Oozie

Complex work flow management

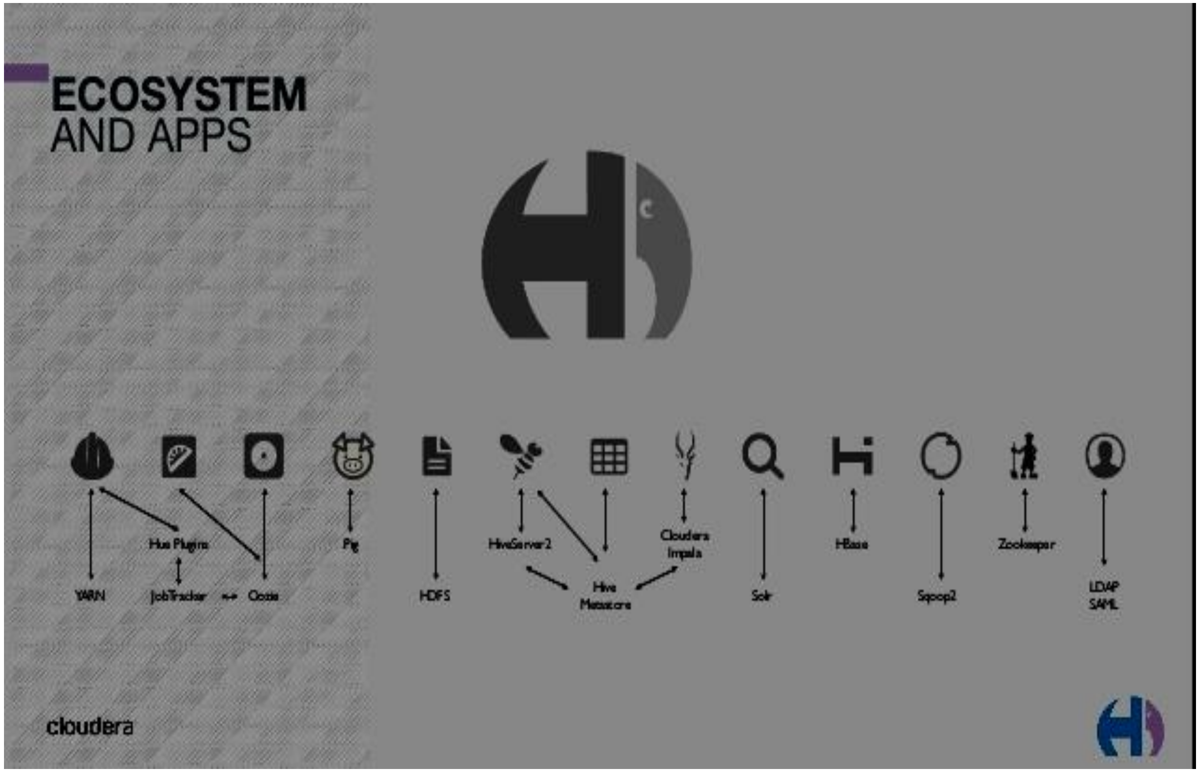
Relief to developers!!

Agenda

- Intro to Hue:
  - What is Hue
  - When to Use it
- Oozie
  - What is Oozie
  - Life with out it
  - When to use it
- Oozie Workflow
  - Actions
  - Control flow
- Actions
  - Sqoop Action
  - Hive Action
  - HDFS Action
- Workflows:
  - Example1
  - Example2
- Scheduling Workflows – Coordinators
  - Bundles

What is HUE

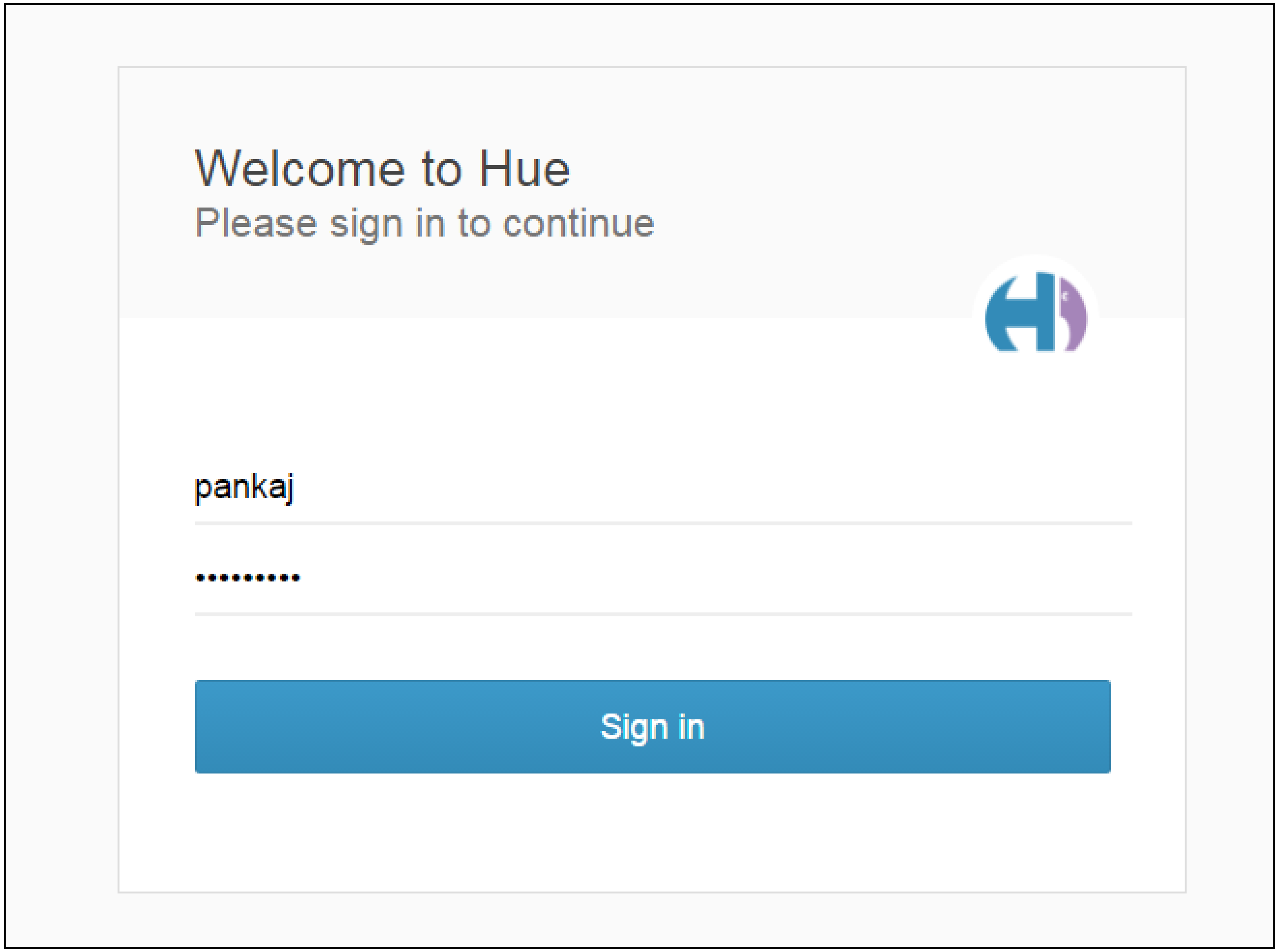
- Cloudera designed it, initially it was a commercial tool, later made it Apache Open Source.
- Single tool, which provides multiple options to developers:
- Hive Editor
- Pig Editor
- Hive Metastore Manager
- Impala
- DB Query
- Oozie
- File Browser – HDFS
- Job Browser – Resource Manager
- One stop for developers



When to use it

- If you are looking for the following:
- Editor to develop programs in Hive/ Pig/ Impala
- HDFS browser similar to Windows File Browser
- Track progress of :
  - Hive jobs
  - Pig Jobs
  - Spark Jobs
  - Map Reduce etc..
- Hive Metastore Manager
- Better Access to Databases and Tables on Hive
- Download results of Hive queries
- Visualize results of Hive Queries

HUE Login

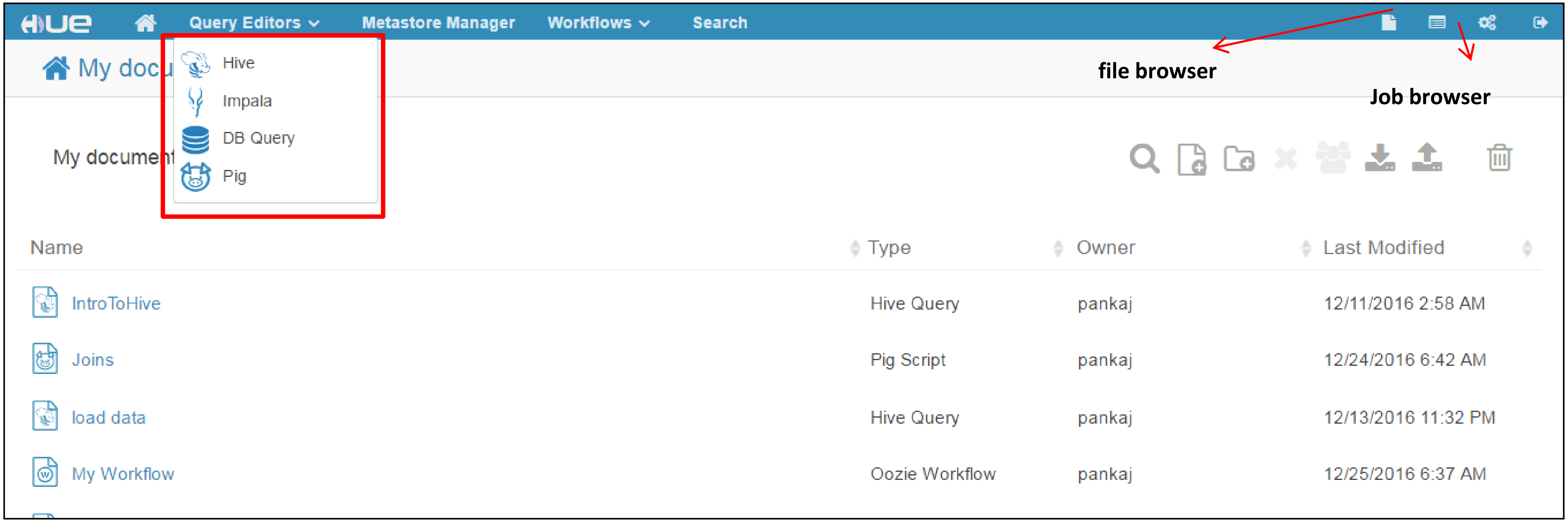


Welcome to Hue
Please sign in to continue

pankaj

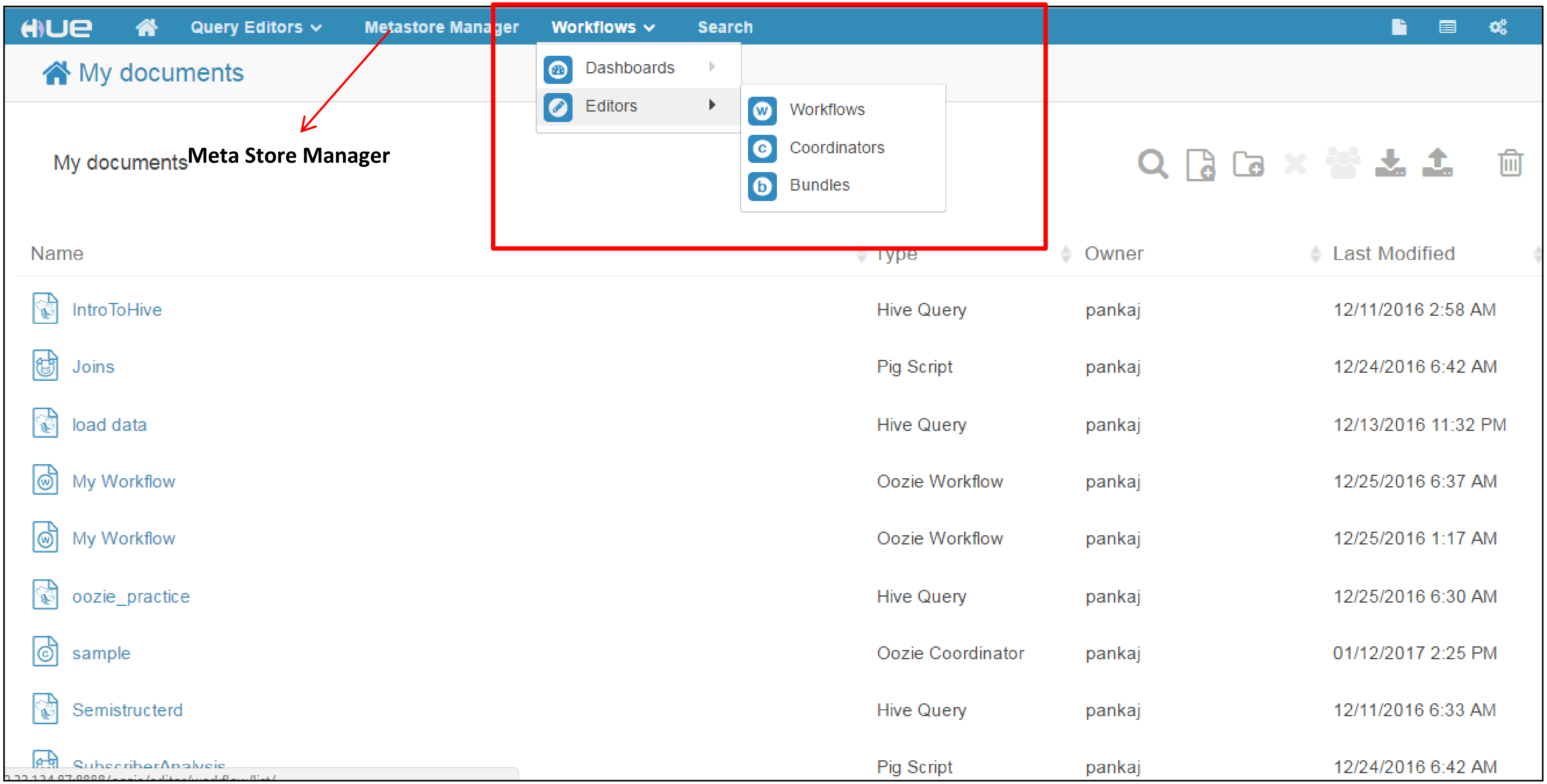
••••••••••

Query Editors, Job and File Browsers



Name	Type	Owner	Last Modified
IntroToHive	Hive Query	pankaj	12/11/2016 2:58 AM
Joins	Pig Script	pankaj	12/24/2016 6:42 AM
load data	Hive Query	pankaj	12/13/2016 11:32 PM
My Workflow	Oozie Workflow	pankaj	12/25/2016 6:37 AM

Oozie Workflow

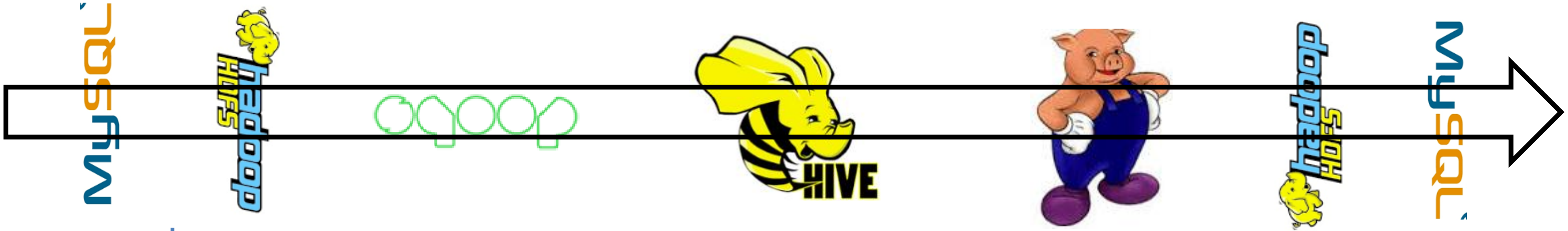


Meta Store Manager

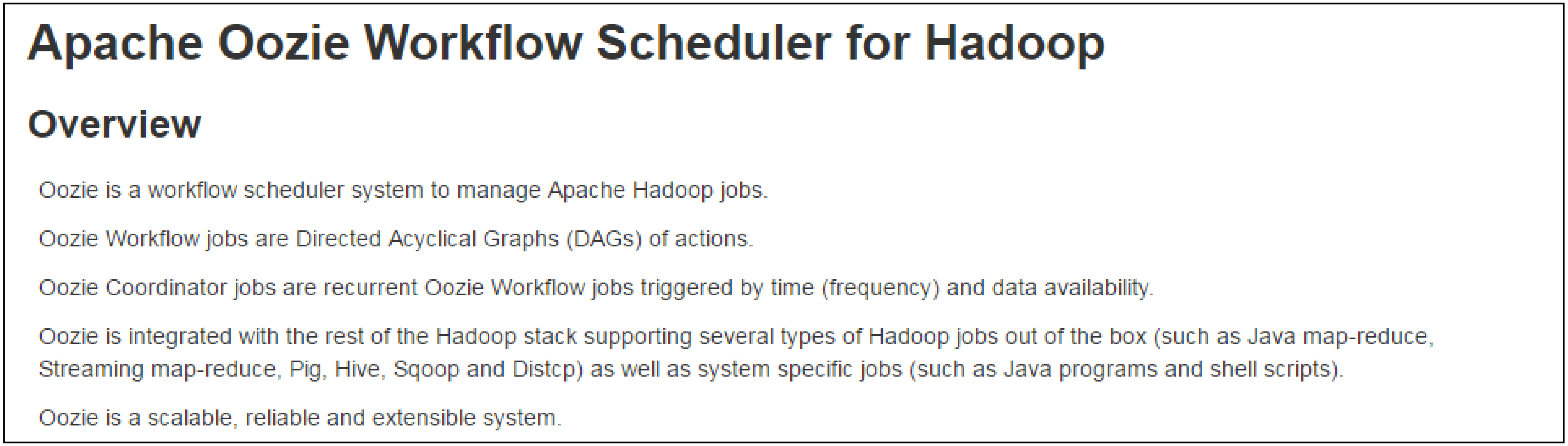
Name	Type	Owner	Last Modified
IntroToHive	Hive Query	pankaj	12/11/2016 2:58 AM
Joins	Pig Script	pankaj	12/24/2016 6:42 AM
load data	Hive Query	pankaj	12/13/2016 11:32 PM
My Workflow	Oozie Workflow	pankaj	12/25/2016 6:37 AM
My Workflow	Oozie Workflow	pankaj	12/25/2016 1:17 AM
oozie_practice	Hive Query	pankaj	12/25/2016 6:30 AM
sample	Oozie Coordinator	pankaj	01/12/2017 2:25 PM
Semistructerd	Hive Query	pankaj	12/11/2016 6:33 AM
SubscriberAnalysis	Pig Script	pankaj	12/24/2016 6:42 AM

Oozie

- Oozie as a tool is known for designing workflows, schedule and track them.
- In many industries, there are technology workflows, where output from one tool is consumed by another tool.
- Consider a example where you want to do the following:
- Clean data in HDFS
- Sqoop Import
- Hive / Pig scripts
- Sqoop export
- Drop a mail up on completion
- Answer to this example is Oozie. If you carefully observe, there is direction for data flow, and these flows are called as DAG(Directed Acyclic Graphs)
- The flow of the graph can be controlled using some control nodes, which helps in setting the start and end of the flow and some decision making nodes based on some intermediate predicate based values



In Simple Words



Apache Oozie Workflow Scheduler for Hadoop

Overview

Oozie is a workflow scheduler system to manage Apache Hadoop jobs.

Oozie Workflow jobs are Directed Acyclical Graphs (DAGs) of actions.

Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availability.

Oozie is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp) as well as system specific jobs (such as Java programs and shell scripts).

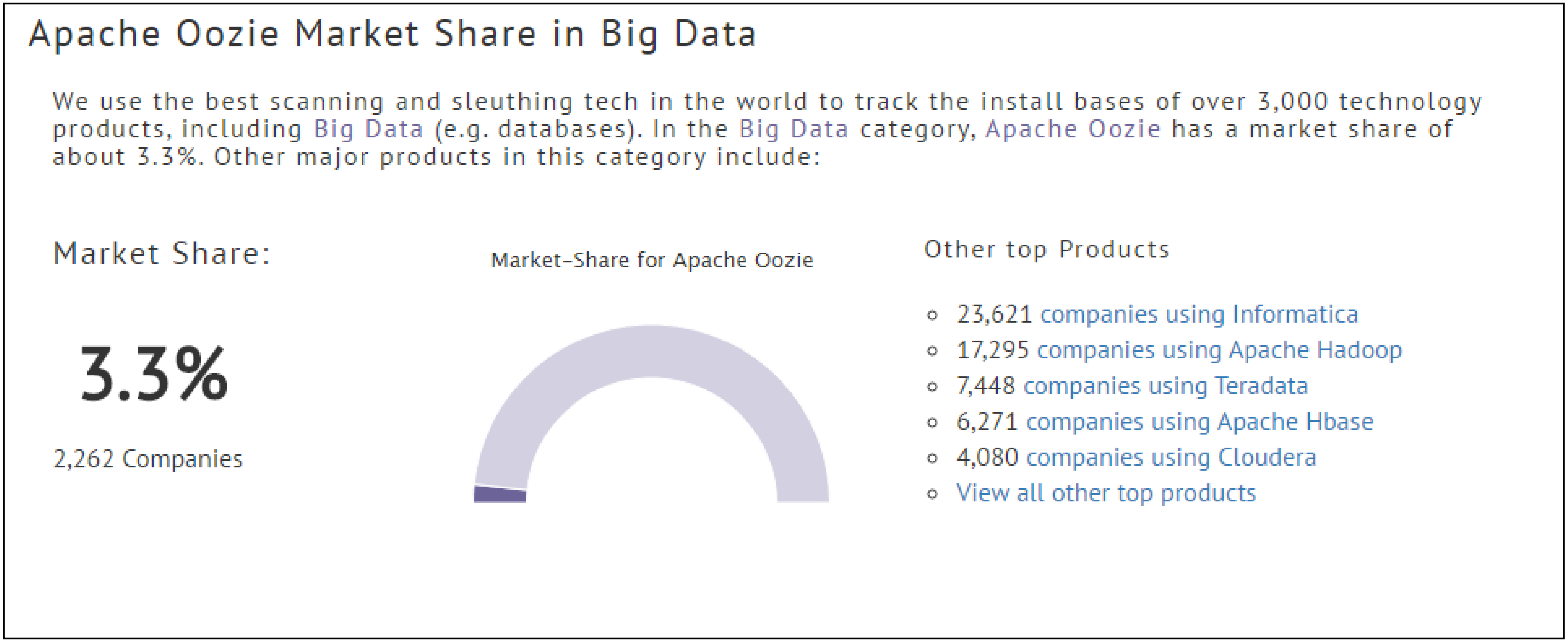
Oozie is a scalable, reliable and extensible system.

