Hands-On Hadoop

This hands-on lab will walk you through your first experience with using Hadoop. The target audience for this lab is users with some database and/or programming experience, but no prior experience with Hadoop. This lab covers a basic introduction to several core components of the Hadoop platform. Some exercises, however, include *Additional Exercises* sections that provide less guided exercises for more advanced users.

Tips and Tricks

As you're working through the lab exercises, here are some tips to help you along the way:

- To copy and paste within the VM, the easiest way may be to right-click and select *Copy* or *Paste* as CTRL-C and CTRL-V may not be interpreted correctly by the VM.
- If you're executing a command or running a job that produces output, and it is failing, check that the output directory doesn't already exist. In most cases you have to delete or move the existing directory for the command to succeed. To delete a directory in HDFS, run the following command:

  ```
  hadoop fs -rm -R <directory_path>
  ```

- If you find that right-clicking in a text area in Hue no longer gives you the option of pasting from the clipboard, trying closing the browser and reopening it.
- In the terminal window, the Hive shell, the Impala shell, and Pig's Grunt shell you can scroll through previously entered commands using the up and down arrows.
- If you find that Impala isn't able to see a table that you know exists in Hive, you may need to force Impala to refresh it's metadata. You do that by running the command:

  ```
  invalidate metadata;
  ```

  in the Impala Query Editor window or the Impala shell.

Exercise 0: Launching Your Virtual Machine

Before starting this lab, you should take a moment to make sure your virtual machine is configured correctly and all software component are working. Please follow these steps:

1. **Install VirtualBox**

   If you have not already done so, please download and install the latest version of Oracle's VirtualBox software. The URL to download the software is here:

   https://www.virtualbox.org/wiki/Downloads

   You can find installation instructions here:

   https://www.virtualbox.org/manual/ch01.html#intro-installing

2. **Download the Cloudera Quickstart VM 5.5**

   If you have not already done so, please download and unpack the latest version of the Cloudera Quickstart VM. The URL to download the VM is here:

   http://www.cloudera.com/content/cloudera-content/cloudera-docs/DemoVMs/Cloudera-

QuickStart-VM/cloudera_quickstart_vm.html

3. **Launch the Virtual Machine**

   Open VirtualBox and from the *File* menu select *Import Appliance*.

   Browse to the location of the .ovf file.

   Select the .ovf file and click the *Open* button.

   Click the *Continue* button.

   Double-click on the *CPU* field and increase the number of CPUs to about half of the number of cores available on your system.

   Double-click on the *Memory* field and increase the memory to as much memory as your system can sensibly allocate to VirtualBox.

   Click the *Import* button.

4. **Download the lab contents**

   Open a terminal window by selecting *Applications* → *System Tools* → *Terminal* from the desktop menu bar.

   In the terminal window, enter:

   ```
   git clone 'https://github.com/templedf/handsonhadoop.git'
   ```

   Then run:

   ```
   ~/handsonhadoop/fix.sh
   ```

   The `fix.sh` script will unpack the data, place a link to the lab guide onto your desktop, and resolve a couple of knows issues that you would otherwise encounter while doing the exercises in this lab.

