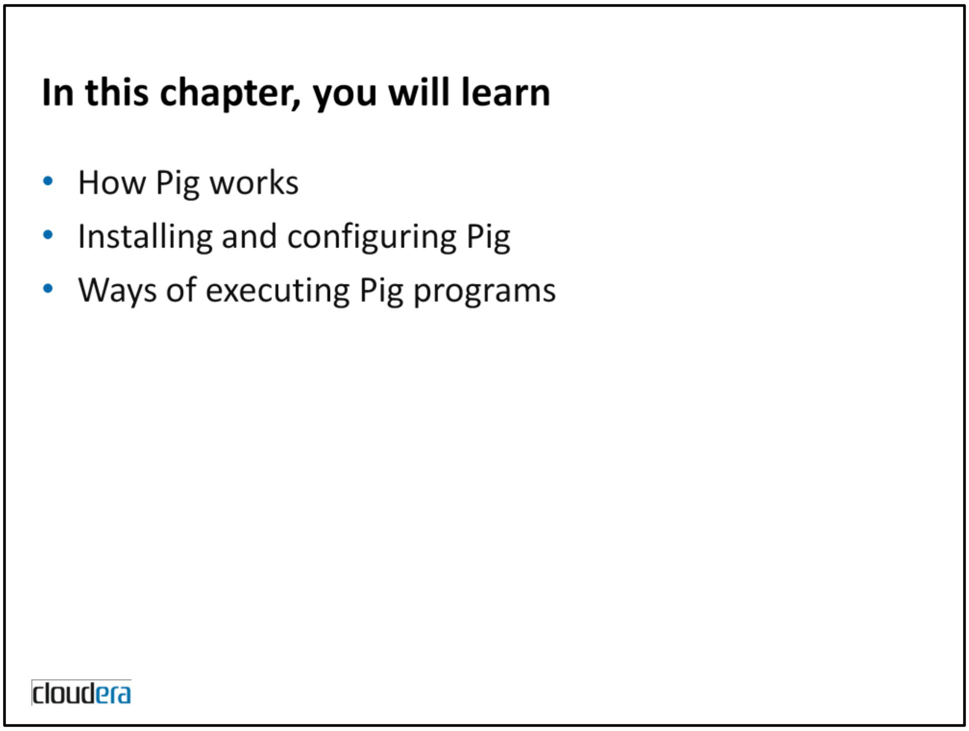
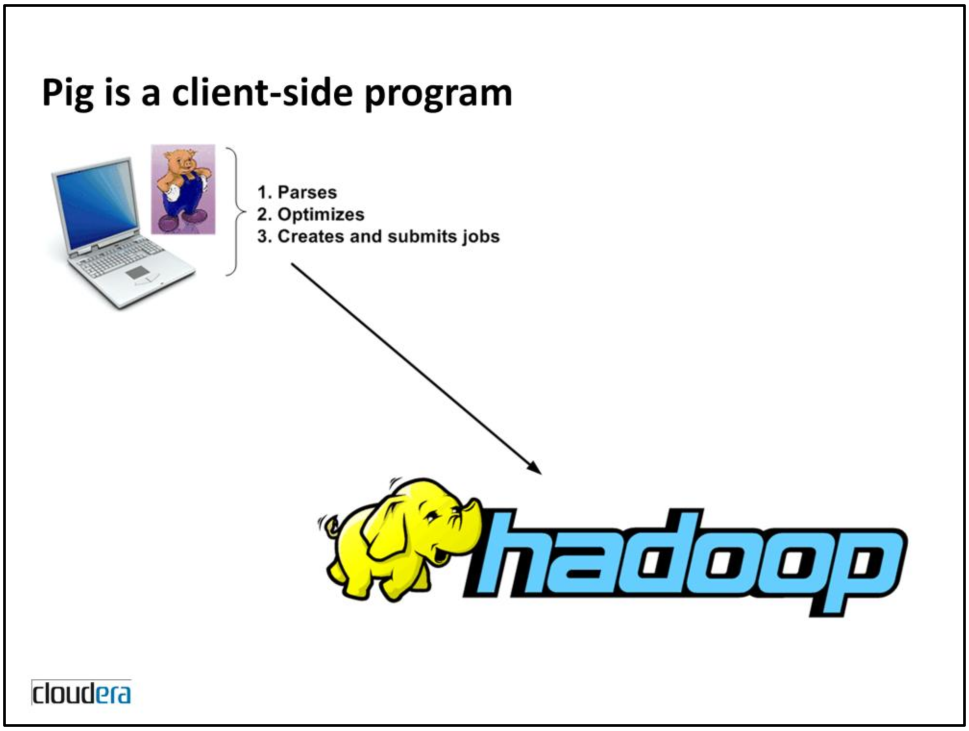
Pig's Architecture