Life with out oozie

- Imagine, you have to do the following manually:
  - Connect hadoop components
  - If something goes wrong, drop a mail to stake holders
  - Up on completion, send a detailed report
  - Launch parallel jobs daily, weekly, monthly..

- All the mentioned points are hard to manage, schedule, coordinate, track the progress.
- Oozie can do this for you!

The best open source software for scheduling jobs on hadoop



Apache Oozie Market Share in Big Data

We use the best scanning and sleuthing tech in the world to track the install bases of over 3,000 technology products, including Big Data (e.g. databases). In the Big Data category, Apache Oozie has a market share of about 3.3%. Other major products in this category include:

Market Share:

3.3%

2,262 Companies

Market-Share for Apache Oozie

Other top Products

- 23,621 companies using Informatica
- 17,295 companies using Apache Hadoop
- 7,448 companies using Teradata
- 6,271 companies using Apache Hbase
- 4,080 companies using Cloudera
- View all other top products

Reference: https://idatalabs.com/tech/products/apache-oozie

Other tools similar to Oozie for Hadoop/ Big Data

talend

Commercial tool, reliable fast and easy to use

informatica
Data powers business™

Not a good tool for Big Data, too costly!

syncsort

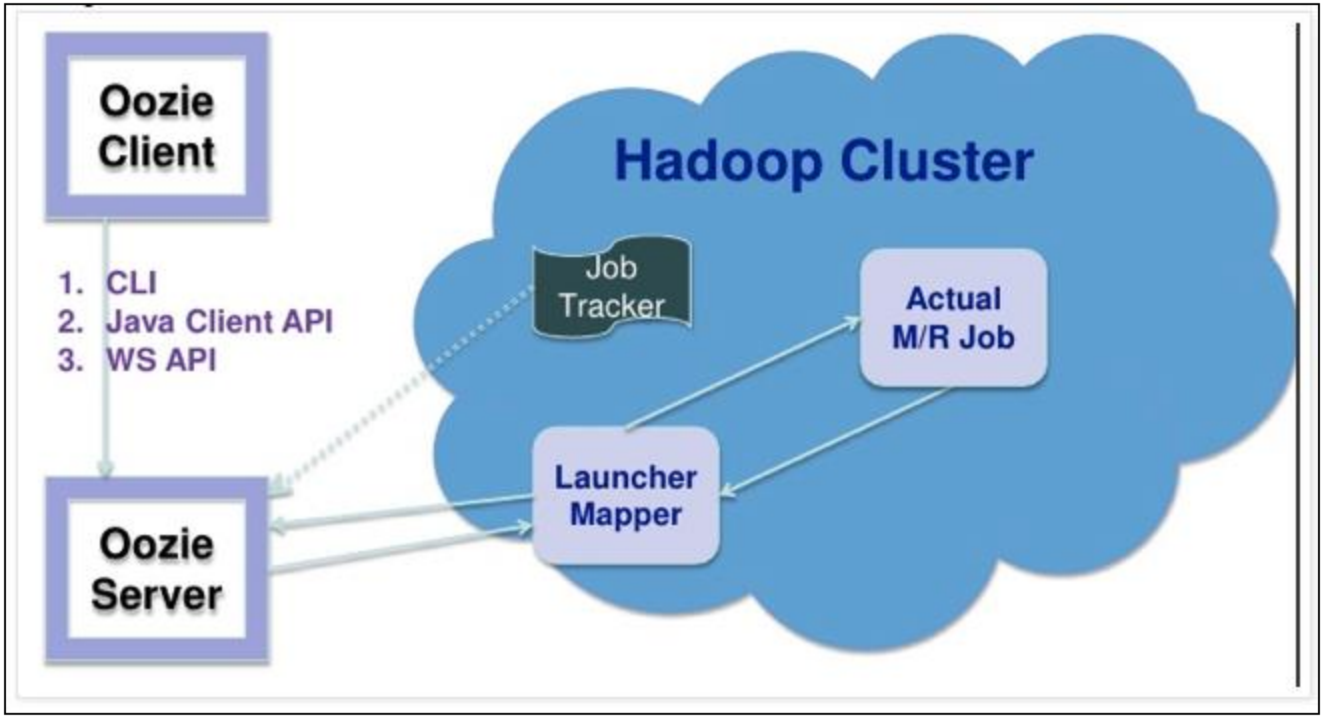
Good Competitor to talend

Azkaban, open source workflow manager, similar to Oozie

OOZIE WORKFLOW

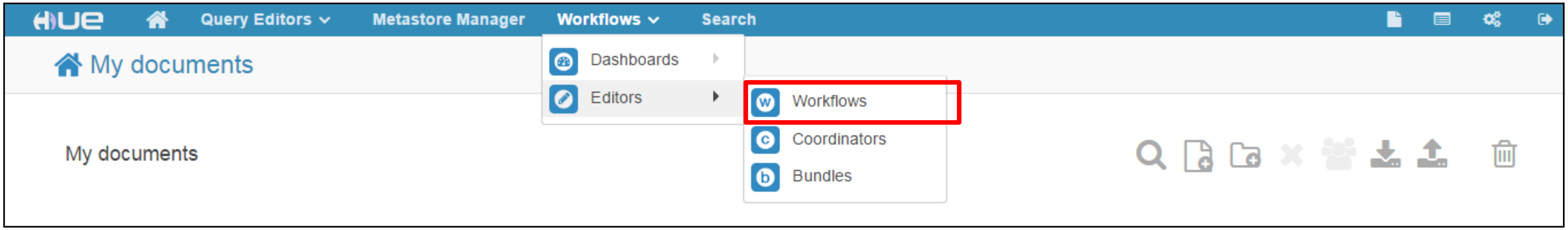
- Oozie workflows contain control flow nodes and action nodes.
- Control flow node define the beginning and the end of the workflow (start, end and fail nodes)
- They also provide a way to control the workflow execution path (decision, fork and join)
- Action nodes can trigger the execution of a computation/processing task
- These tasks include – map-reduce, HDFS commands, Pig, SSH, HTTP, eMail and Oozie sub-workflow

Oozie Architecture



Reference: yahoo

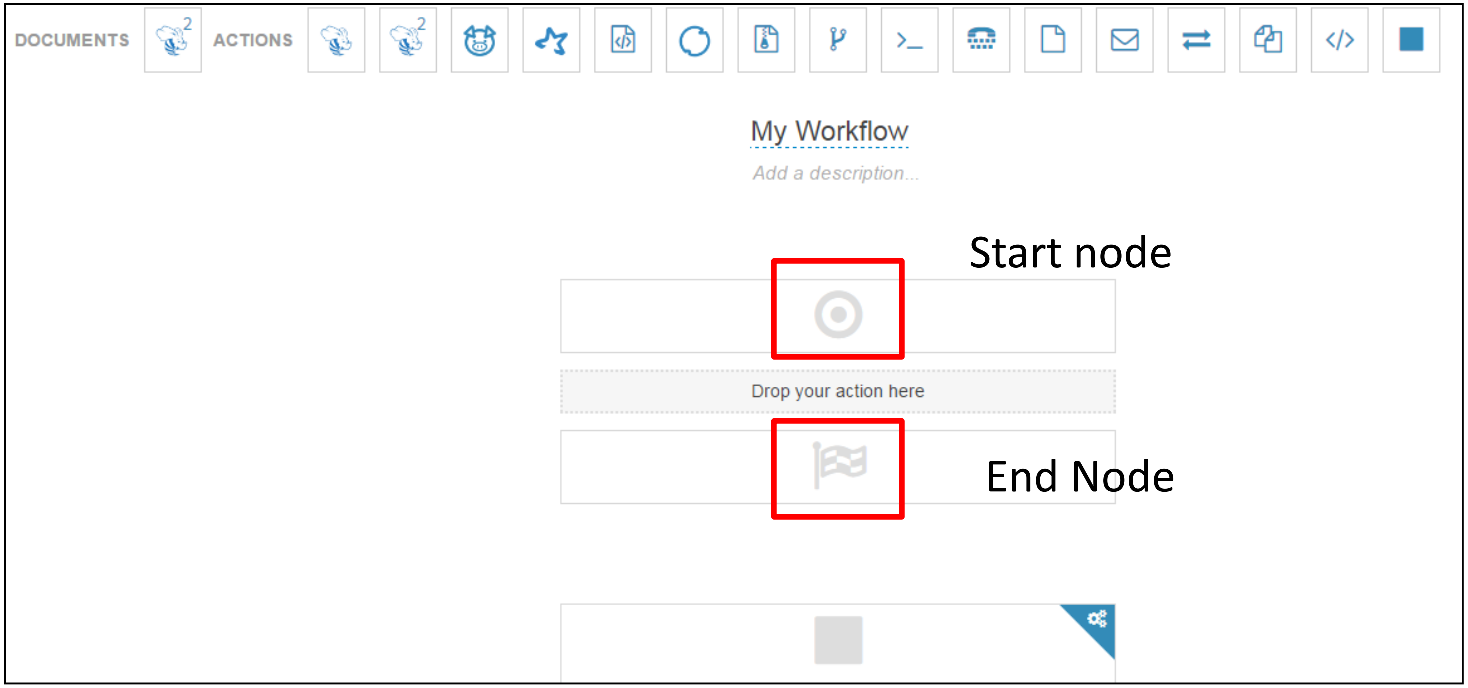
Oozie on Hue



- Under workflow, you see two options:
  - Dashboards for monitoring jobs.
  - Editors for designing, scheduling and bundling jobs.
- Click on workflows to design workflows.
- Click on coordinators to schedule workflows.
- Bundle to batch a set of coordinator applications.

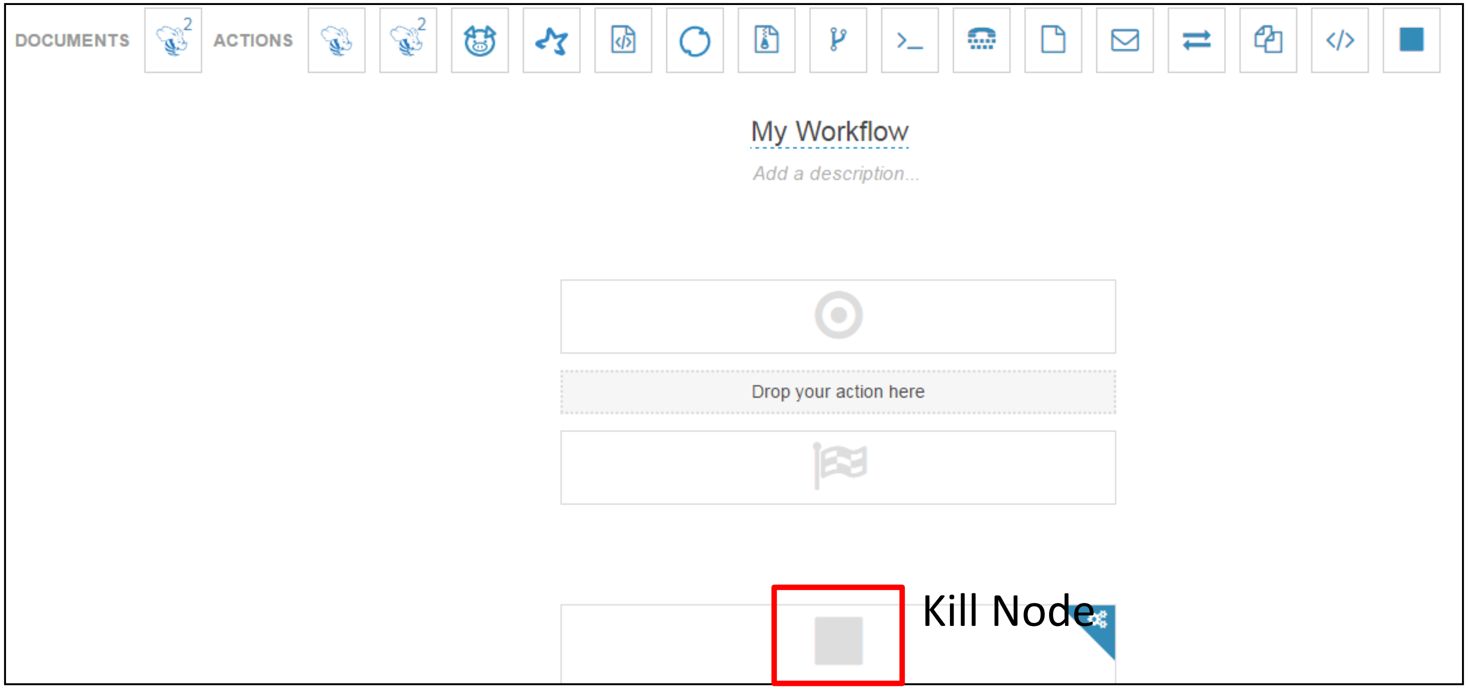
Control Flow Nodes

- **Start node**
  It indicates the first workflow node, through which the workflow job will start. This is the starting point for the workflow job. The “to” attribute points to the node where the job starts
- **End node**
  It is the end of the workflow job. When a workflow job reaches its end, it has completed successfully. Even if some other workflow jobs are in running state, when an end node is reached, these are killed forcefully and the program still exits successfully



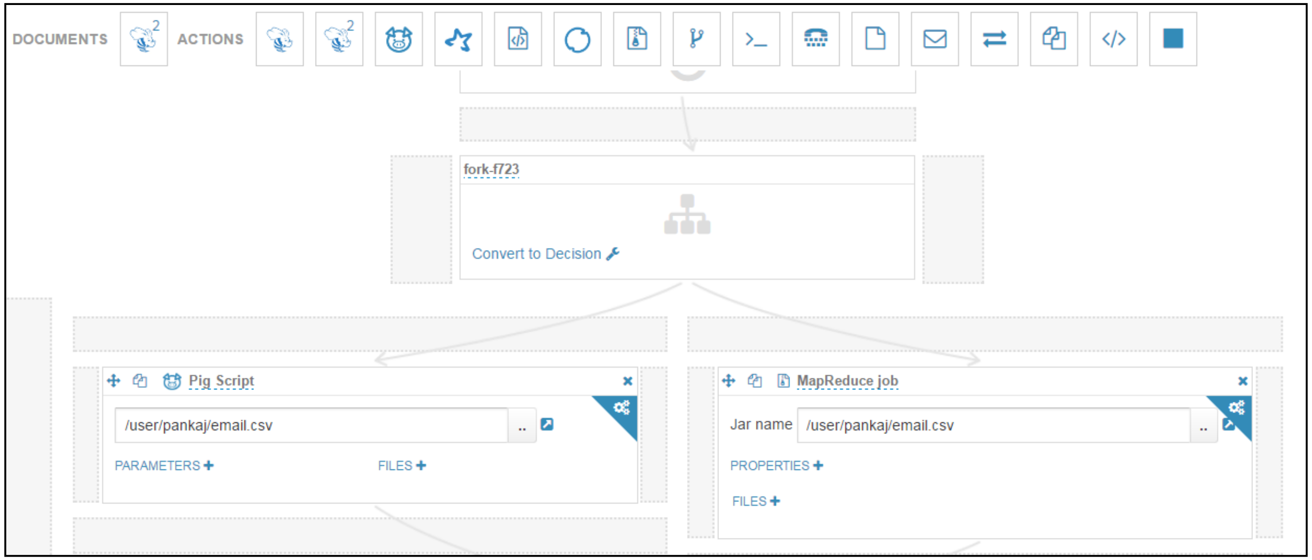
Control Flow Nodes

- **Kill node**
  Kill node allows a workflow job to kill itself. All the running actions of the workflow job would also be killed and a message as mentioned in the tags will be entered in the log file



Control Flow nodes

- **Fork and join control nodes**
  Fork node is used to split a path into multiple concurrent nodes. It allows tasks to be run in parallel. Join nodes then waits for every concurrent execution paths to reach to it. Fork and Join should be used in pairs
- **Decision node**
  Decision node allows a workflow job to make a selection between its execution paths based on a list of predicate. The predicates are evaluated on order of appearance until one of them comes true. In case all return false, default transition is taken. These predicate can contain logics such as the size of file being greater than some threshold, or the file being completely loaded or the exit status of an action node



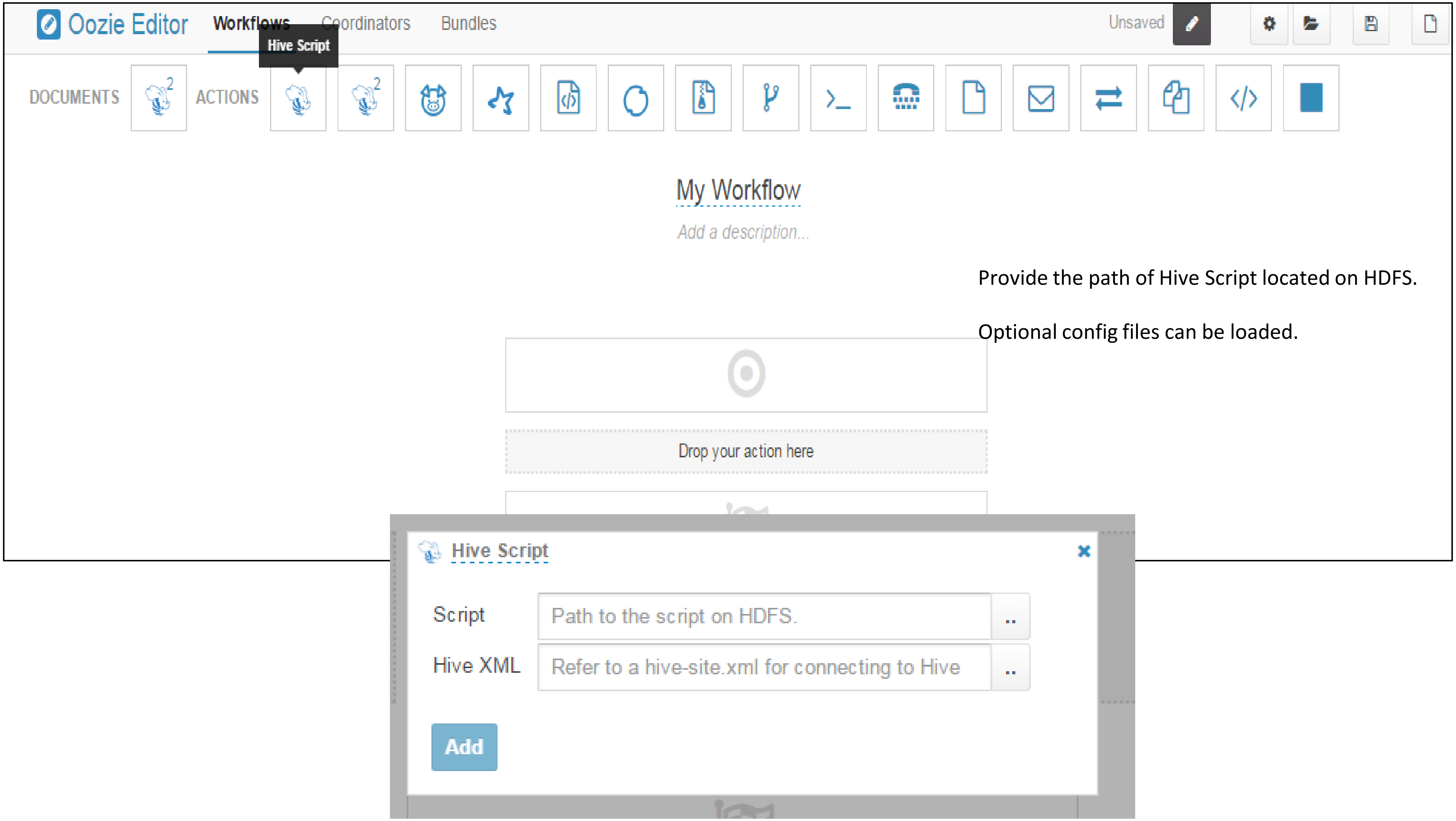
/user/pankaj/email.csv

Jar name /user/pankaj/email.csv

Oozie Actions

- Action nodes are generally used for performing computation tasks
- No Computation performed by Action nodes takes within oozie, all compution are performed remotely
- All computations performed by action node are asynchronous, but for most of the computations the workflow jobs waits for the action task to complete by polling and callbacks
- Ok attribute provides the path to follow on successful completion
- Error attribute provides the path to follow in case of error

Hive Action



My Workflow

Add a description...

Hive Script

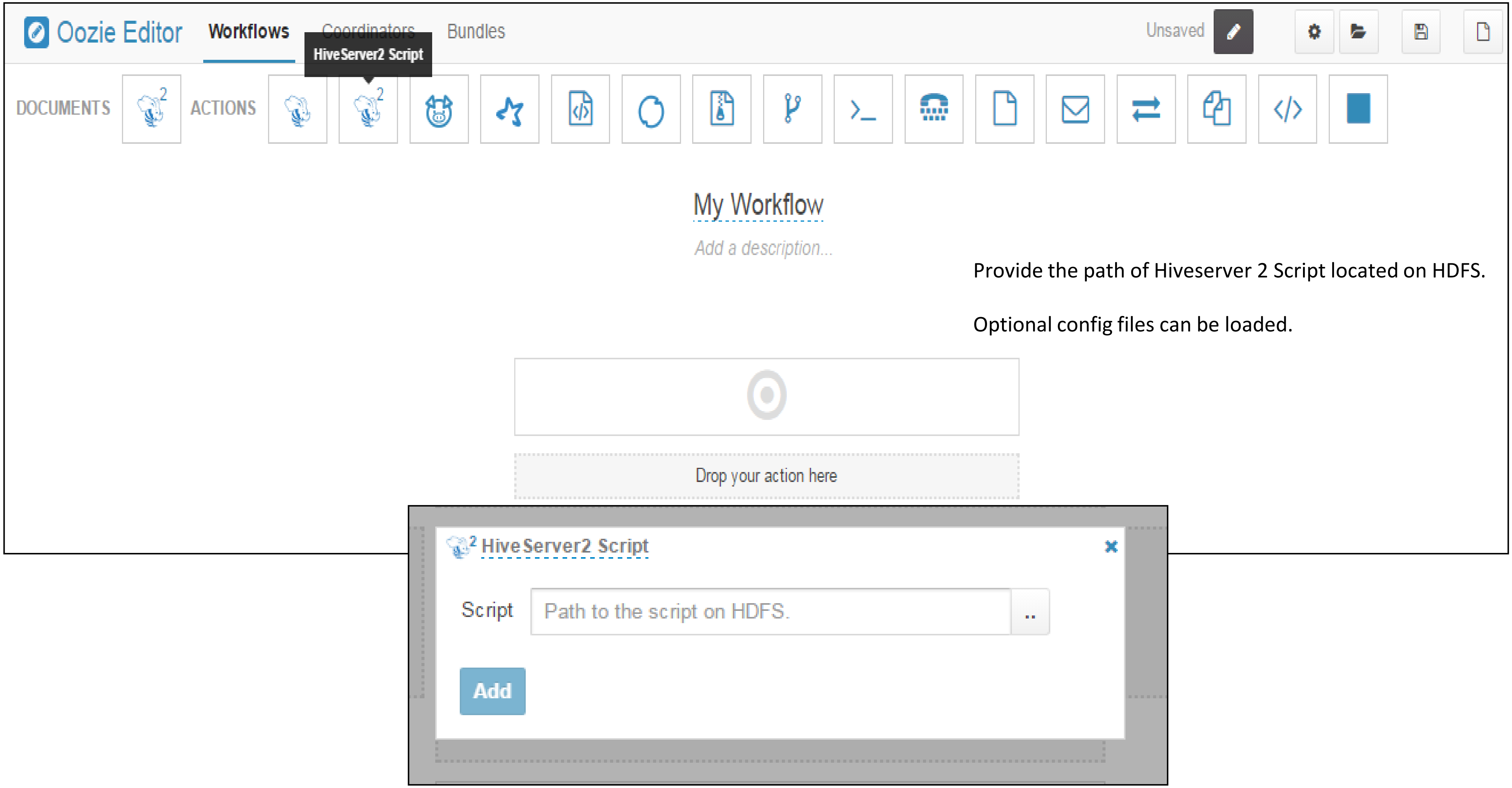
Script: Path to the script on HDFS.