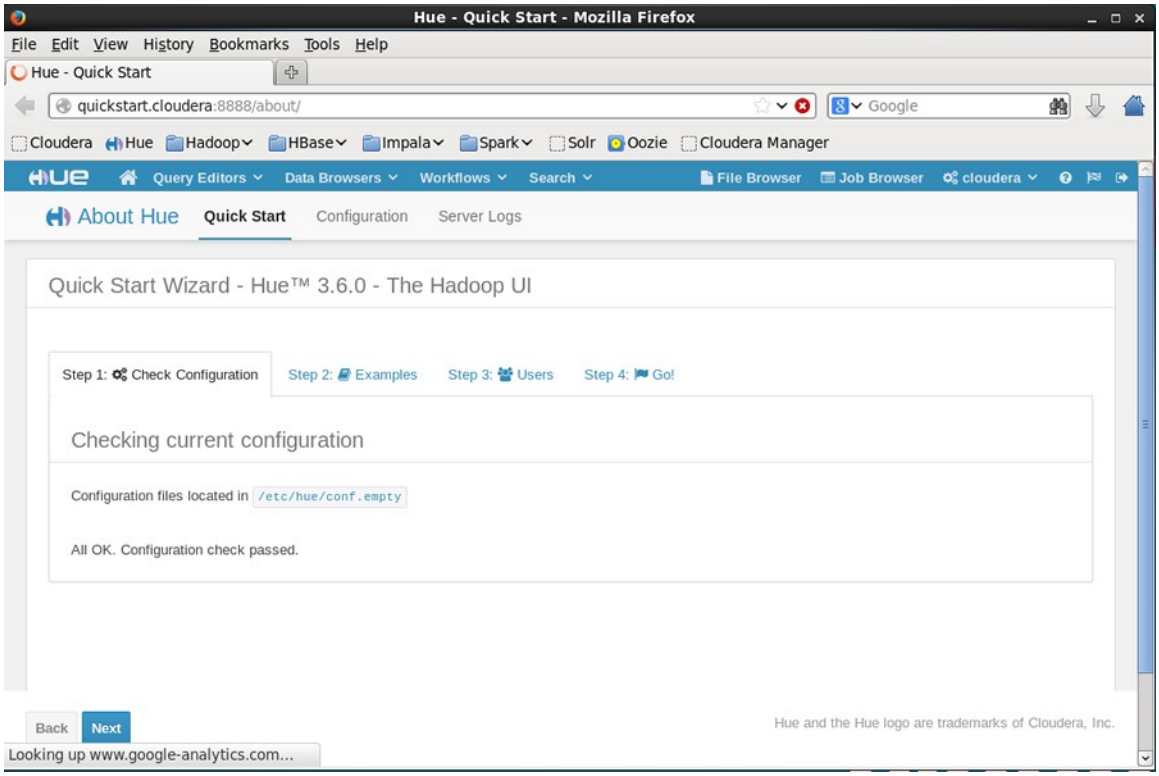
5. **Open a terminal window**

   If you don't already have a terminal window open (e.g. from the previous step), open a terminal window by selecting *Applications* → *System Tools* → *Terminal* from the desktop menu bar.

6. **Open a Firefox browser window**

   If you don't already have a Firefox browser window open, open a Firefox browser window by selecting *Applications* → *Internet* → *Firefox Web Browser* from the desktop menu bar.

   The browser will open to the Hue Quickstart Wizard. You **do not** need to follow the steps in the Quickstart Wizard.



Exercise 1: Exploring HDFS [15 minutes]

In this lab, you will be interacting with Hadoop through the command line in a terminal and through the Hue web user interface using the Firefox browser. If you do not already have a terminal window and a Firefox browser window open, open them now by following steps 4 and 5 in Exercise 0 above.

Introduction

HDFS is the distributed file system around which Hadoop is built. HDFS allows for the reliable storage of data at massive scales and is optimized for the distributed computations that are the staple of MapReduce programs (typically called *jobs*).

HDFS is separate from the local file system, and the data it stores is spread out across the nodes in the Hadoop cluster. A single file in HDFS may be stored as thousands of *blocks* that are distributed and replicated across the entire cluster.

HDFS, like your local file system, has a notion of your user's home directory. In this VM environment, your user name is *cloudera*. Your HDFS home directory is `/user/cloudera`.

Using HDFS

1. Interacting with HDFS is done through the `hadoop` or `hdfs` command. In this lab we will use the hadoop command. To view the files in your HDFS home directory, from the terminal window run:

   ```
   hadoop fs -ls
   ```

You should notice that the command returns with no output. The reason is that your home directory is currently empty. Even though there are files in your local file system's home directory, there are currently none in HDFS.

2. As with the Linux ls command, you can view the files in a specific location by specifying a directory path:

   ```
   hadoop fs -ls /
   ```

   The output above shows you the contents of the HDFS root directory, which includes the `/user` directory where the home directories are stored. Notice that the output from the hadoop `fs -ls` command looks similar to the output from the Linux `ls -l` command. File listings in HDFS are always in the long format, meaning that the file mode, ownership, and size is always displayed.