In this chapter, you will learn

- How Pig works
- Installing and configuring Pig
- Ways of executing Pig programs

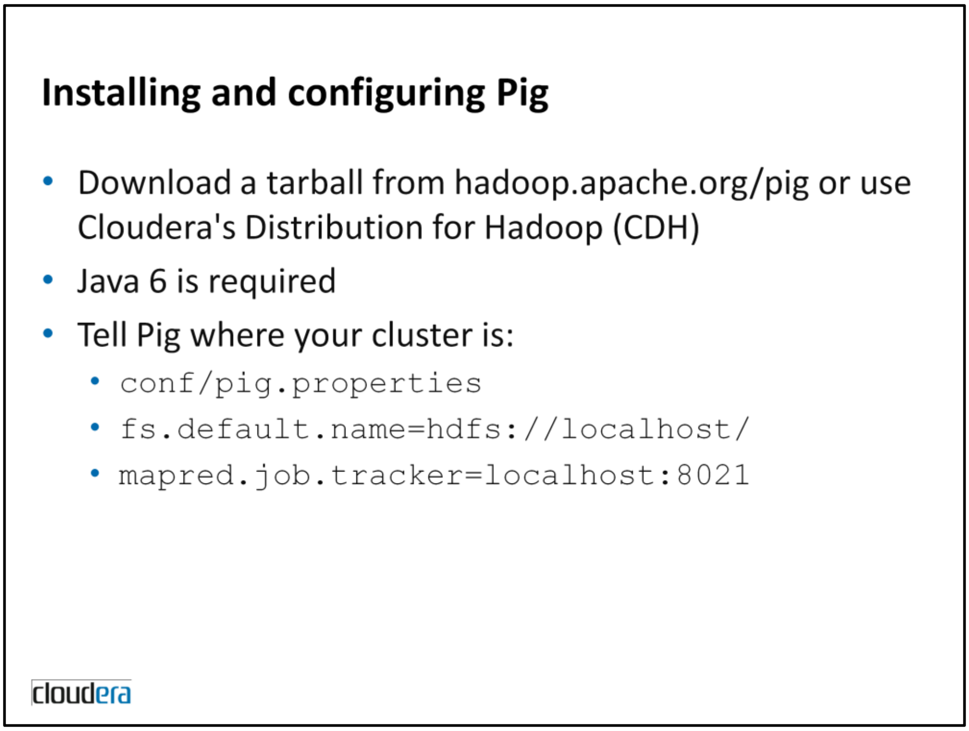
cloudera



Pig is a client-side program

1. Parses
2. Optimizes
3. Creates and submits jobs

Pig runs as a client-side application. It is not necessary to install any special software on the Hadoop cluster. When the developer writes a Pig program, it is parsed, optimized and finally executed. Pig does the work of turning the Pig program into a series of MapReduce jobs which are then sent to the Hadoop cluster (or the LocalJobRunner).