Hive XML: Refer to a hive-site.xml for connecting to Hive

Provide the path of Hive Script located on HDFS.

Optional config files can be loaded.

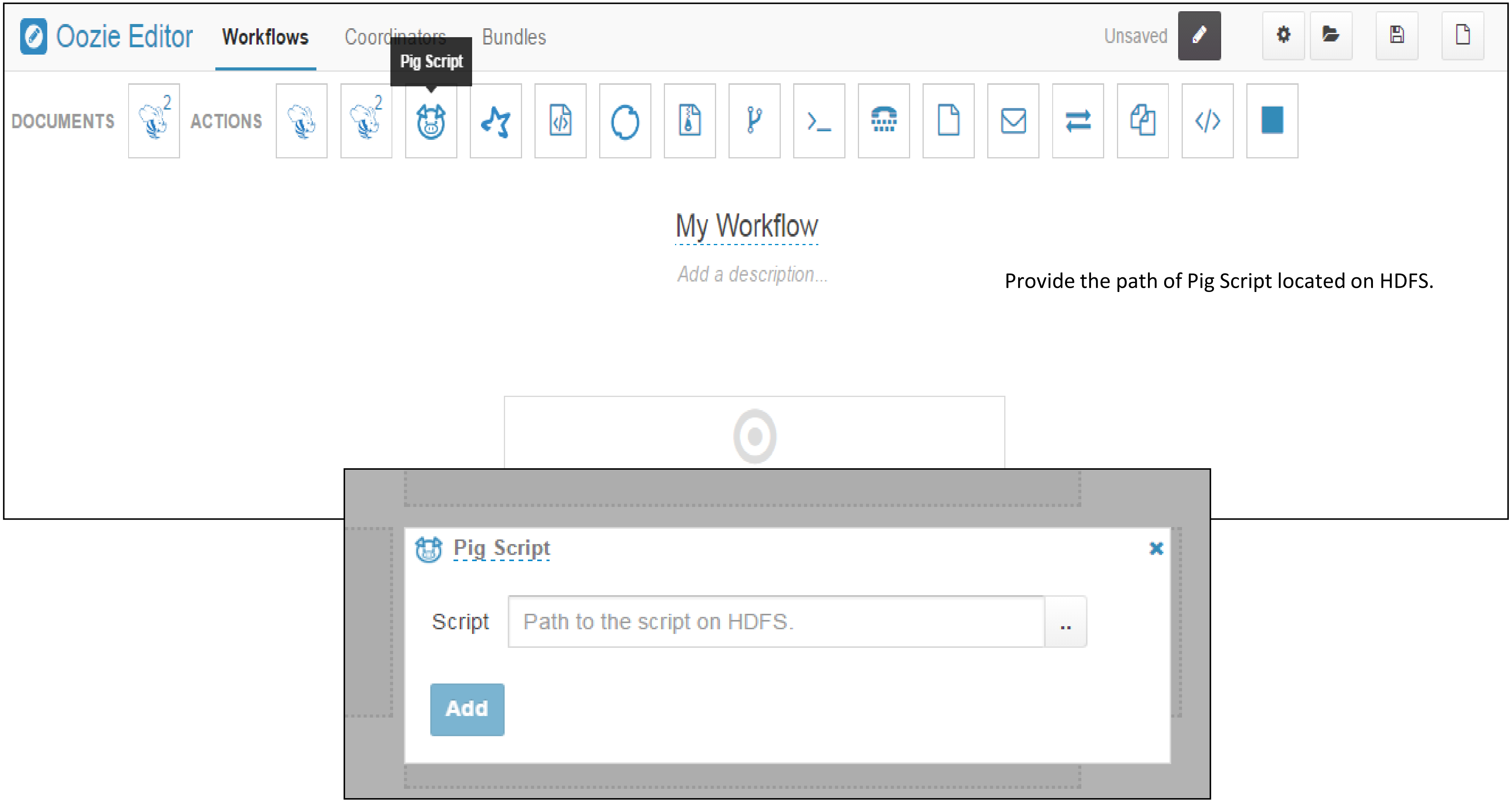
Hive Server2 Action



Provide the path of Hiveserver 2 Script located on HDFS.

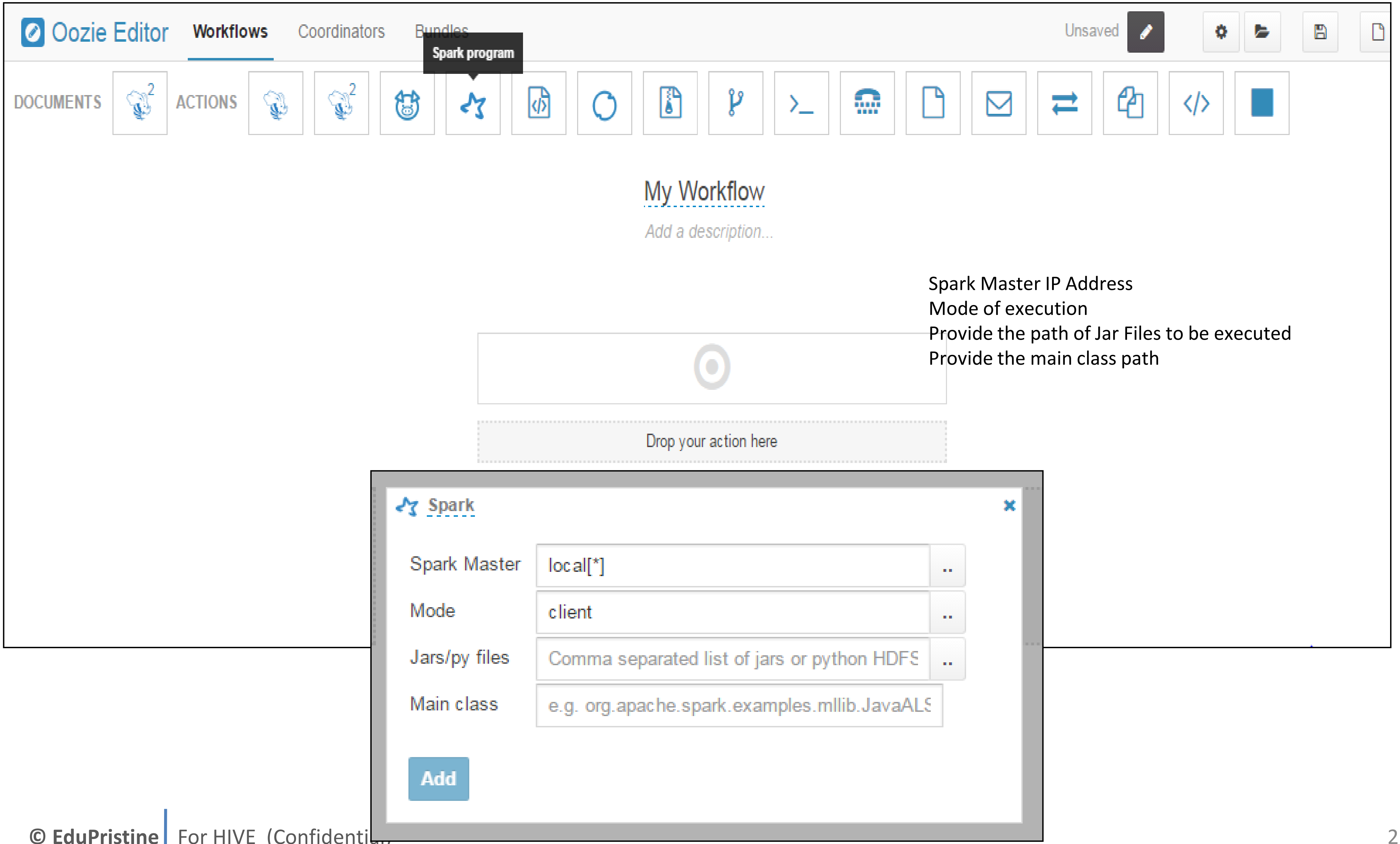
Optional config files can be loaded.

Pig Action



Provide the path of Pig Script located on HDFS.

Spark Action



Spark Master: local[*]

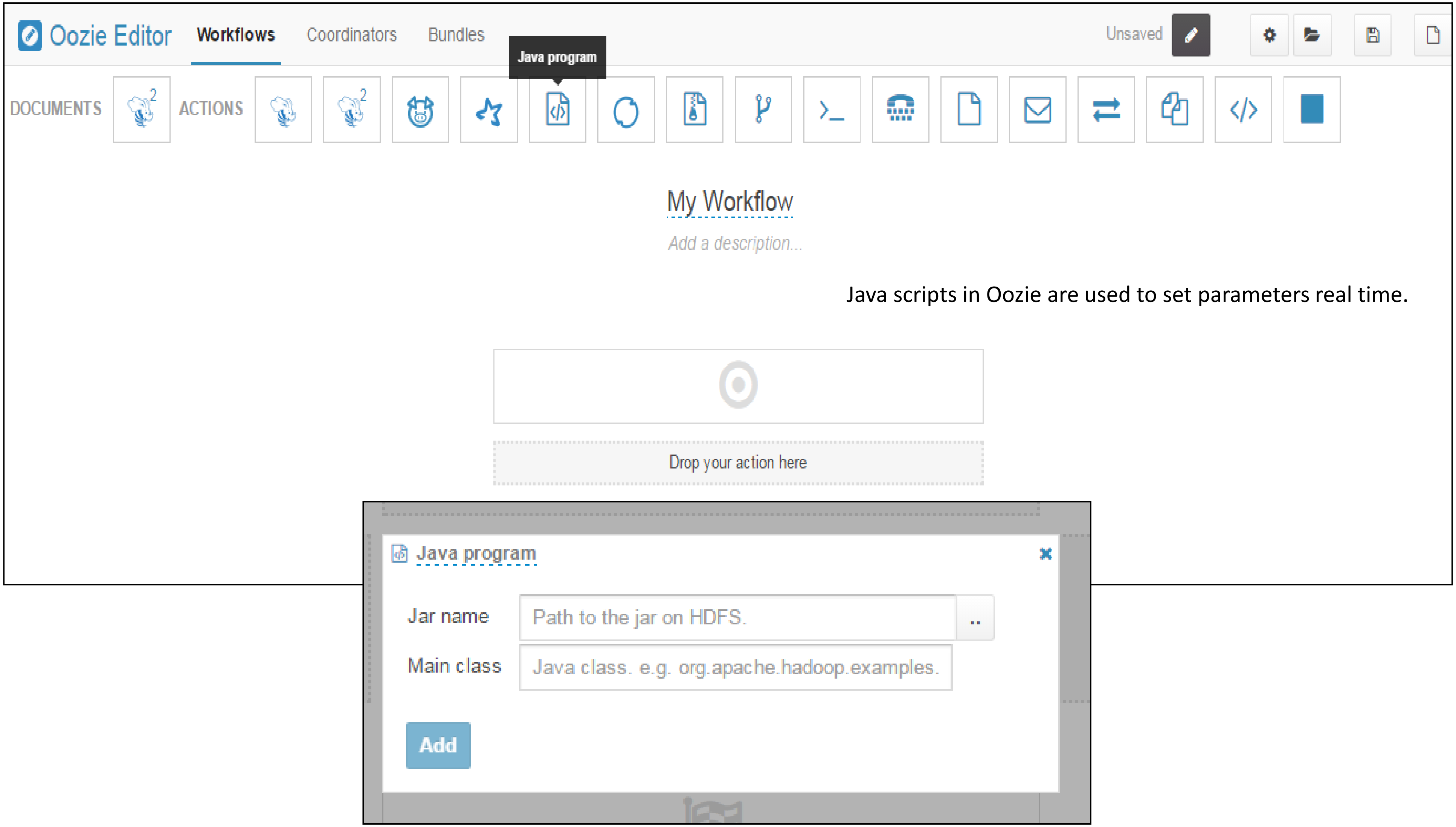
Mode: client

Jars/py files: Comma separated list of jars or python HDFS

Main class: e.g. org.apache.spark.examples.mllib.JavaALS

Spark Master IP Address
Mode of execution
Provide the path of Jar Files to be executed
Provide the main class path

Java Action



Java scripts in Oozie are used to set parameters real time.

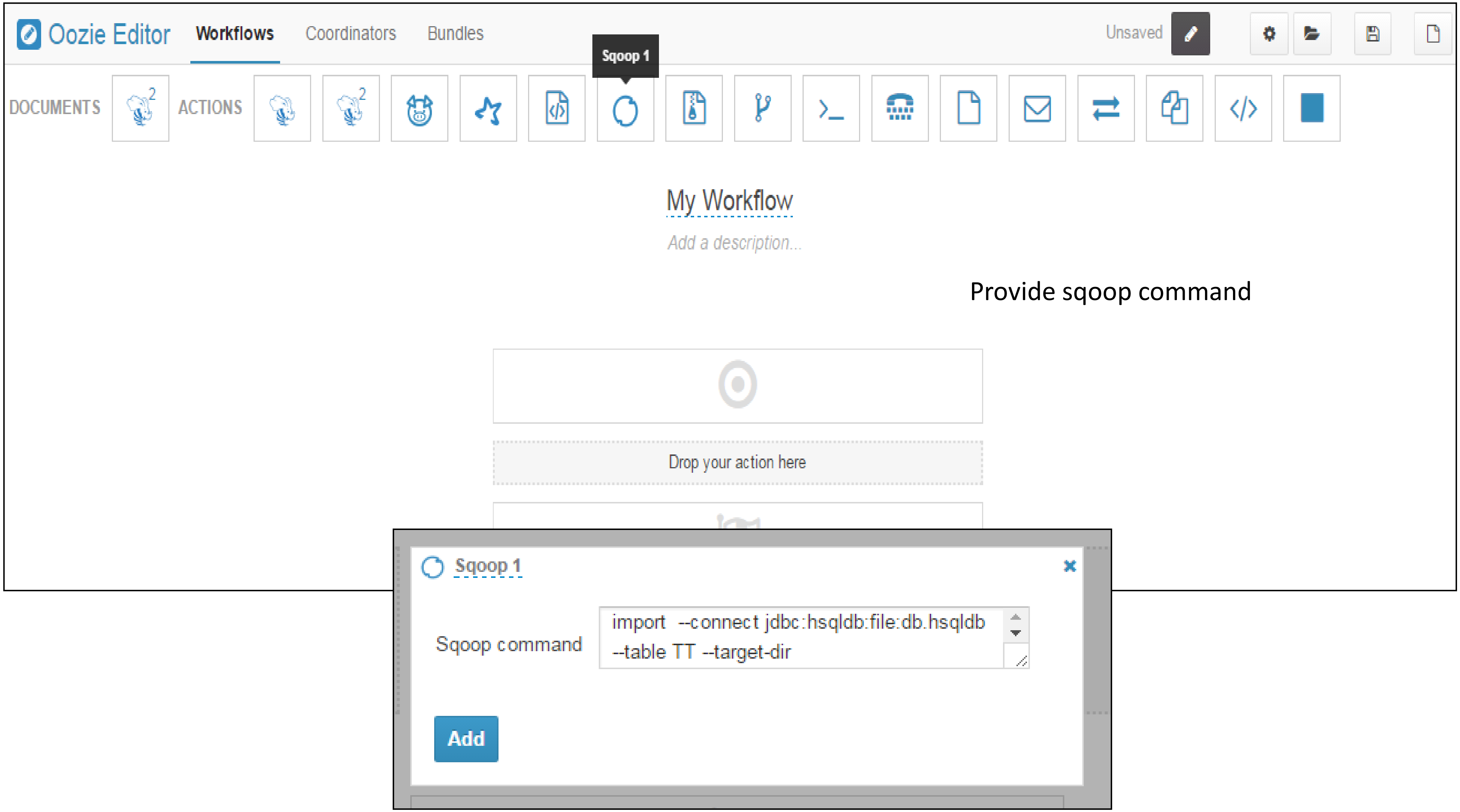
Java program

Jar name: Path to the jar on HDFS.

Main class: Java class. e.g. org.apache.hadoop.examples.

Add

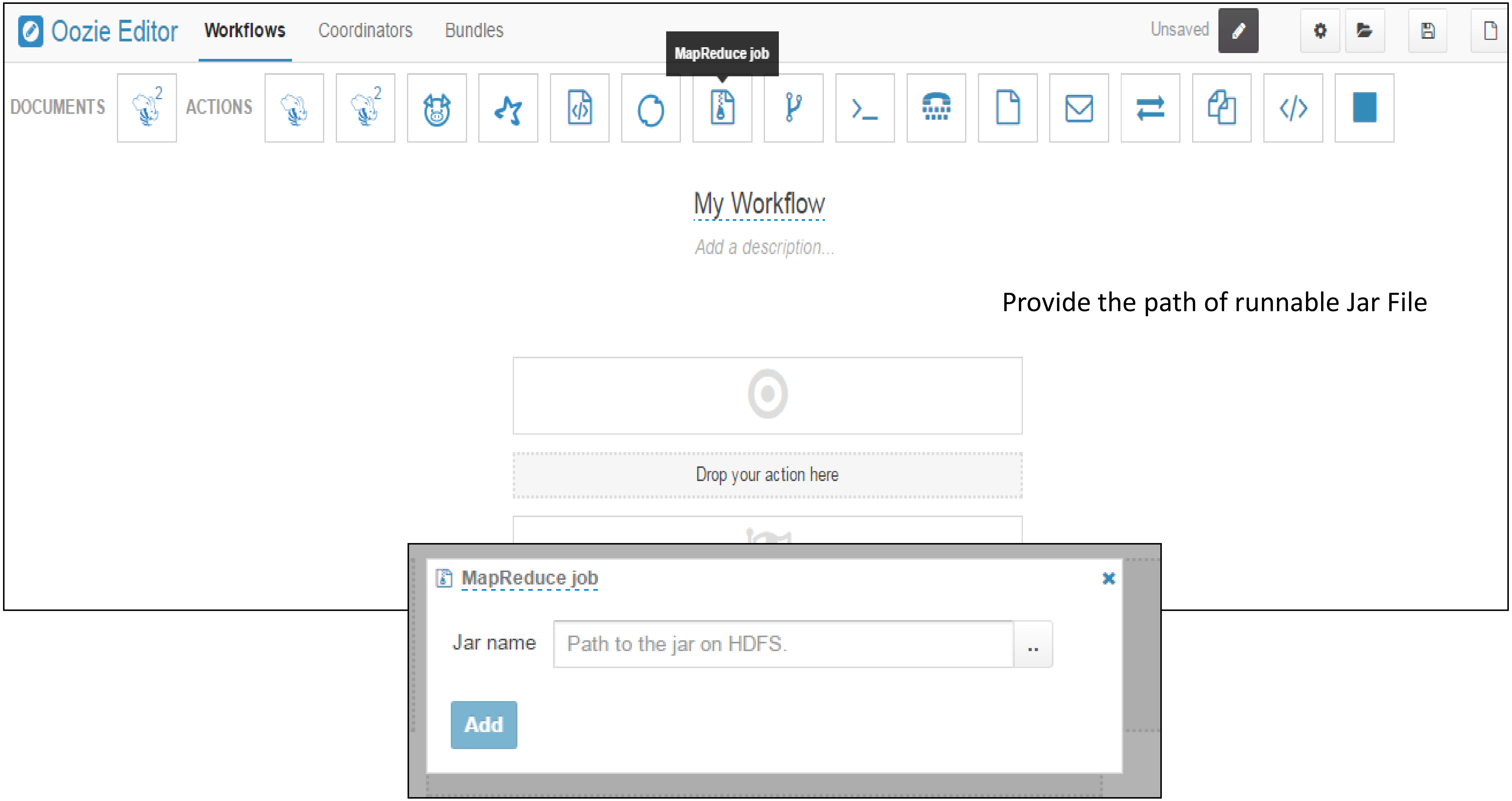
Sqoop Action



Provide sqoop command

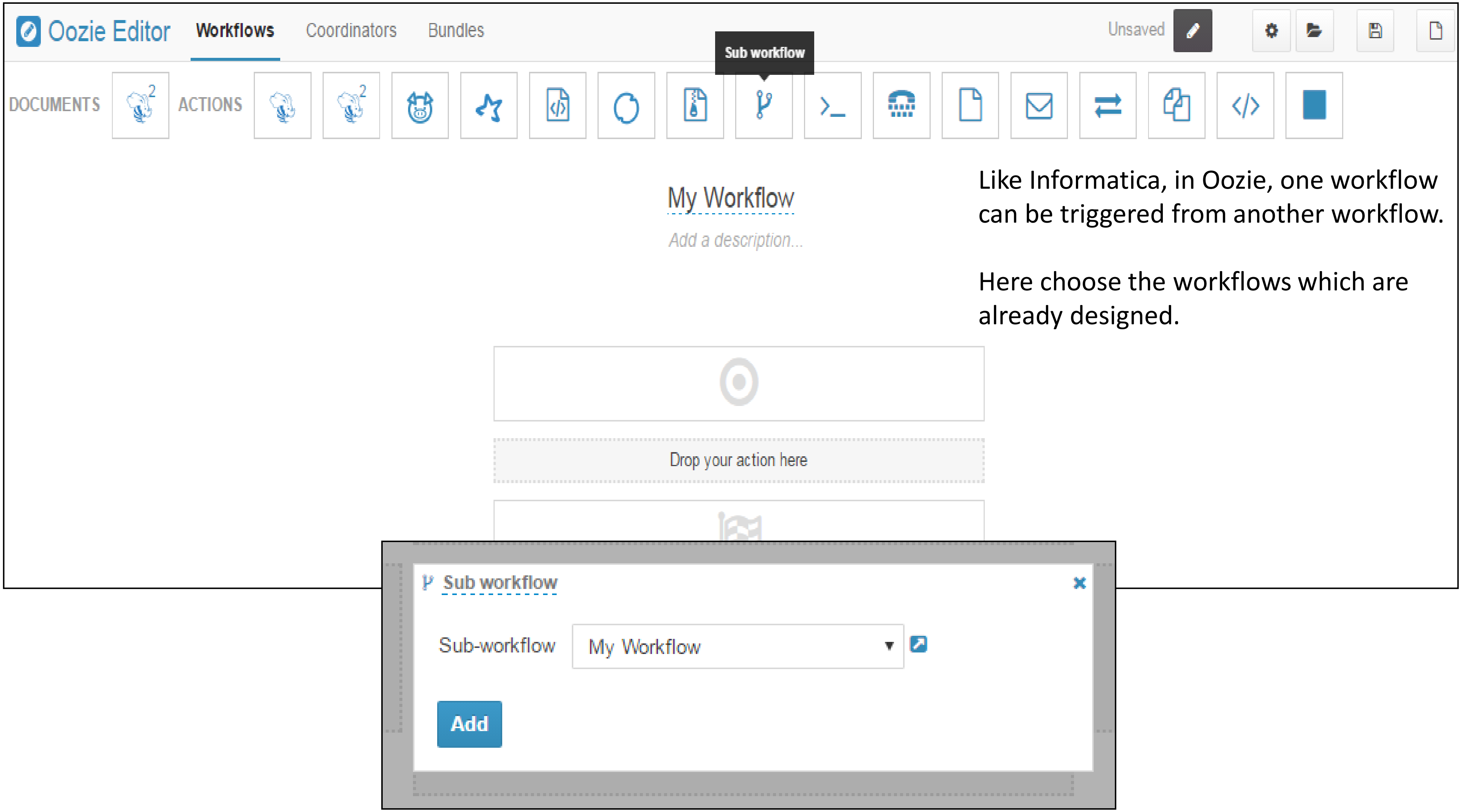
Sqoop command: `import --connect jdbc:hsqldb:file:db.hsqldb --table TT --target-dir`

