

## Oozie

Complex work flow management

Relief to developers!!

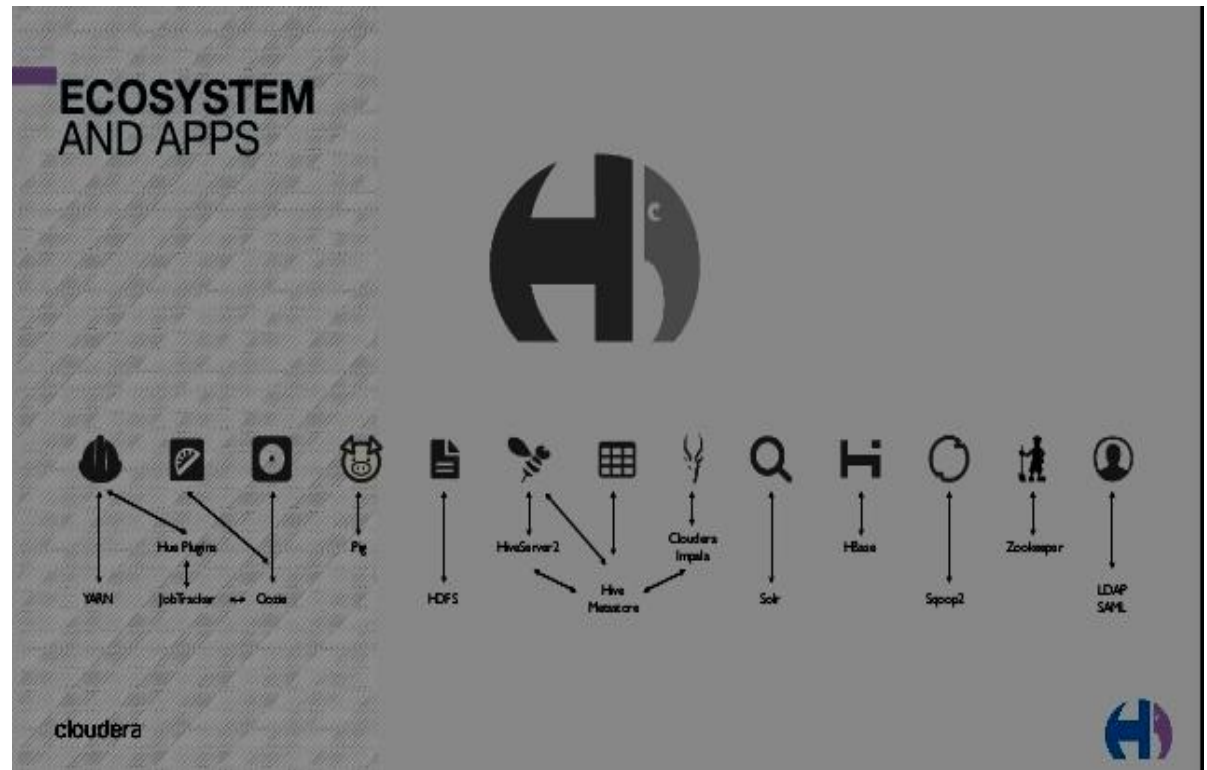


# Agenda

- 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

Welcome to Hue  
Please sign in to continue

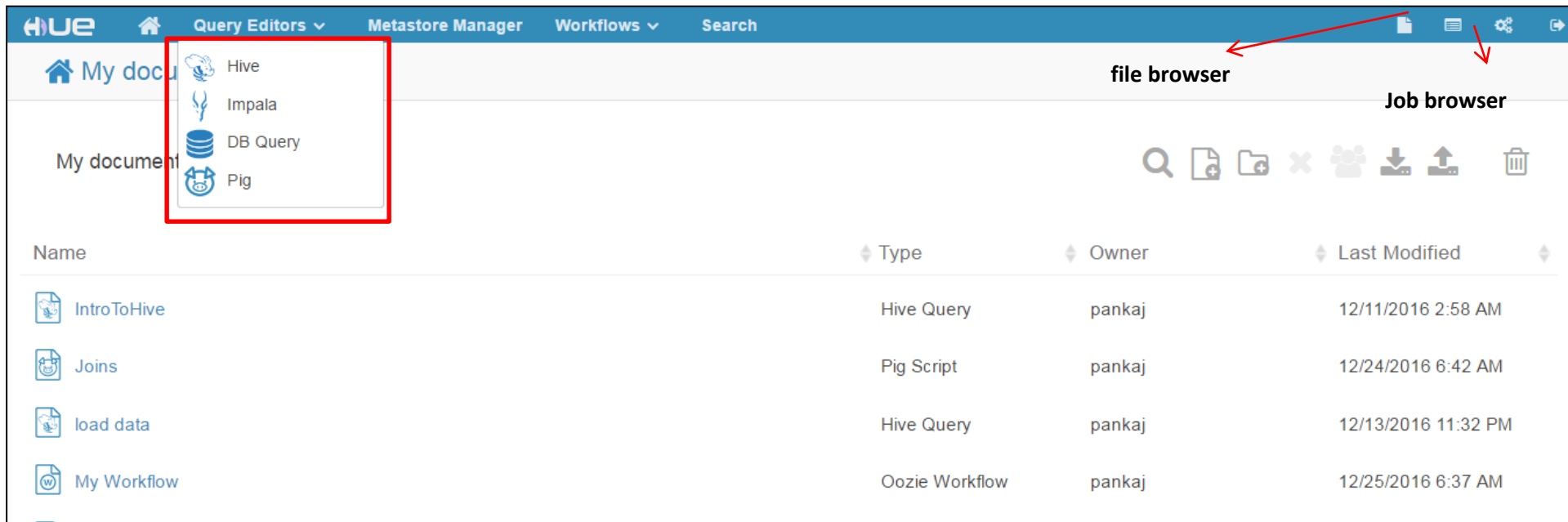


pankaj

.....

Sign in

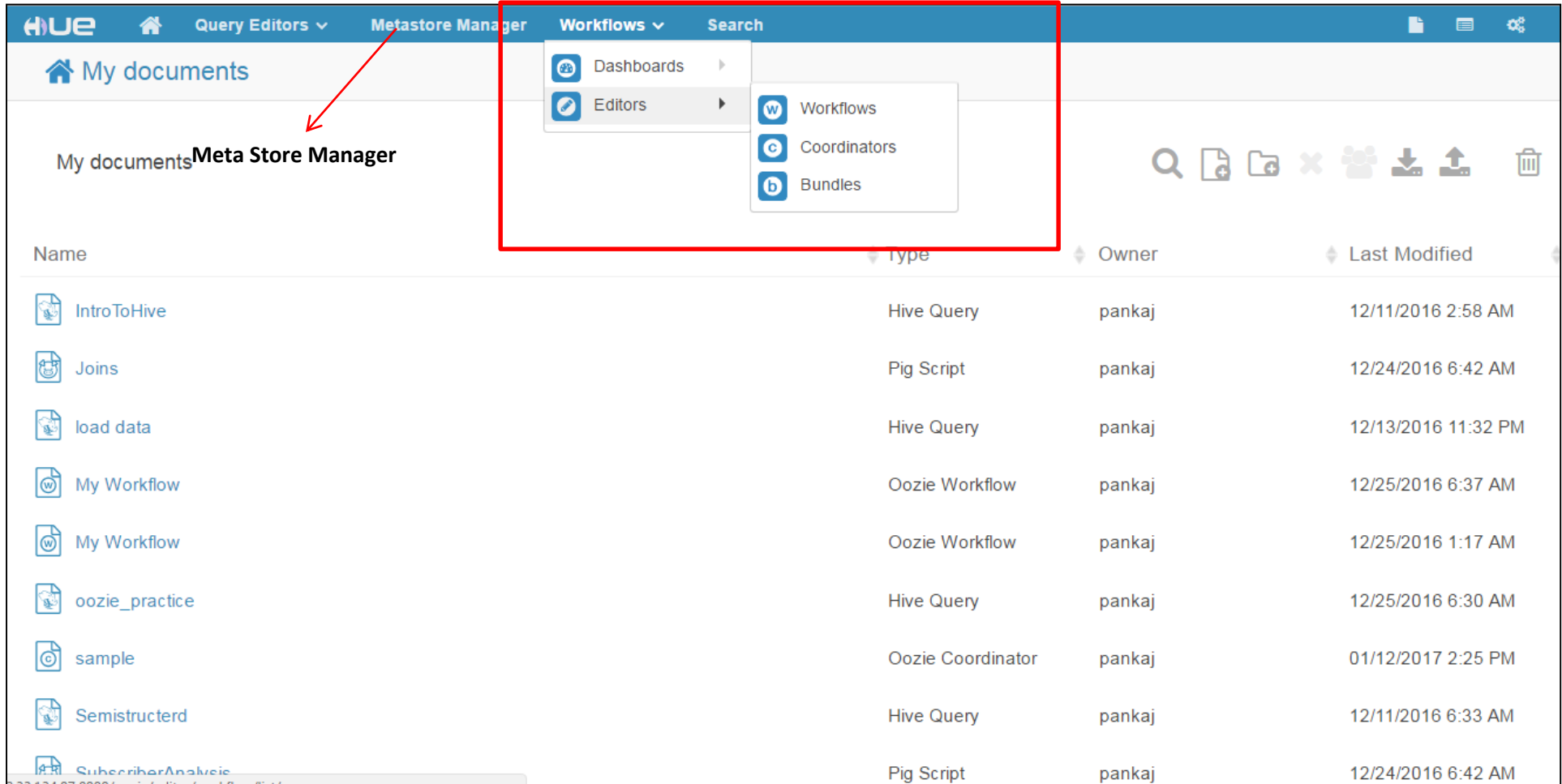
# Query Editors, Job and File Browsers



The screenshot shows the HUE web interface. The top navigation bar includes 'HUE', a home icon, 'Query Editors', 'Metastore Manager', 'Workflows', and 'Search'. A dropdown menu for 'Query Editors' is open, showing options: Hive, Impala, DB Query, and Pig. Below the navigation bar, there are two tabs: 'file browser' and 'Job browser'. The 'file browser' tab is active, displaying a table of documents. The table has columns: Name, Type, Owner, and Last Modified. The documents listed are 'IntroToHive' (Hive Query), 'Joins' (Pig Script), 'load data' (Hive Query), and 'My Workflow' (Oozie Workflow). All documents are owned by 'pankaj'.

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

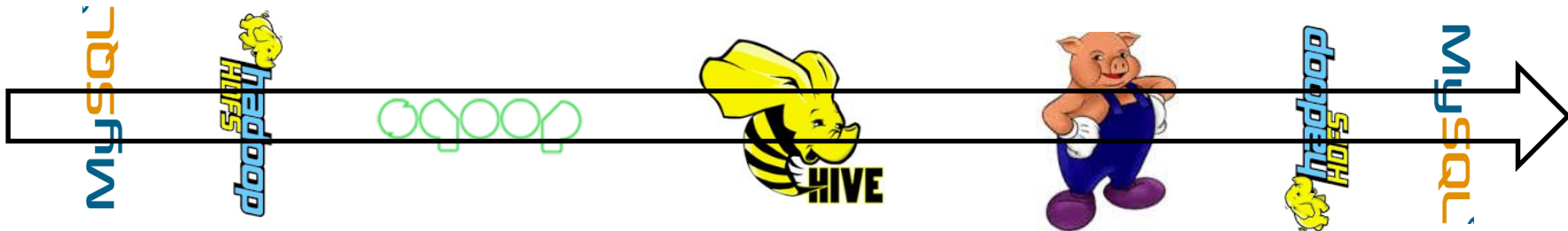


The screenshot displays the Hue web interface. The top navigation bar includes 'Query Editors', 'Metastore Manager', 'Workflows', and 'Search'. The 'Workflows' dropdown menu is expanded, showing options for 'Dashboards', 'Editors', 'Workflows', 'Coordinators', and 'Bundles'. A red box highlights this menu. A red arrow points to the 'Meta Store Manager' link in the 'My documents' section. Below the navigation bar, there is a table listing documents with columns for Name, Type, Owner, and Last Modified.

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





## 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!

## 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



**Commercial tool, reliable fast and easy to use**



**Not a good tool for Big Data, too costly!**



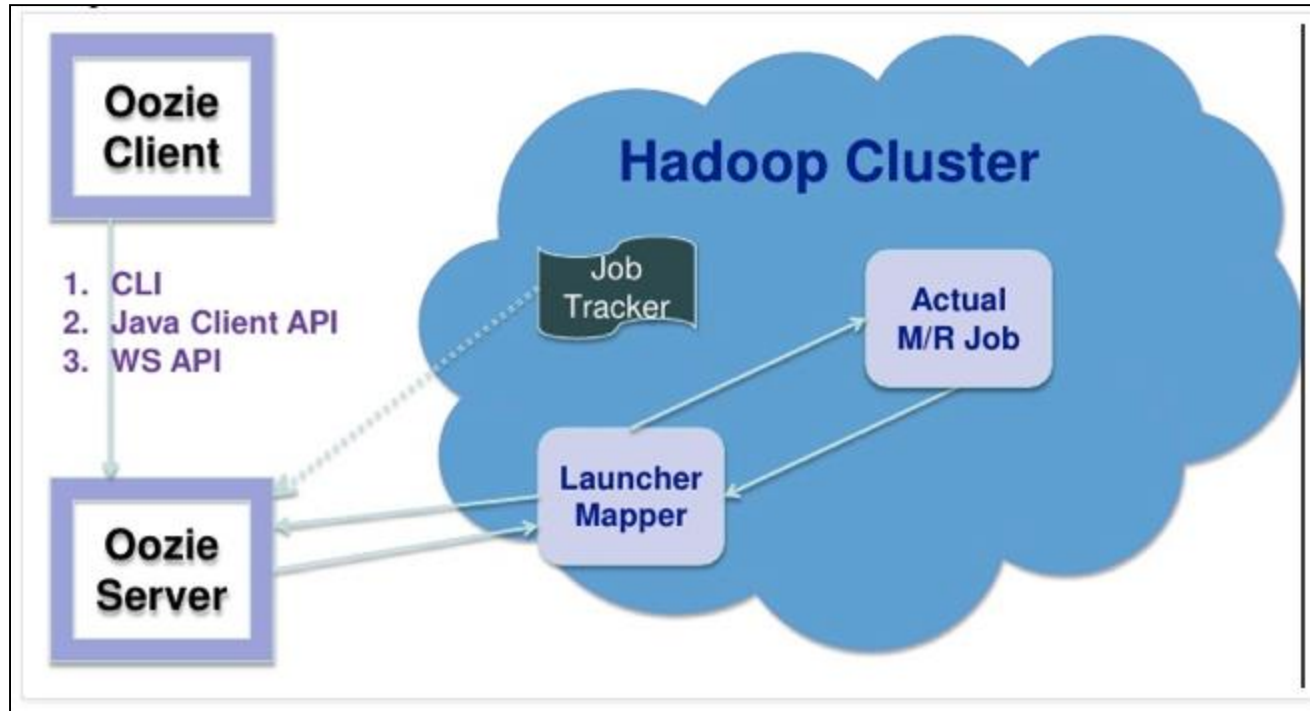
**Good Competitor to talend**



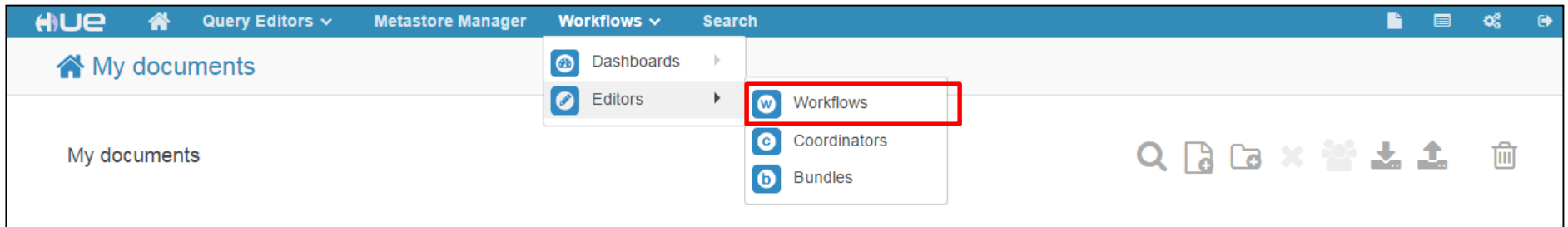
**Azkaban, open source workflow manager, similar to Oozie**