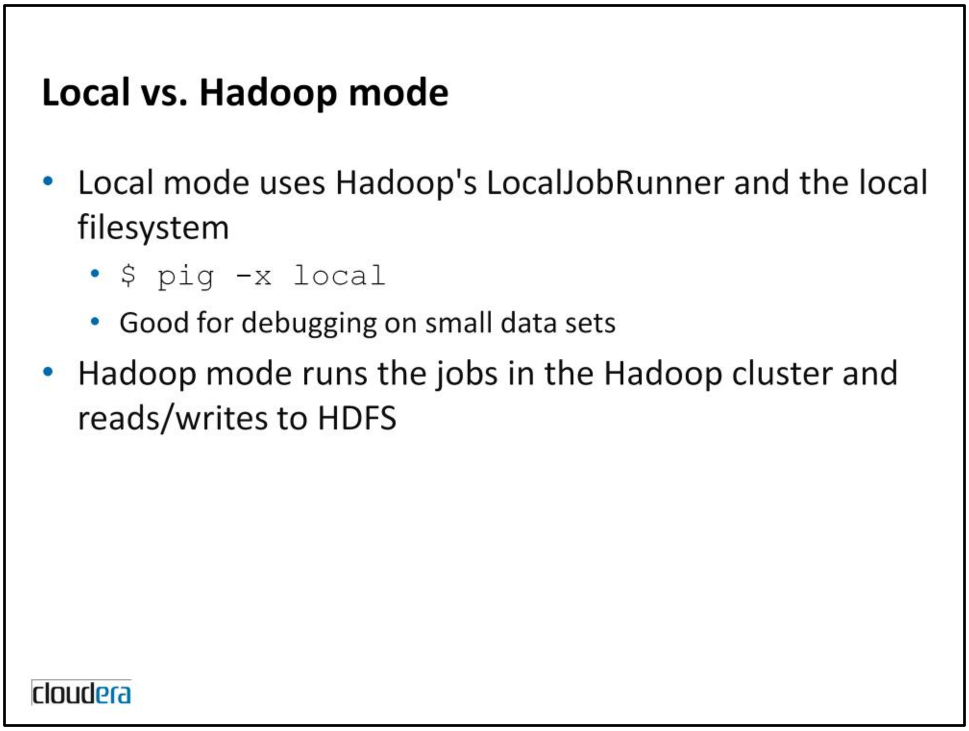
Installing and configuring Pig

- Download a tarball from hadoop.apache.org/pig or use Cloudera's Distribution for Hadoop (CDH)
- Java 6 is required
- Tell Pig where your cluster is:
  - `conf/pig.properties`
  - `fs.default.name=hdfs://localhost/`
  - `mapred.job.tracker=localhost:8021`

Installing Pig is very easy. Either download the tarball from http://hadoop.apache.org/pig/releases.html or use Cloudera's Distribution for Hadoop which provides RPMs and DEBs. The client will also need to have Java 6. If the client is a Windows environment, Cygwin is also required.

To configure Pig, you must tell it where your Hadoop cluster is. This can be done by editing the `pig.properties` file in Pig's `conf` directory. The necessary settings are called `fs.default.name` and `mapred.job.tracker`. For a pseudo-distributed cluster, these would be set as follows:

```
fs.default.name=hdfs://localhost/
mapred.job.tracker=localhost:8021
```



