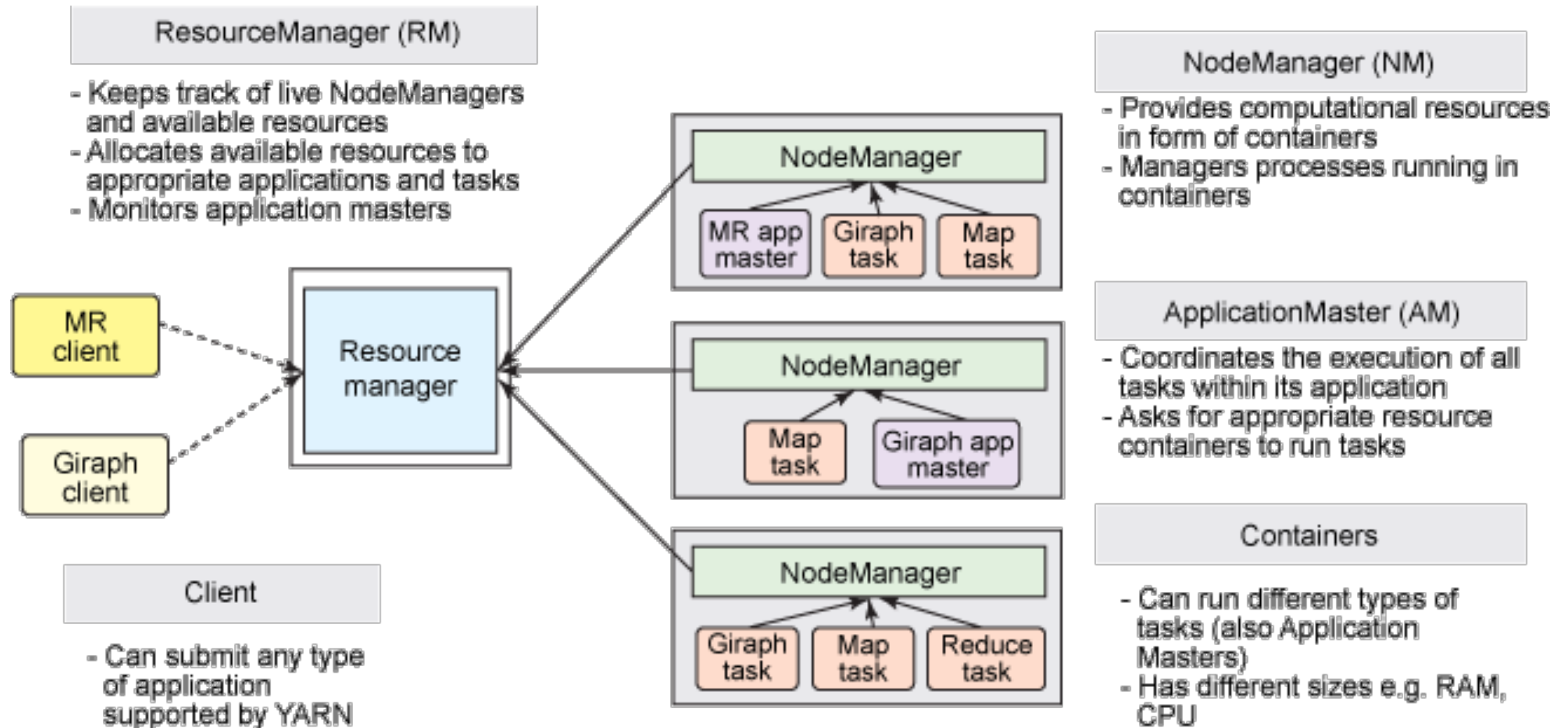
**Running Jobs**

none

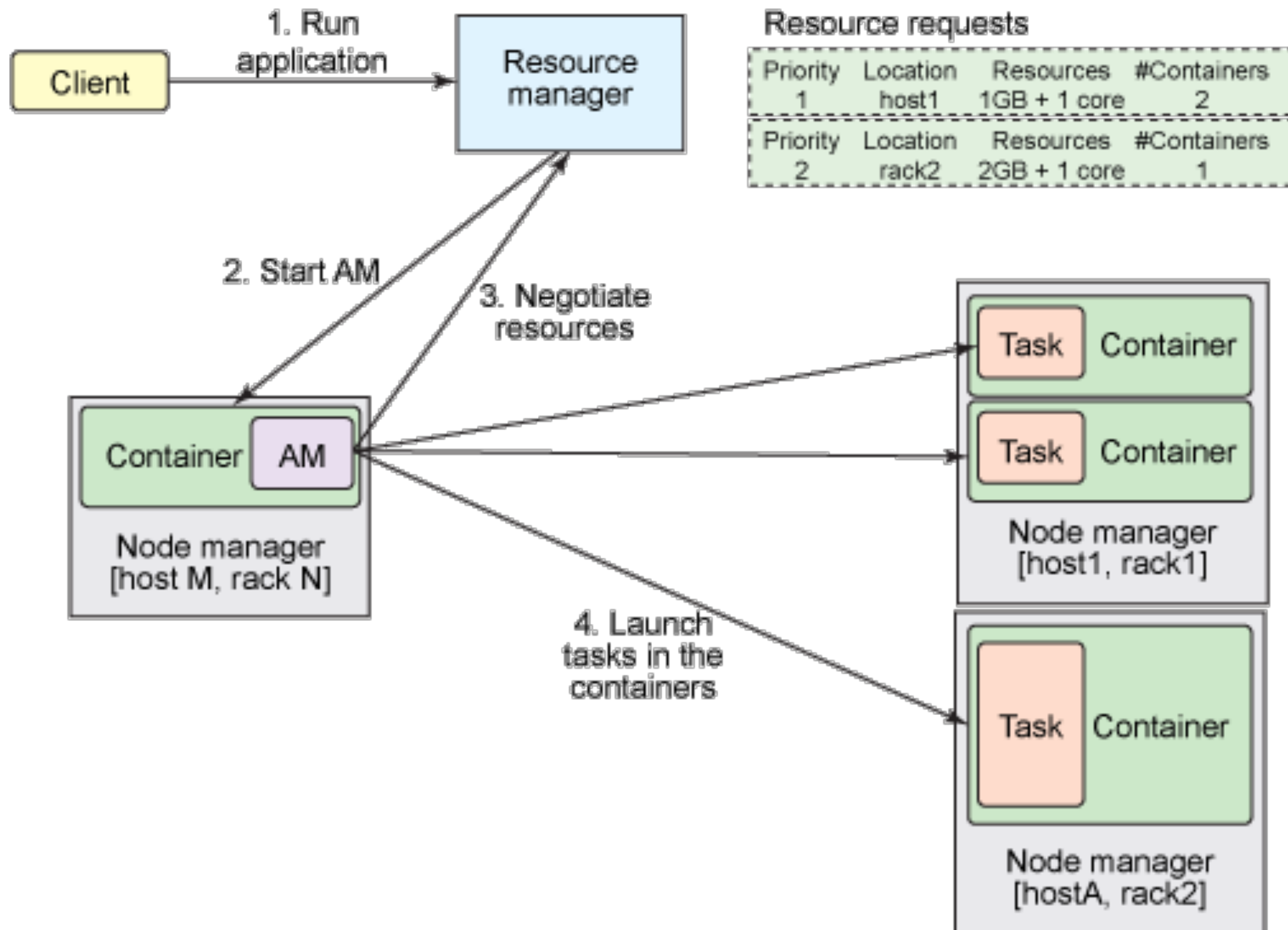
**Completed Jobs**

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total
<a href="#">job_201201231826_0006</a>	NORMAL	oracle	DataGenCSV	<div>100.00%</div>	128	128	<div>100.00%</div>	0
<a href="#">job_201201231826_0007</a>	NORMAL	oracle	OracleLoader_fivdti_i7_hash128_dtext_dl_500000000_128_tasks128	<div>100.00%</div>	384	384	<div>100.00%</div>	128
<a href="#">job_201201231826_0008</a>	NORMAL	oracle	OracleLoader_fivdti_i7_hash128_dtext_dl_500000000_128_tasks128	<div>100.00%</div>	3837	3837	<div>100.00%</div>	128
<a href="#">job_201201231826_0009</a>	NORMAL	oracle	OracleLoader_fivdti_i7_hash128_dtext_dl_500000000_128_tasks128	<div>100.00%</div>	3837	3837	<div>100.00%</div>	128
<a href="#">job_201201231826_0010</a>	NORMAL	oracle	OracleLoader_fivdti_i7_hash128_dtext_dl_500000000_128_tasks128	<div>100.00%</div>	3837	3837	<div>100.00%</div>	128