# 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



Reference: yahoo



- 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.

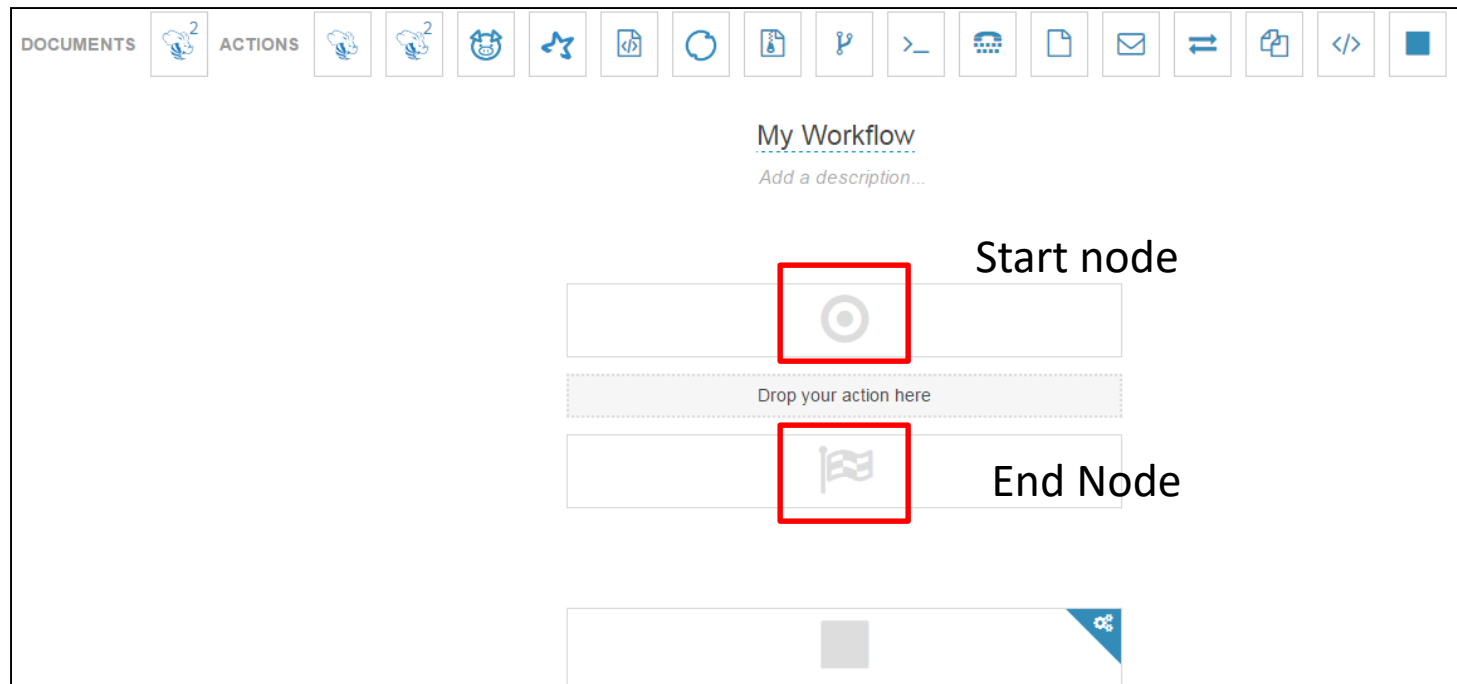
# Control Flow Nodes

- **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

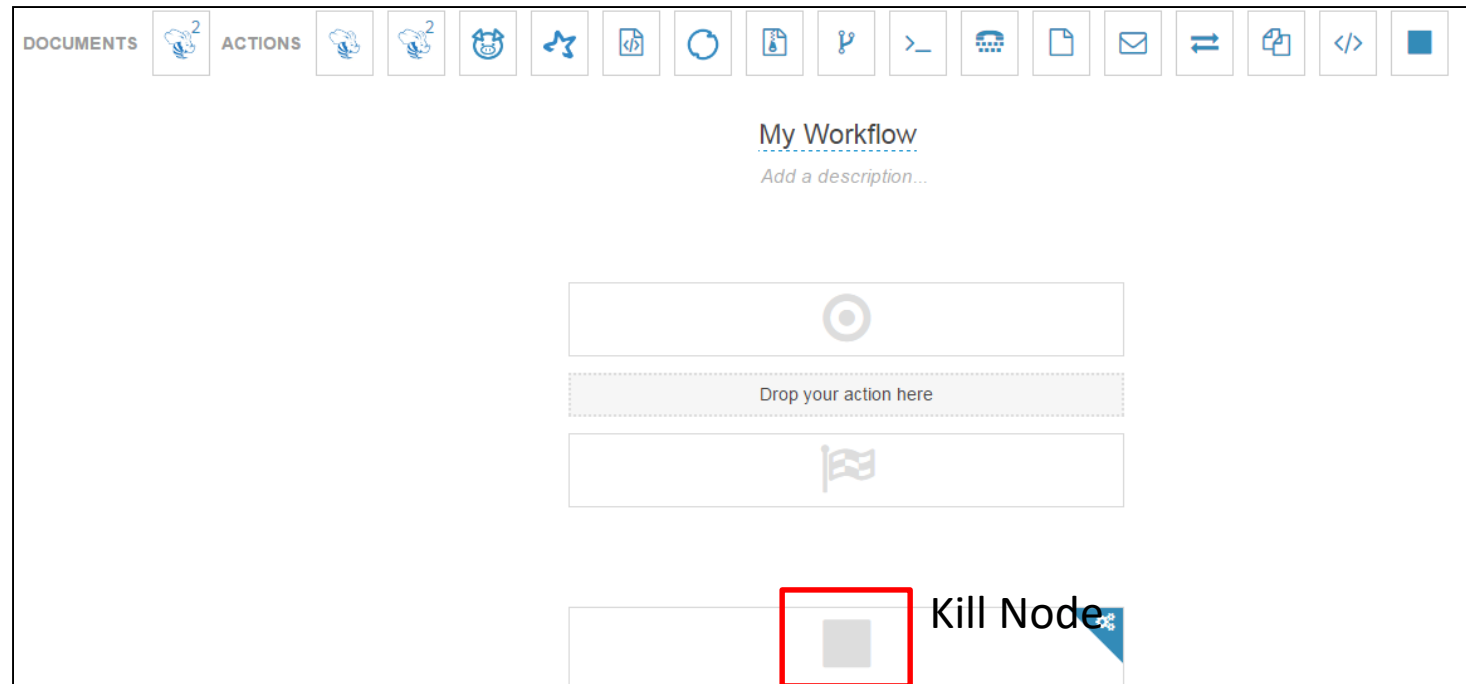




# Control Flow Nodes

- **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



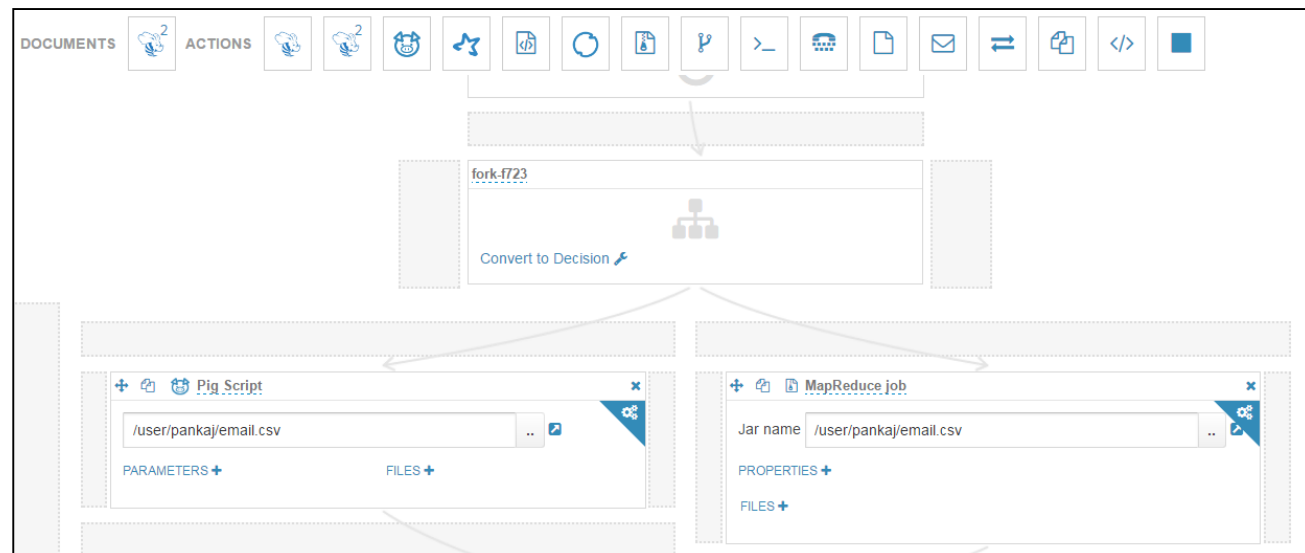
# Control Flow nodes

## ■ 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

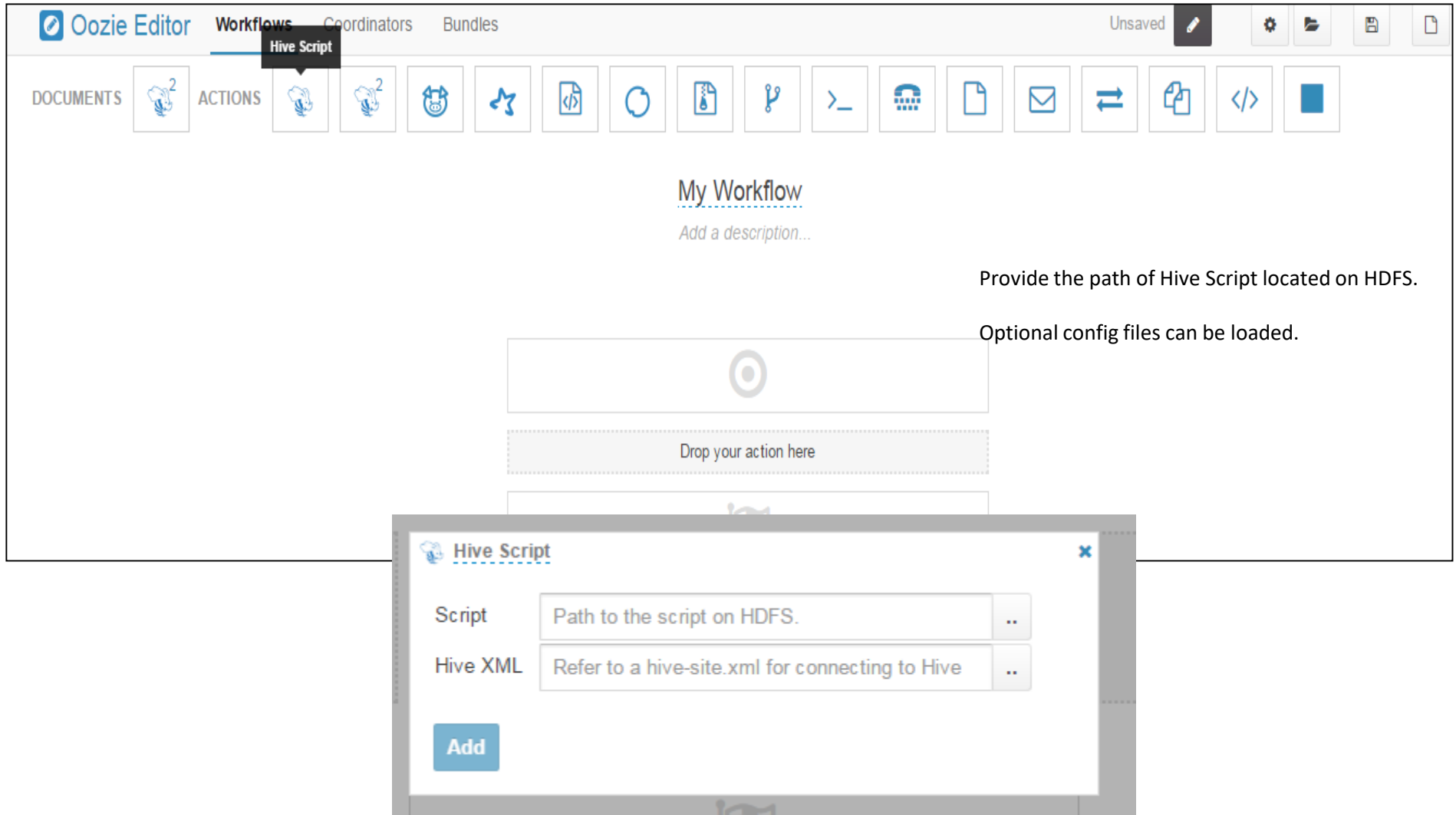
Decision node allows a workflow job to make a selection between its execution paths based on a list of predicate. The predicates are evaluated on order of appearance until one of them comes true. In case all return false, default transition is taken. These predicate can contain logics such as the size of file being greater than some threshold, or the file being completely loaded or the exit status of an action node