Map reduce Action



Provide the path of runnable Jar File

Sub Workflow, Fork and Joins control nodes

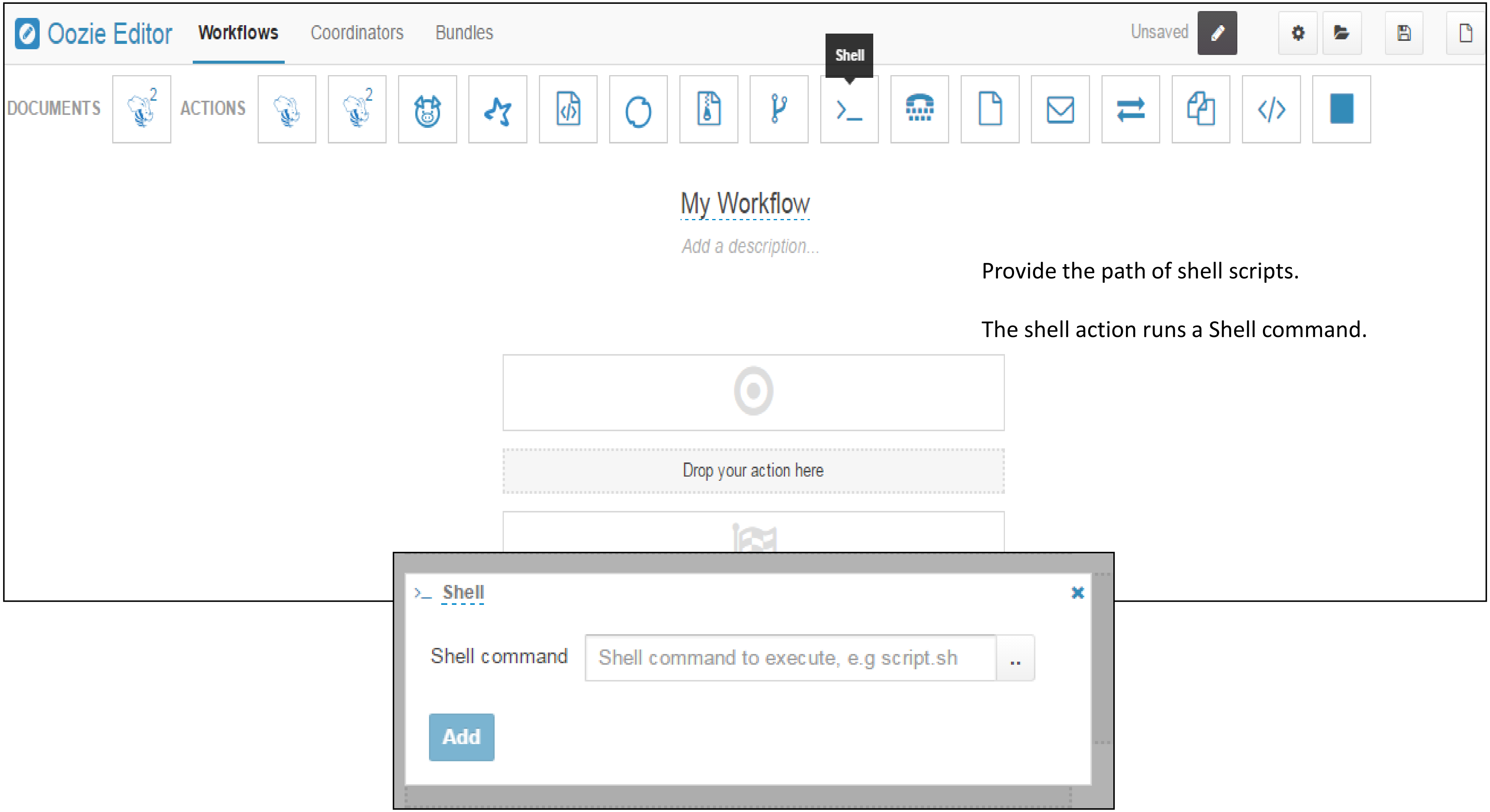


Sub-workflow: My Workflow

Like Informatica, in Oozie, one workflow can be triggered from another workflow.

Here choose the workflows which are already designed.

Shell Action

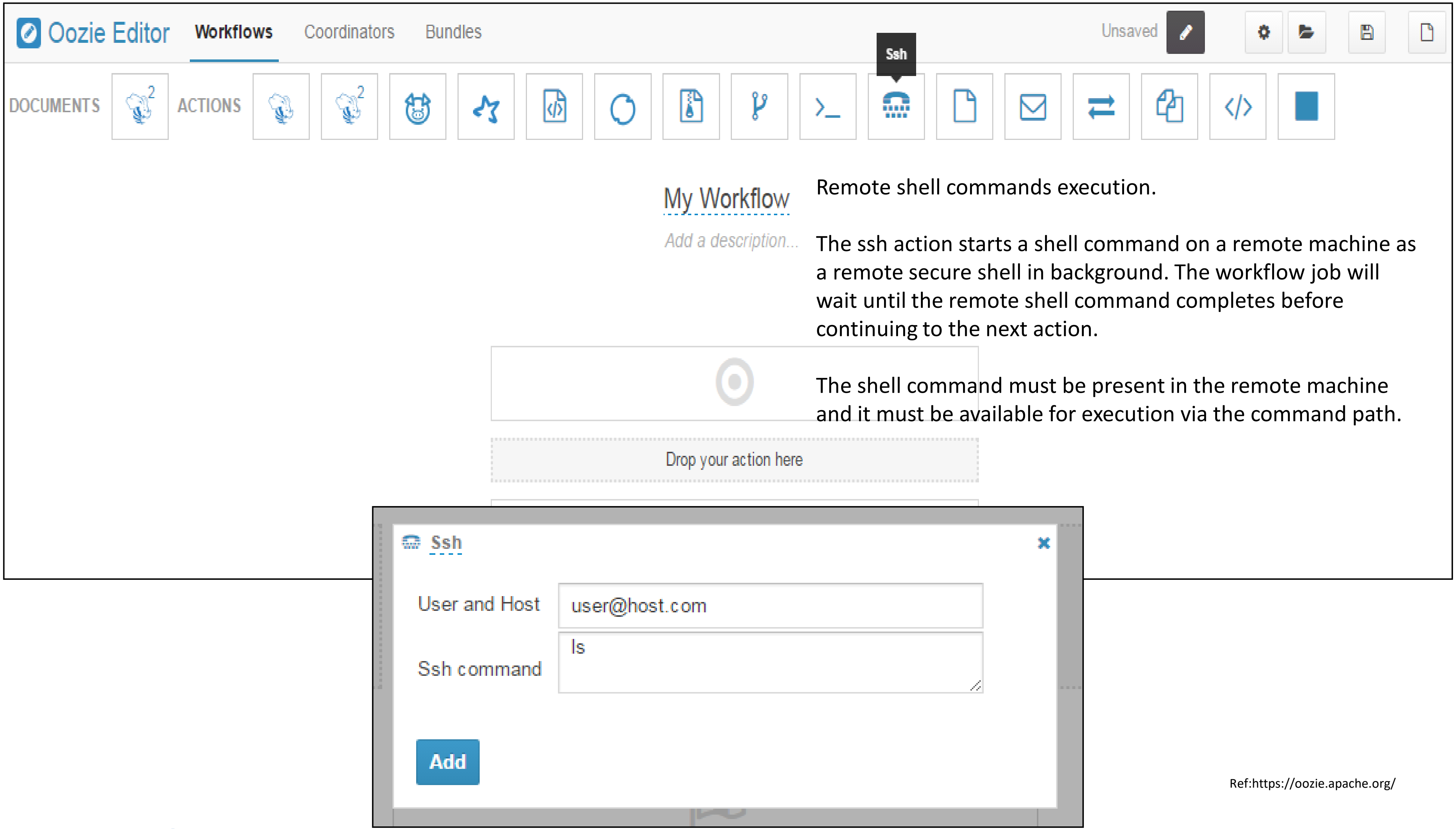


Shell command: Shell command to execute, e.g script.sh

Provide the path of shell scripts.

The shell action runs a Shell command.

SSH action



Ssh

User and Host: user@host.com

Ssh command: ls

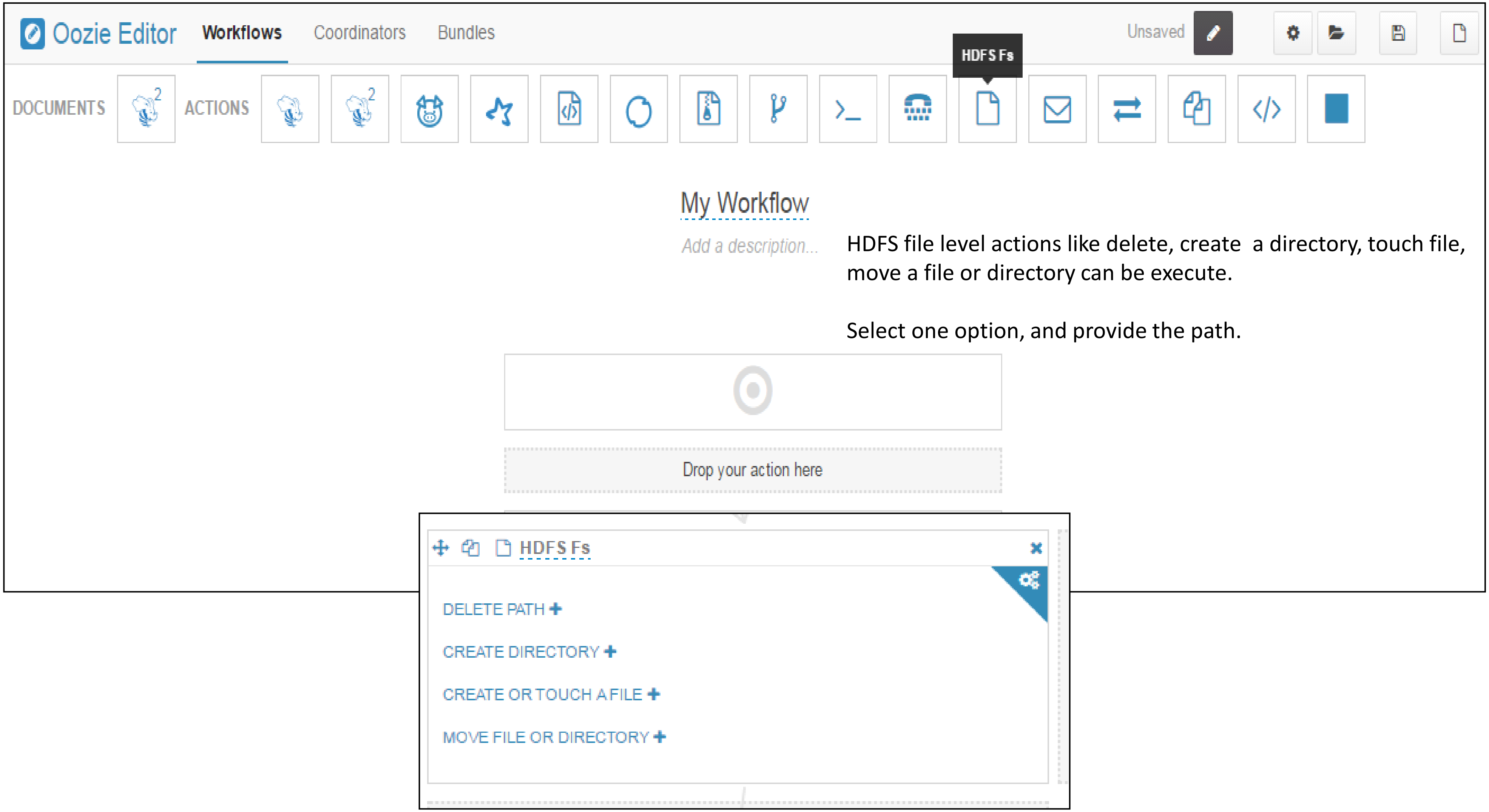
Remote shell commands execution.

The ssh action starts a shell command on a remote machine as a remote secure shell in background. The workflow job will wait until the remote shell command completes before continuing to the next action.

The shell command must be present in the remote machine and it must be available for execution via the command path.

Ref:https://oozie.apache.org/

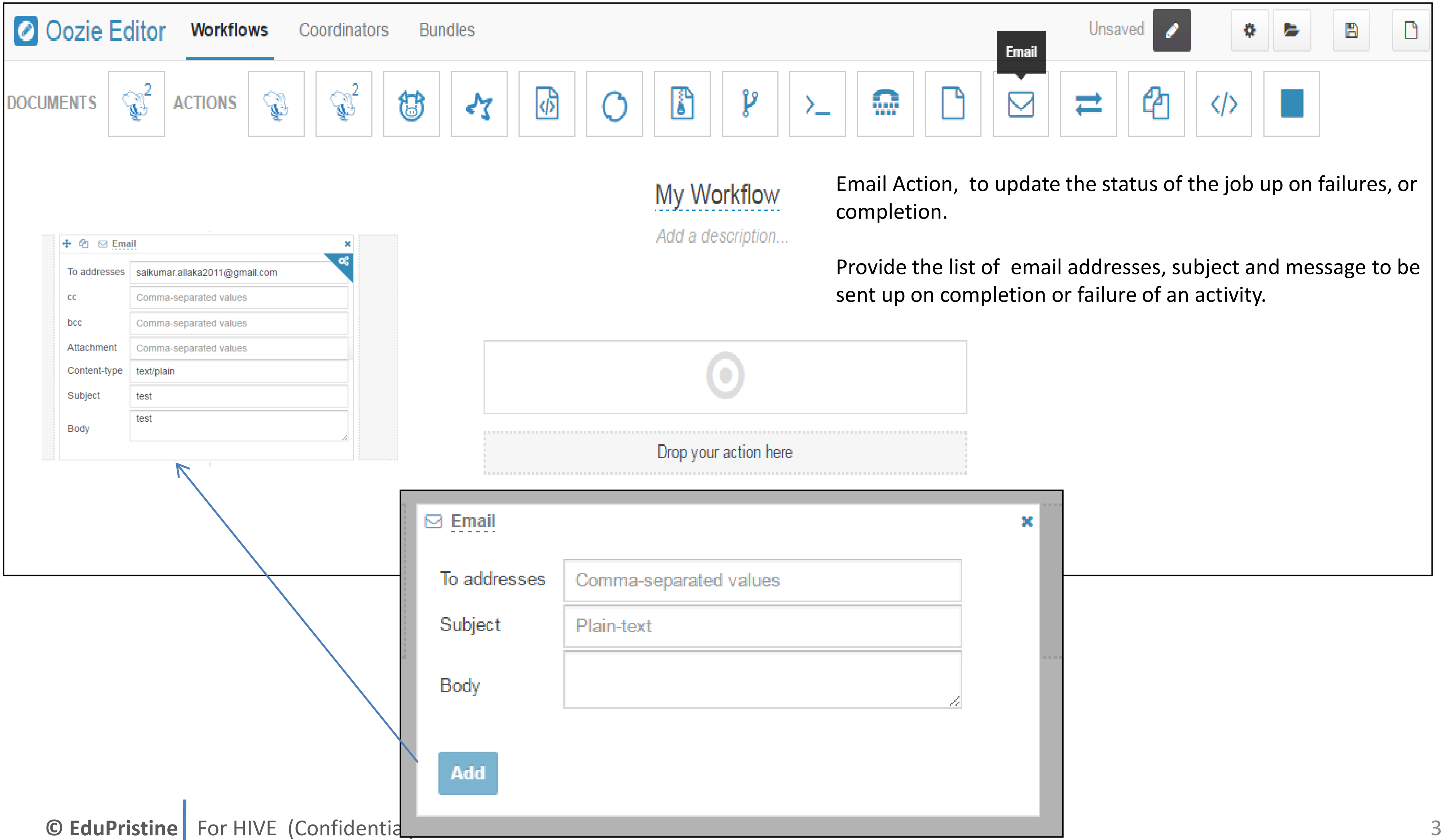
HDFS file action



HDFS file level actions like delete, create a directory, touch file, move a file or directory can be execute.

Select one option, and provide the path.

Email Action



Email Action, to update the status of the job up on failures, or completion.

Provide the list of email addresses, subject and message to be sent up on completion or failure of an activity.

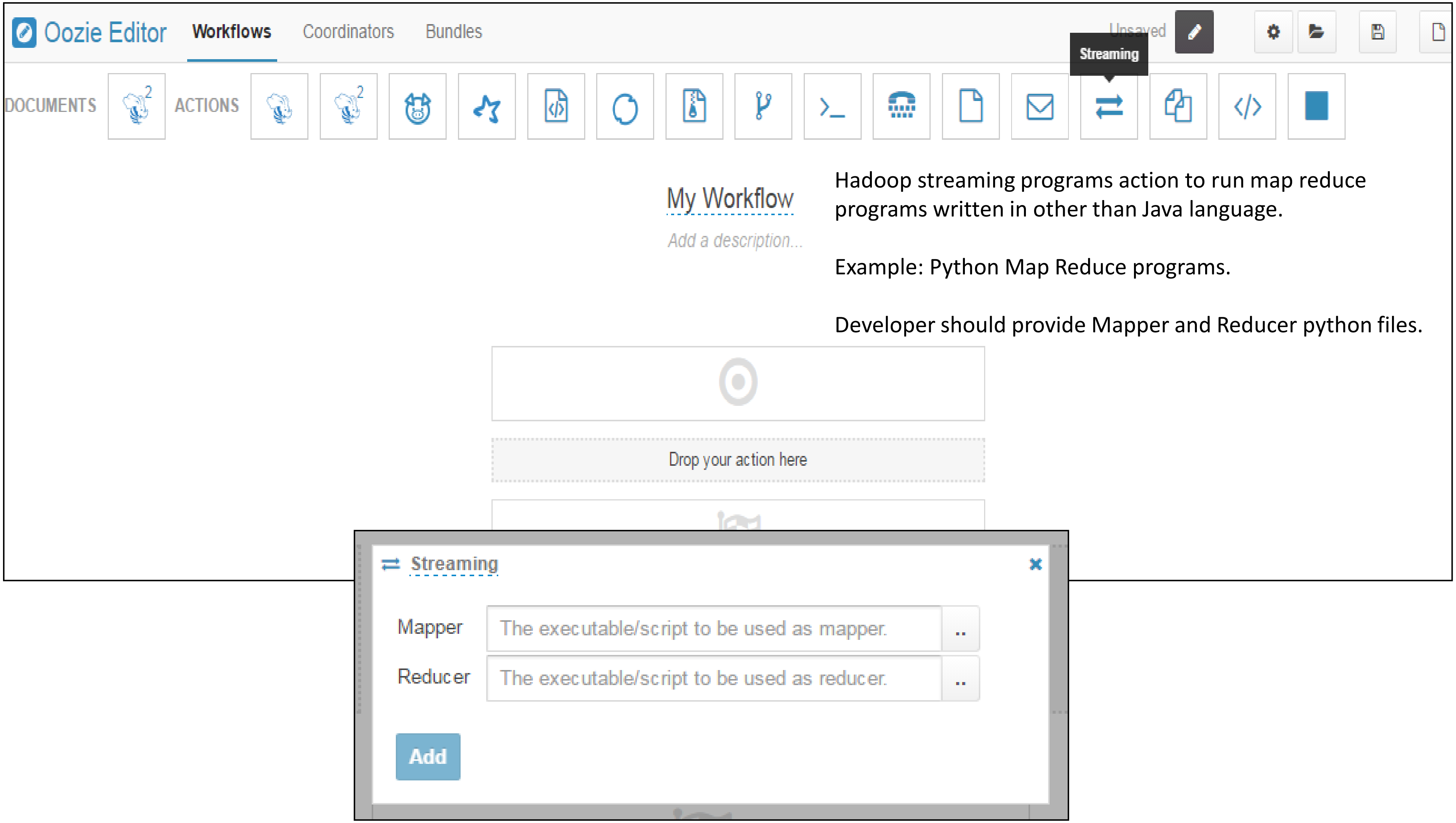
Email

To addresses	saikumar.allaka2011@gmail.com
cc	Comma-separated values
bcc	Comma-separated values
Attachment	Comma-separated values
Content-type	text/plain
Subject	test
Body	test

Email

To addresses	Comma-separated values
Subject	Plain-text
Body	

Hadoop Streaming Action



Mapper: The executable/script to be used as mapper.