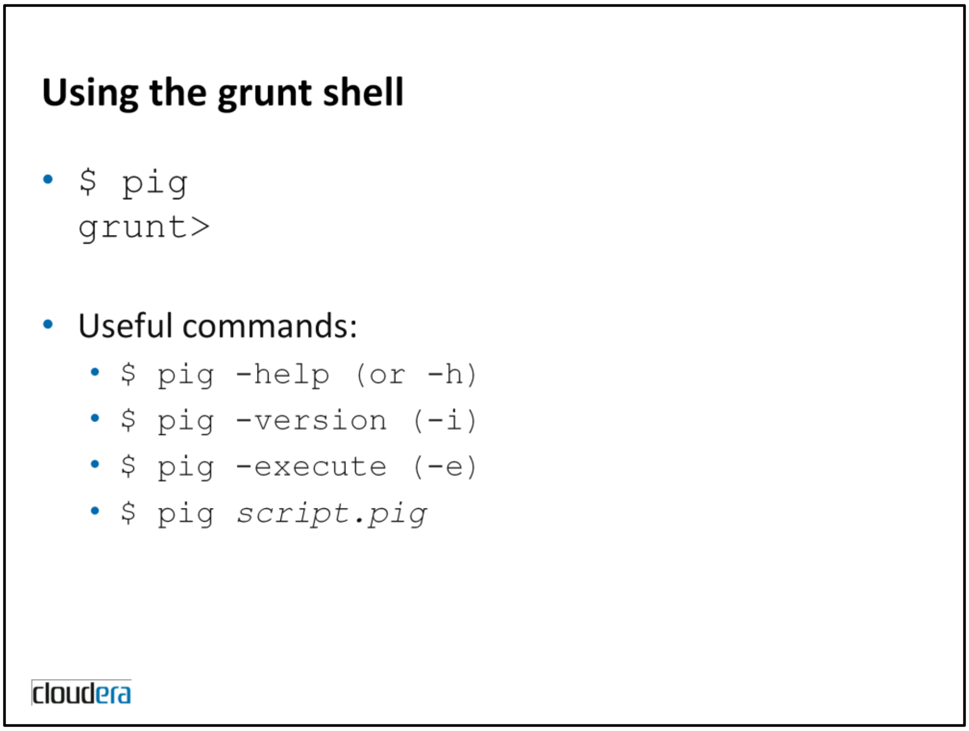
Local vs. Hadoop mode

- Local mode uses Hadoop's LocalJobRunner and the local filesystem
  - `$ pig -x local`
  - Good for debugging on small data sets
- Hadoop mode runs the jobs in the Hadoop cluster and reads/writes to HDFS

cloudera

Pig has a useful option to run in "local" mode or on a Hadoop cluster. Local mode uses Hadoop's LocalJobRunner which executes a MapReduce program in a single JVM on the client's machine. The local filesystem is used for reading/writing data. To invoke Pig in local mode, use the "`-x local`" option. This is useful for testing or debugging Pig programs without involving the cluster. Obviously only small data sets should be used in local mode.

In the default mode, Pig submits the job(s) to the Hadoop cluster and reads/writes data in Hadoop's Distributed File System (HDFS).



Using the grunt shell

- `$ pig`
  `grunt>`
- Useful commands:
  - `$ pig -help (or -h)`
  - `$ pig -version (-i)`
  - `$ pig -execute (-e)`
  - `$ pig` *`script.pig`*

Invoking `pig` in a terminal will start the grunt shell. This is an interactive shell that takes commands from the user. Other useful commands from the terminal are:

`$ pig -help` (or for short, `$ pig -h`)

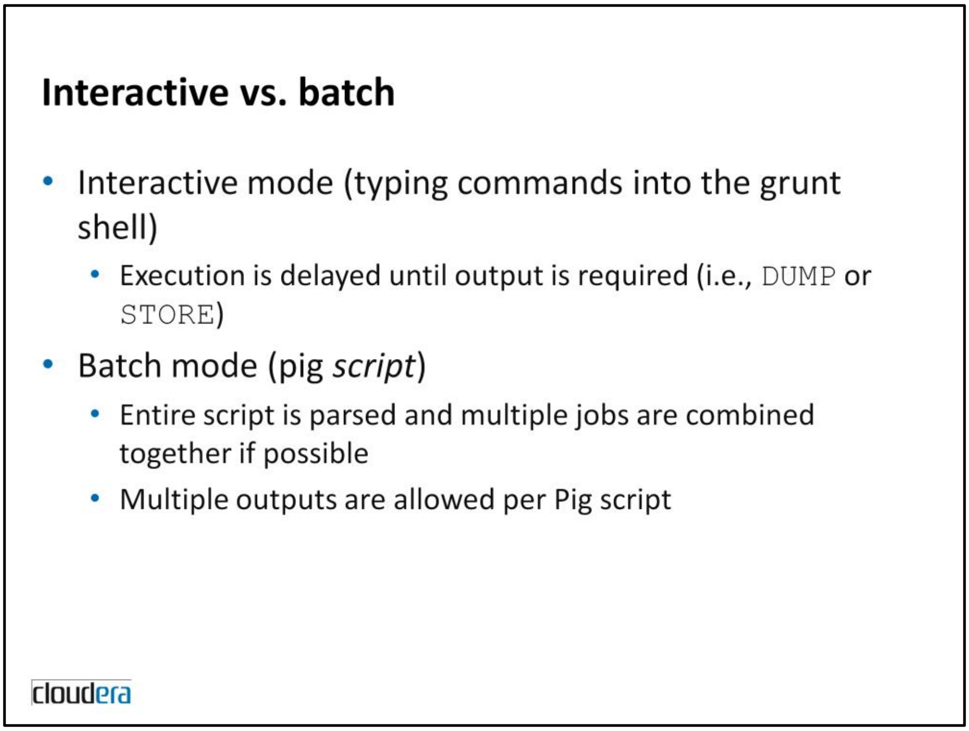
```
$ pig -version
$ pig -execute
```

The last one takes a command or commands to execute. They should be in quotes like this:

```
$ pig -e "fs -ls"
```

