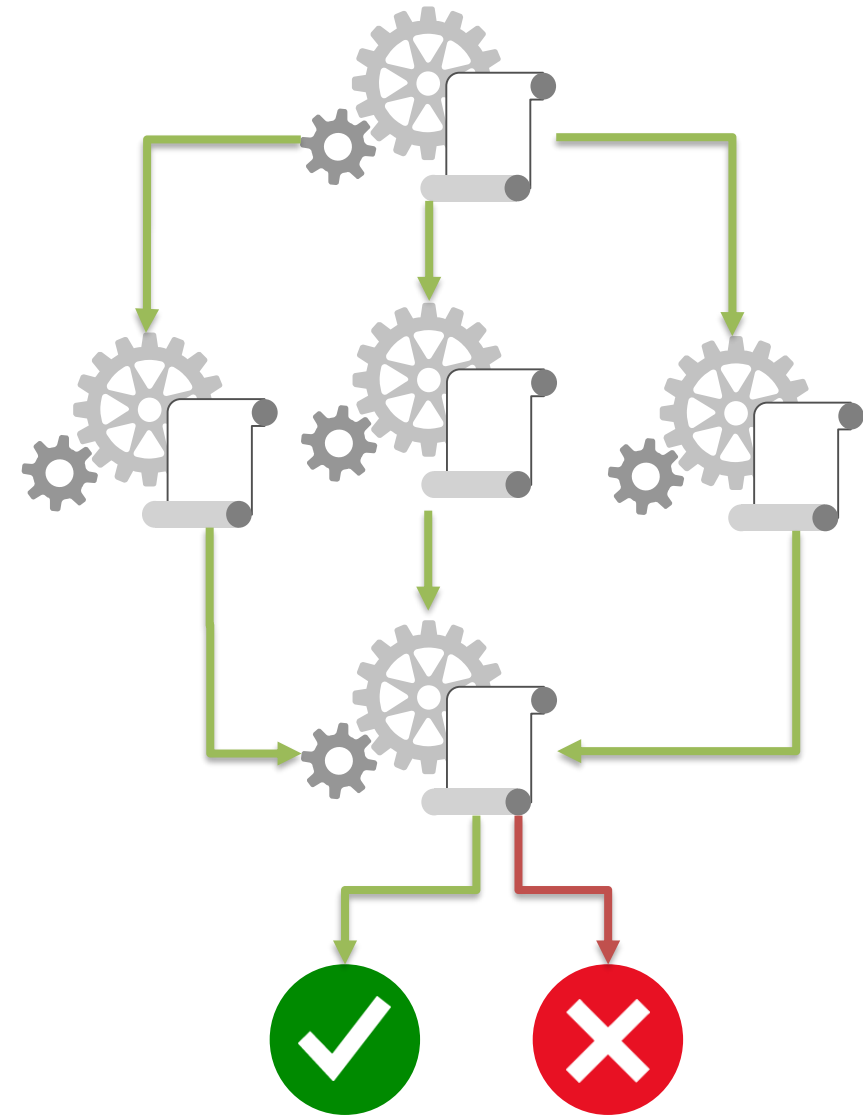
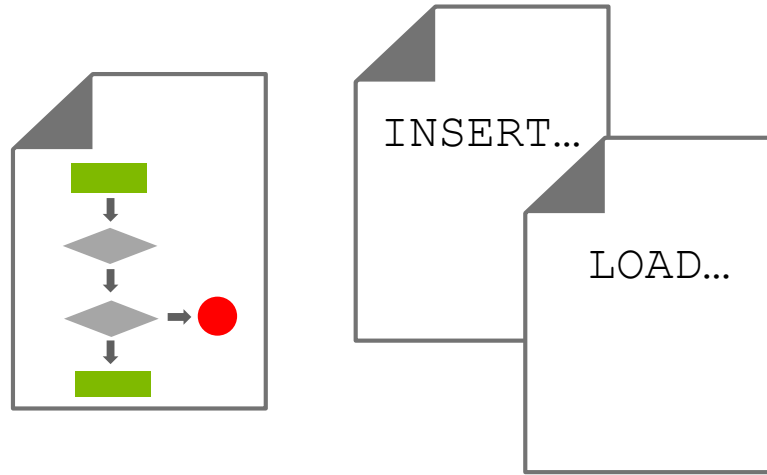


Big Data Workflows with Oozie and Sqoop

What is Oozie?