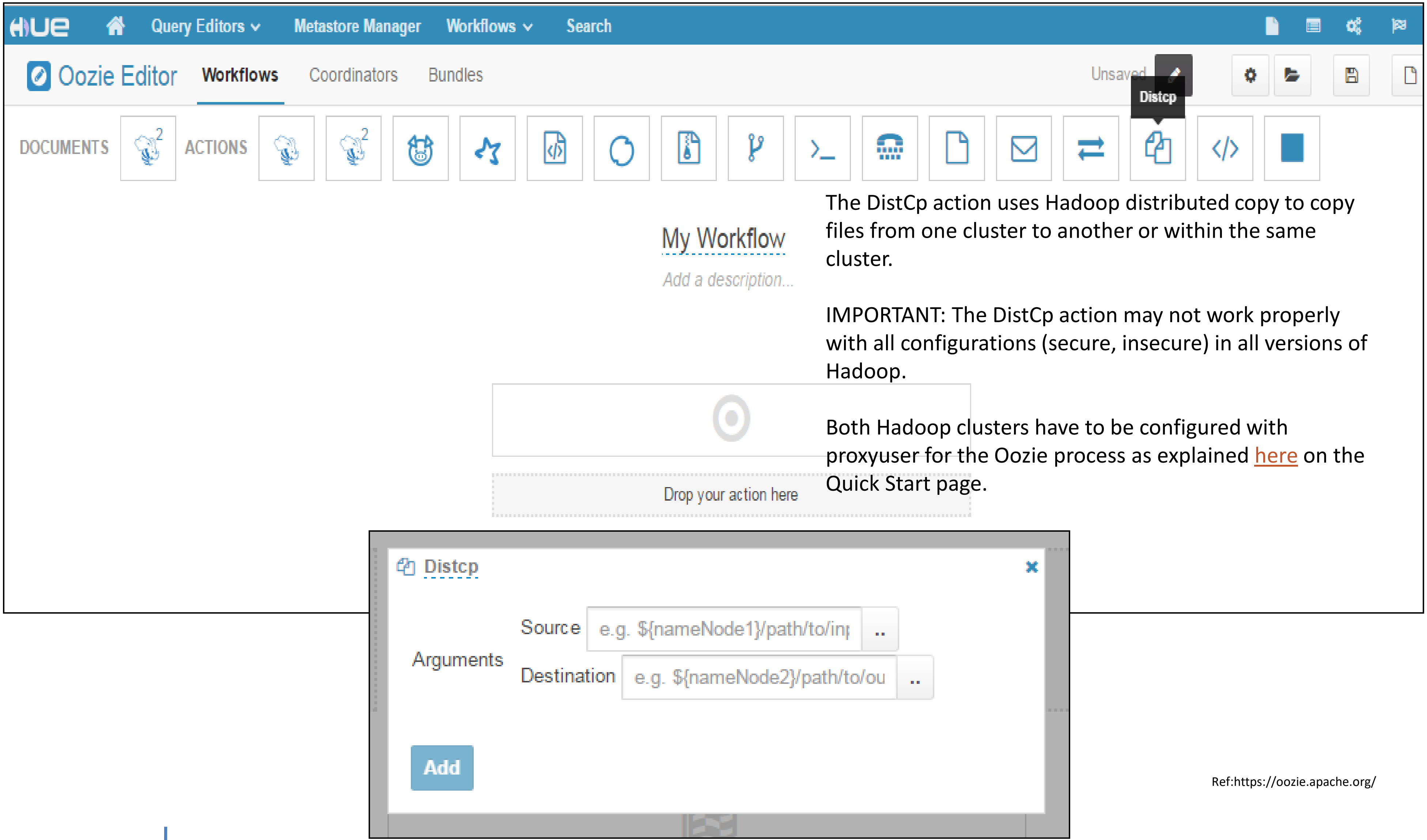
Reducer: The executable/script to be used as reducer.

Hadoop streaming programs action to run map reduce programs written in other than Java language.

Example: Python Map Reduce programs.

Developer should provide Mapper and Reducer python files.

Distcp Action



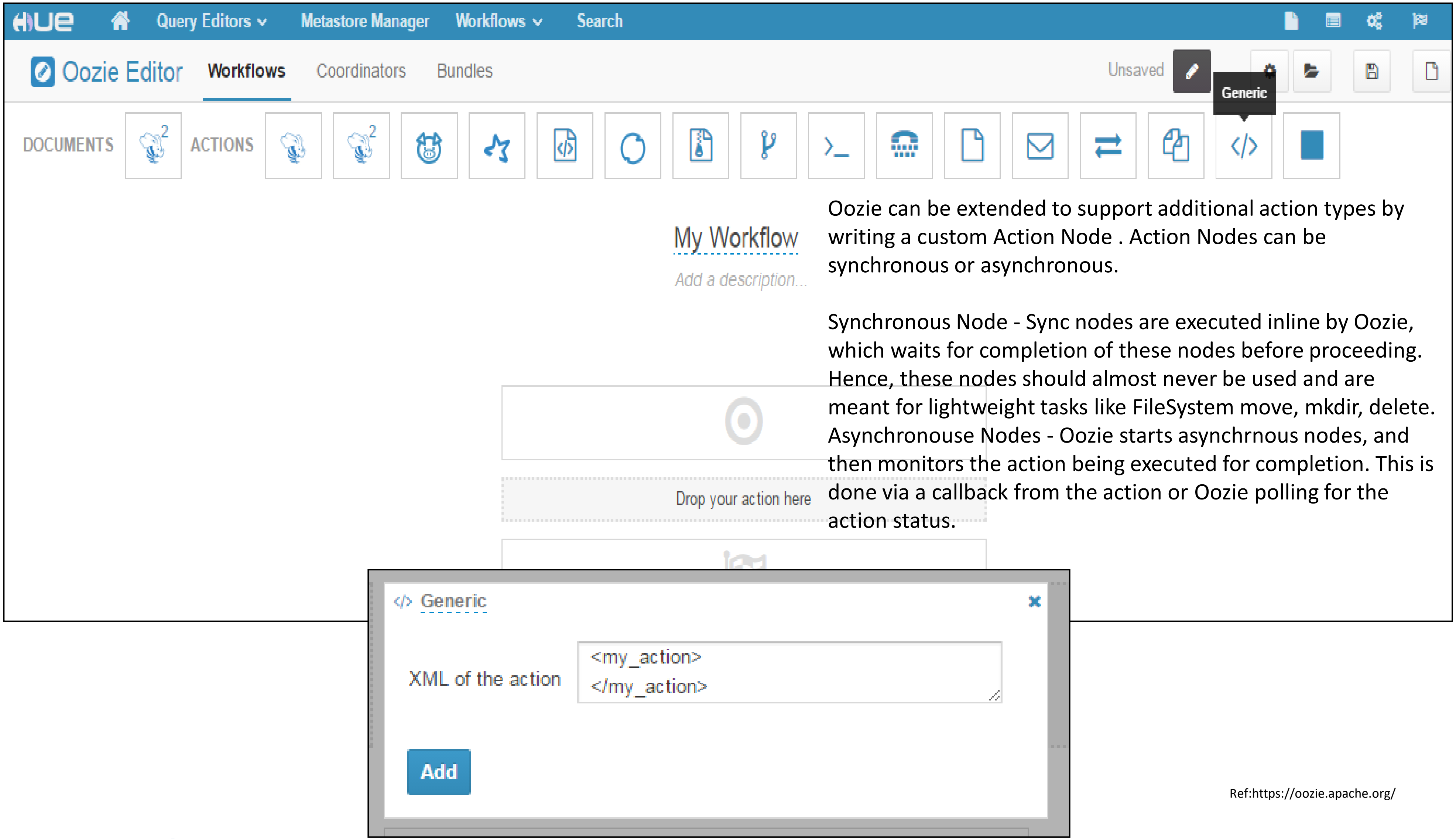
The DistCp action uses Hadoop distributed copy to copy files from one cluster to another or within the same cluster.

IMPORTANT: The DistCp action may not work properly with all configurations (secure, insecure) in all versions of Hadoop.

Both Hadoop clusters have to be configured with proxyuser for the Oozie process as explained here on the Quick Start page.

Ref:https://oozie.apache.org/

Generic action



XML of the action

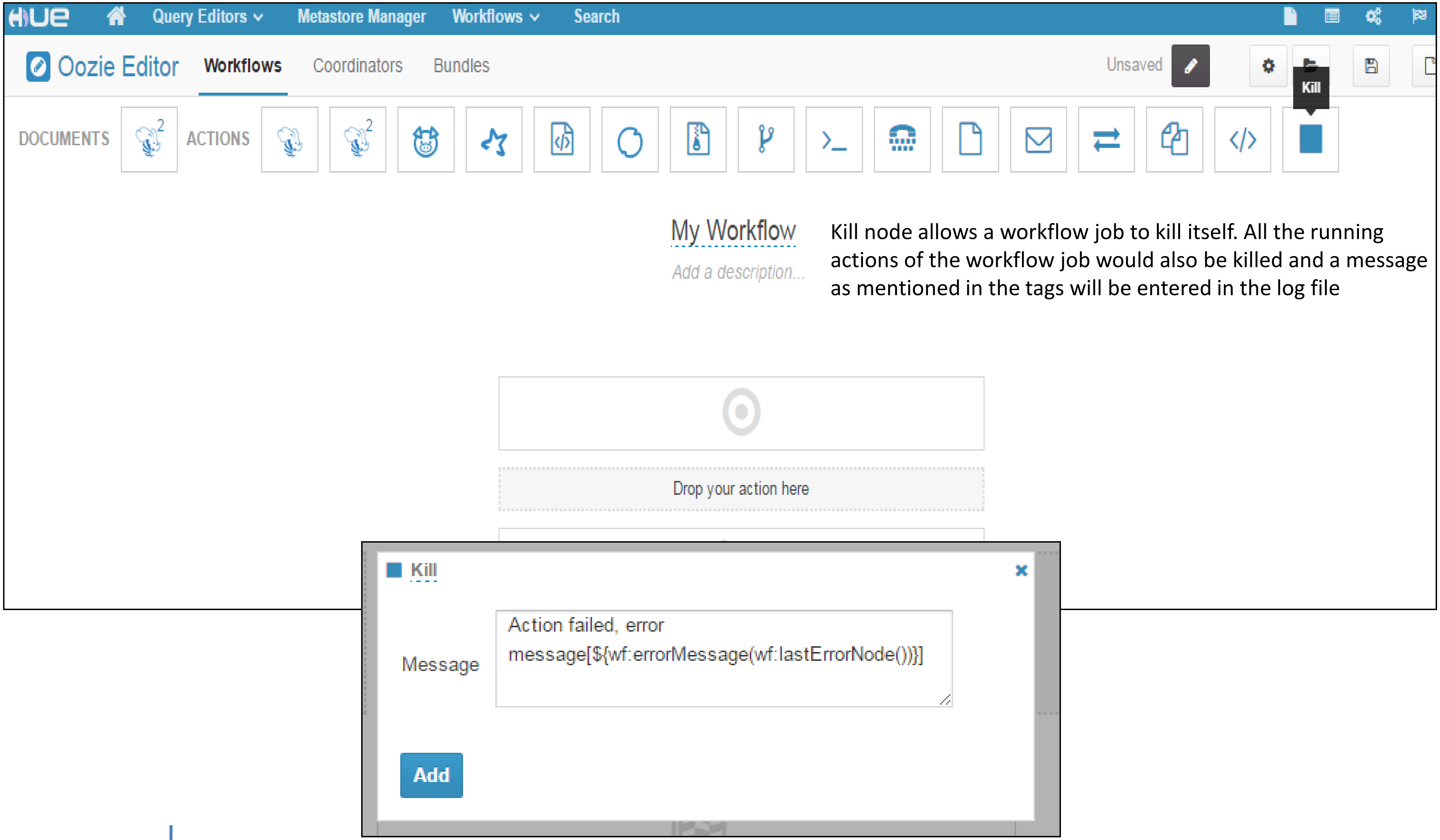
```
<my_action>
</my_action>
```

Oozie can be extended to support additional action types by writing a custom Action Node . Action Nodes can be synchronous or asynchronous.

Synchronous Node - Sync nodes are executed inline by Oozie, which waits for completion of these nodes before proceeding. Hence, these nodes should almost never be used and are meant for lightweight tasks like FileSystem move, mkdir, delete. Asynchronous Nodes - Oozie starts asynchrnous nodes, and then monitors the action being executed for completion. This is done via a callback from the action or Oozie polling for the action status.

Ref:https://oozie.apache.org/

Kill Node



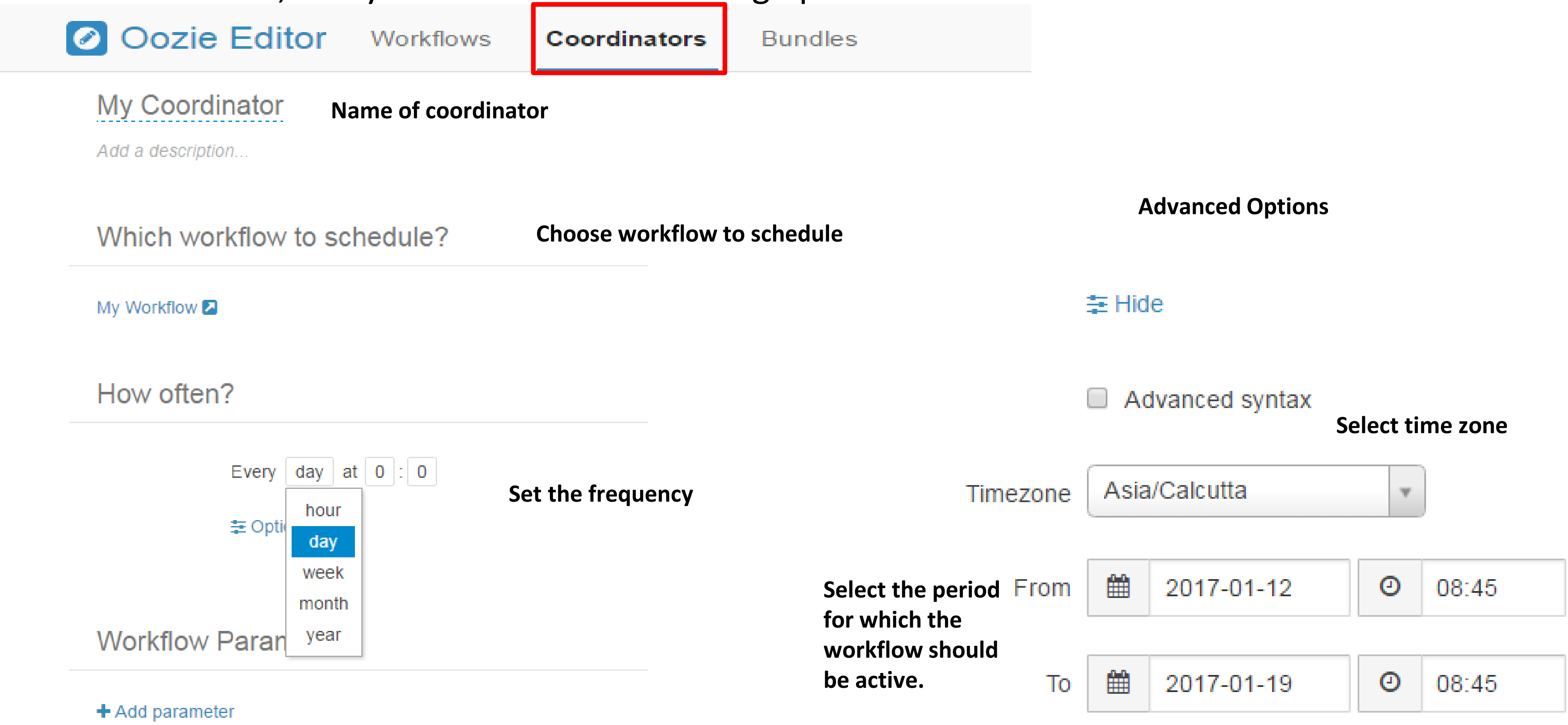
Kill node allows a workflow job to kill itself. All the running actions of the workflow job would also be killed and a message as mentioned in the tags will be entered in the log file

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

Coordinate Functional Specification

The Oozie **Coordinator** system allows the user to define and execute recurrent and interdependent workflow jobs (data application pipelines).

Choose a workflow, and you will find the following options:



My Coordinator

Name of coordinator

Add a description...

Which workflow to schedule?

Choose workflow to schedule

My Workflow

How often?

Every day at 0 : 0

hour
day
week
month
year

Set the frequency

Workflow Paran

Advanced Options

□ Advanced syntax

Select time zone

Timezone Asia/Calcutta

Select the period for which the workflow should be active.

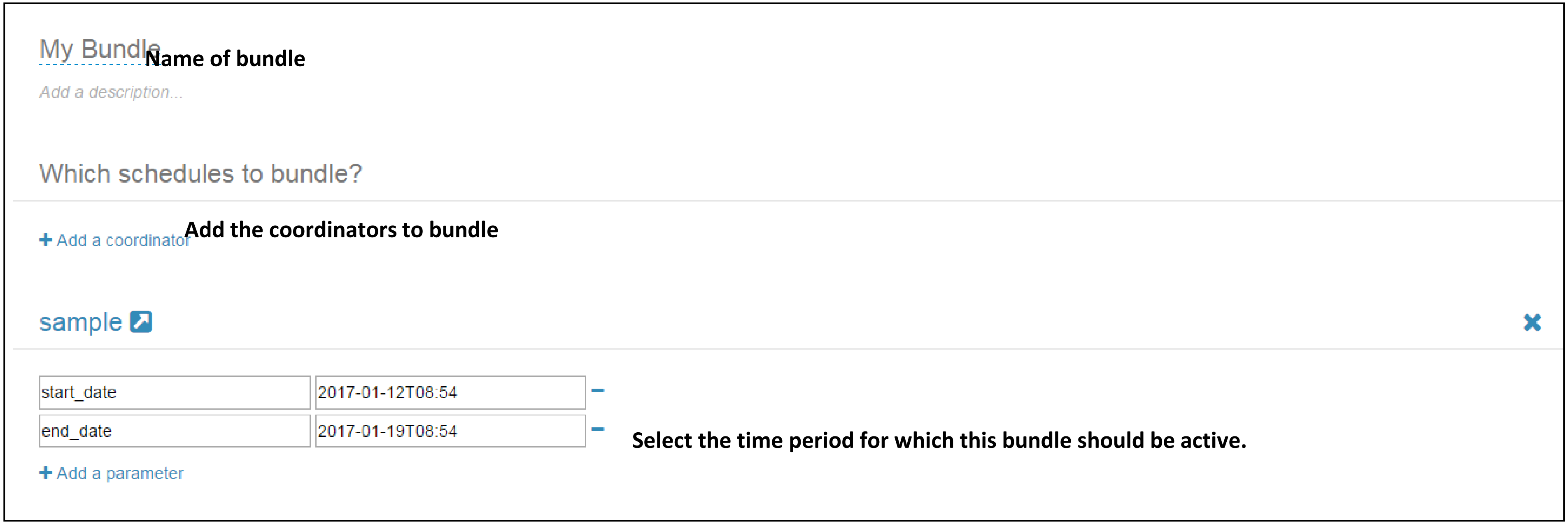
From 2017-01-12 08:45

To 2017-01-19 08:45

Bundle

Bundle is a higher-level oozie abstraction that will batch a set of coordinator applications. The user will be able to start/stop/suspend/resume/rerun in the bundle level resulting a better and easy operational control.

More specififcally, the oozie Bundle system allows the user to define and execute a bunch of coordinator applications often called a data pipeline.



My Bundle

Name of bundle

Add a description...

Which schedules to bundle?

+ Add a coordinator

Add the coordinators to bundle

sample

start_date	2017-01-12T08:54
end_date	2017-01-19T08:54

+ Add a parameter

Select the time period for which this bundle should be active.

Lab

In this lab, you will learn the follwing:

1. Creating sqoop action
2. Creating HDFS action
3. Creating Hive action
4. Creating Pig Action
5. Creating Workflows
6. Integrating sub-workflows
7. Coordinating workflows

Steps

1. Sqoop data to HDFS
2. Pig consumes this data, and generates top10Cust. Sqoop exports the results back to MySQL.
3. Hive creates external tables, and generates chain_stats. Sqoop exports the results back to MySQL.

Create required tables in MySQL

Login:

```
mysql -h 54.149.41.179 -u username -p --local-infile

use database;

CREATE TABLE transactions(id varchar(20),chain varchar(20), dept varchar(20),
category varchar(20), company varchar(20), brand varchar(20), date1 varchar(10),
productsize int, productmeasure varchar(10), purchasequantity int, purchaseamount FLOAT);

LOAD DATA LOCAL INFILE '/home/training/Desktop/transactions.csv'
INTO TABLE transactions FIELDS TERMINATED BY ',' ENCLOSED BY '"'
LINES TERMINATED BY '\r\n';

-- This table will be imported to HDFS


-- Results will be exported back to mysql

create table chain_stats(chain varchar(3), deptcin int,categorycin int,
companycin int,brandcin int,totalspent float);

create table chainTop10Cust(chain varchar(3), id varchar(10),totalSales float);
```

Copy the programs to HDFS

Login to Hive, and create database oozie_username. Replace your name under username.

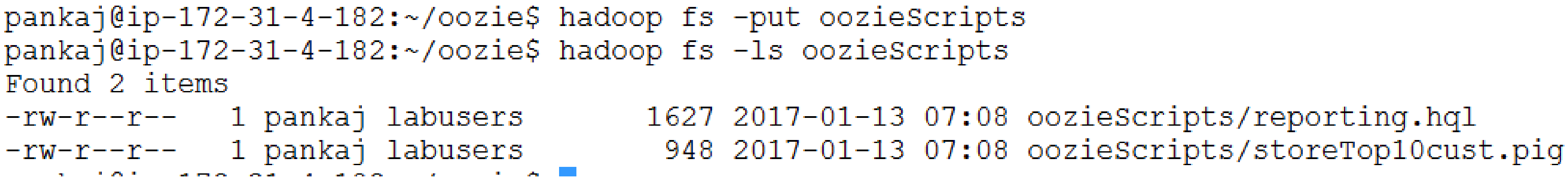
In oozieScripts folder, change the DBName in hive script to oozie_username.

Up on making relavent changes:

Copy oozieScripts folder to home folder using winscp.