Pig can also execute a script containing pig commands:

```
$ pig script.pig
```



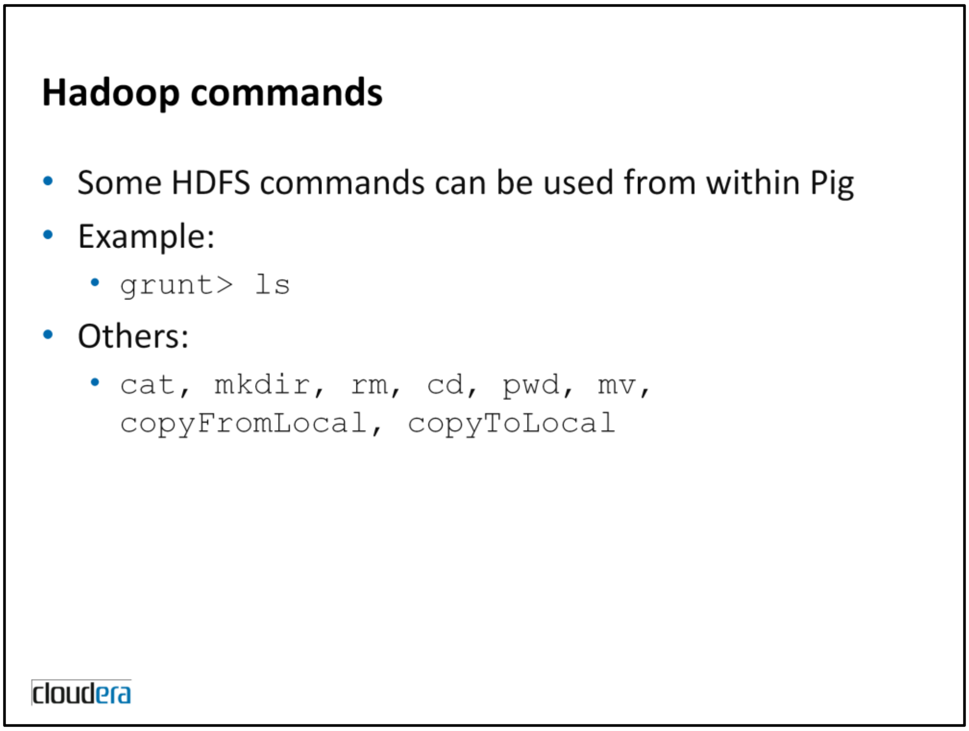
Interactive vs. batch

- Interactive mode (typing commands into the grunt shell)
  - Execution is delayed until output is required (i.e., `DUMP` or `STORE`)
- Batch mode (pig *script*)
  - Entire script is parsed and multiple jobs are combined together if possible
  - Multiple outputs are allowed per Pig script

Running Pig in interactive mode (i.e., typing commands into the grunt shell) behaves slightly differently than executing a script (batch mode).

Interactive mode: Commands are parsed as they are entered, but execution of the Pig program happens when a `DUMP` or `STORE` command is given. These commands tell Pig that you are ready for output.

Batch mode: The *entire* script is parsed even if there are multiple `DUMP` or `STORE` commands. This allows Pig to optimize operations that can be combined.



Hadoop commands

- Some HDFS commands can be used from within Pig
- Example:
  - `grunt> ls`
- Others:
  - `cat, mkdir, rm, cd, pwd, mv, copyFromLocal, copyToLocal`

Pig will accept certain Hadoop commands. For example,

```
grunt> ls
```

is equivalent to:

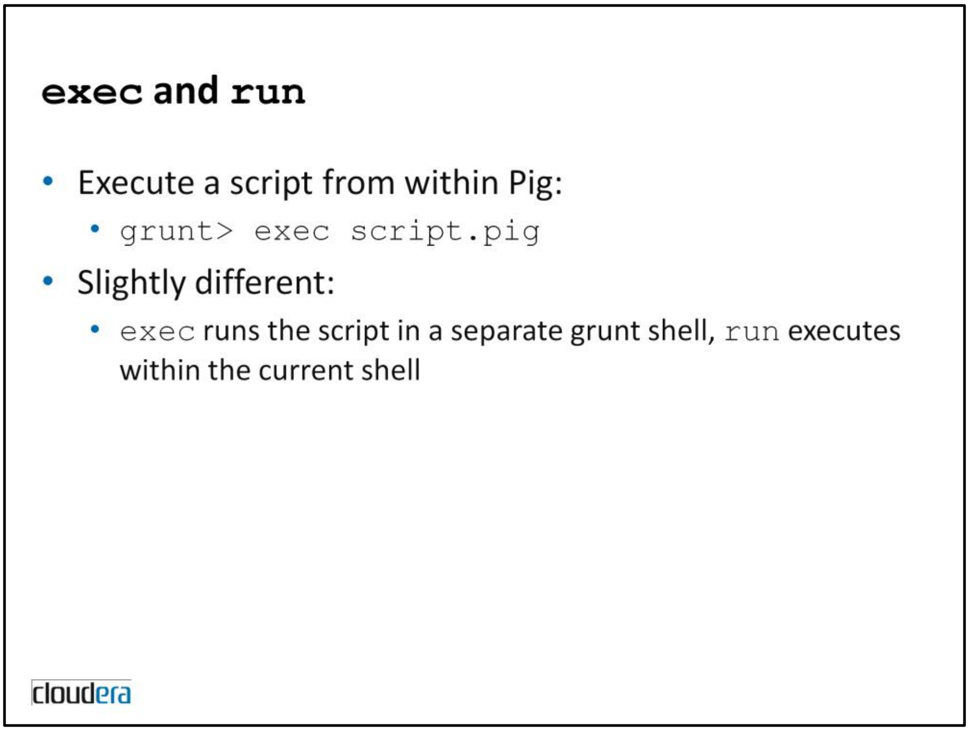
```
$ hadoop fs -ls
```

Several of them act just like their Unix equivalents such as `cat, mkdir, rm, cd, pwd and mv.` There is also copyFromLocal and copyToLocal which are used to copy files from the client's local filesystem to HDFS and vice versus.

Usage:

```
grunt> copyFromLocal /tmp/myfile /user/training/myfile
```

The above command will copy the file `/tmp/myfile` from the client's machine to the Hadoop cluster.