# YARN



# Submit to YARN



# YARN queues



Logged in as: dr.who

## NEW,NEW\_SAVING,SUBMITTED,ACCEPTED,RUNNING Applications

Cluster

- About
- Nodes
- Node Labels
- Applications
  - NEW
  - NEW\_SAVING
  - SUBMITTED
  - ACCEPTED
  - RUNNING
  - FINISHED
  - FAILED
  - KILLED
- Scheduler

Tools

### Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	2.20 GB	0 B	0	8	0	1	0	0	0	0

### Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:250, vCores:1>	<memory:2250, vCores:8>

Dump scheduler logs 1 min

### Application Queues

Legend: Capacity Used Used (over capacity) Max Capacity

Queue: root	0.0% used
Queue: Engineering	0.0% used
Queue: Engineering.Development	0.0% used
Queue: Engineering.QE	0.0% used
Queue: Marketing	0.0% used
Queue: Marketing.Advertising	0.0% used
Queue: Marketing.Sales	0.0% used
Queue: Support	0.0% used
Queue: Support.Services	0.0% used
Queue: Support.Training	0.0% used

Show 20 entries

Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCore	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
----	------	------	------------------	-------	----------------------	-----------	------------	-------	-------------	--------------------	---------------------	---------------------	------------	--------------	----------	-------------	-------------------

No data available in table

Showing 0 to 0 of 0 entries

First Previous Next Last

# Schedulers









- **Capacity**
  - **FIFO**
- **Fair**
  - **FAIR**







# YARN frameworks

- **Distributed-Shell**
- **Hadoop MapReduce**
- **Apache Tez**
- **Apache Giraph**
- **Apache Spark**
- **Apache Storm**

# Oozie

  Query Editors ▾ Data Browsers ▾ Workflows ▾ Search  File Browser  Job Browser  admin ▾   

 **Oozie Dashboard** Workflows Coordinators Bundles SLA Oozie

171530

SucceededRunningKilled

ManuallyCoordinator

## Running

☐

Submission

Status

Name

Progress

Submitter

Last Modified

Id

Parent

No data available

Showing 0 to 0 of 0 entries

← Previous

Next →

## Completed

 Completion

Status

Name

Duration

Submitter

Id

Parent

No data available

Showing 0 to 0 of 0 entries

← Previous

Next →

ACTIONS

My Workflow

Add a description...