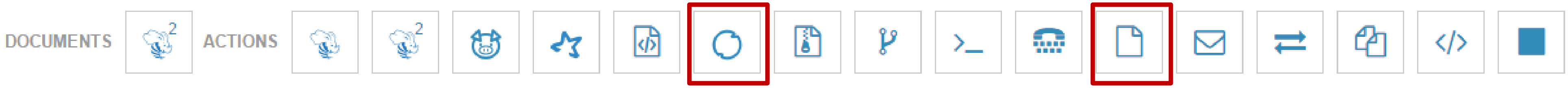
Push the folder to hadoop:

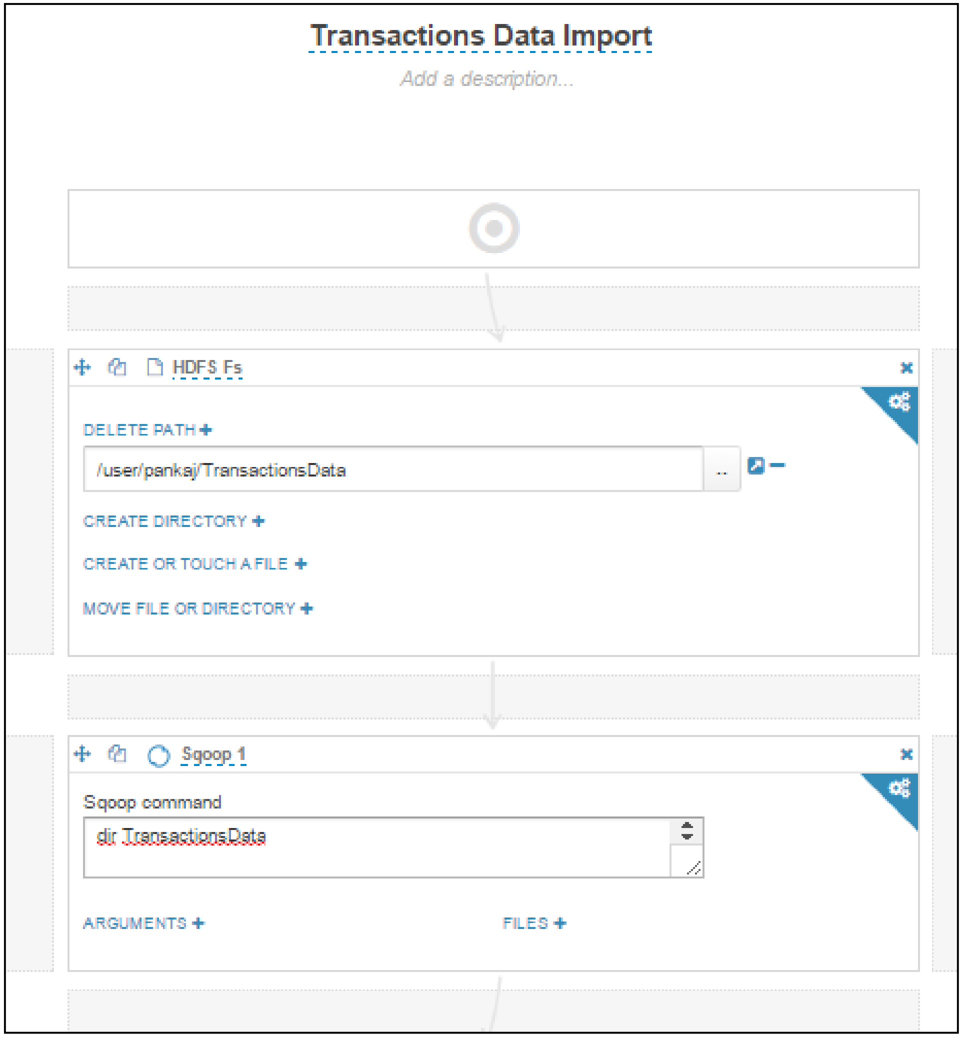
hadoop fs -put oozieScripts



```
pankaj@ip-172-31-4-182:~/oozie$ hadoop fs -put oozieScripts
pankaj@ip-172-31-4-182:~/oozie$ hadoop fs -ls oozieScripts
Found 2 items
-rw-r--r--   1 pankaj labusers       1627 2017-01-13 07:08 oozieScripts/reporting.hql
-rw-r--r--   1 pankaj labusers        948 2017-01-13 07:08 oozieScripts/storeTop10cust.pig
```

Sqoop Import workflow




Transactions Data Import

Add a description...

HDFS Fs

DELETE PATH +

/user/pankaj/TransactionsData

CREATE DIRECTORY +

CREATE OR TOUCH A FILE +

MOVE FILE OR DIRECTORY +

Sqoop 1

Sqoop command

dir TransactionsData

ARGUMENTS +

FILES +

Name of the workflow

Drag and drop HDFS fs action.

Delete the directory, if exists before sqoop starts.

Drag and drop sqoop action.

Add the below command:

import --connect jdbc:mysql://54.149.41.179/pankaj --username username--password pwd--table transactions --split-by id --target-dir TransactionsData

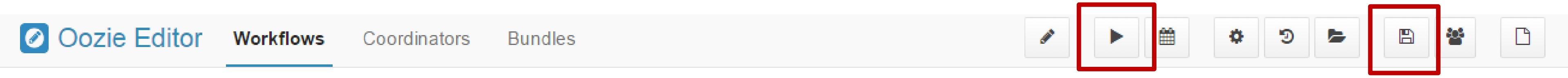
Add username and pwd.

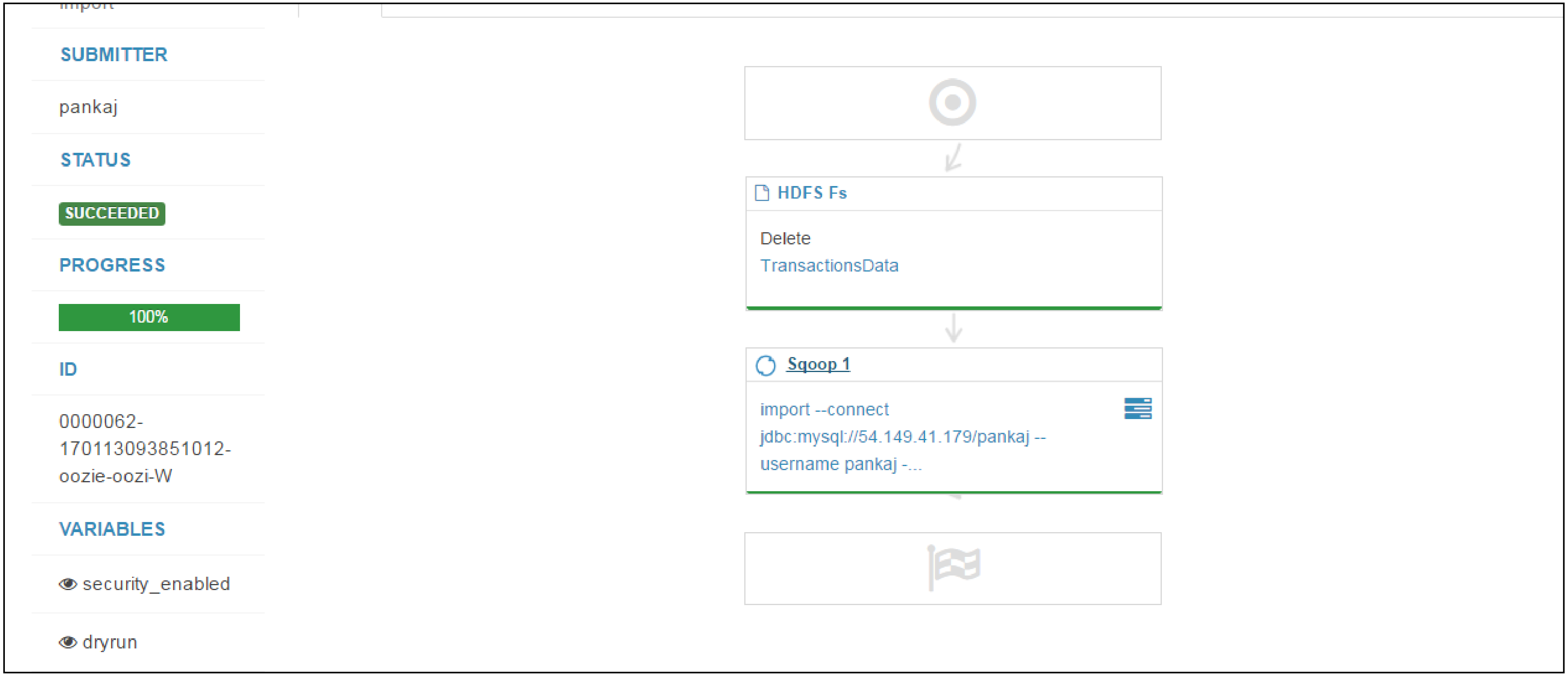
Sqoop Import workflow - Submit

Save and submit.

Submit

Save





import

SUBMITTER

pankaj

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000062-170113093851012-oozie-oozi-W

VARIABLES

security_enabled

dryrun

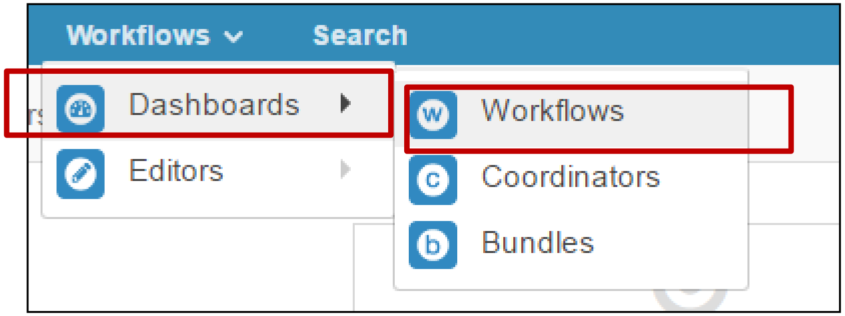
HDFS Fs

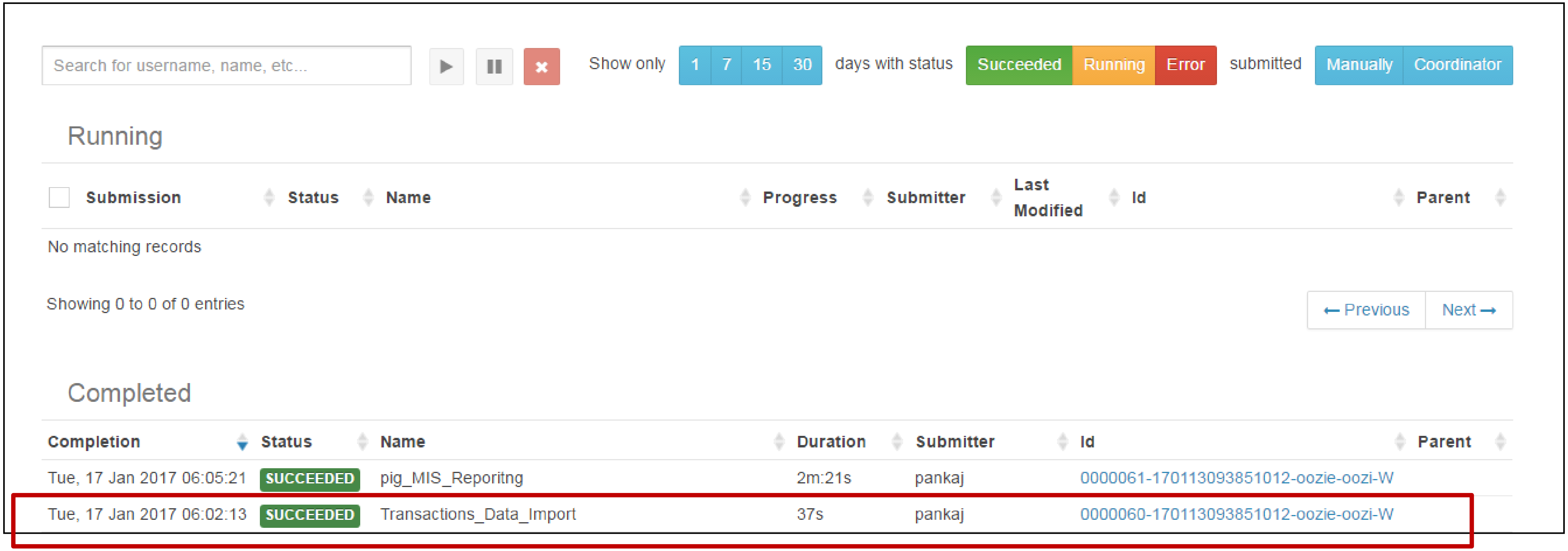
Delete TransactionsData

Sqoop 1

import --connect jdbc:mysql://54.149.41.179/pankaj --username pankaj -...

Sqoop Import workflow – Track progress





Running

□	Submission	Status	Name	Progress	Submitter	Last Modified	Id	Parent

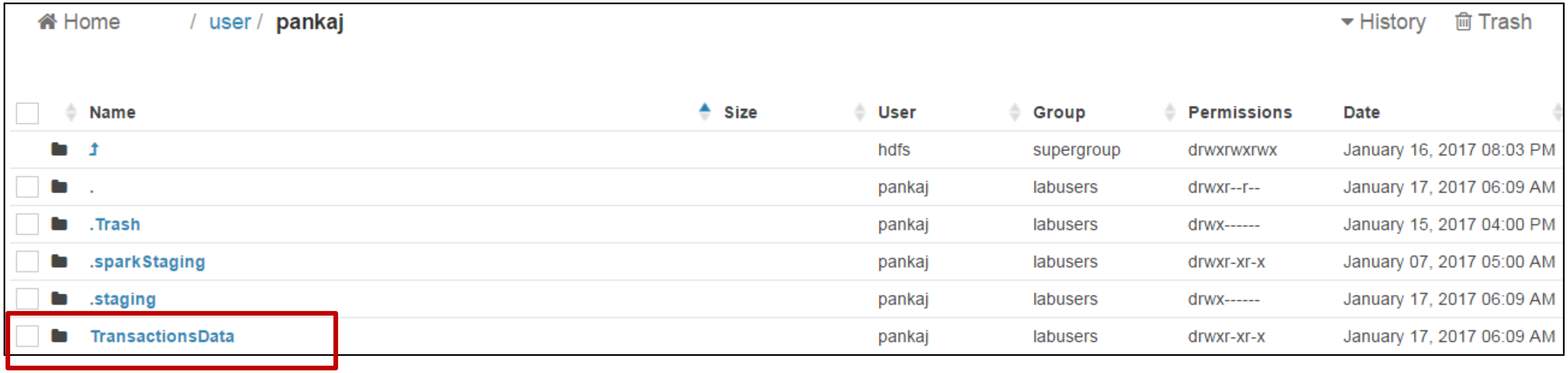
No matching records

Showing 0 to 0 of 0 entries

Completed

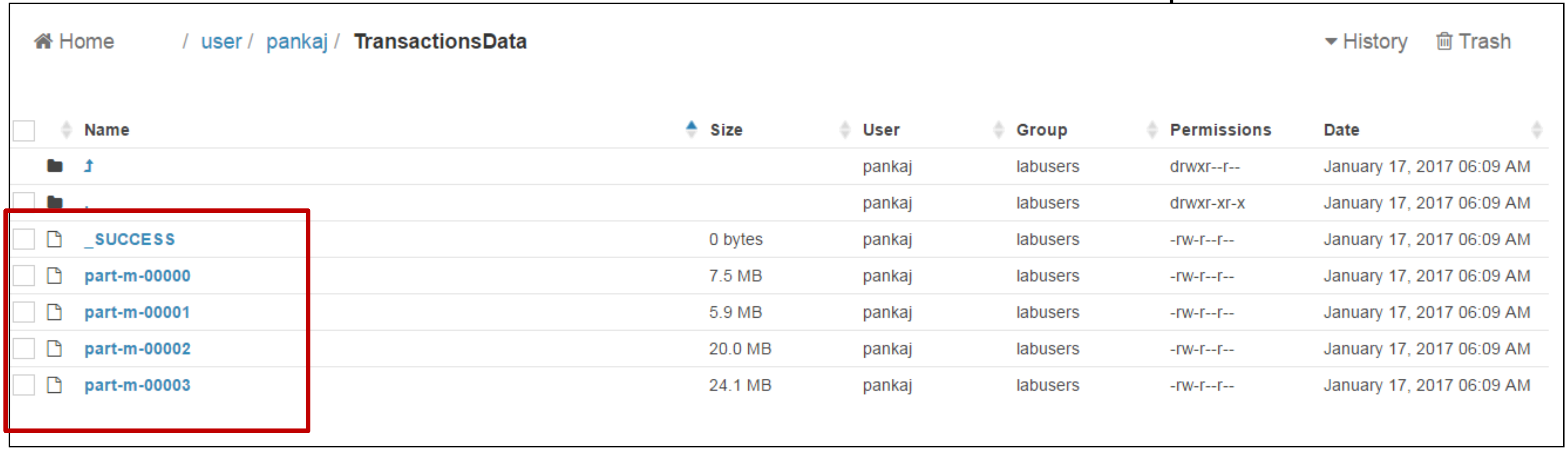
Completion	Status	Name	Duration	Submitter	Id	Parent
Tue, 17 Jan 2017 06:05:21	SUCCEEDED	pig_MIS_Reporitng	2m:21s	pankaj	0000061-170113093851012-oozie-oozi-W	
Tue, 17 Jan 2017 06:02:13	SUCCEEDED	Transactions_Data_Import	37s	pankaj	0000060-170113093851012-oozie-oozi-W	

Data Check up on completion



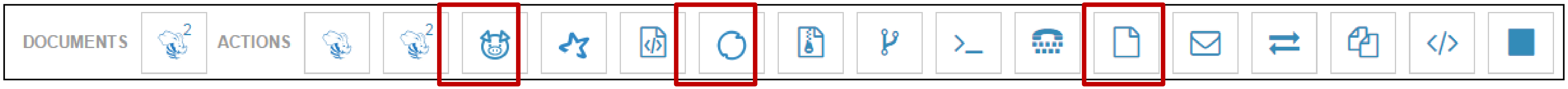
	Name	Size	User	Group	Permissions	Date
	↥		hdfs	supergroup	drwxrwxrwx	January 16, 2017 08:03 PM
□	.		pankaj	labusers	drwxr--r--	January 17, 2017 06:09 AM
□	.Trash		pankaj	labusers	drwx------	January 15, 2017 04:00 PM
□	.sparkStaging		pankaj	labusers	drwxr-xr-x	January 07, 2017 05:00 AM
□	.staging		pankaj	labusers	drwx------	January 17, 2017 06:09 AM
□	TransactionsData		pankaj	labusers	drwxr-xr-x	January 17, 2017 06:09 AM

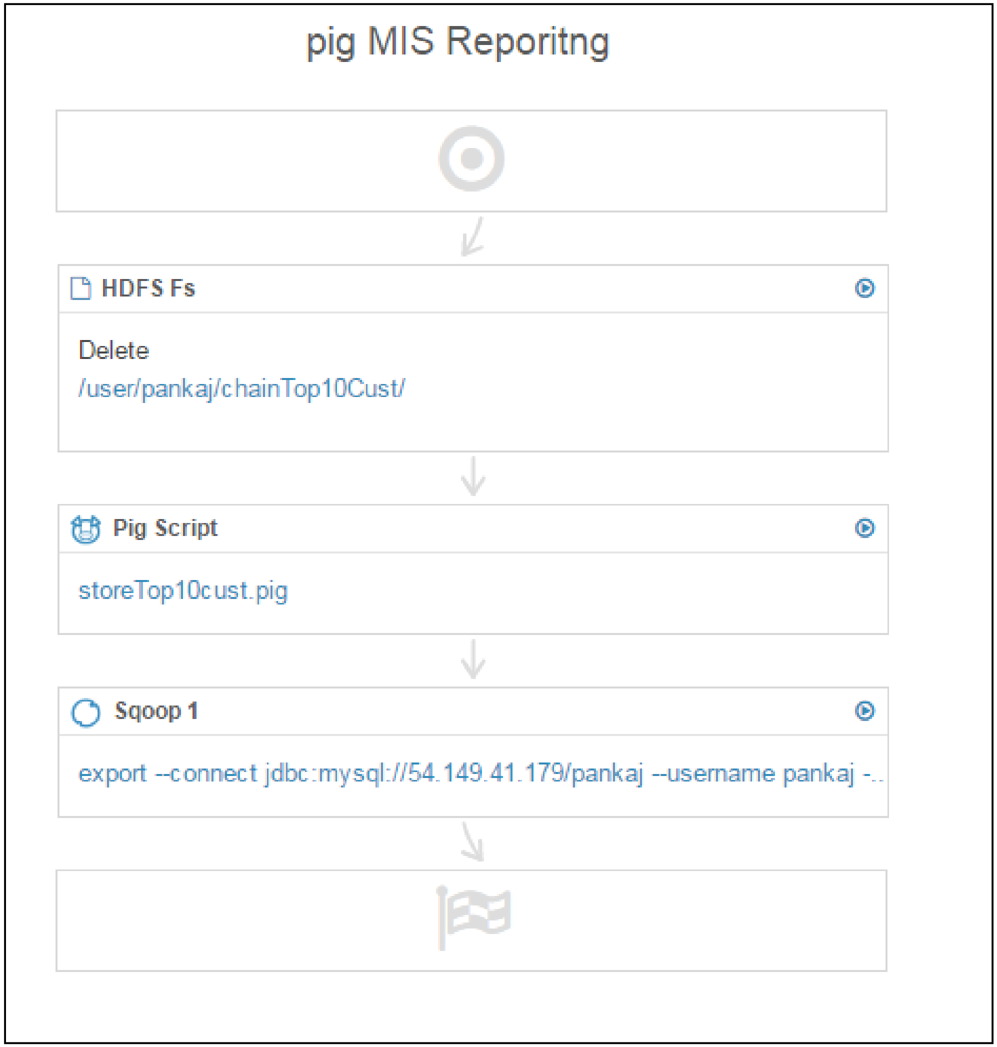
You should see data imported to HDFS



	Name	Size	User	Group	Permissions	Date
	↥		pankaj	labusers	drwxr--r--	January 17, 2017 06:09 AM
□	.		pankaj	labusers	drwxr-xr-x	January 17, 2017 06:09 AM
□	_SUCCESS	0 bytes	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM
□	part-m-00000	7.5 MB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM
□	part-m-00001	5.9 MB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM
□	part-m-00002	20.0 MB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM
□	part-m-00003	24.1 MB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM

Pig <--> Sqoop <--> MySQL – top 10 customers in each chain - report





pig MIS Reporitng

HDFS Fs

Delete
/user/pankaj/chainTop10Cust/

Pig Script

storeTop10cust.pig

Sqoop 1

```
export --connect jdbc:mysql://54.149.41.179/pankaj --username pankaj -..
```

Name of the workflow

Clean the directories.

Delete /user/username/chainTop10Cust

Replace username with your username

Path of pig script:

Browse and select the pig file in HDFs.