# Oozie Actions

- Action nodes are generally used for performing computation tasks
- No Computation performed by Action nodes takes within oozie, all computation 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

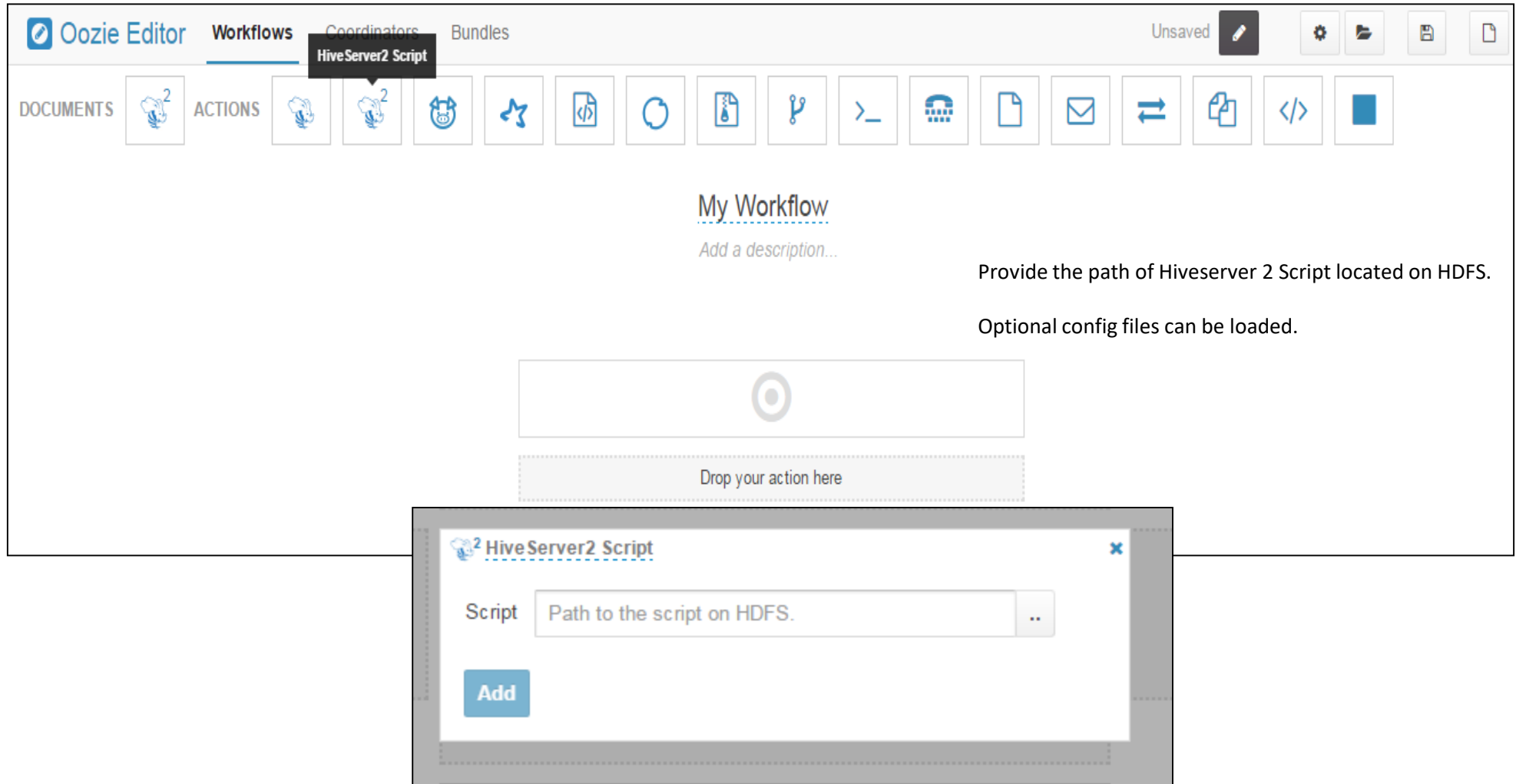


The screenshot shows the Oozie Editor interface with the 'Workflows' tab selected. The 'ACTIONS' panel on the left contains a 'Hive Script' icon. The main workspace displays 'My Workflow' with a description field. A 'Hive Script' dialog box is open, prompting for the script path and Hive XML configuration files.

Provide the path of Hive Script located on HDFS.

Optional config files can be loaded.

# Hive Server2 Action



The screenshot shows the Oozie Editor interface with the 'Workflows' tab selected. A tooltip for 'HiveServer2 Script' is visible over the action icon in the 'ACTIONS' palette. The main workspace displays 'My Workflow' with a description field and a target icon. A modal dialog titled 'HiveServer2 Script' is open, showing a 'Script' field with the placeholder text 'Path to the script on HDFS.' and an 'Add' button.

Oozie Editor Workflows Coordinators Bundles Unsaved

DOCUMENTS ACTIONS

My Workflow

Add a description...

Drop your action here

HiveServer2 Script

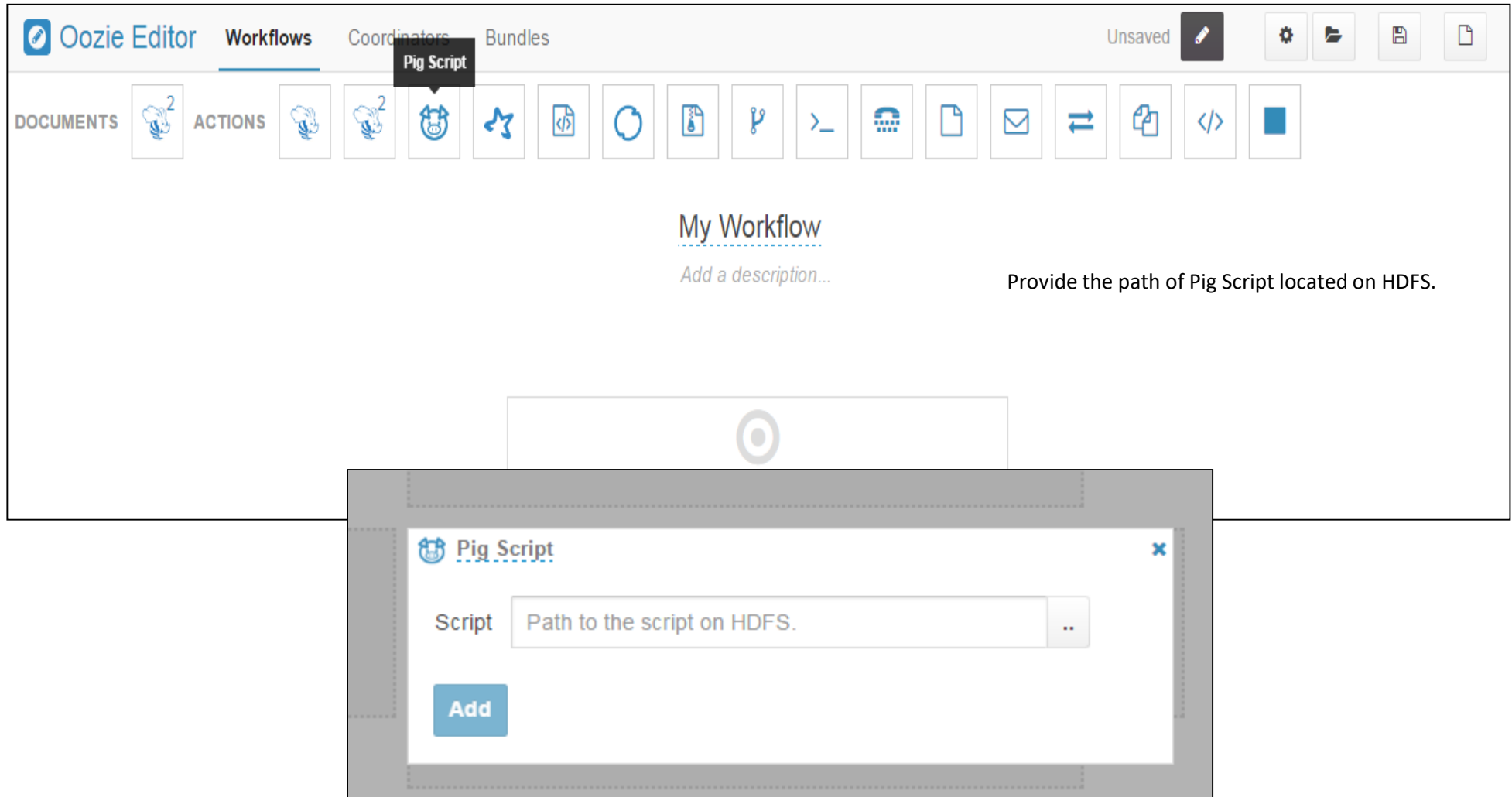
Script Path to the script on HDFS.

Add

Provide the path of Hiveserver 2 Script located on HDFS.

Optional config files can be loaded.

# Pig Action



The screenshot shows the Oozie Editor interface with the 'Workflows' tab selected. The 'Pig Script' action is highlighted in the 'ACTIONS' toolbar. The main workspace displays 'My Workflow' with a description field and a text prompt: 'Provide the path of Pig Script located on HDFS.' A modal dialog titled 'Pig Script' is open, showing a 'Script' field with the placeholder text 'Path to the script on HDFS.' and an 'Add' button.

Oozie Editor Workflows Coordinators Bundles Unsaved

DOCUMENTS ACTIONS

**Pig Script**

My Workflow  
Add a description...

Provide the path of Pig Script located on HDFS.

**Pig Script**

Script Path to the script on HDFS. ..

Add

# Spark Action

Oozie Editor Workflows Coordinators Bundles Unsaved

DOCUMENTS ACTIONS

Spark program

My Workflow

Add a description...

Drop your action here

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

Spark

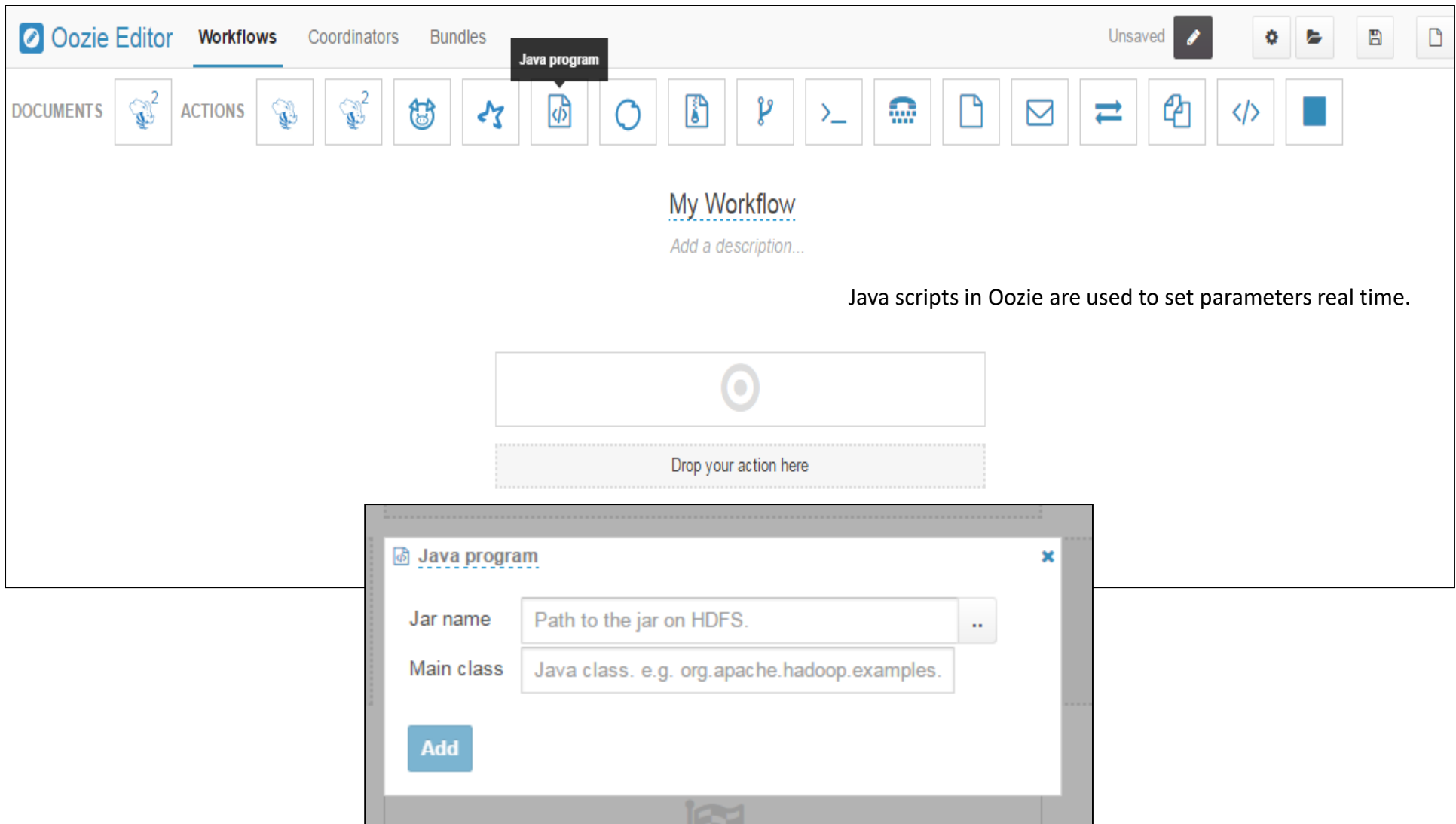
Spark Master local[\*]