- A workflow engine for actions in a Hadoop cluster
 - MapReduce
 - Hive
 - Pig
 - Others
- Support parallel workstreams and conditional branching





- Oozie workflow file
 - XML file defining workflow actions
- Script files
 - Files used by workflow actions - for example, HiveQL or Pig Latin

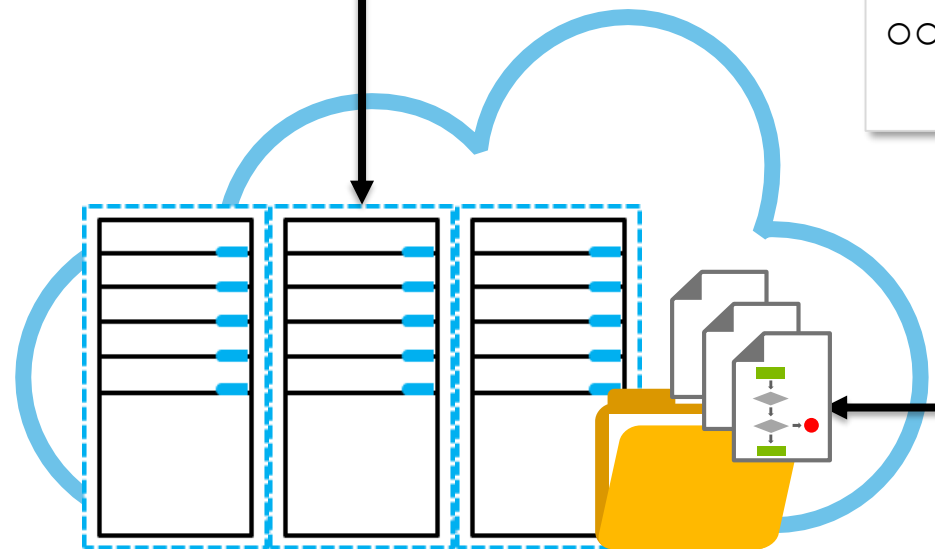
```
<workflow-app xmlns="uri:oozie:workflow:0.2" name="MyWorkflow">
  <start to="FirstAction"/>
  <action name="FirstAction">
    <hive xmlns="uri:oozie:hive-action:0.2">
      <script>CreateTable.hql</script>
    </hive>
    <ok to="SecondAction"/>
    <error to="fail"/>
  </action>
  <action name="SecondAction">
    ...
  </action>
  <kill name="fail">
    <message>Workflow failed. [${wf:errorMessage(wf:lastErrorNode())}]</message>
  </kill>
  <end name="end"/>
</workflow-app>
```

The diagram illustrates the workflow structure with three callout boxes:

- Start here**: Points to the `<start to="FirstAction"/>` tag.
- This action runs a Hive script file in the workflow folder**: Points to the `<script>CreateTable.hql</script>` tag.
- Workflow branches based on action outcome**: Points to the `<ok to="SecondAction"/>` and `<error to="fail"/>` tags.

```
oozie job -oozie http://localhost:11000/oozie -config job.properties -run
```

```
nameNode=wasb://hdfiles@hdstore.blob.core.windows.net  
jobTracker=jobtrackerhost:9010  
queueName=default  
oozie.use.system.libpath=true  
  
oozie.wf.application.path=/data/workflow/
```



How can I parameterize actions?

```
nameNode=wasb://my_container@my_storage_account.blob.core.windows.net
jobTracker=jobtrackerhost:9010
queueName=default
oozie.use.system.libpath=true
oozie.wf.application.path=/data/workflow/
```

```
tableName=mytable
```

```
tableFolder=/data/mytable
```

```
<action name="FirstAction">
  <hive xmlns="uri:oozie:hive-action:0.2">
    <script>CreateTable.hql</script>
    <param>TABLE_NAME=${tableName}</param>
    <param>LOCATION=${tableFolder}</param>
  </hive>
</action>
```

```
DROP TABLE IF EXISTS ${TABLE_NAME};
CREATE EXTERNAL TABLE ${TABLE_NAME}
(Col1 STRING,
 Col2 FLOAT,
 Col3 FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '${LOCATION}';
```


What is Sqoop?

Sqoop is a database integration service

- Built on open source Hadoop technology
- Enables bi-directional data transfer between Hadoop clusters and databases via JDBC



How do I run Sqoop commands?

- Basic syntax:

sqoop command --arg1, --arg2, ...--argN

- Commands:

- | | |
|---------------------|------------------|
| ■ import | ■ list-databases |
| ■ export | ■ list-tables |
| ■ help | ■ eval |
| ■ import-all-tables | ■ codegen |
| ■ create-hive-table | ■ version |

sqoop import

--connect *jdbc_connection_string*

--username *user_name* --password *password* | -P

--table *table_name* --columns *col1,...colN* | --query '*SELECT...*'

--warehouse-dir | --target-dir *path*

--fields-terminated-by *char* --lines-terminated-by *char*

--hive-import [--hive-overwrite]

-m | --num-mappers *number_of_mappers*

sqoop export

--connect *jdbc_connection_string*

--username *user_name* --password *password* | -P

--table *table_name*

--export-dir *path*

--fields-terminated-by *char* --lines-terminated-by *char*

-m | --num-mappers *number_of_mappers*



Microsoft

©2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, Office, Azure, System Center, Dynamics and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.