Sqoop Export o MYSQL

Add the following command to export results to MySQL:

```
export --connect jdbc:mysql://54.149.41.179/pankaj --
username username--password pwd --table chainTop10Cust --
export-dir chainTop10Cust --input-fields-terminated-by "\t“
```

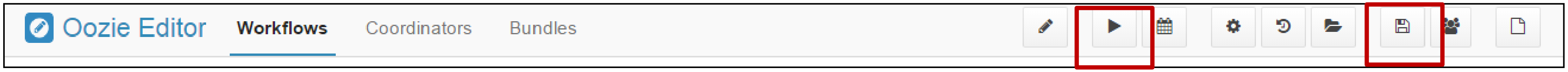
Add username and pwd
After completion, check the table in MySQL

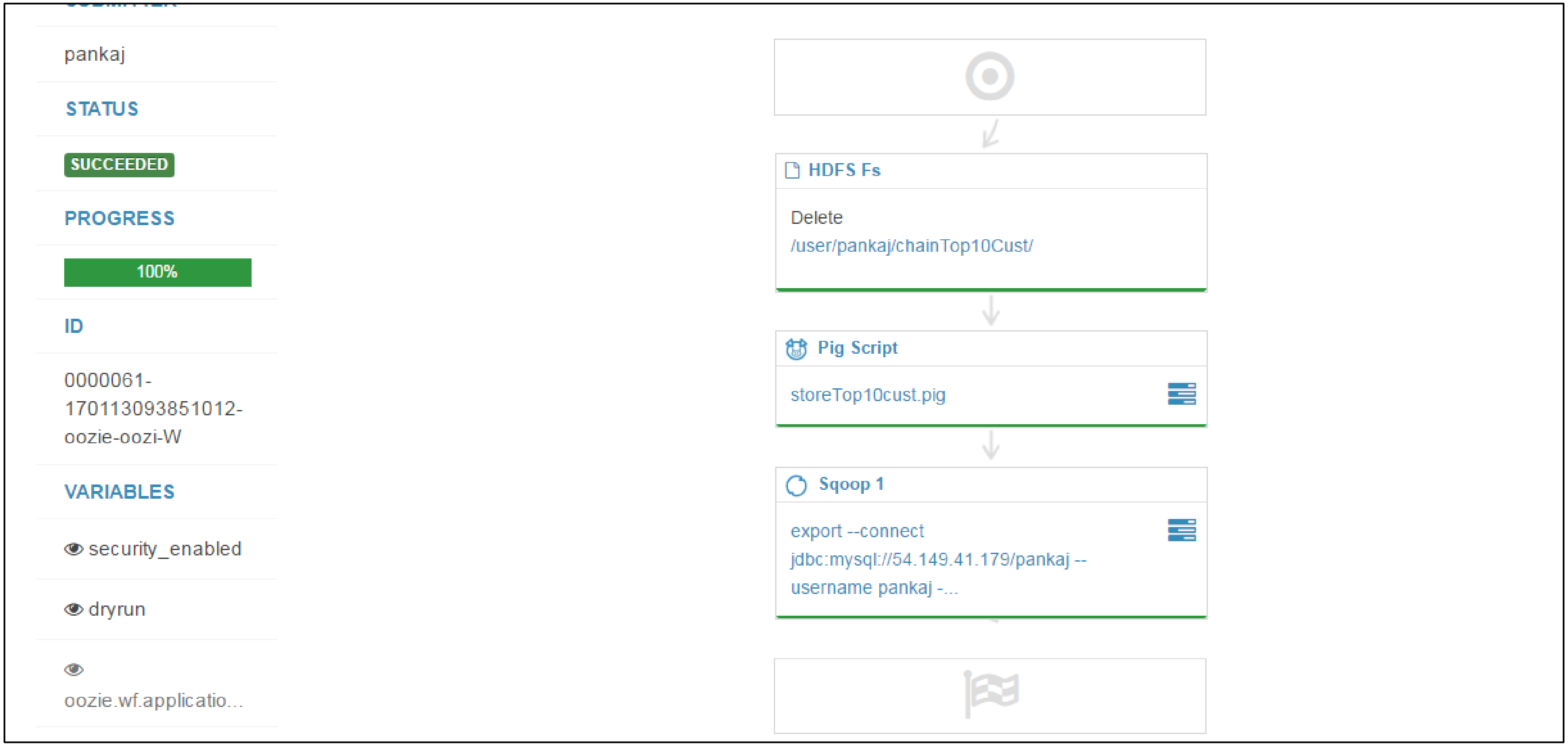
Pig <--> Sqoop <--> MySQL– top 10 customers in each chain - Submit

Save and submit.

Submit

Save





pankaj

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000061-170113093851012-oozie-oozi-W

VARIABLES

security_enabled

dryrun

oozie.wf.applicatio...

HDFS Fs

Delete /user/pankaj/chainTop10Cust/

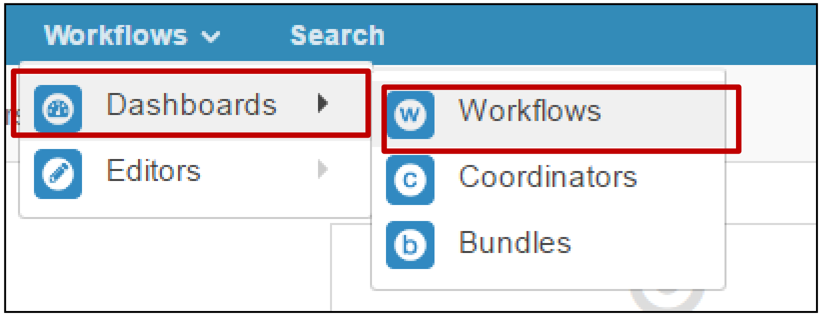
Pig Script

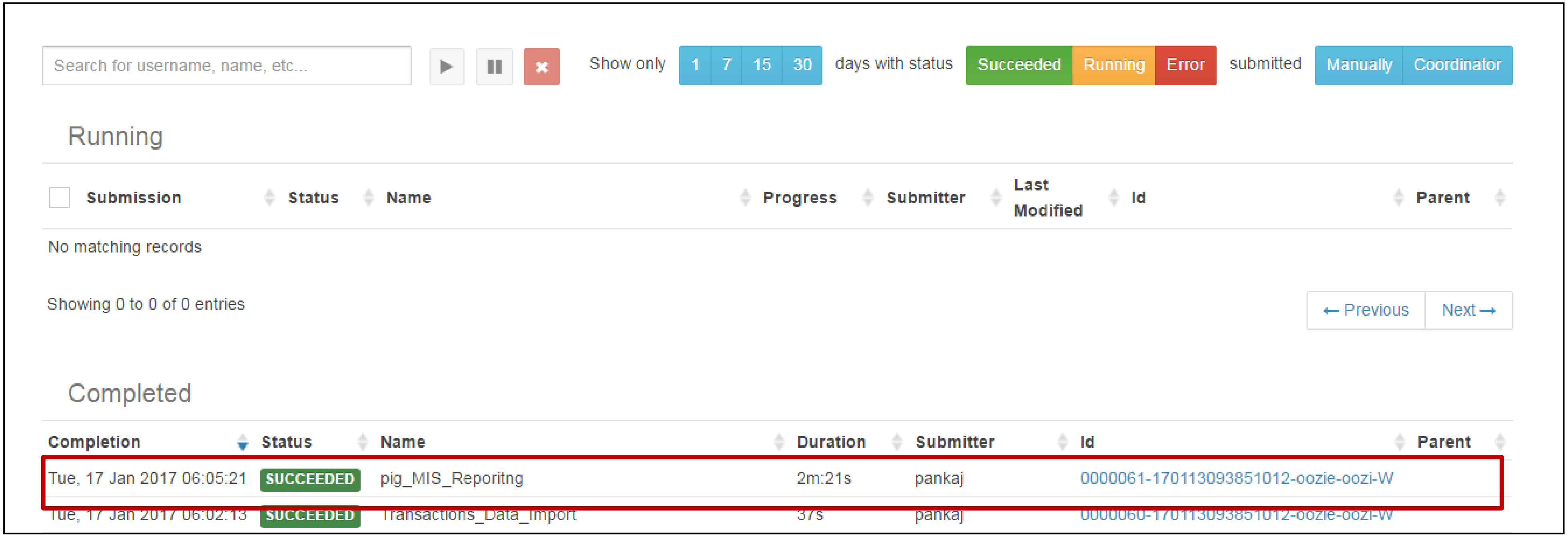
storeTop10cust.pig

Sqoop 1

export --connect jdbc:mysql://54.149.41.179/pankaj --username pankaj -...

Pig <--> Sqoop <--> MySQL– top 10 customers in each chain – Track progress





Running

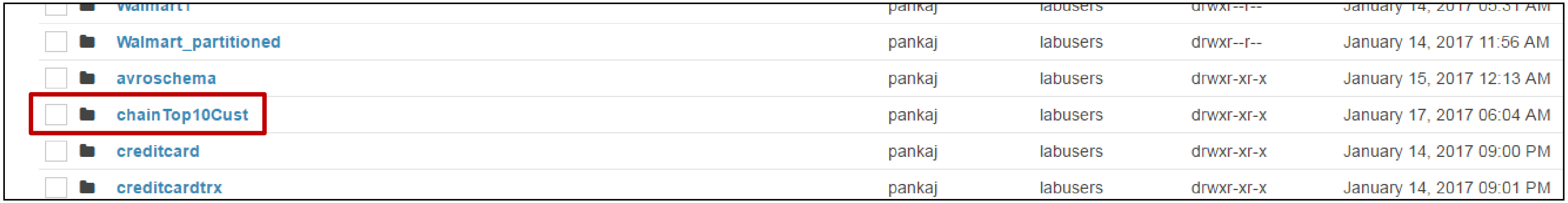
□ Submission	Status	Name	Progress	Submitter	Last Modified	Id	Parent
No matching records							

Showing 0 to 0 of 0 entries

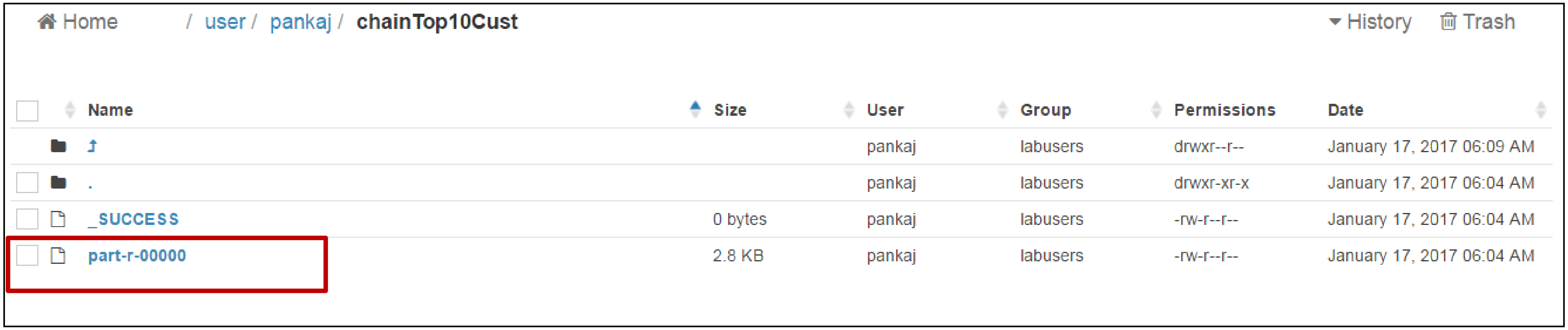
Completed

Completion	Status	Name	Duration	Submitter	Id	Parent
Tue, 17 Jan 2017 06:05:21	SUCCEEDED	pig_MIS_Reporitng	2m:21s	pankaj	0000061-170113093851012-oozie-oozi-W	
Tue, 17 Jan 2017 06:02:13	SUCCEEDED	Transactions_Data_Import	37s	pankaj	0000060-170113093851012-oozie-oozi-W	

Pig <--> Sqoop <--> MySQL, pig output check



	Name	User	Group	Permissions	Date
□	Walmart1	pankaj	labusers	drwxr--r--	January 14, 2017 05:31 AM
□	Walmart_partitioned	pankaj	labusers	drwxr--r--	January 14, 2017 11:56 AM
□	avroschema	pankaj	labusers	drwxr-xr-x	January 15, 2017 12:13 AM
□	chainTop10Cust	pankaj	labusers	drwxr-xr-x	January 17, 2017 06:04 AM
□	creditcard	pankaj	labusers	drwxr-xr-x	January 14, 2017 09:00 PM
□	creditcardtrx	pankaj	labusers	drwxr-xr-x	January 14, 2017 09:01 PM



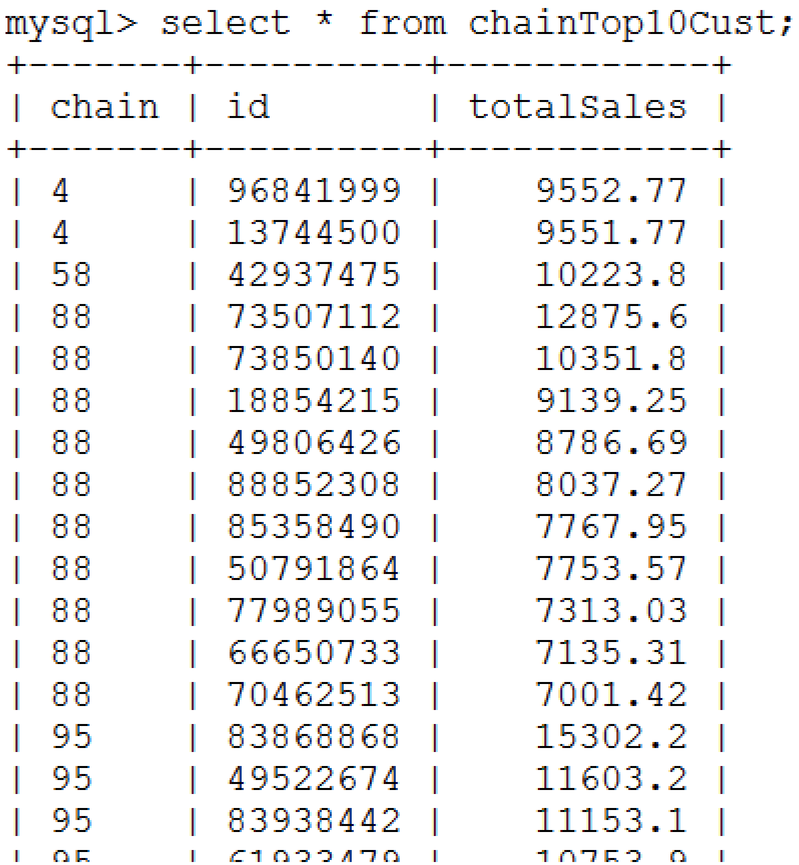
Home / user / pankaj / **chainTop10Cust**

□	Name	Size	User	Group	Permissions	Date
	↥		pankaj	labusers	drwxr--r--	January 17, 2017 06:09 AM
□	.		pankaj	labusers	drwxr-xr-x	January 17, 2017 06:04 AM
□	_SUCCESS	0 bytes	pankaj	labusers	-rw-r--r--	January 17, 2017 06:04 AM
□	part-r-00000	2.8 KB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:04 AM

Pig <--> Sqoop <--> MySQL, Sqoop export check

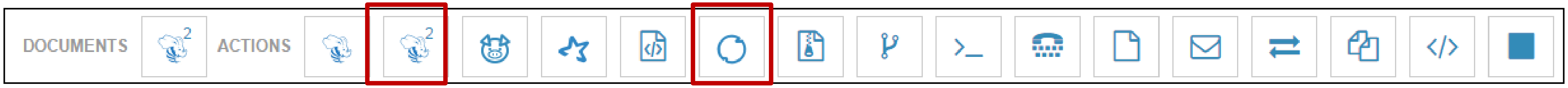
Login to MySQL :

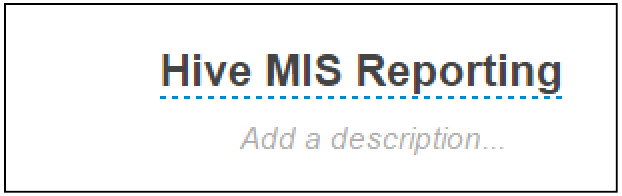
mysql -h 54.149.41.179 -u username-p



```
mysql> select * from chainTop10Cust;
+-------+----------+-----------+
| chain | id       | totalSales |
+-------+----------+-----------+
| 4     | 96841999 |    9552.77 |
| 4     | 13744500 |    9551.77 |
| 58    | 42937475 |    10223.8 |
| 88    | 73507112 |    12875.6 |
| 88    | 73850140 |    10351.8 |
| 88    | 18854215 |    9139.25 |
| 88    | 49806426 |    8786.69 |
| 88    | 88852308 |    8037.27 |
| 88    | 85358490 |    7767.95 |
| 88    | 50791864 |    7753.57 |
| 88    | 77989055 |    7313.03 |
| 88    | 66650733 |    7135.31 |
| 88    | 70462513 |    7001.42 |
| 95    | 83868868 |    15302.2 |
| 95    | 49522674 |    11603.2 |
| 95    | 83938442 |    11153.1 |
| 95    | 61933479 |    10753.9 |
```

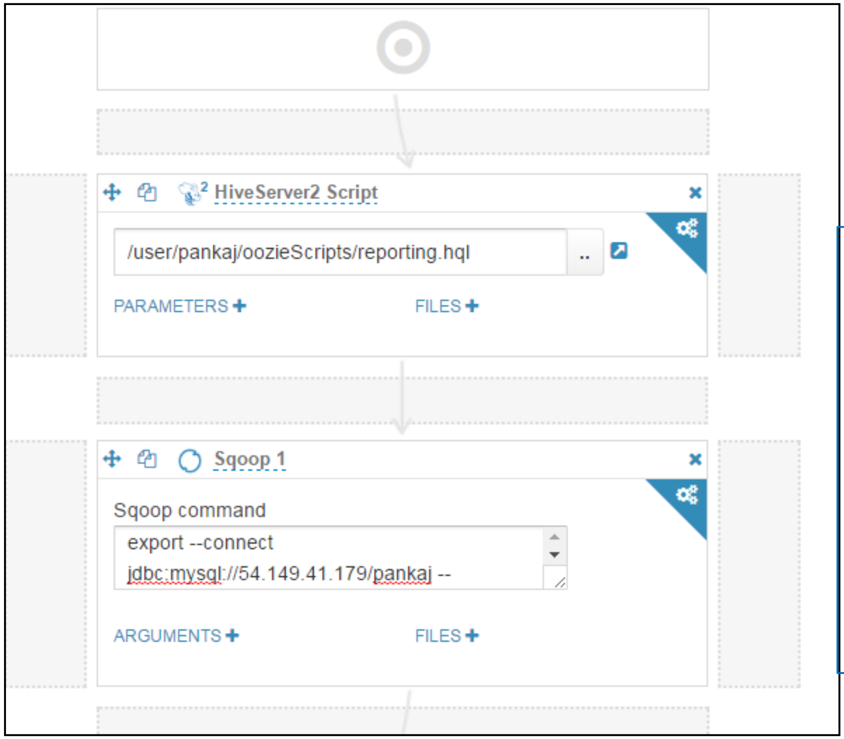
Hive<--> Sqoop <--> MySQL – Chain Report




Hive MIS Reporting

Add a description...

Name of the workflow


HiveServer2 Script

/user/pankaj/oozieScripts/reporting.hql

Sqoop 1

Sqoop command

```
export --connect
jdbc:mysql://54.149.41.179/pankaj --
```

Path of hive script:

Browse and select the hive file in HDFs.

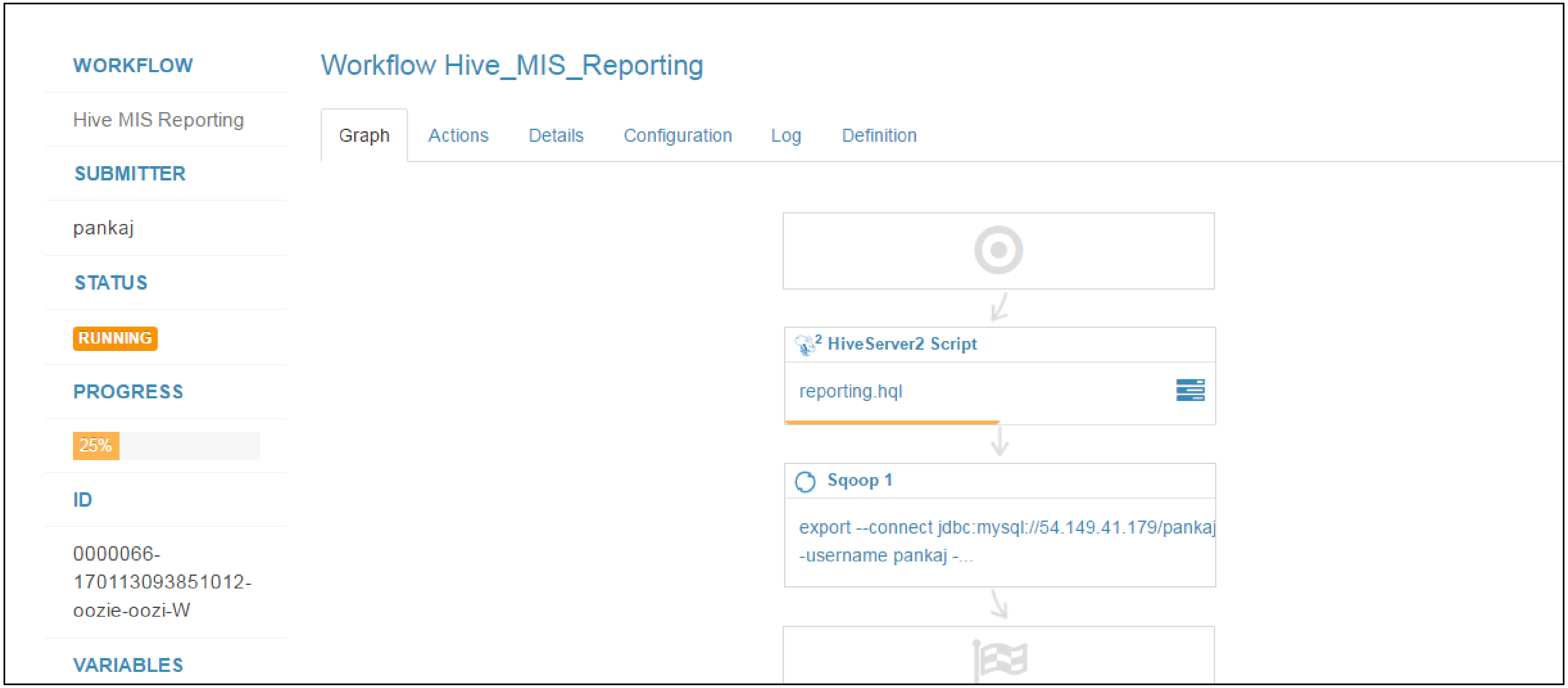
Sqoop Export o MYSQL

Add the following command to export results to MySQL:

export --connect jdbc:mysql://54.149.41.179/pankaj --username username -password pwd --table chain_stats --export-dir /user/hive/warehouse/oozie_test.db/chain_stats/ --input-fields-terminated-by "\001“

Add username and pwd
After completion, check the table in MySQL

Hive – Chain Report - Submit



WORKFLOW

Hive MIS Reporting

SUBMITTER

pankaj

STATUS

RUNNING

PROGRESS

25%

ID

0000066-170113093851012-oozie-oozi-W

VARIABLES

Workflow Hive_MIS_Reporting

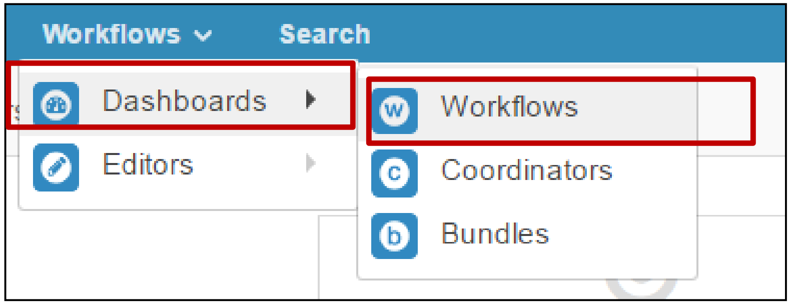
HiveServer2 Script

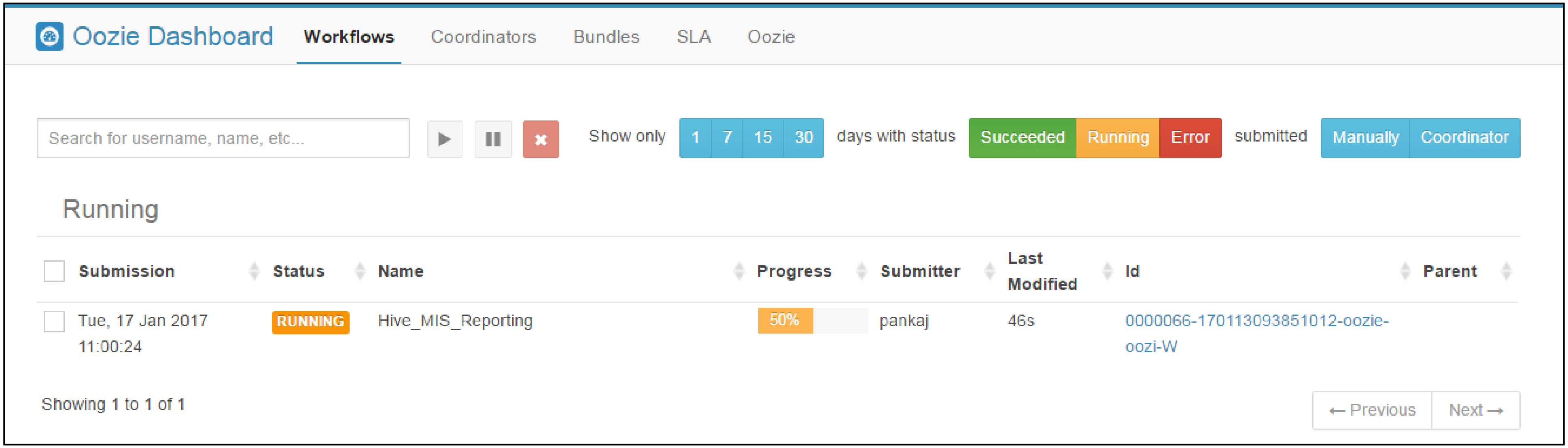
reporting.hql

Sqoop 1

export --connect jdbc:mysql://54.149.41.179/pankaj -username pankaj -...