Mode client

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

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

Add

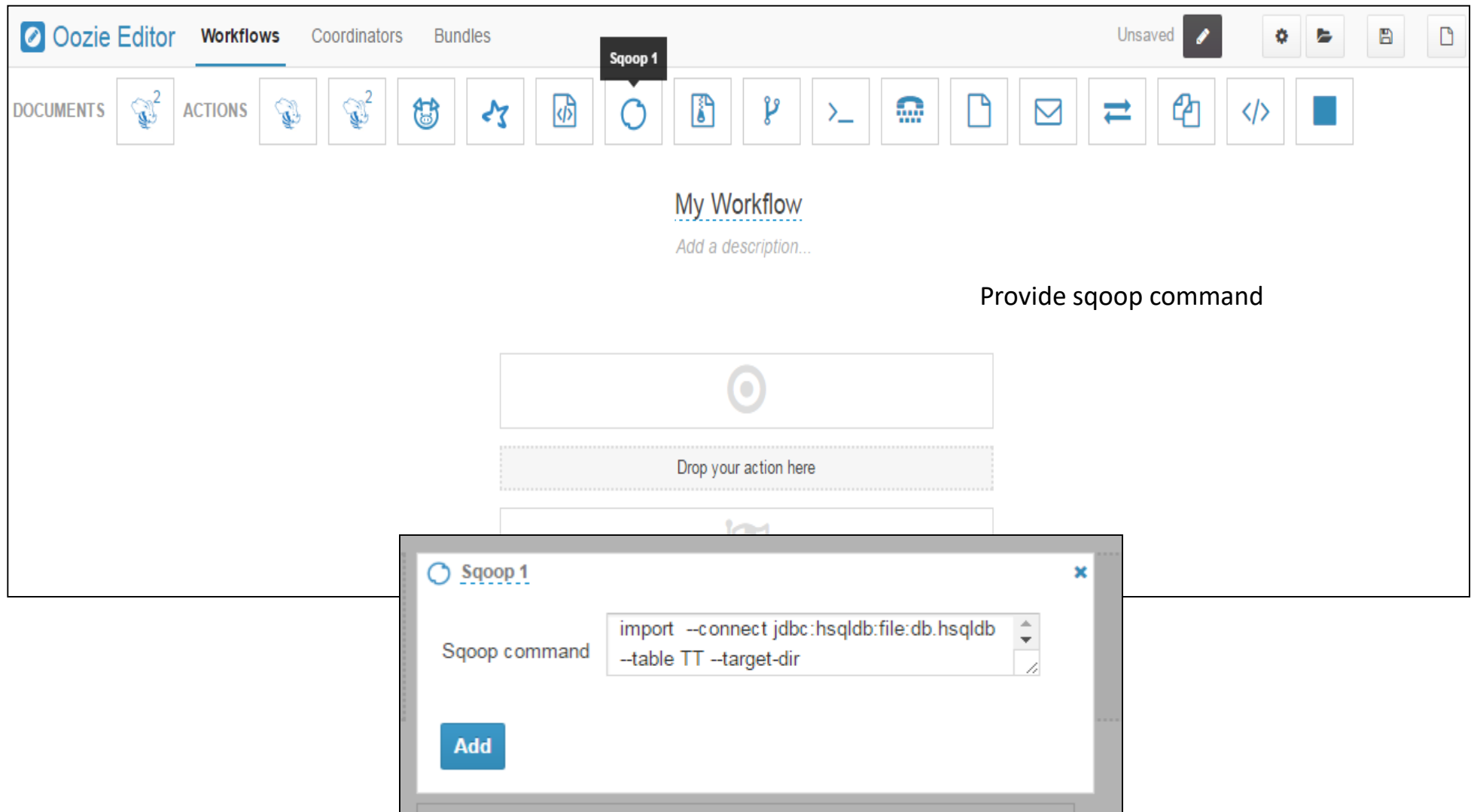


The screenshot displays the Oozie Editor interface. At the top, there are tabs for 'Workflows', 'Coordinators', and 'Bundles'. The 'Workflows' tab is active. Below the tabs, there is a toolbar with various icons. A tooltip labeled 'Java program' points to the icon representing a Java program. The main workspace shows a workflow diagram with a box labeled 'My Workflow' and a description 'Add a description...'. Below this, there is a large rectangular area with a target icon and the text 'Drop your action here'. A modal dialog box titled 'Java program' is open in the foreground, containing two input fields: 'Jar name' with the placeholder text 'Path to the jar on HDFS.' and 'Main class' with the placeholder text 'Java class. e.g. org.apache.hadoop.examples.'. An 'Add' button is located at the bottom of the dialog.

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



# Sqoop Action



The screenshot displays the Oozie Editor interface. At the top, there are tabs for 'Workflows', 'Coordinators', and 'Bundles'. The 'Workflows' tab is active. Below the tabs, there is a toolbar with various icons. A tooltip labeled 'Sqoop 1' is visible over one of the icons. The main workspace is titled 'My Workflow' and contains a text prompt 'Add a description...'. Below this, there is a large rectangular area with a target icon and the text 'Drop your action here'. A modal window titled 'Sqoop 1' is open in the foreground, showing a text input field for the 'Sqoop command'. The command entered is 'import --connect jdbc:hsqldb:file:db.hsqldb --table TT --target-dir'. An 'Add' button is located at the bottom of the modal window.

Oozie Editor Workflows Coordinators Bundles Unsaved

DOCUMENTS ACTIONS

Sqoop 1

My Workflow

Add a description...

Provide sqoop command

Drop your action here

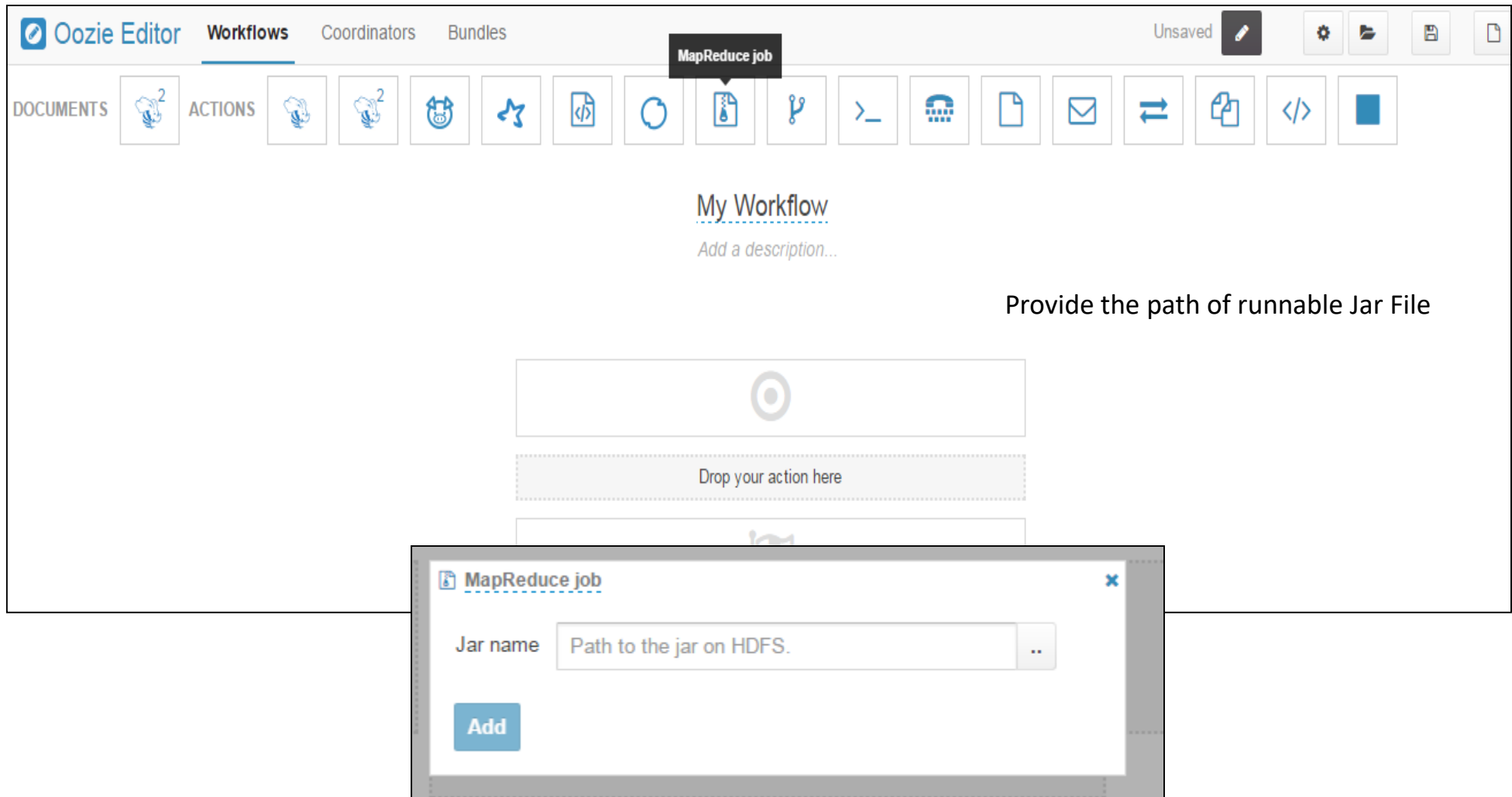
Sqoop 1

Sqoop command

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

Add

# Map reduce Action



The screenshot shows the Oozie Editor interface with the 'Workflows' tab selected. The main workspace displays 'My Workflow' with a description field. A 'MapReduce job' dialog box is open, prompting the user to provide the path of the runnable Jar File. The dialog includes a text input field for the 'Jar name' and an 'Add' button.

Oozie Editor Workflows Coordinators Bundles Unsaved

DOCUMENTS ACTIONS

MapReduce job

My Workflow

Add a description...

Provide the path of runnable Jar File

Drop your action here

MapReduce job

Jar name Path to the jar on HDFS.

Add

# Sub Workflow, Fork and Joins control nodes

Oozie Editor

Workflows

Coordinators

Bundles

Unsaved

DOCUMENTS

2

ACTIONS

2

Sub workflow

My Workflow

Add a description...

Drop your action here

Sub workflow

Sub-workflow

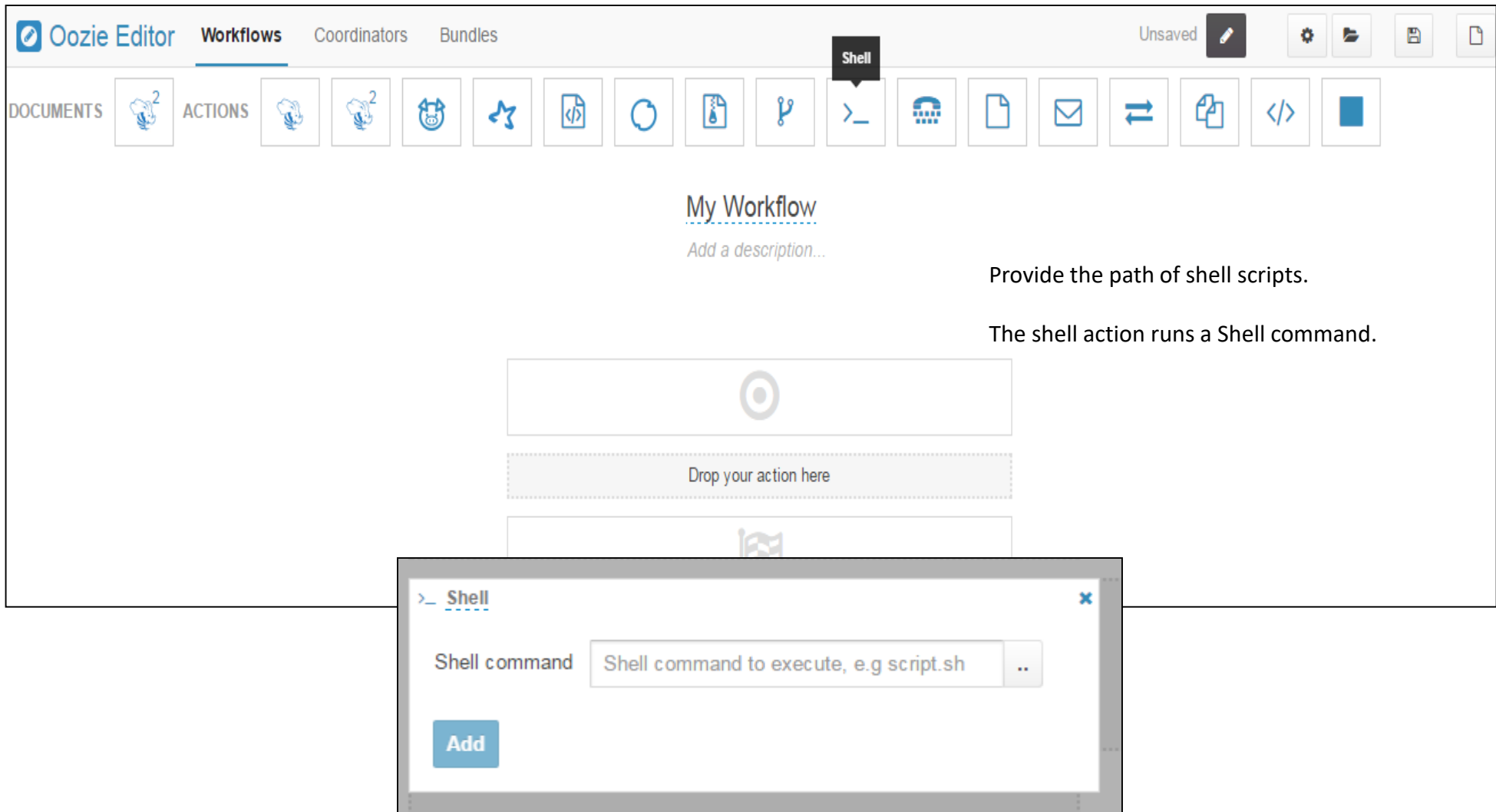
My Workflow

Add

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

Here choose the workflows which are already designed.