Data Ingest

In the later sections of this lab, you'll be performing transformations and queries on data in HDFS. Since HDFS is currently empty, you'll need to load some data into HDFS for those later sections. In the `/home/cloudera/handsonhadoop/data` directory you will find a data file called `users.csv`. This file contains data that will be used through the rest of this lab.

1. Before you upload the file into HDFS, you should first examine the file a little. Run the following commands:

   ```
   ls -lh ~/handsonhadoop/data/users.csv

   wc -l ~/handsonhadoop/data/users.csv

   head ~/handsonhadoop/data/users.csv
   ```

   From the output of these commands, you can see that this file is 12MB and contains 200,000 records of fairly simple comma-delimited data points. You can see from first few rows that the data appears to be some kind of user records.

2. To upload the file into HDFS, use the `hadoop fs -put` or `hadoop fs -copyFromLocal` command. Both are equivalent:

   ```
   hadoop fs -put ~/handsonhadoop/data/users.csv
   ```

3. Run the command to list the files in your home directory again:

   ```
   hadoop fs -ls
   ```

   You'll find that the `users.csv` file is now there. Notice that the file size reported by HDFS is the same as the file size reported by the local file system.

   Notice that the file was placed into your HDFS home directory even though the path on the local file system includes a directory path. Because you did not specify where to place the file, HDFS automatically places the file directly into your home directory. The additional directories in the local file path are not created or transferred.

4. While it's excellent that you now have your data loaded into HDFS, it's generally a good idea to keep your home directory tidy, just as with your local file system. You should create a directory in HDFS for this lab and store the data file there. To create a directory in HDFS, run the following command:

```
hadoop fs -mkdir data
```

5. List the files in your home directory again:

   ```
   hadoop fs -ls
   ```

   You will see both the data file and the directory you just created.

NOTE: HDFS has no notion of the current working directory, so whenever you issue an HDFS command, you must either give a fully qualified path or a path that is relative to your home directory.

---

6. You can now move the file from your HDFS home directory into the new directory by running:

   ```
   hadoop fs -mv users.csv data
   ```

7. In the `/home/cloudera/handsonhadoop/data` directory you will find a second data file called `logins.log` to be uploaded into HDFS. Run the following commands:

   ```
   wc -l ~/handsonhadoop/data/logins.log

   head ~/handsonhadoop/data/logins.log
   ```

   From the output, you can see that the `logins.log` file has very different contents from the `users.csv` file.

8. Upload this second data file into your newly created directory by running:

   ```
   hadoop fs -put ~/handsonhadoop/data/logins.log data
   ```

Data Access

Now that you have your data in HDFS, you may be wondering what you can do with it. The rest of the exercises in this lab will focus on using the core tools in the Hadoop platform to work with the data you've uploaded. There are, however, some other common ways to access the data in HDFS that belongs in this exercise.

1. The simplest way is using the `hadoop fs -get` command:

   ```
   hadoop fs -get data/users.csv
   ```

   After running the above command, you will have a copy of the `users.csv` file in your shell's working directory on the local filesystem. As with the `hadoop fs -put` command, if you want to place the file into another directory or rename it during the transfer, you can provide a second argument to the command.

2. One common pattern in Hadoop is to treat a directory like a single file. HDFS is a write-once file system, meaning that once a file has been written, it cannot be later modified. (In some cases, appending to an existing file is supported.) For this reason (and a couple others), users will typically treat a directory like a file. If you want to modify the “file” represented by the directory, you can add or delete files within the directory. You will see this pattern used in later exercises. To facilitate this pattern, HDFS includes a command to download a directory as a single file, `hadoop fs -getmerge`. To download the `data` directory as a single file, run the following command:

```
hadoop fs -getmerge data ~/data.csv
```

3. View the line counts of the files in your local filesystem's home directory:

   ```
   wc -l ~/handsonhadoop/data/* ~/data.csv
   ```

   You will see that the data.csv file is the size of the users.csv file and login.log file combined.

NOTE: The merged file has exactly the contents of all the files in the data directory merged together. There are no headings to mark file boundaries, and the individual file names are not preserved. The data directory is treated as a single data set that just happens to be stored in more than one file.