Hive<--> Sqoop <--> MySQL Track Progress





Running

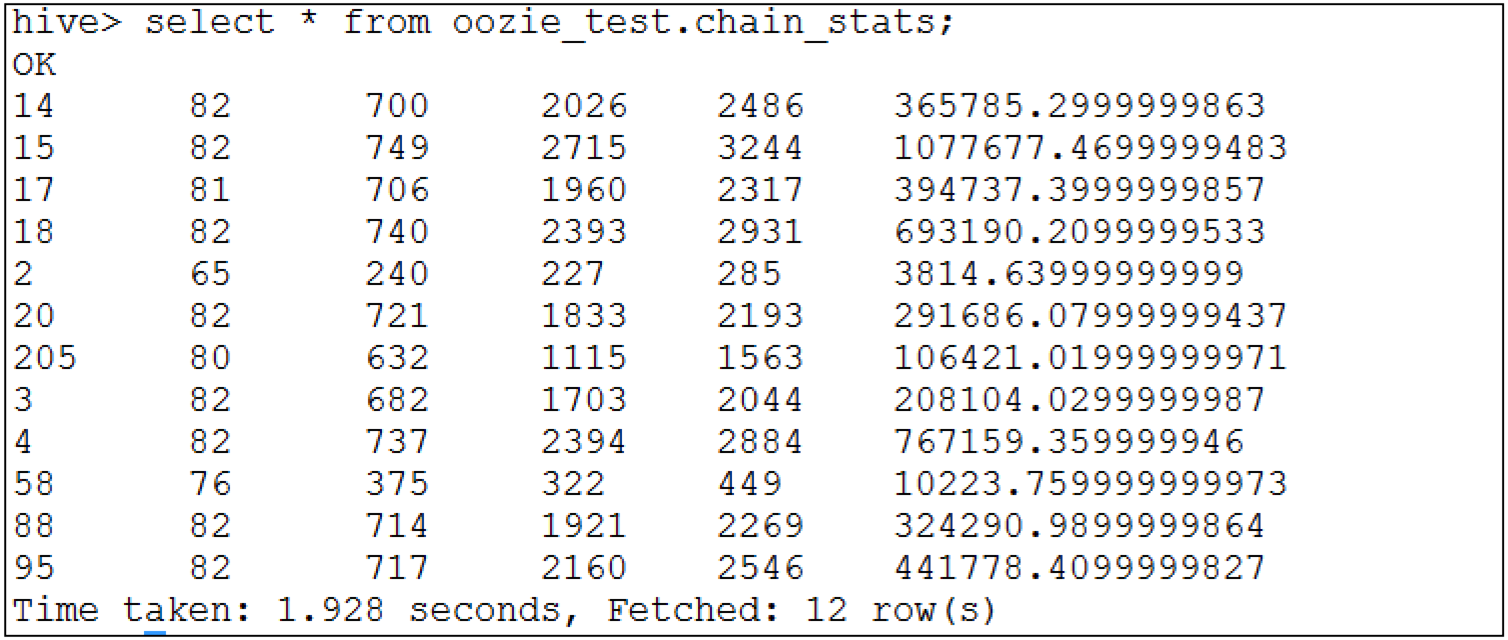
□	Submission	Status	Name	Progress	Submitter	Last Modified	Id	Parent
□	Tue, 17 Jan 2017 11:00:24	RUNNING	Hive_MIS_Reporting	50%	pankaj	46s	0000066-170113093851012-oozie-oozi-W	

Showing 1 to 1 of 1

Hive<--> Sqoop <--> MySQL Hive report check

Login to hive, or use Hue.

Check in the oozie database created.

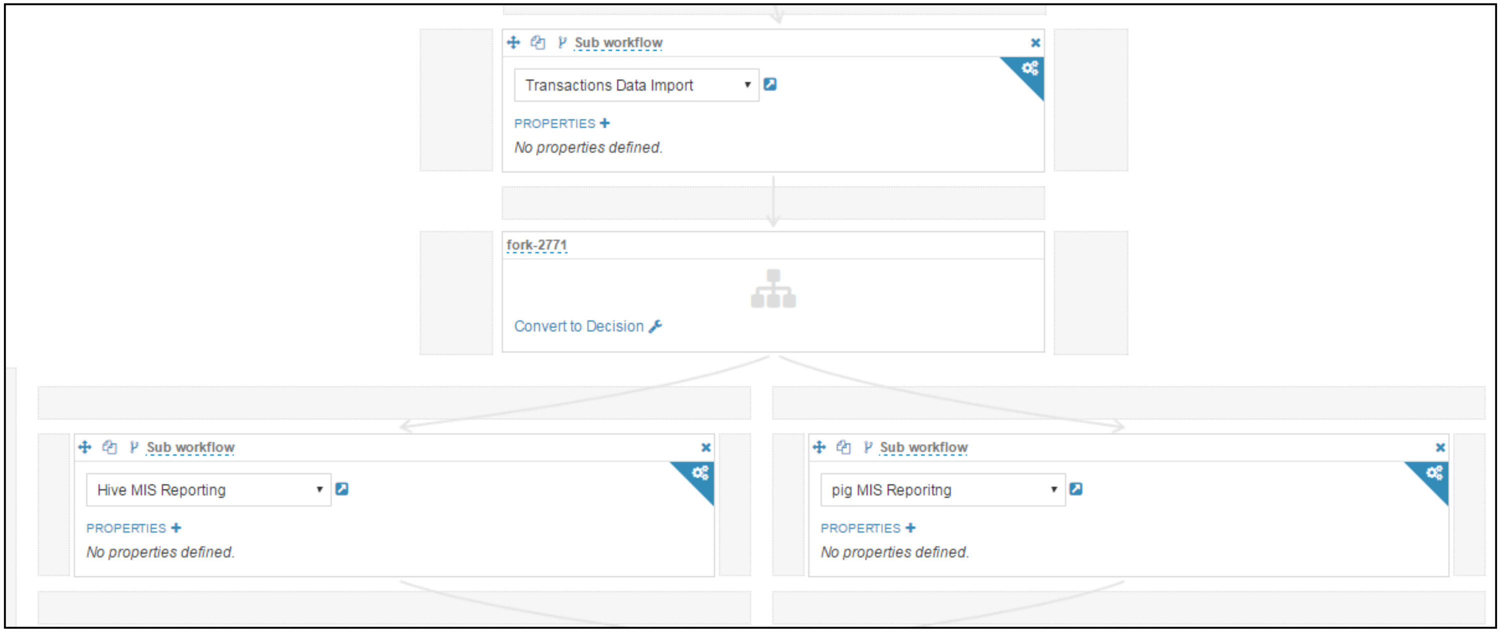


```
hive> select * from oozie_test.chain_stats;
OK
14      82      700     2026    2486    365785.2999999863
15      82      749     2715    3244    1077677.4699999483
17      81      706     1960    2317    394737.3999999857
18      82      740     2393    2931    693190.2099999533
2       65      240     227     285     3814.6399999999
20      82      721     1833    2193    291686.07999999437
205     80      632     1115    1563    106421.01999999971
3       82      682     1703    2044    208104.0299999987
4       82      737     2394    2884    767159.359999946
58      76      375     322     449     10223.759999999973
88      82      714     1921    2269    324290.9899999864
95      82      717     2160    2546    441778.4099999827
Time taken: 1.928 seconds, Fetched: 12 row(s)
```

Hive<--> Sqoop <--> MySQL Sqoop MySQL export check

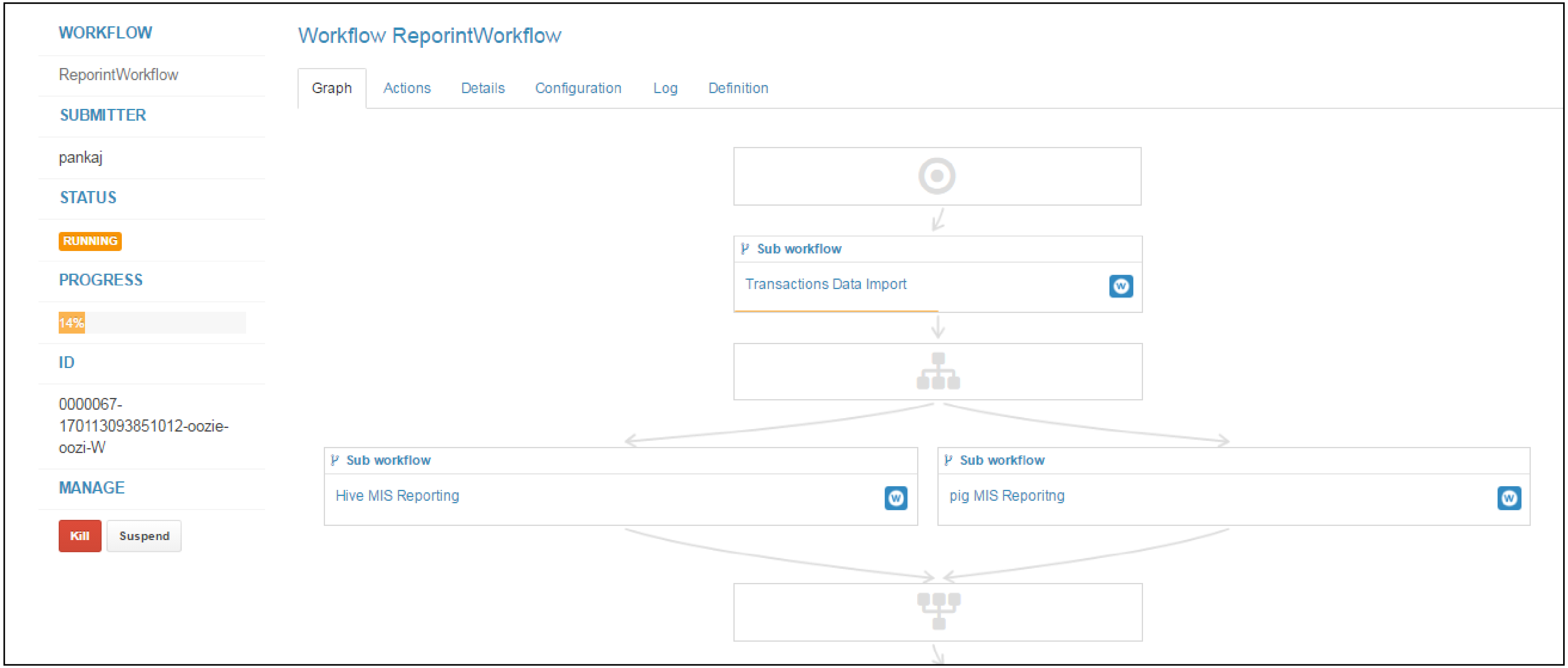
Combining all three workflows in to single workflow.

Create a new workflow.
Use sub workflow action.
Step1: Sqoop import work flow.
Step2: Hive and Pig reporting workflows in parallel.



Save it as reporting workflow.
Run and see what happens!

Submit



WORKFLOW

ReporintWorkflow

SUBMITTER

pankaj

STATUS

RUNNING

PROGRESS

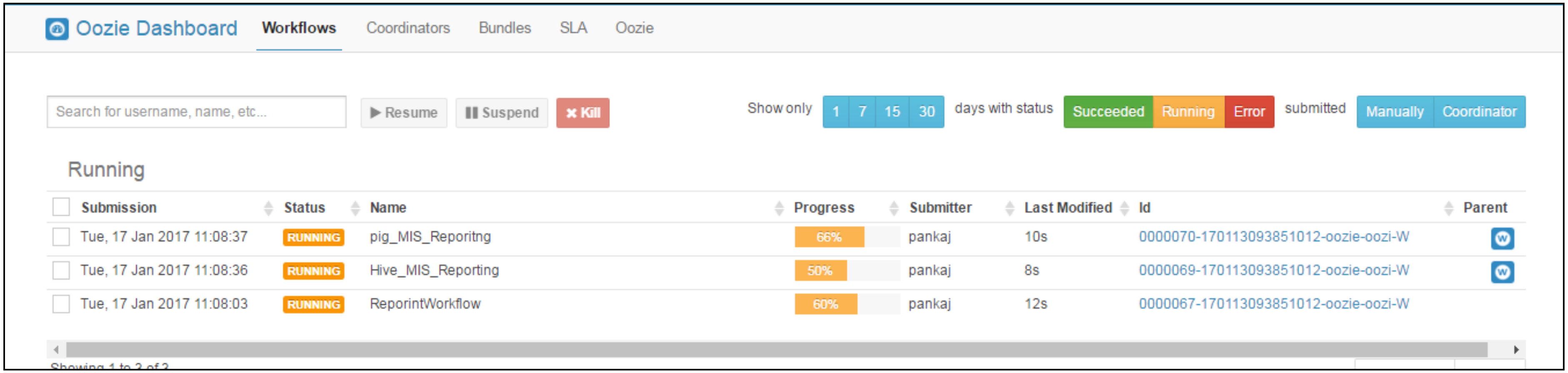
14%

ID

0000067-170113093851012-oozie-oozi-W

MANAGE

Progress



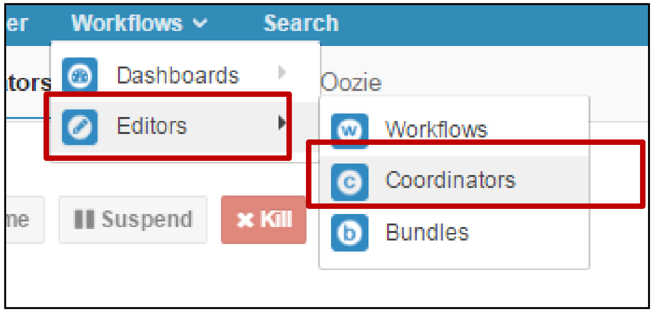
Running

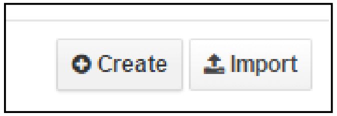
□	Submission	Status	Name	Progress	Submitter	Last Modified	Id	Parent
□	Tue, 17 Jan 2017 11:08:37	RUNNING	pig_MIS_Reporitng	66%	pankaj	10s	0000070-170113093851012-oozie-oozi-W	W
□	Tue, 17 Jan 2017 11:08:36	RUNNING	Hive_MIS_Reporting	50%	pankaj	8s	0000069-170113093851012-oozie-oozi-W	W
□	Tue, 17 Jan 2017 11:08:03	RUNNING	ReporintWorkflow	60%	pankaj	12s	0000067-170113093851012-oozie-oozi-W	

Showing 1 to 3 of 3

Coordinating workflow

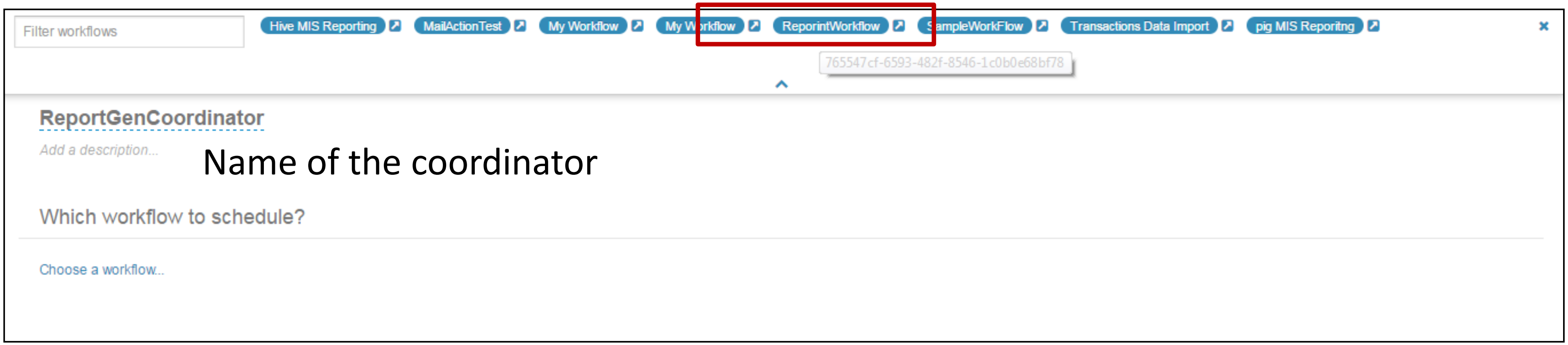
Add the previous workflow and schedule it.





Create a new one

Select this workflow



ReportGenCoordinator

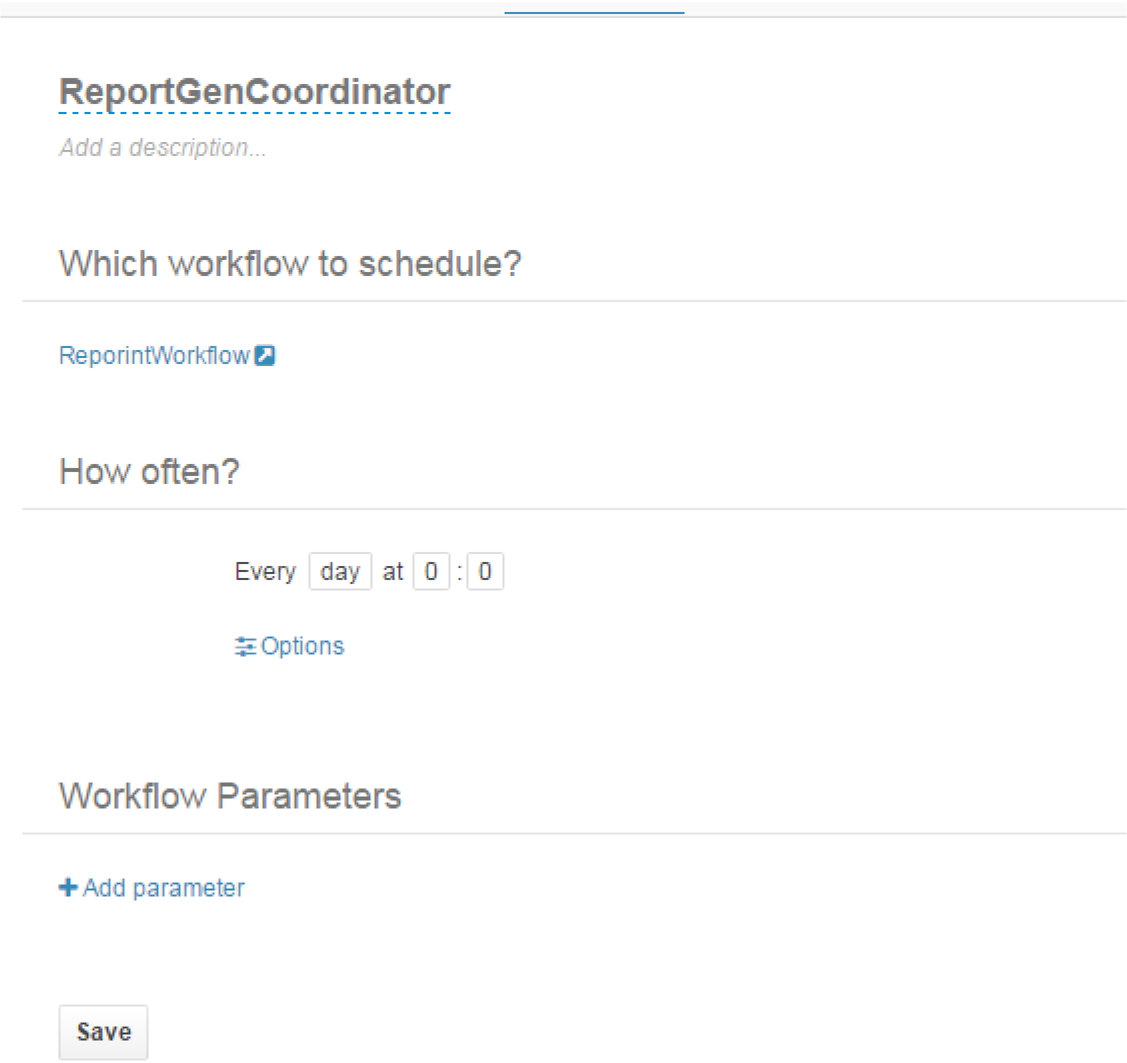
Add a description...

Name of the coordinator

Which workflow to schedule?

Choose a workflow...

Schedule it



ReportGenCoordinator

Add a description...

Which workflow to schedule?

ReporintWorkflow

How often?

Every day at 0 : 0

Options

Workflow Parameters

Conclusion

You learnt the following:

1. Creating sqoop action
2. Creating HDFS action
3. Creating Hive action
4. Creating Pig Action
5. Creating Workflows
6. Integrating sub-workflows
7. Coordinating workflows