# Shell Action

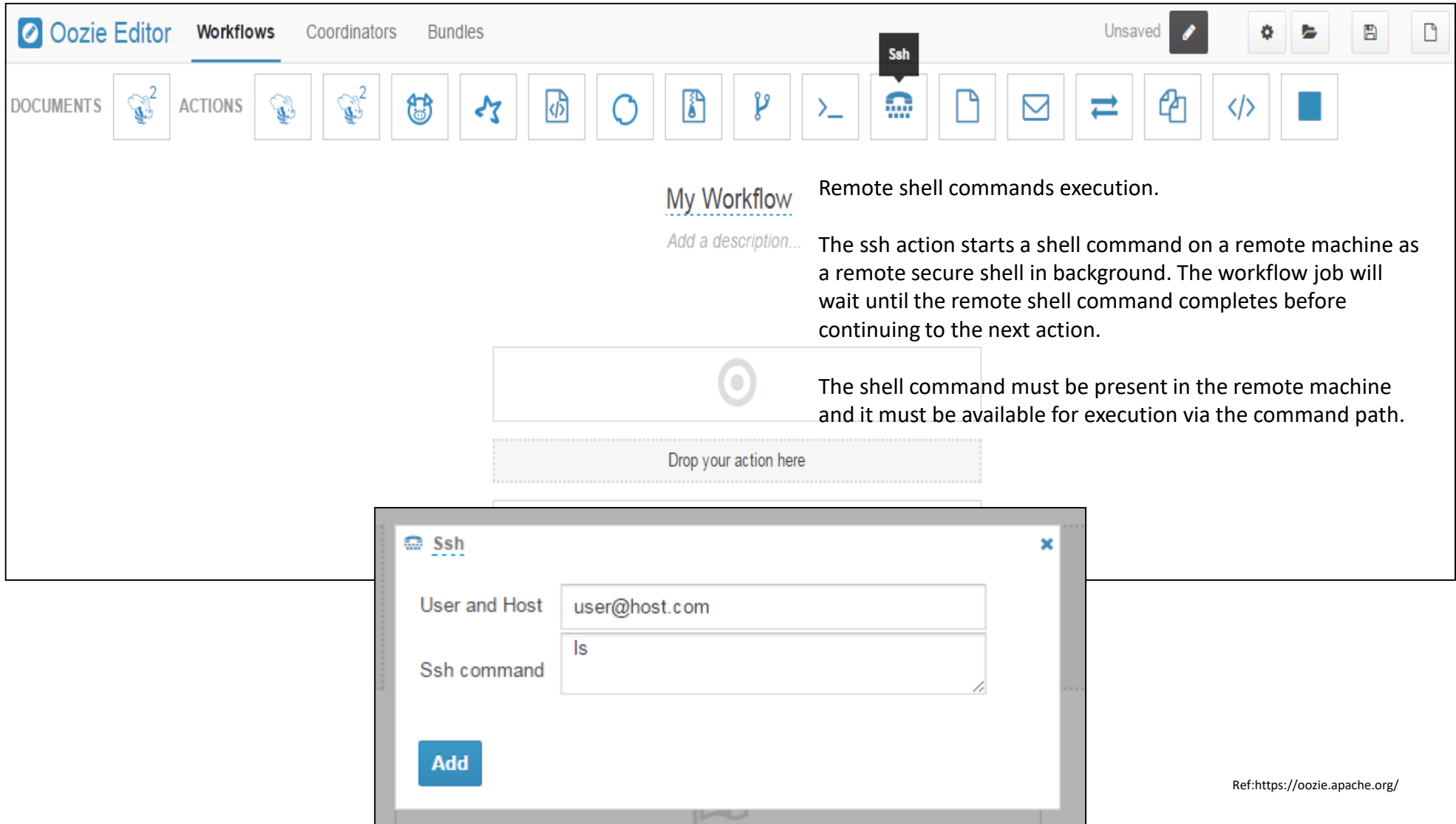


The screenshot displays the Oozie Editor interface. At the top, there are tabs for 'Workflows', 'Coordinators', and 'Bundles'. Below these is a toolbar with various icons, including a 'Shell' icon (a terminal prompt). The main workspace shows a workflow diagram with a box labeled 'My Workflow' and a dashed box below it labeled 'Drop your action here'. A modal window titled 'Shell' is open in the foreground, showing a text input field for the 'Shell command' with the placeholder text 'Shell command to execute, e.g script.sh'. An 'Add' button is visible at the bottom of the modal.

Provide the path of shell scripts.

The shell action runs a Shell command.

# SSH action



The screenshot shows the Oozie Editor interface with the 'Workflows' tab selected. The top toolbar includes icons for documents, actions, and various workflow symbols. The 'SSH' icon is highlighted. Below the toolbar, the 'My Workflow' section is visible, with a description: '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.' A dashed box labeled 'Drop your action here' is shown. An inset window titled 'Ssh' displays the configuration fields: 'User and Host' with the value 'user@host.com' and 'Ssh command' with the value 'ls'. An 'Add' button is at the bottom of the inset.

**Oozie Editor** Workflows Coordinators Bundles Unsaved

DOCUMENTS ACTIONS

My Workflow  
Add a description...

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.

Drop your action here

**Ssh**

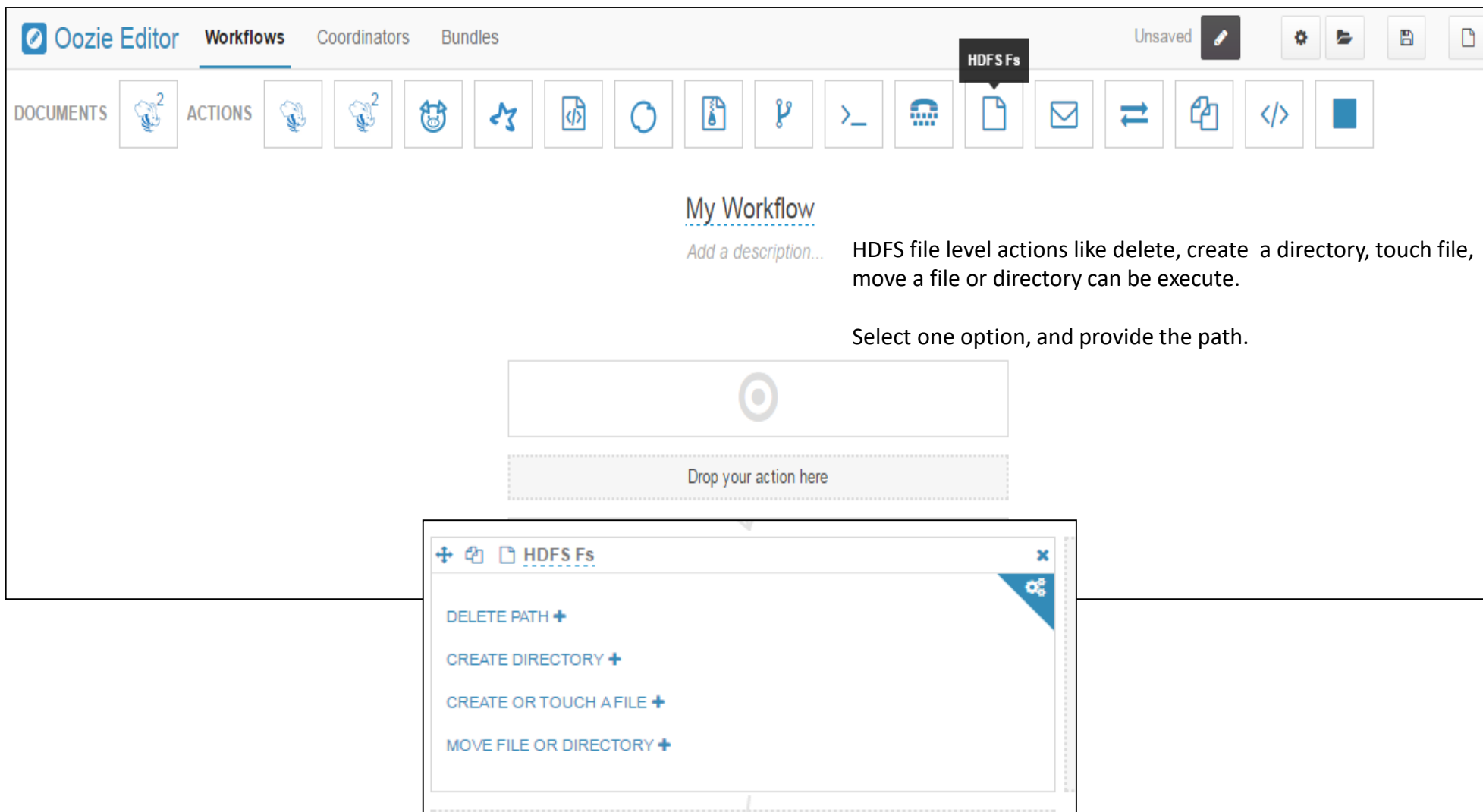
User and Host: user@host.com

Ssh command: ls

Add

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

# HDFS file action



The screenshot shows the Oozie Editor interface with the 'Workflows' tab selected. The top toolbar includes icons for documents, actions, and various workflow operations. A tooltip labeled 'HDFS Fs' is visible over the file icon in the toolbar. The main workspace is titled 'My Workflow' and contains a target icon and a dashed box labeled 'Drop your action here'. A dropdown menu is open, listing four HDFS file actions: 'DELETE PATH +', 'CREATE DIRECTORY +', 'CREATE OR TOUCH A FILE +', and 'MOVE FILE OR DIRECTORY +'. The menu also shows a settings icon in the top right corner.

Oozie Editor Workflows Coordinators Bundles Unsaved

DOCUMENTS ACTIONS

My Workflow

Add a description...

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

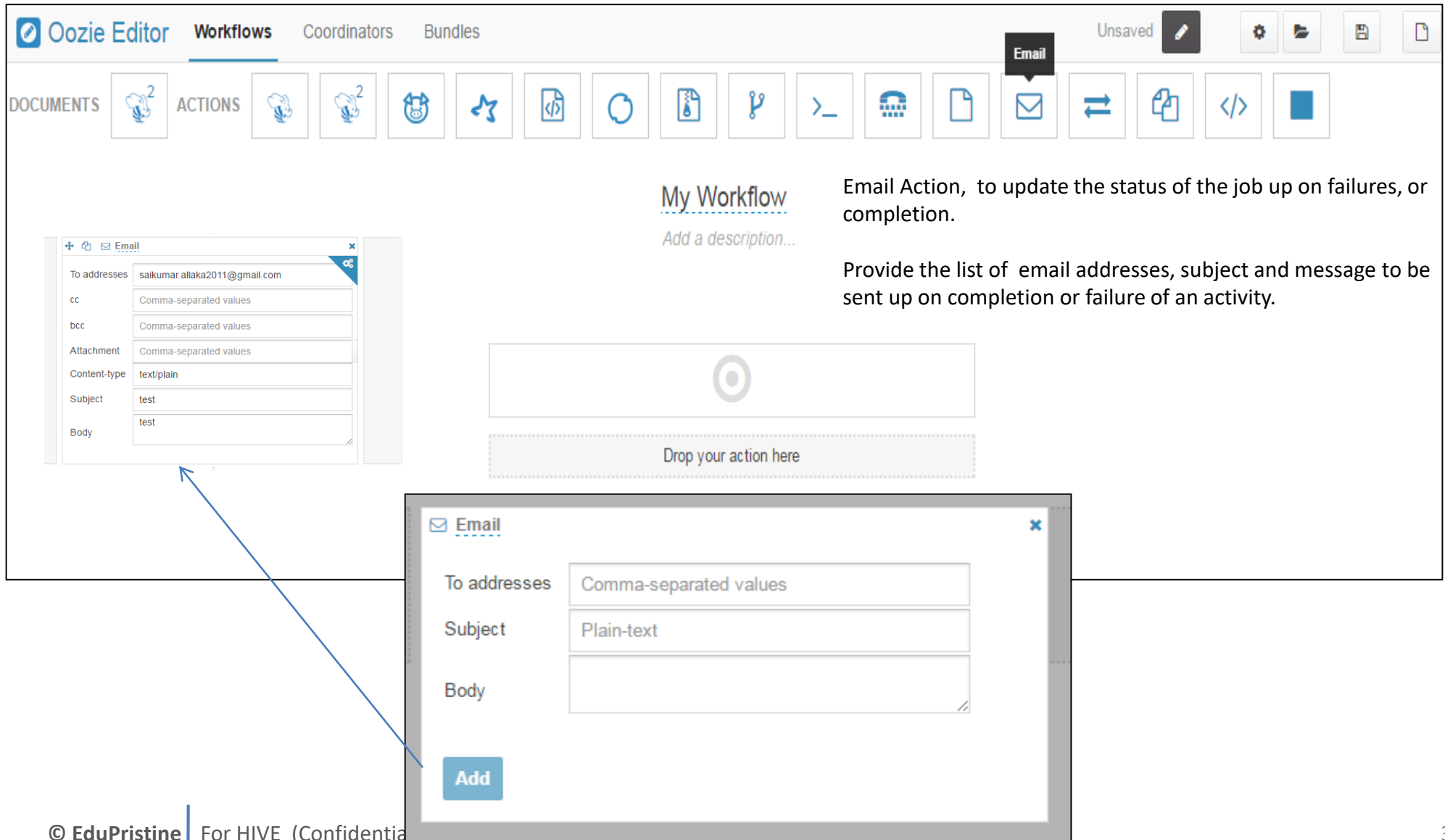
Select one option, and provide the path.

Drop your action here

HDFS Fs

- DELETE PATH +
- CREATE DIRECTORY +
- CREATE OR TOUCH A FILE +
- MOVE FILE OR DIRECTORY +

# Email Action



The screenshot displays the Oozie Editor interface with the 'Workflows' tab selected. The top toolbar includes icons for documents, actions, and a specific 'Email' action icon. The main workspace shows a workflow named 'My Workflow' with a target icon and a 'Drop your action here' area. On the left, a configuration panel for the 'Email' action is visible, showing fields for 'To addresses', 'cc', 'bcc', 'Attachment', 'Content-type', 'Subject', and 'Body'. A blue arrow points from this panel to a larger, detailed view of the 'Email' action configuration shown in a separate window at the bottom. This detailed view includes fields for 'To addresses', 'Subject', and 'Body', along with an 'Add' button.

**Oozie Editor** Workflows Coordinators Bundles

Unsaved [Edit] [Settings] [New] [Save] [Print]

DOCUMENTS ACTIONS [Email]

My Workflow  
Add a description...