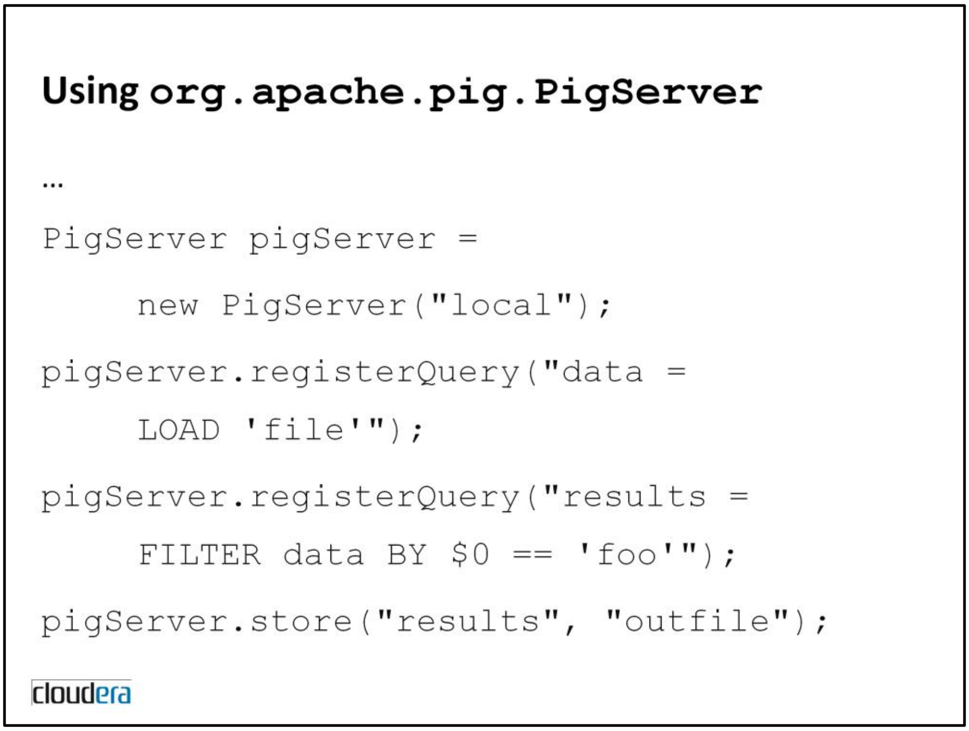


`exec` and `run`

- Execute a script from within Pig:
  - `grunt> exec script.pig`
- Slightly different:
  - `exec` runs the script in a separate grunt shell, `run` executes within the current shell

cloudera

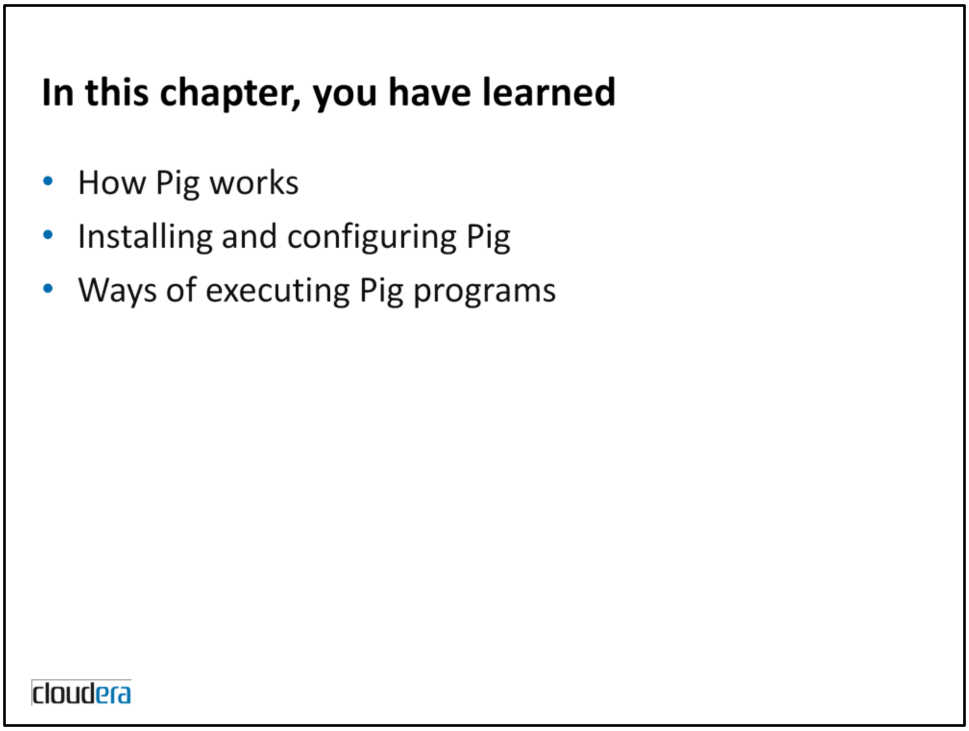
While in the grunt shell, you can use `exec` or `run` to execute a script of commands. The difference between these commands is that `exec` runs the script in a *separate* grunt shell. Any aliases that the script makes are not accessible to the user's shell. However, `run` executes the script *within* the current grunt shell. There are some optimizations (see "multiquery execution" later on) that are only used by `exec`, not `run`.



Using `org.apache.pig.PigServer`

```
…
PigServer pigServer =
     new PigServer("local");
pigServer.registerQuery("data =
     LOAD 'file'");
pigServer.registerQuery("results =
     FILTER data BY $0 == 'foo'");
pigServer.store("results", "outfile");
```

The Pig API comes with a class called `PigServer` which allows Java programs to invoke Pig commands. When instantiating the `PigServer` object, use "local" or "mapreduce" to indicate local mode or Hadoop mode. Then use the `registerQuery` method to include Pig commands and `store` to save the output.



In this chapter, you have learned

- How Pig works
- Installing and configuring Pig
- Ways of executing Pig programs