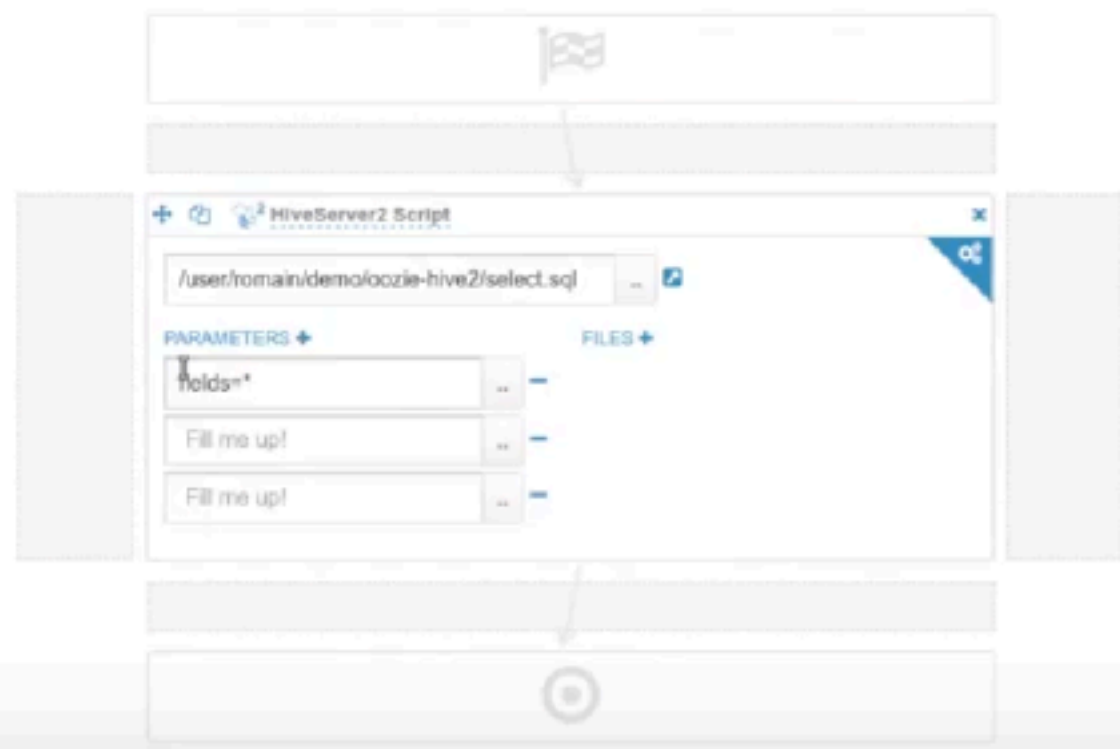
Drop your action here

ACTIONS



My Workflow

Add a description...



WORKFLOW

My Workflow

SUBMITTER

romain

STATUS

RUNNING

PROGRESS

33%

ID

0000008-150319151313008-oozie-oozi-W

VARIABLES

oozie.wf.application.path

MANAGE

Kill

Suspend

Back

Workflow My\_Workflow

Graph

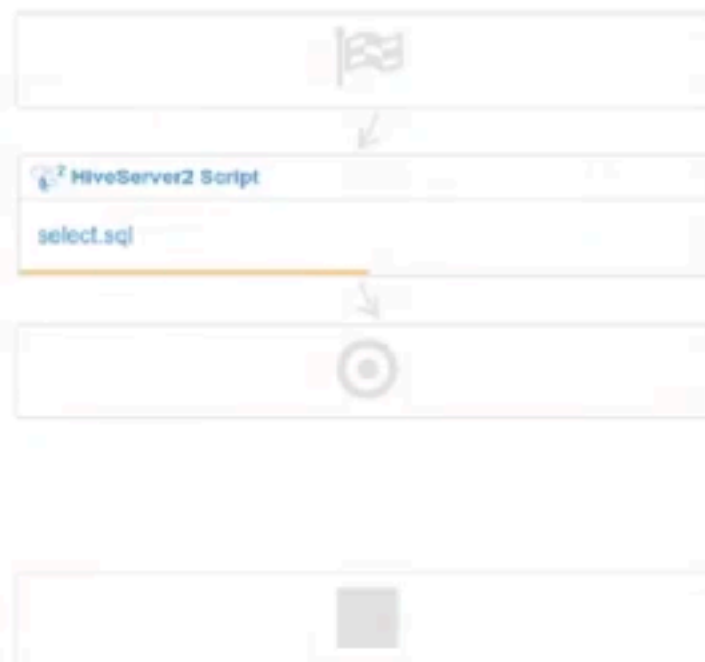
Actions

Details

Configuration

Log

Definition



WORKFLOW

My Workflow

SUBMITTER

romain

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000008-150319151313008-  
oozie-oozi-W

VARIABLES

oozie.wf.application.path

MANAGE

Rerun

Workflow My\_Workflow

Graph Actions Details Configuration Log Definition

```
1 <workflow-app name="My_Workflow" xmlns="uri:oozie:workflow:0.5">
2   <start to="hive2-099c"/>
3   <kill name="Kill">
4     <message>Action failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>
5   </kill>
6   <action name="hive2-099c">
7     <hive2 xmlns="uri:oozie:hive2-action:0.1">
8       <job-tracker>${jobTracker}</job-tracker>
9       <name-node>${nameNode}</name-node>
10      <jdbc-url>jdbc:hive2://localhost:10000/default</jdbc-url>
11      <script>/user/romain/demo/oozie-hive2/select.sql</script>
12      <param>fields=*</param>
13      <param>tablename=sample_07</param>
14      <param>n=10</param>
15    </hive2>
16    <ok to="End"/>
17    <error to="Kill"/>
18  </action>
19  <end name="End"/>
20 </workflow-app>
21
```



