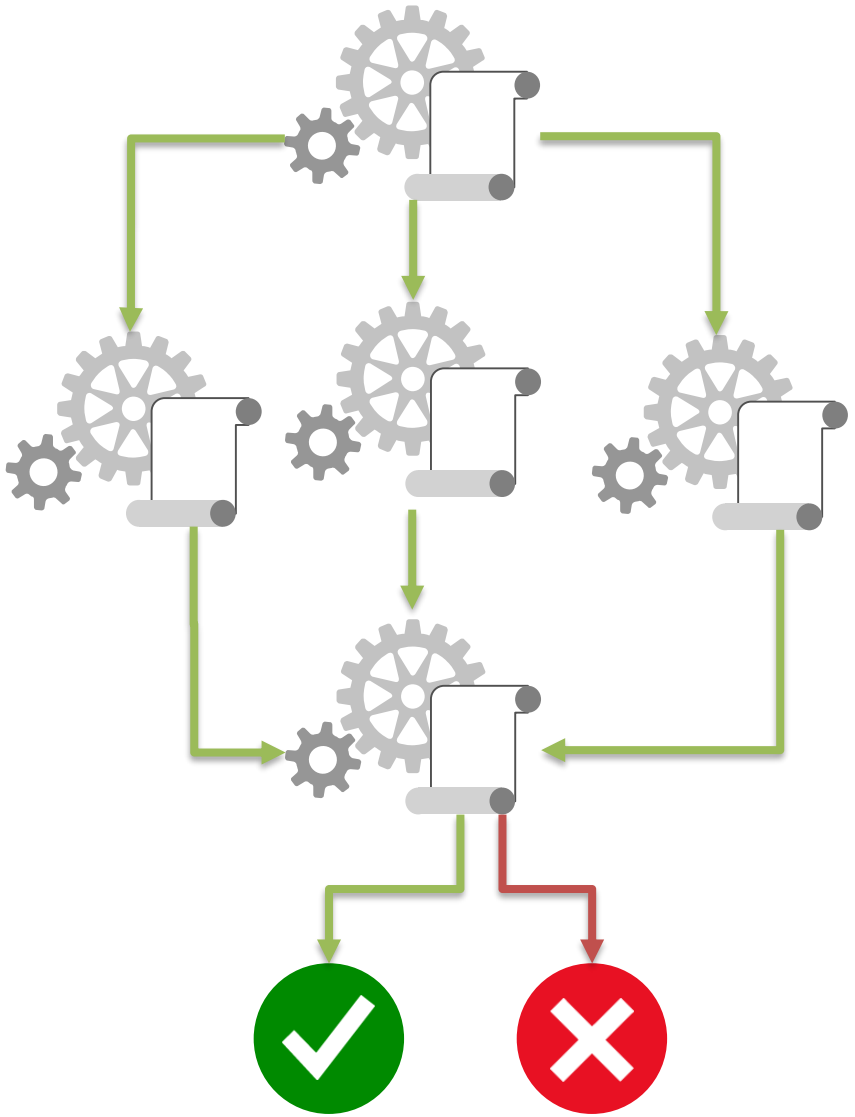
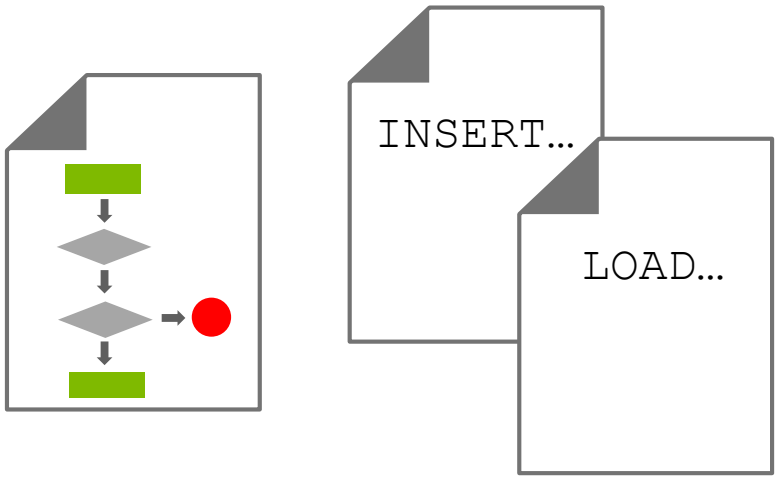
Big Data Workflows with Oozie and Sqoop