---

4. In this case, having both files in the same directory is probably a bad idea since the two files have different schemas. To fix that issue, run the following commands to create two new subdirectories and move the files into them:

   ```
   hadoop fs -mkdir data/users

   hadoop fs -mv data/users.csv data/users

   hadoop fs -mkdir data/logins

   hadoop fs -mv data/logins.log data/logins

   hadoop fs -ls -R data
   ```

   From the output of the last command, you can see that you now have a tidy directory structure for your data.

5. What if your data is so large, that downloading it just to look at it is inconvenient or impossible? Hadoop gives you several methods for access your data in HDFS without downloading it.

   If you just want to see what kind of data is in a file, the `hadoop fs -tail` command is useful. Just like the Linux `tail` command, `hadoop fs -tail` shows you the end of the file. (Note that in some versions of Hadoop, the `hadoop fs -tail` command has a bug that causes the beginning of the first line of output to be left off.) Run the following command:

   ```
   hadoop fs -tail 'data/users/users.csv'
   ```

   The output shows you the last lines of the user data file.

6. A more generally useful way to access your data in HDFS without downloading it to the local filesystem is the `hadoop fs -cat` command, which behaves the same as the Linux `cat` command. The `hadoop fs -cat` command is particularly helpful when you want to perform some operation on the file contents, such as using `grep` to look for a particular bit of text or using `wc -l` to count the lines. To count the lines in your data files, run the following commands:

   ```
   hadoop fs -cat 'data/users/*' | wc -l

   hadoop fs -cat 'data/logins/*' | wc -l
   ```

The output should be the same as what you saw when counting the lines of the files on your local filesystem earlier.

Notice that you used two commands with a wildcard for the file name, rather than of one command with wildcards for the subdirectory and file name, i.e. `data/*/*`. The reason is that if you had also used a wildcard for the subdirectory, both data files would have been concatenated together, and the line count would have been the total for both files.

Notice that the expressions with wildcards are in single quotes. You must quote (single or double) any wildcards that you want to pass in the command line to HDFS. If you do not, the local shell will interpret them according to the local filesystem, which in most cases not what you want. (Alternately, you can escape wildcards with a backslash (\))

Also notice that in these two command you used the wildcard to effectively treat the subdirectories as files. Were there more than one file in each, you'd have gotten the total line count for all the files in each subdirectory. This use of wildcards is very common in Hadoop, and we'll use it repeatedly in this lab.

Additional Exercises

1. In the Firefox browser window, click on the *Manage HDFS* button in the upper right corner of the Hue UI. Use the Hue file browser to examine the directories and files in your home directory.
2. Write a command to count the total number of lines in both data files.
3. Write a command to count the number of users named Daniel.
4. Write a command to print the first and last names of all users in California.
5. Write a command to print the list of unique first names.

Summary

You have now explored getting data into and out of HDFS. You understand how to issue HDFS commands, how the HDFS directory structure is similar to the Linux directory structure, and how to use wildcards.

For more information about Hive, these resources are recommended:

- Hadoop: the Definitive Guide: http://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1449311520
- The Hadoop File System Shell documentation: http://archive.cloudera.com/cdh5/cdh/5/hadoop/hadoop-project-dist/hadoop-common/FileSystemShell.html
- The HDFS User's Guide: http://archive.cloudera.com/cdh5/cdh/5/hadoop/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html

Exercise 2: Working With Hive [30 minutes]

In this lab, you will be interacting with Hadoop through the command line in a terminal and through the Hue web user interface using the Firefox browser. If you do not already have a terminal window and a Firefox browser window open, open them now by following steps 4 and 5 in Exercise 0 above.

Introduction

Hive is a component of the Hadoop ecosystem that lets users operate on data stored in HDFS using a

SQL-like language called HiveQL. HiveQL is almost SQL-92 compliant, so most of HiveQL will be familiar to users who have experience with SQL.