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.

Drop your action here

**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**

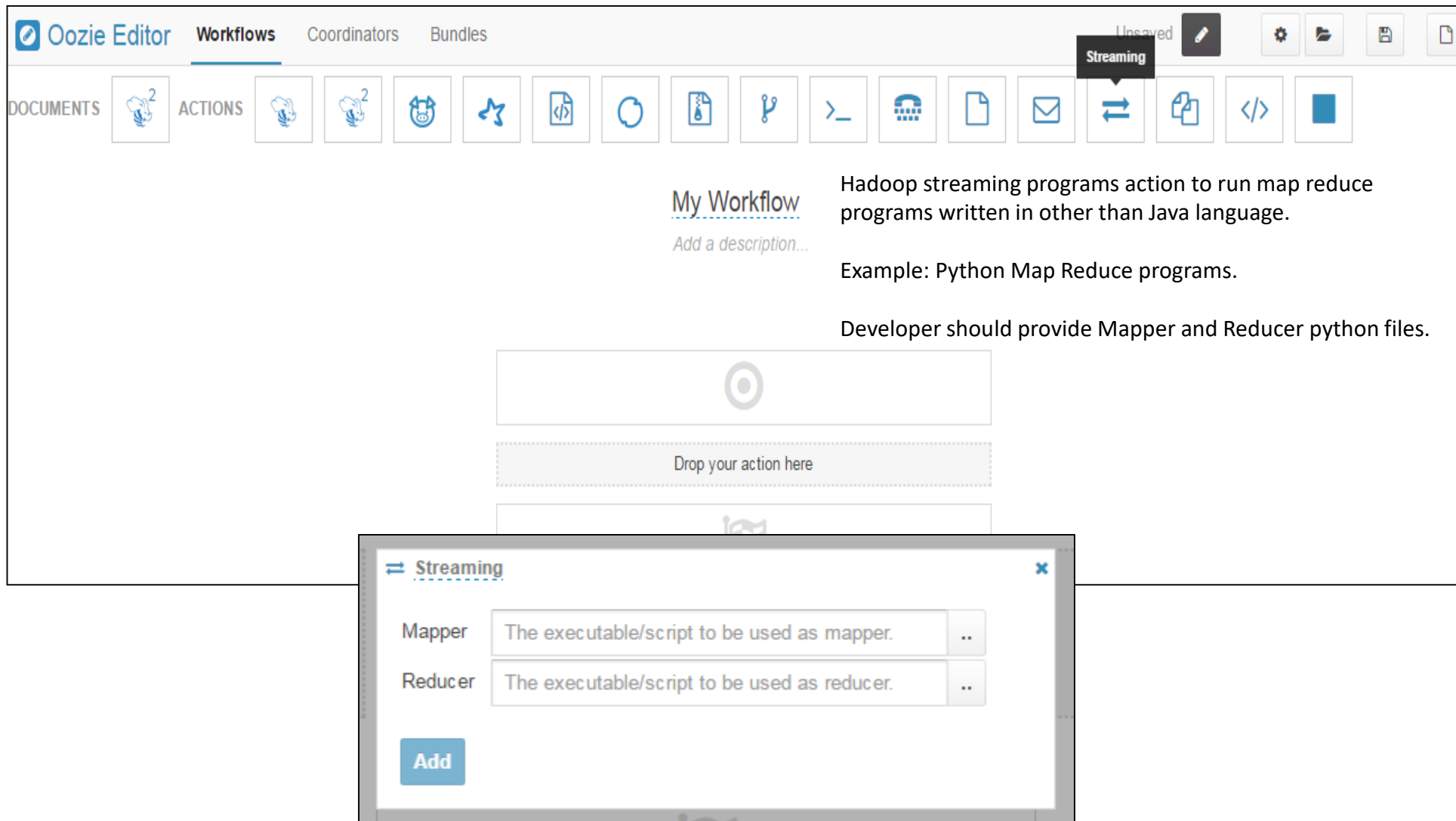
To addresses: Comma-separated values

Subject: Plain-text

Body: [Text Area]

Add

# Hadoop Streaming Action



The screenshot shows the Oozie Editor interface with the 'Workflows' tab selected. The top toolbar includes icons for documents, actions, and various workflow symbols. A 'Streaming' tooltip is visible over the corresponding icon in the toolbar. The main workspace displays a workflow diagram with a start node, a dashed box labeled 'Drop your action here', and an end node. A modal dialog titled 'Streaming' is open in the foreground, containing fields for 'Mapper' and 'Reducer' (both with placeholder text 'The executable/script to be used as mapper.' and 'The executable/script to be used as reducer.' respectively) and an 'Add' button.

Oozie Editor Workflows Coordinators Bundles

DOCUMENTS ACTIONS

My Workflow  
Add a description...

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.

Drop your action here

Streaming

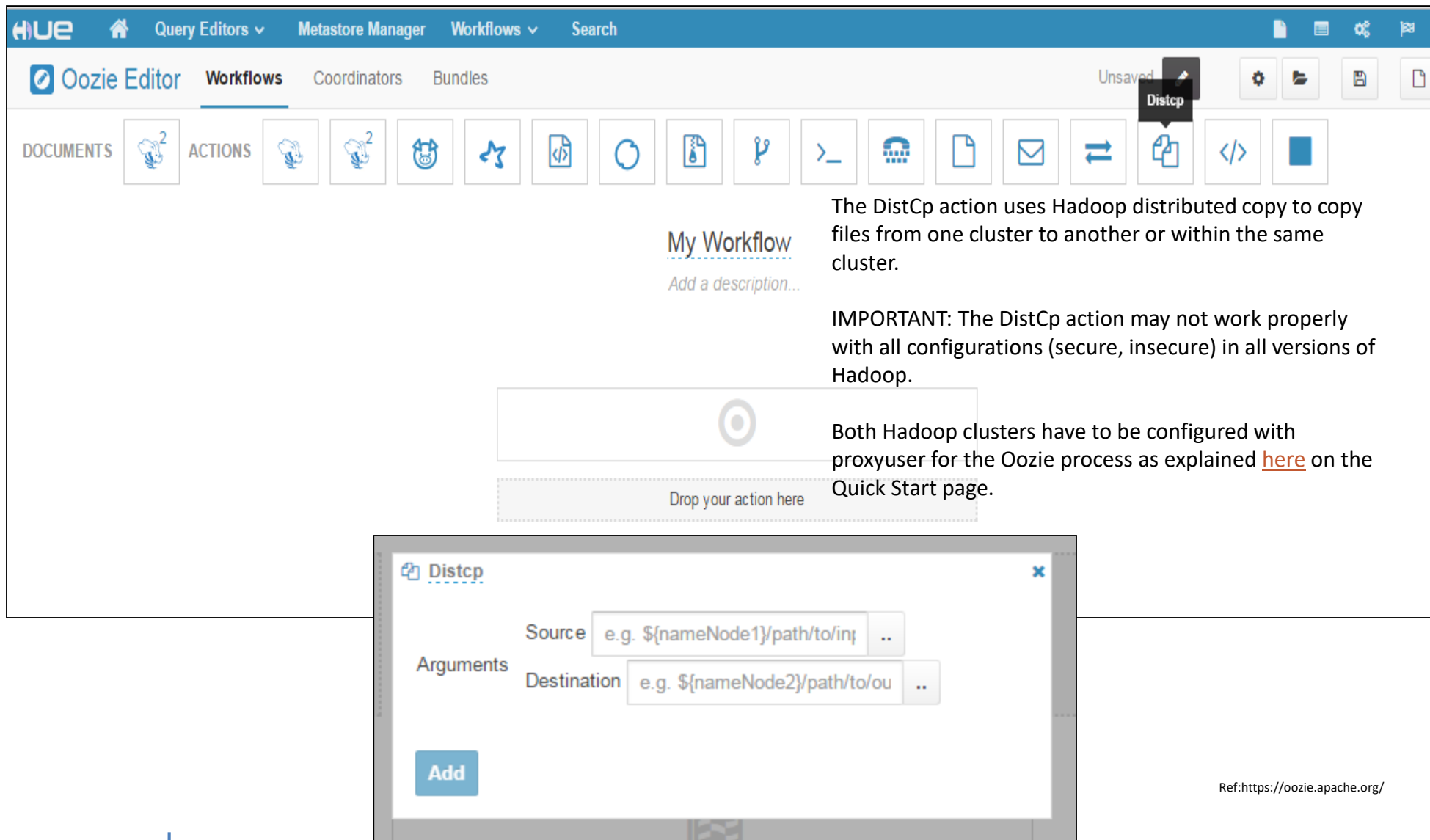
Mapper The executable/script to be used as mapper. ..

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

Add



# Distcp Action



The screenshot shows the Hue Oozie Editor interface. The top navigation bar includes 'Query Editors', 'Metastore Manager', 'Workflows', and 'Search'. The 'Oozie Editor' tab is active, showing 'Workflows', 'Coordinators', and 'Bundles'. A toolbar at the top contains various icons, including a 'Distcp' icon. Below the toolbar, a workflow named 'My Workflow' is shown with a description 'Add a description...'. A dashed box labeled 'Drop your action here' is positioned over the workflow area. A tooltip for the 'Distcp' icon is visible, stating: '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.'

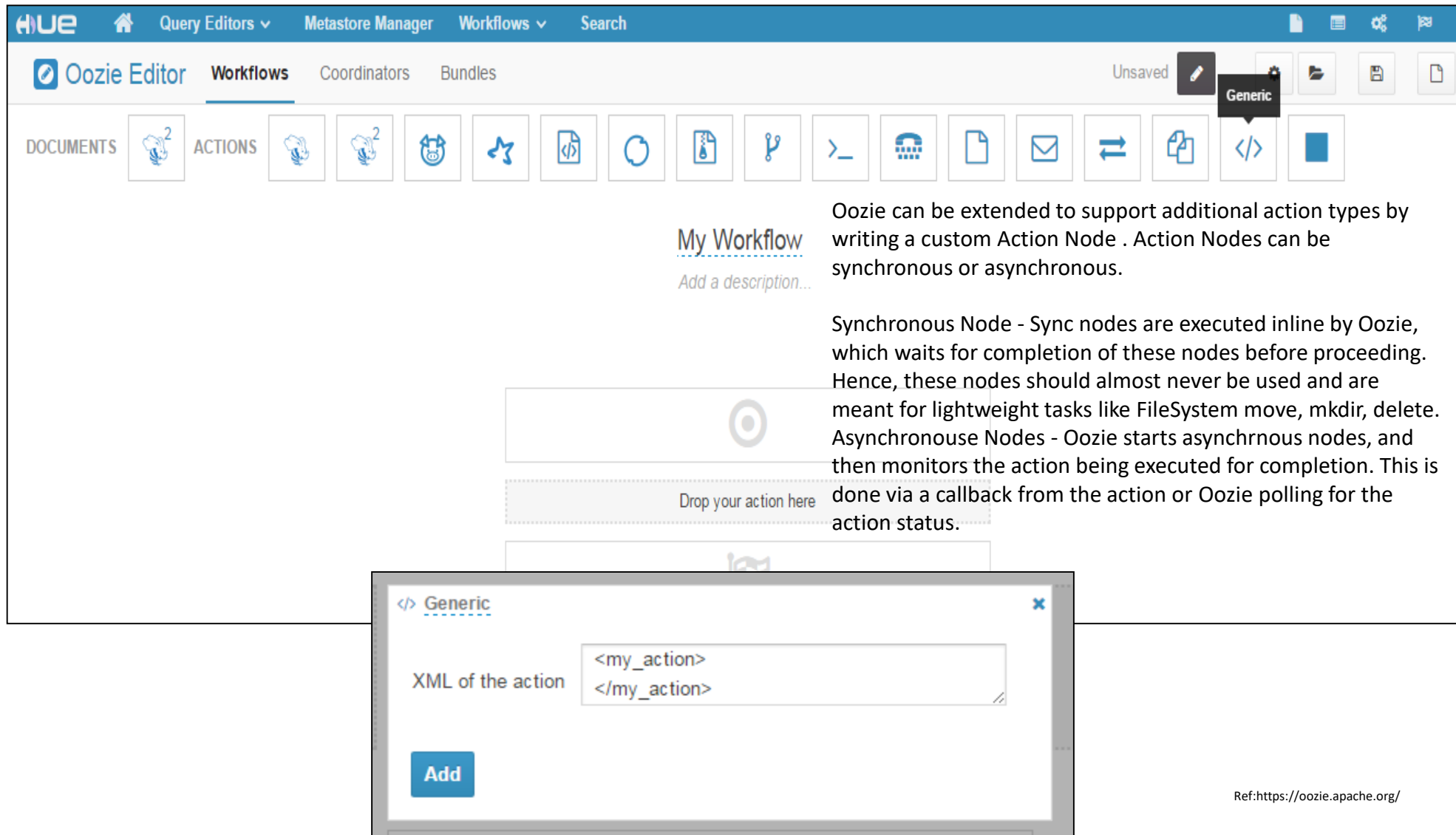
The 'Distcp' action configuration dialog is shown, with the following fields:

- Source: e.g. `${nameNode1}/path/to/in/` ..
- Destination: e.g. `${nameNode2}/path/to/ou` ..

An 'Add' button is located at the bottom of the dialog.

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

# Generic action



The screenshot shows the Hue Oozie Editor interface. The top navigation bar includes 'Query Editors', 'Metastore Manager', 'Workflows', and 'Search'. The 'Oozie Editor' tab is active, with sub-tabs for 'Workflows', 'Coordinators', and 'Bundles'. The 'Workflows' tab is selected, and the 'My Workflow' section is visible. A 'Drop your action here' box is present. A 'Generic' action node is being added, and a dialog box is open for its configuration. The dialog box has a title 'Generic' and a text area for the XML of the action, containing the code: `<my_action>` and `</my_action>`. An 'Add' button is at the bottom of the dialog.