WORKFLOW

My Workflow

SUBMITTER

omain

STATUS

RUNNING

PROGRESS

33%

ID

0000010-150319151313000-oozie-oozi-W

VARIABLES

oozie.wf.application.path

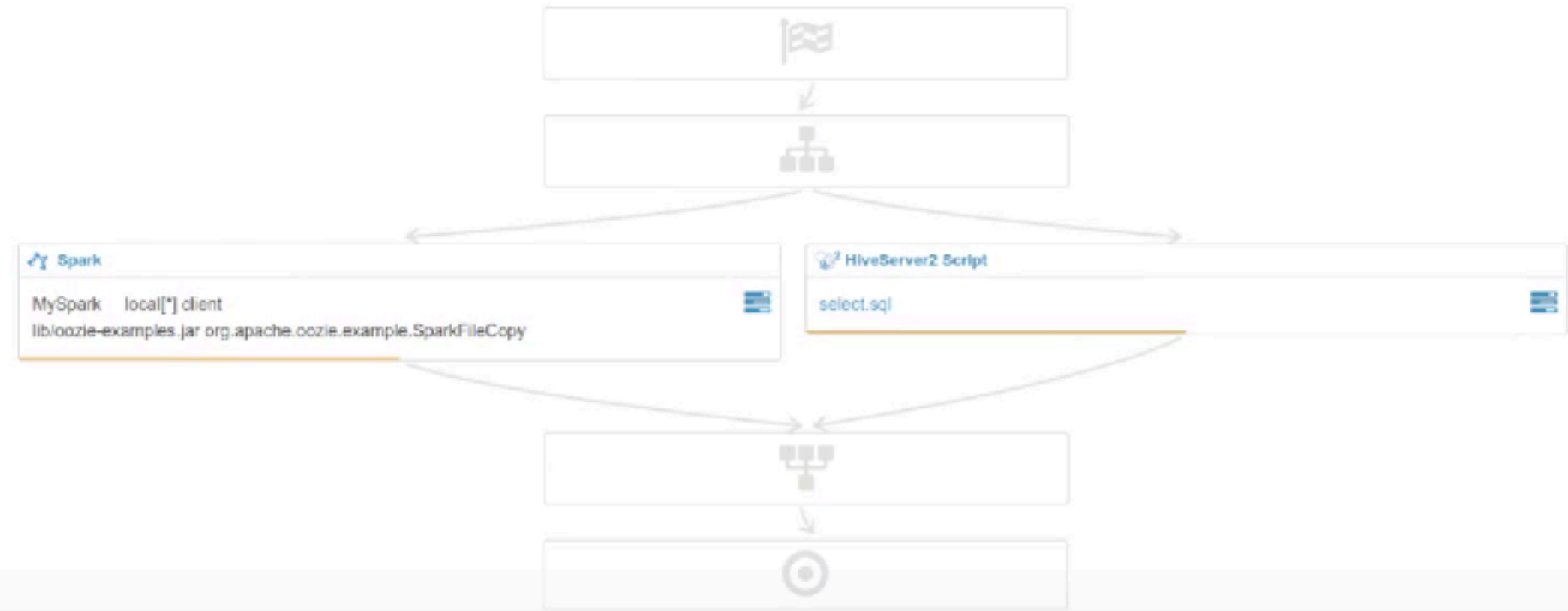
output

MANAGE

Kill Suspend

Workflow My\_Workflow

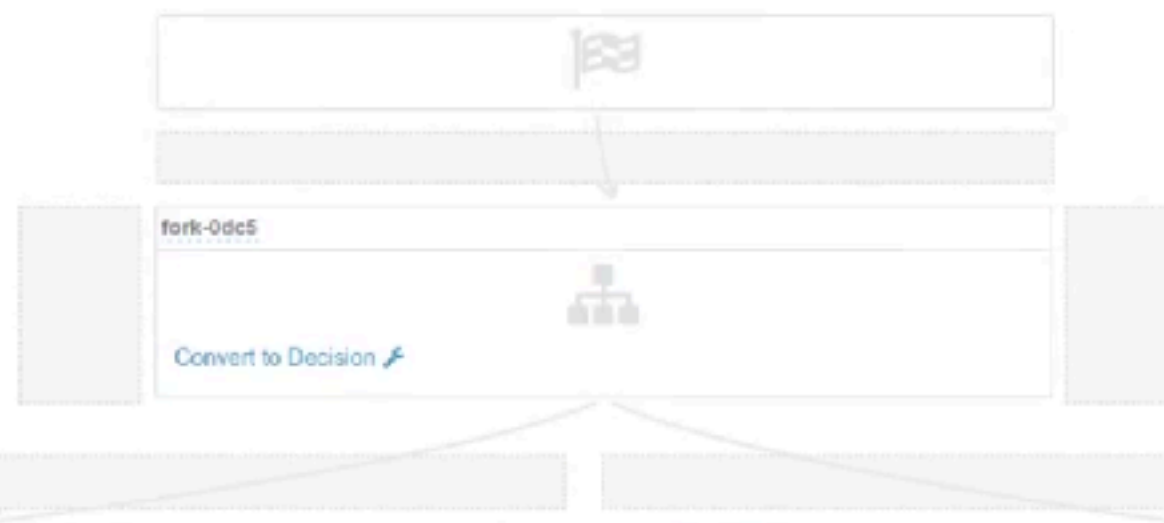
Graph Actions Details Configuration Log Definition





## Fork Example

Run a Pig script and a Sub-workflow in parallel



**Pig Script**

dump.pig

PARAMETERS +

input=/user/hue/oozie/workspaces/	..
limit=10	..

FILES +

**Sub workflow**

Spark

PROPERTIES +

input	/user/hue/oozie/workspaces/data/sonnets.txt	..
output	\${output}	..

# Luigi task status

Luigi Task Status

Task ListDependency GraphWorkersResources

TASK MAP LISTS

13 ActiveDataSet

13 ActiveUsersDataSet

1 BSSIDEventHadoop

2 BSSIDHadoopTarget

2 BSSIDMergedProfilesHadoop

1 BSSIDProfilesHadoop

1 BSSIDSitesProfilesHadoop

1 ClickStreamLogs2

1 ClickStreamSessions

1 ClickStreamTopProfiles

1 ConvertActiveDataToTrainingFormat

10 ConvertDataToTrainingFormat

62 DataSet

31 DemographicFeaturesDataSet

1 DeviceIDPairs

1 DistributedGdApply

1 DistributedGdMr

⏸

PENDING TASKS

72

▶

RUNNING TASKS

19

▶

BATCH RUNNING TASKS

0

✓

DONE TASKS

744

✗

FAILED TASKS

0

⚠

UPSTREAM FAILURE

0

⊖

DISABLED TASKS

0

⚠

UPSTREAM DISABLED

0

Displaying PENDING tasks.