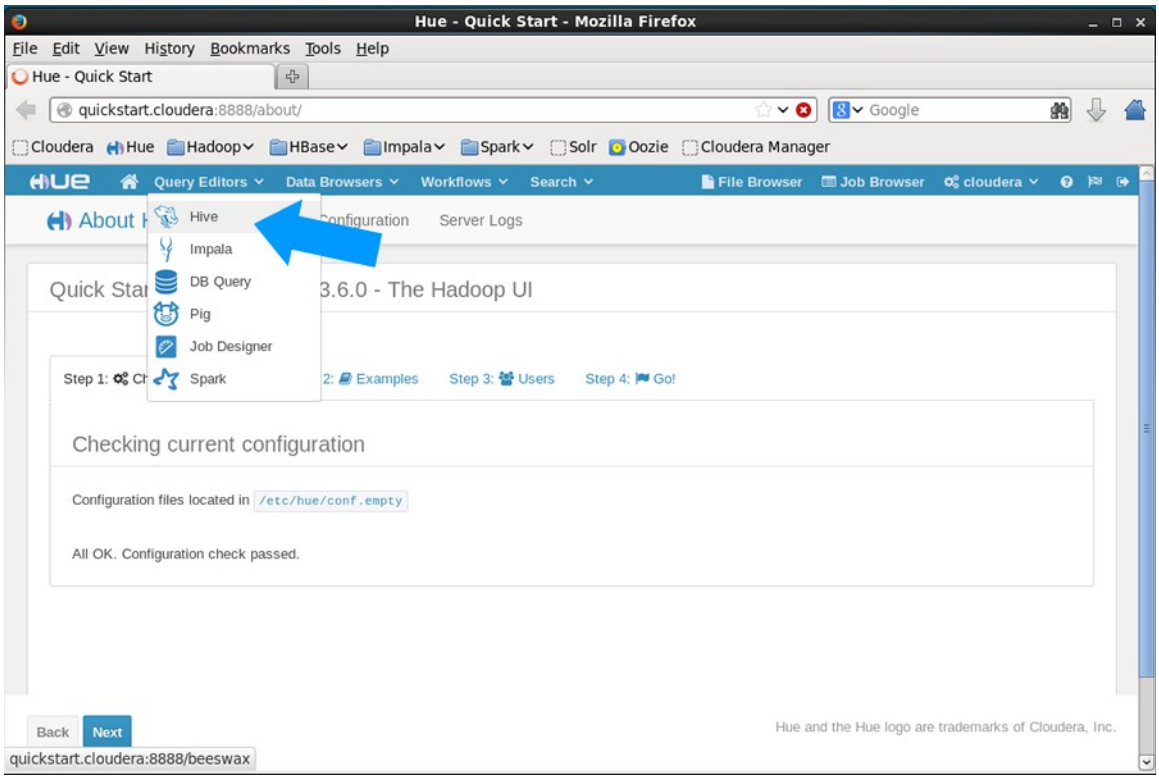
Before going any further, it's worth taking a moment to discuss how Hive can let users query "unstructured" data. If there is no structure, how can it be queried? Except in a few (generally uninteresting) cases, all data has some structure to it. The term "unstructured" therefore generally means data that has structure but does not have a fixed schema. A CSV file is unstructured by that definition, as there's nothing enforcing the types or sizes (or number!) of fields in a CSV file. If a line in a CSV file contains bad data, then that line either requires human intervention to fix, or it just isn't readable. Hive allows users to overlap a schema on top of data and then query that data using the schema. The powerful thing about Hive is that if that schema turns out not to be useful, then the schema can be discarded or replaced without changing the data at all. This concept is often known as "schema on read." The schema is not applied to the data until you try to query it.

Creating Metadata

For this exercise, we're going to use the Hue user interface rather than the terminal. You can launch the Hive shell from the command line by running the `hive` command, but the user experience with Hue is much nicer for many purposes.

In Hive, before we can work with our data, we first have to associate a schema with it. The schema is most often referred to as the "metadata." The way to associate metadata with our data is to create a new Hive table. Unlike in a traditional database, in Hive a table is nothing more than metadata. The data itself is stored independently in HDFS and has a separate life cycle from the metadata, e.g. you can delete the metadata without impacting the data, and vice versa.