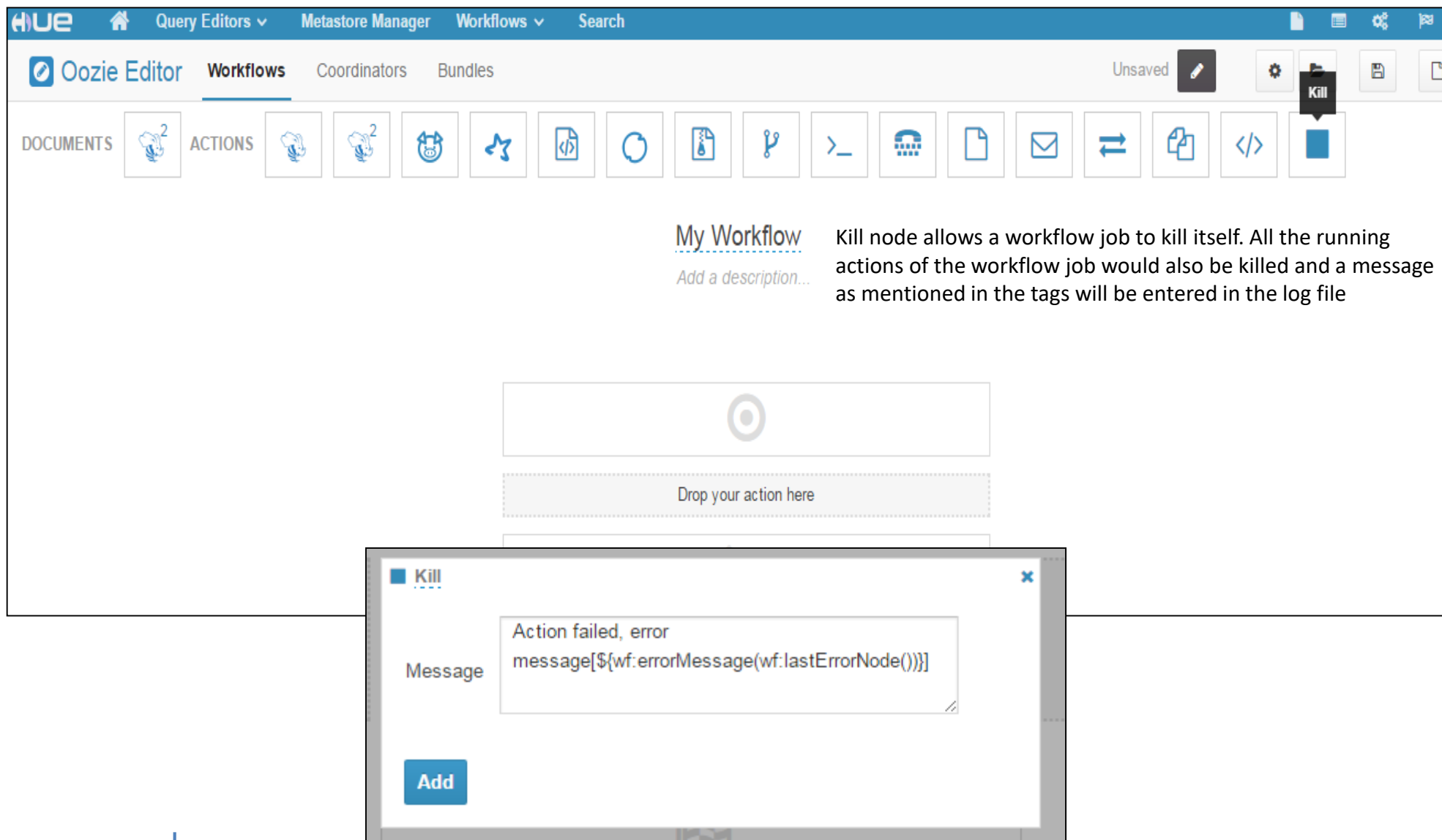
Oozie can be extended to support additional action types by writing a custom Action Node . Action Nodes can be synchronous or asynchronous.

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 asynchronous 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



The screenshot shows the Hue Oozie Editor interface. The top navigation bar includes 'Query Editors', 'Metastore Manager', 'Workflows', and 'Search'. The 'Oozie Editor' tab is active, with sub-tabs for 'Workflows', 'Coordinators', and 'Bundles'. The 'Workflows' tab is selected. The 'DOCUMENTS' section shows a 'My Workflow' with a description 'Add a description...'. The 'ACTIONS' section shows a 'Kill' node. The 'Kill' node configuration dialog is open, showing the 'Message' field with the text: `message[${wf.errorMessage(wf.lastErrorNode())}]`. The 'Add' button is visible at the bottom of the dialog.

**My Workflow**  
Add a description...

**Kill**


Message  
`message[${wf.errorMessage(wf.lastErrorNode())}]`

Add

# 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:


**Oozie Editor**

[Workflows](#)
[Coordinators](#)
[Bundles](#)

Name of coordinator

Which workflow to schedule?
Choose workflow to schedule

How often?

☐ Advanced syntax

☐ Advanced syntax

# 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 specifically, 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.

In this lab, you will learn 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

# 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:

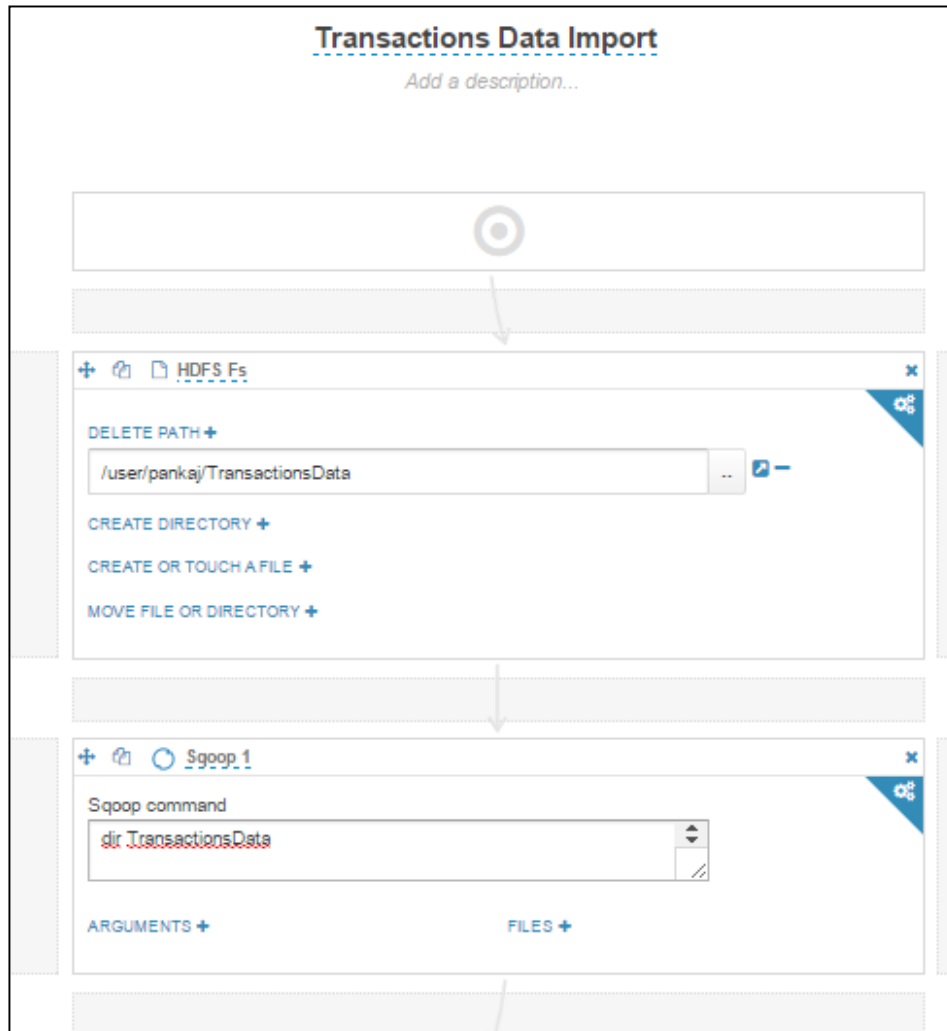
Copy oozieScripts folder to home folder using winscp.

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



## 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/p  
ankaj --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


Workflows
Coordinators
Bundles

import

SUBMITTER

pankaj

STATUS

SUCCEEDED

PROGRESS

100%



ID

0000062-170113093851012-oozie-oozi-W

VARIABLES


security\_enabled

dryrun


HDFS Fs

Delete TransactionsData

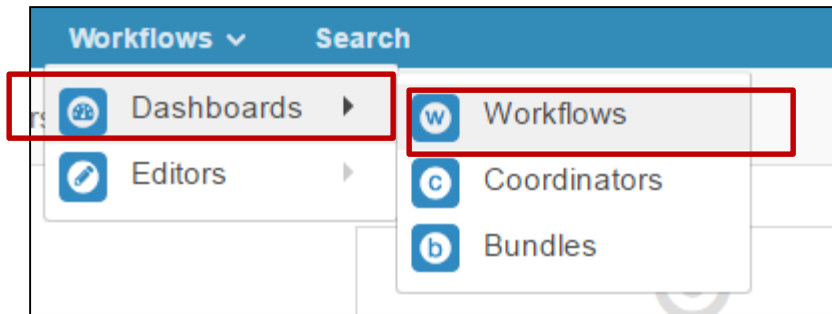


Sqoop 1

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



# Sqoop Import workflow – Track progress



▶ ⏸ ✖

Show only 1 7 15 30 days with status Succeeded Running Error submitted Manually Coordinator

### Running

<input type="checkbox"/> Submission	Status	Name	Progress	Submitter	Last Modified	Id	Parent
No matching records							

Showing 0 to 0 of 0 entries

← Previous
Next →

### Completed

Completion	Status	Name	Duration	Submitter	Id	Parent
Tue, 17 Jan 2017 06:05:21	SUCCEEDED	pig_MIS_Reportitng	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

Home / user / pankaj History Trash

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<a href="#">↑</a>		hdfs	supergroup	drwxrwxrwx	January 16, 2017 08:03 PM
<input type="checkbox"/>	<a href="#">.</a>		pankaj	labusers	drwxr--r--	January 17, 2017 06:09 AM
<input type="checkbox"/>	<a href="#">.Trash</a>		pankaj	labusers	drwx-----	January 15, 2017 04:00 PM
<input type="checkbox"/>	<a href="#">.sparkStaging</a>		pankaj	labusers	drwxr-xr-x	January 07, 2017 05:00 AM
<input type="checkbox"/>	<a href="#">.staging</a>		pankaj	labusers	drwx-----	January 17, 2017 06:09 AM
<input type="checkbox"/>	<a href="#">TransactionsData</a>		pankaj	labusers	drwxr-xr-x	January 17, 2017 06:09 AM


You should see data imported to HDFS

Home / user / pankaj / TransactionsData History Trash


<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<a href="#">↑</a>		pankaj	labusers	drwxr--r--	January 17, 2017 06:09 AM
<input type="checkbox"/>	<a href="#">.</a>		pankaj	labusers	drwxr-xr-x	January 17, 2017 06:09 AM
<input type="checkbox"/>	<a href="#">_SUCCESS</a>	0 bytes	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM
<input type="checkbox"/>	<a href="#">part-m-00000</a>	7.5 MB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM
<input type="checkbox"/>	<a href="#">part-m-00001</a>	5.9 MB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM
<input type="checkbox"/>	<a href="#">part-m-00002</a>	20.0 MB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM
<input type="checkbox"/>	<a href="#">part-m-00003</a>	24.1 MB	pankaj	labusers	-rw-r--r--	January 17, 2017 06:09 AM


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


DOCUMENTS





ACTIONS










































pig MIS Reporting

  
↓  

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

Oozie Editor Workflows Coordinators Bundles

Icons: Edit, Run, Calendar, Settings, Refresh, Folder, Save, Share, Document

pankaj

**STATUS**

**SUCCEEDED**

**PROGRESS**

100%

**ID**

0000061-170113093851012-oozie-oozi-W

**VARIABLES**

security\_enabled

dryrun

oozie.wf.applicatio...

Target

HDFS Fs

Delete

/user/pankaj/chainTop10Cust/

Pig Script

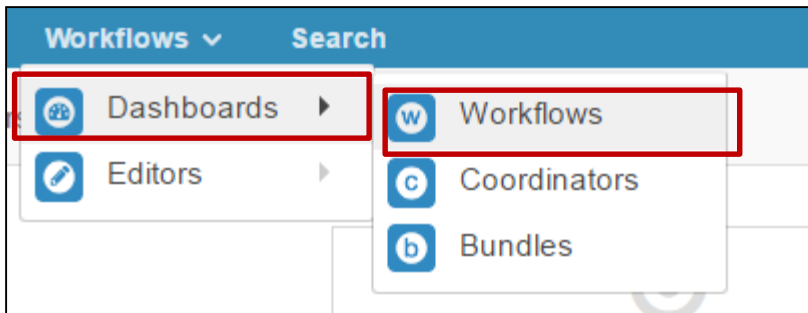
storeTop10cust.pig

Sqoop 1

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

Flag

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



Search for username, name, etc...