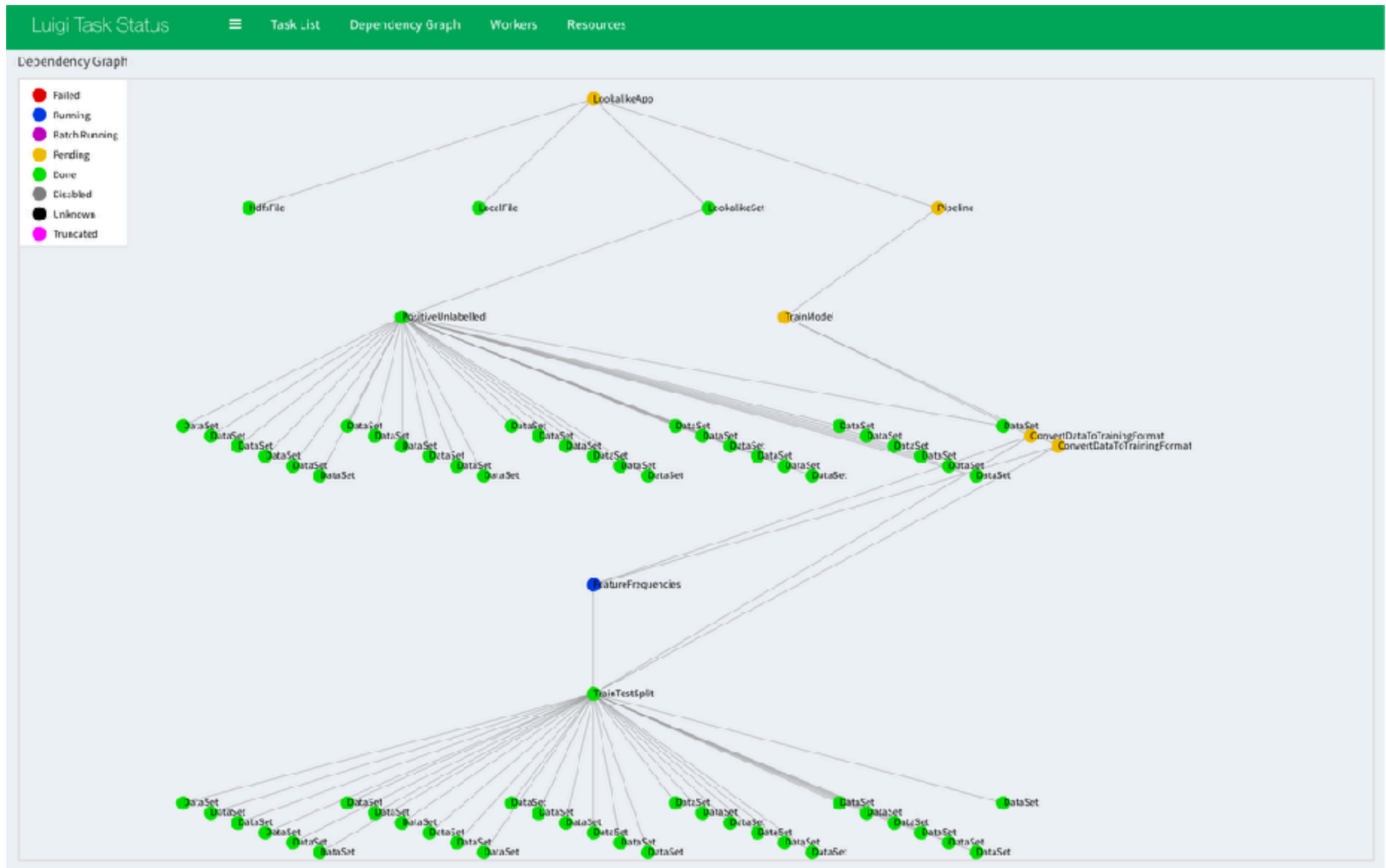
Show 10 entries

Filter table:

Filter on Server

Name	Details	Priority	Time	Actions
<div><div>⏸</div><div>PENDING</div></div> <div>Local k8sApp</div>	<div>prediction_subdir=prediction, local_k8s_config={"data_set_types":["demo"],"min_users_per_feature":1000,"min_features_per_user":3}, input_data_hdfs_subdir=input, input_data_subdir=input, app_name=LocalTest5, mining_hdfs_apps_dir=data/dm/apps, data_sets_hdfs_subdir=datasets, users_location=quality_hdfs_subdir=quality, date=2017-03-27, unlabelled_category=2, mining_apps_dir=data/dm/apps, quality_subdir=quality, data_sets_subdir=datasets, training_hdfs_subdir=train, prediction_config_subdir=prediction_config, prediction_config_hdfs_subdir=prediction_config, training_subdir=training, model_config={"data_set_types":["train"],"min_users_per_feature":400,"min_features_per_user":10,"grid":[{"model_type":"sklearn_binomial_multinomial_nb","params":{"inner_model__preprocessor__threshold":(0,1,3),"inner_model__model__alpha":(0.1,1),"inner_model__model__fit_prior":(true,false)}}}, prediction_hdfs_subdir=prediction</div>	0	Invalid Date	<div>⚙</div>
<div><div>⏸</div><div>PENDING</div></div> <div>Local k8sApp</div>	<div>prediction_subdir=prediction, local_k8s_config={"data_set_types":["demo"],"min_users_per_feature":1000,"min_features_per_user":3}, input_data_hdfs_subdir=input, input_data_subdir=input,</div>	0	Invalid Date	<div>⚙</div>

# Luigi task graph



# ОСНОВНЫЕ МЕТОДЫ ДЛЯ MR

- `def requires(self):`
  - определяет зависимости
- `def output(self):`
  - определяет результат
- `def jar(self):`
  - определяет путь до jar файла
- `def main(self):`
  - определяет класс с `main` методом
- `def args(self):`
  - определяет аргументы при вызове

# Дополнительные методы