Microsoft

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>
```

Start here

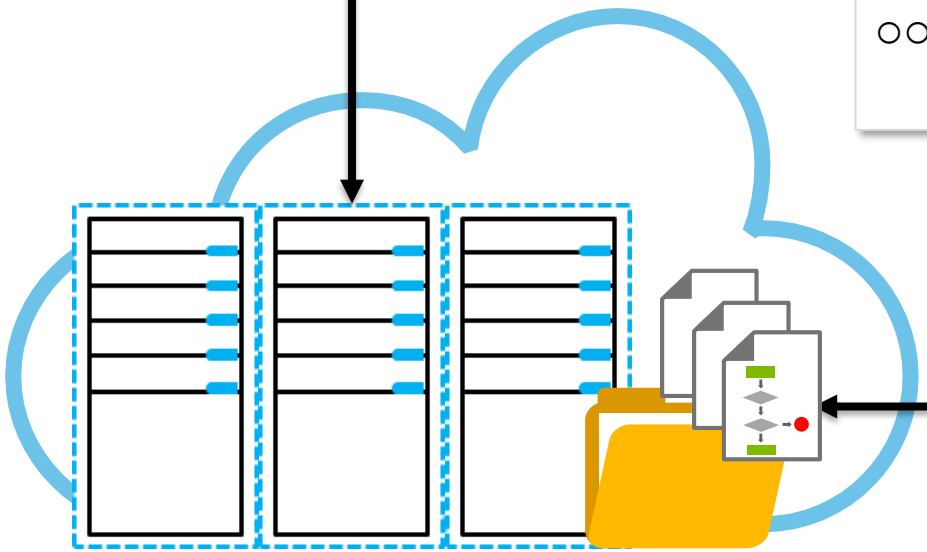
This action runs a Hive script file in the workflow folder

Workflow branches based on action outcome

```
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
```

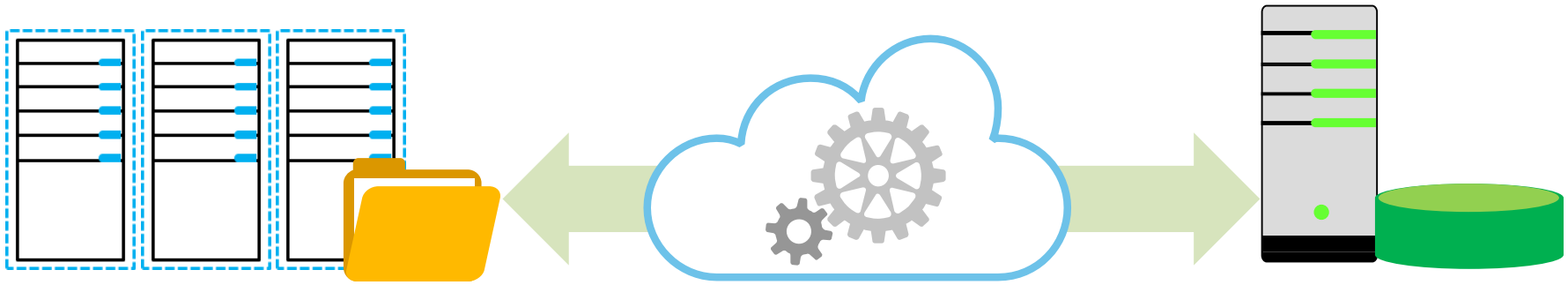
```
<action name="FirstAction">
    <hive xmlns="uri:oozie:hive-action:0.2">
      <script>CreateTable.hql</script>
        <param>TABLE_NAME=${tableName}</param>
        <param>LOCATION=${tableFolder}</param>
    </hive>
</action>
```

```
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
  - export
  - help
  - import-all-tables
  - create-hive-table
  - list-databases
  - list-tables
  - eval
  - codegen
  - 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