▶ || ✕ Show only 1 7 15 30 days with status Succeeded Running Error submitted Manually Coordinator

### Running

<input type="checkbox"/> Submission	Status	Name	Progress	Submitter	Last Modified	Id	Parent
No matching records							
Showing 0 to 0 of 0 entries							

← Previous Next →

### Completed

Completion	Status	Name	Duration	Submitter	Id	Parent
Tue, 17 Jan 2017 06:05:21	SUCCEEDED	pig_MIS_Reportitng	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

<input type="checkbox"/>	Walmart	pankaj	labusers	drwxr--r--	January 14, 2017 05:31 AM
<input type="checkbox"/>	Walmart_partitioned	pankaj	labusers	drwxr--r--	January 14, 2017 11:56 AM
<input type="checkbox"/>	avroschema	pankaj	labusers	drwxr-xr-x	January 15, 2017 12:13 AM
<input type="checkbox"/>	chainTop10Cust	pankaj	labusers	drwxr-xr-x	January 17, 2017 06:04 AM
<input type="checkbox"/>	creditcard	pankaj	labusers	drwxr-xr-x	January 14, 2017 09:00 PM
<input type="checkbox"/>	creditcardtrx	pankaj	labusers	drwxr-xr-x	January 14, 2017 09:01 PM

Home / user / pankaj / chainTop10Cust						History	Trash
<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date	
<input type="checkbox"/>	↑		pankaj	labusers	drwxr--r--	January 17, 2017 06:09 AM	
<input type="checkbox"/>	.		pankaj	labusers	drwxr-xr-x	January 17, 2017 06:04 AM	
<input type="checkbox"/>	_SUCCESS	0 bytes	pankaj	labusers	-rw-r--r--	January 17, 2017 06:04 AM	
<input type="checkbox"/>	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	61032170	10753.0

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



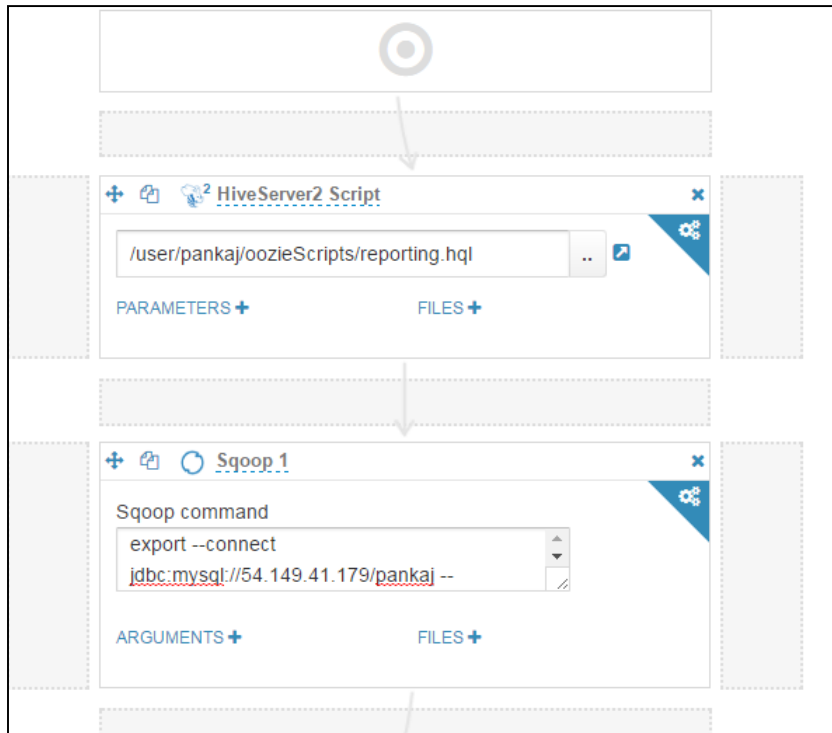
## Hive MIS Reporting

Add a description...

### Name of the workflow

### 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


Graph Actions Details Configuration Log Definition







 <sup>2</sup> HiveServer2 Script


reporting.hql




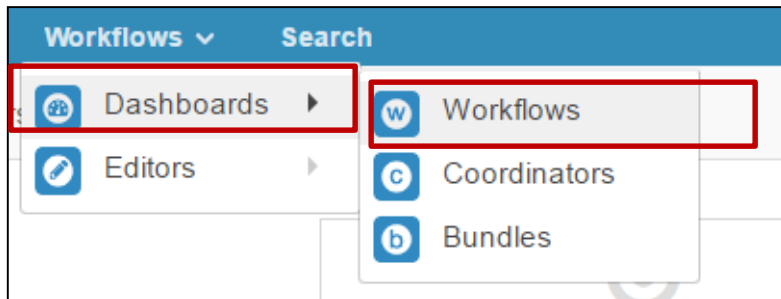


 Sqoop 1

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







**Oozie Dashboard** | Workflows | Coordinators | Bundles | SLA | Oozie

Search for username, name, etc... [Play] [Pause] [Stop] Show only [1] [7] [15] [30] days with status [Succeeded] [Running] [Error] submitted [Manually] [Coordinator]

### Running

<input type="checkbox"/> Submission	Status	Name	Progress	Submitter	Last Modified	Id	Parent
<input type="checkbox"/> Tue, 17 Jan 2017 11:00:24	<b>RUNNING</b>	Hive_MIS_Reporting	50%	pankaj	46s	0000066-170113093851012-oozie-oozi-W	

Showing 1 to 1 of 1

← Previous    Next →

# 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      3814.639999999999
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

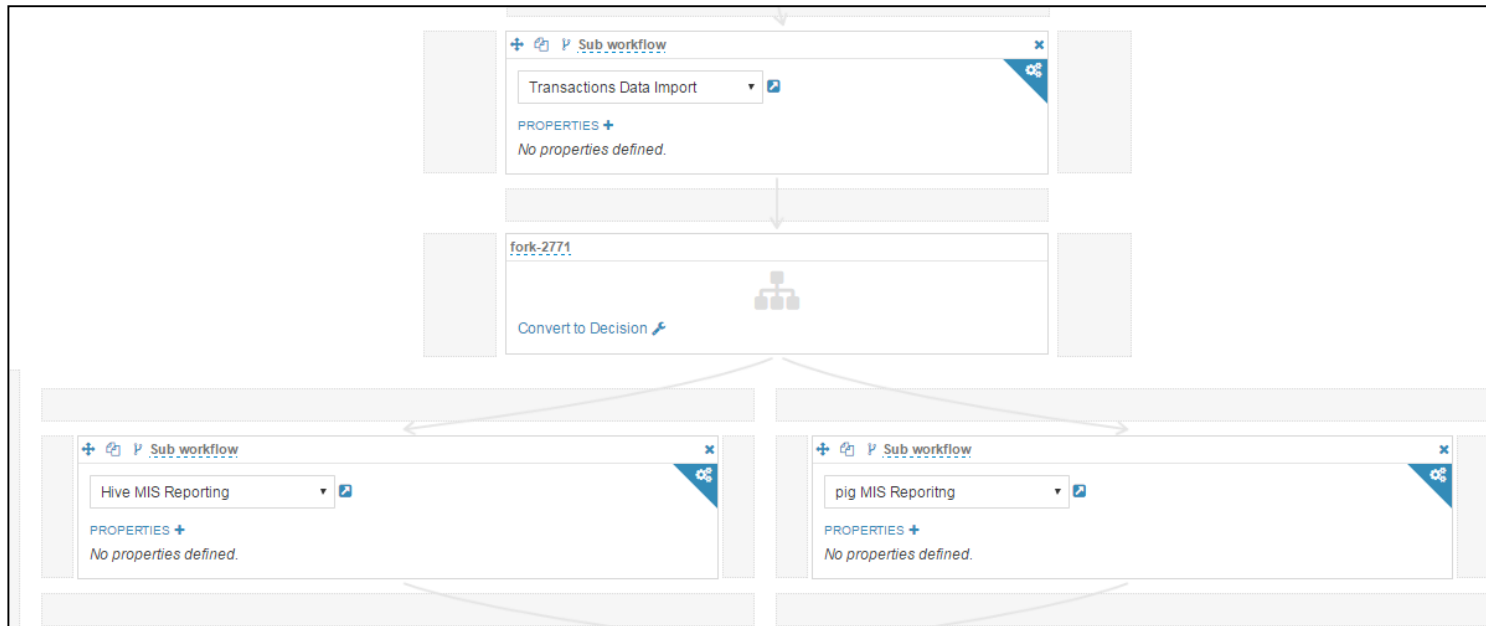
# Combining all three workflows in to single workflow.

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!



WORKFLOW

ReporintWorkflow

SUBMITTER

pankaj

STATUS

RUNNING

PROGRESS

1.4%

ID

0000067-170113093851012-oozie-oozi-W

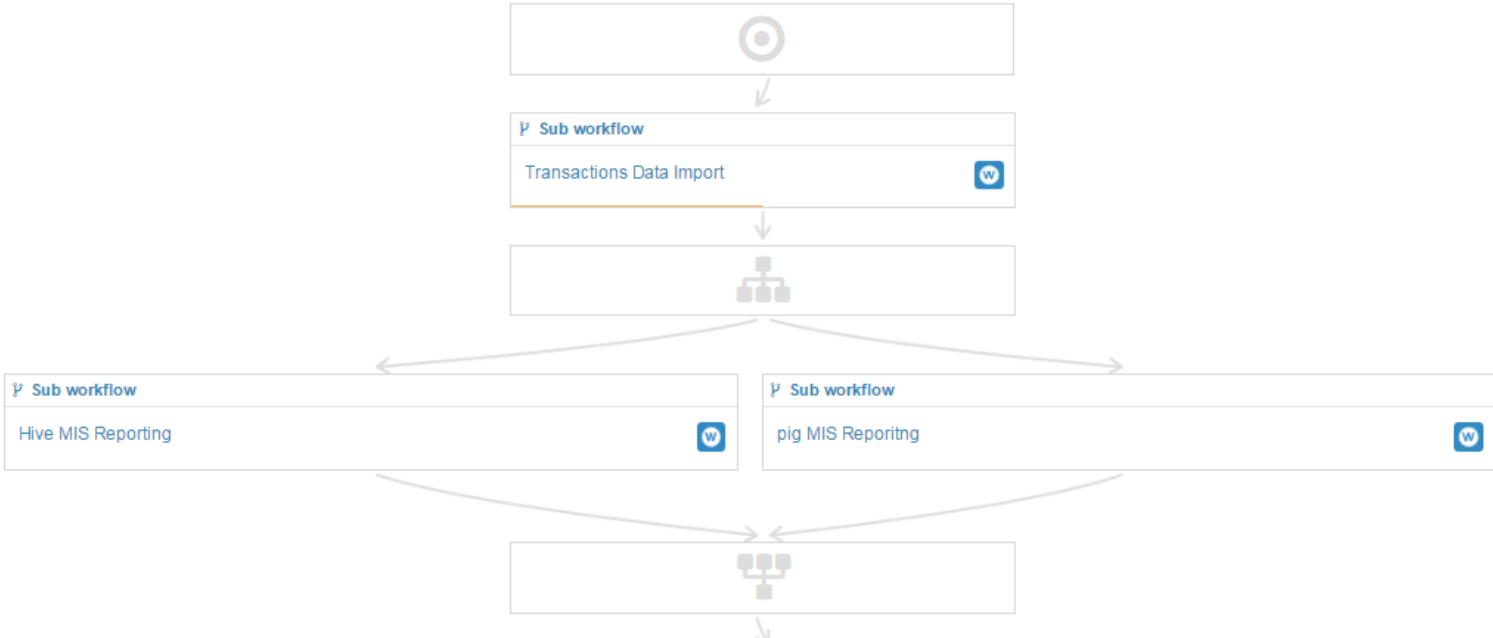
MANAGE

Kill

Suspend

Workflow ReporintWorkflow

Graph Actions Details Configuration Log Definition



```

graph TD
    Start(( )) --> Sub1[Sub workflow  
Transactions Data Import]
    Sub1 --> Join1(( ))
    Join1 --> Sub2[Sub workflow  
Hive MIS Reporting]
    Join1 --> Sub3[Sub workflow  
pig MIS Reporting]
    Sub2 --> Join2(( ))
    Sub3 --> Join2
    Join2 --> End(( ))
        
```

Oozie Dashboard

Workflows

Coordinators

Bundles

SLA

Oozie

▶ Resume

⏸ Suspend

✖ Kill

Show only

1

7

15

30

days with status

Succeeded

Running

Error

submitted

Manually

Coordinator

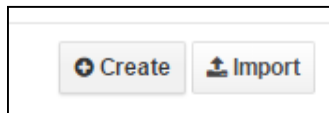
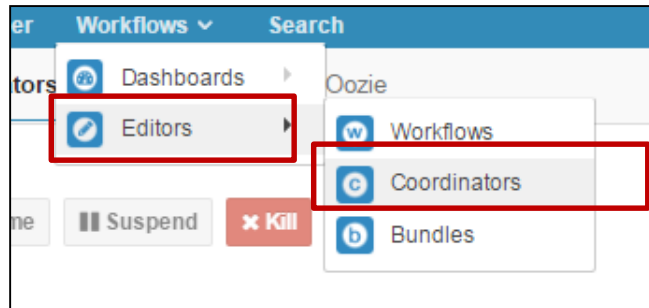
Running

<input type="checkbox"/> Submission	Status	Name	Progress	Submitter	Last Modified	Id	Parent
<input type="checkbox"/> Tue, 17 Jan 2017 11:08:37	RUNNING	pig_MIS_Reportitng	66%	pankaj	10s	0000070-170113093851012-oozie-oozi-W	
<input type="checkbox"/> Tue, 17 Jan 2017 11:08:36	RUNNING	Hive_MIS_Reporting	50%	pankaj	8s	0000069-170113093851012-oozie-oozi-W	
<input type="checkbox"/> Tue, 17 Jan 2017 11:08:03	RUNNING	ReporintWorkflow	60%	pankaj	12s	0000067-170113093851012-oozie-oozi-W	

Showing 1 to 3 of 3

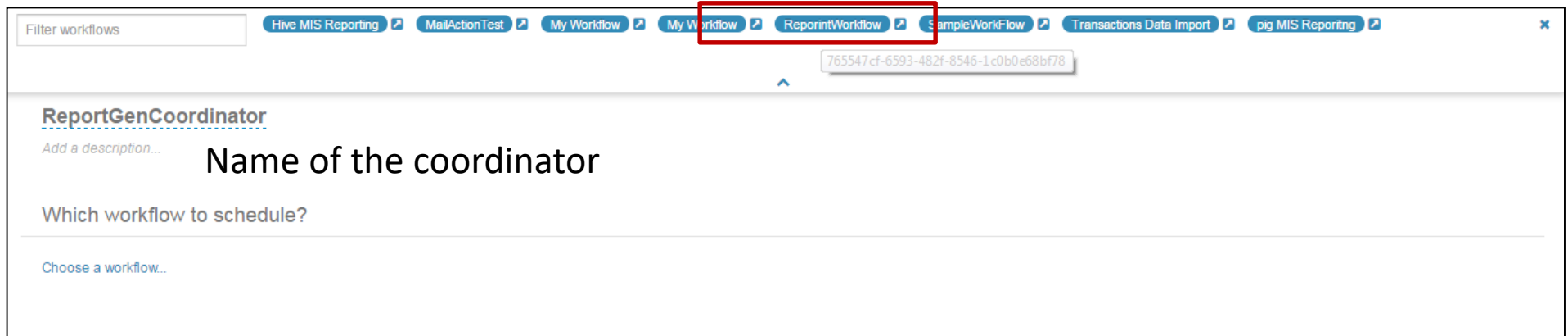
# Coordinating workflow

Add the previous workflow and schedule it.



Create a new one

Select this workflow



## ReportGenCoordinator

Add a description...

### Which workflow to schedule?

ReporintWorkflow 

### How often?

Every  at  :

 Options

### Workflow Parameters

[+ Add parameter](#)

Save

# 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