1. In the Hue browser window, click on *Query Editors* → *Hive* on the Hue menu bar.



2. In the *Query Editor* box, enter:

   ```
   show tables;
   ```

3. Click on the *Execute* button.

The output will be displayed when the command completes. What you should see is that there are currently two tables defined in Hive, `sample_07` and `sample_08`. Those are two sample tables that come with the Quickstart VM. You won't be using them in this lab. (If you did not follow the steps of the Hue Quickstart Wizard, you may not see these two tables, which is fine.)

4. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
5. In the *Query Editor* box, enter:

```
create external table users (
 id int, fname string, lname string, address string,
 city string, state string, zip int
)
location '/user/cloudera/data/users';
```

   In the `create` statement, the keyword `external` tells Hive that the metadata should be decoupled from the data. If the `external` keyword is left out, deleting the table will also delete the data from HDFS.

   The `location` keyword tells Hive where to find the data to which this metadata should be applied. The location must be a directory; it cannot be a single file. If you do not specify a location, Hive will create a directory for the table under `/user/hive/warehouse`.

6. Click on the *Execute* button.

   The output will be empty as this SQL statement does not return any results.

7. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
8. In the *Query Editor* box, enter:

```
select * from users;
```

9. Click on the *Execute* button.

   The output will show all users in the table. Notice that there's something wrong with the data. All columns in all rows are “NULL”. In the previous exercise you verified that the data in HDFS is complete and correct, so the issue is most likely with the metadata, i.e. the table definition.

   You may recall that the data file is comma-delimited. By default, Hive uses 0x01 (or CTRL-A) as the field delimiter, which would explain why all the columns are “NULL”: Hive is using the wrong delimiters to parse the data.

   To resolve the issue, first you should remove the bad table definition. Recall that you created the table with the external keyword, which means you can delete the table without impacting the data.

10. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
11. In the *Query Editor* box, enter:

```
drop table users;
```

12. Click on the *Execute* button.

    The output will be empty as this SQL statement does not return any results.

13. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
14. In the *Query Editor* box, enter:

```
create external table users (
 id int, fname string, lname string, address string,
 city string, state string, zip int
)
row format delimited fields terminated by ','
location '/user/cloudera/data/users';
```

The `row format` portion of the statement tells Hive to use commas as the field delimiter. By default, Hive uses 0x01 (or CTRL-A) as the field delimiter.

15. Click on the *Execute* button.

    The output will be empty as this SQL statement does not return any results.

16. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
17. In the *Query Editor* box, enter:

    ```
    select * from users;
    ```

18. Click on the *Execute* button.

    The output will show all users. Notice that the data appears to be correct now.

Querying

Now that table queries are working correctly, you can explore the data using some more complex queries.

1. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
2. In the *Query Editor* box, enter:

   ```
   select id, fname, lname from users where state='CA';
   ```

3. Click on the *Execute* button.

   The output will show all of the users located in California. Notice that the processing time for this query is significantly more than for the previous query. The reason this query is slower to process is that it requires running a MapReduce job to complete. The previous query was the special case of a sequential read of the data, which is handled by reading directly from HDFS.

4. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
5. In the *Query Editor* box, enter:

   ```
   select state, count(*) as count from users group by state;
   ```

6. Click on the *Execute* button.

   The output will show each state along with the number of users in that state.

7. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
8. In the *Query Editor* box, enter:

   ```
   select state, collect_set(id) as ids from users group by state;
   ```

9. Click on the *Execute* button.

   The output will show each state along with a list of the user IDs in that state. One feature of Hive that is not commonly available in traditional databases is the use of complex data types, like lists and maps. Performing this query with a traditional database can be quite challenging.

   The `collect_set` operation combines values from a column across different rows into an array of unique elements. Hive supports three non-primitive data types: arrays, maps, and structs. Here you've named the collection `ids.`

Non-tabular data

The previous steps worked with tabular user data. As long as the delimiters and field types are set reasonably, tabular data is quite easy to use in Hive. Fortunately, non-tabular data, such as the

`logins.log` file, can also be processed in Hive easily.

1. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
2. In the *Query Editor* box, enter:

```
create external table logins (
 id string, state string, time string, day string
)
row format serde 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
with serdeproperties (
  "input.regex" =
"(\\d+) in ([A-Z]{2}) at (\\d{2}:\\d{2}:\\d{2}) on (\\d{2}/\\d{2}/\\d{2})"
)
location '/user/cloudera/data/logins';
```

   This time the row format is `serde`, which means you'll be using a specified class for serializing and deserializing the data. The SerDe you're using is the regular expression SerDe, with the regular expression used to parse the data set specified in the `input.regex` property. [Note that the regular expression string must be given all on one line, even though it wraps onto a second line in the text above. When copying and pasting, it should copy as a single line.]

3. Click on the *Execute* button.

   The output will be empty as this SQL statement does not return any results.

4. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
5. In the *Query Editor* box, enter:

```
select * from logins;
```

6. Click on the *Execute* button.

   The output will show all logins in the table. With minimal effort, you have now turned a text log file into a queryable table.

Joins

One of the most powerful features of SQL is the ability to join two or more data sets together. Hive extends that capability to function across different data formats. You'll now demonstrate those capabilities by joining together the two tables you just created.

1. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
2. In the *Query Editor* box, enter:

```
select id, collect_set(state) as states from logins group by id;
```

   The `collect_set` operation combines values from a column across different rows into an array of unique elements. Hive supports three non-primitive data types: arrays, maps, and structs. Here you've named the collection `states.`

3. Click on the *Execute* button.

   The output will be every user id along with an array of all the states from which that user logged in.

4. Click on the *Query Editor* tab in the menu bar to return to the query editor page.
5. In the *Query Editor* box, enter:

```
select users.id, users.fname, users.lname, states.states from users
join (
 select id, collect_set(state) as states from logins group by id
) states on users.id = states.id;
```

6. Click on the *Execute* button.

   This query does two things that are interesting. The first is that it performs a join between two tables. The second is that one of the joined tables is a nested query.

   First the nested query is executed. It's the same query you entered in the previous steps. The results of that query are stored in a temporary table that you've named `states`.

   Second, the `users` table is joined to the `states` table by combining rows where the ID columns are the same. The final results are the ID, names, and login states for all users.

Additional Exercises

1. Place a second copy of the `logins.log` file in the data/`logins` directory and query the `logins` table. What happens? Delete both copies of the `logins.log` file from the data/`logins` directory and query the table again. What happens? Be sure to put a single copy of the `logins.log` file back in the data/`logins` directory for later exercises to use.
2. Write a query to return the list of unique states where users live.
3. Write a query to return the Ids of all users who logged in between 00:00 (midnight) and 01:00.
4. Write a query that lists the names of the users who logged in from each state, with one row per state. For example, if Bob Jones and Sally Jean logged in from CA, then the CA line would contain:

   ```
   CA    ["Bob Jones", "Sally Jean"]
   ```

5. Write a single query statement to calculate the mean number of states from which users have logged in. For example, if Bob logged in from 1 state, and Sally logged in from 2 states, and there are no other users, the mean is (1+2)/2 = 1.5.

Summary

In this exercise, you've only just scratched the surface of what's possible with Hive. You have seen how to overlay metadata over existing data to make it queryable, both with tabular and non-tabular data. You've seen how to work with the data behind the tables. You've seen how to use both joins and nested queries to perform complex queries in Hive.

To read more about Hive, these resources are recommended:

- Hadoop: the Definitive Guide: http://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1449311520
- Hive Language Manual: https://cwiki.apache.org/confluence/display/Hive/LanguageManual

Exercise 3: Working With Impala [10 minutes]

In this lab, you will be interacting with Hadoop through the command line in a terminal and through the Hue web user interface using the Firefox browser. If you do not already have a terminal window and a Firefox browser window open, open them now by following steps 4 and 5 in Exercise 0 above.

Introduction

Impala is a technology for executing high-performance queries on data in HDFS. Impala is built to be fully compatible with Hive, including support for HiveQL and the sharing of Hive's metadata. The primary difference between Impala and Hive is that Impala is a proper parallel database engine, whereas Hive is a toolkit that translates HiveQL statements into MapReduce jobs. The result is that Impala will typically perform HiveQL queries significantly faster than Hive can, using the same data

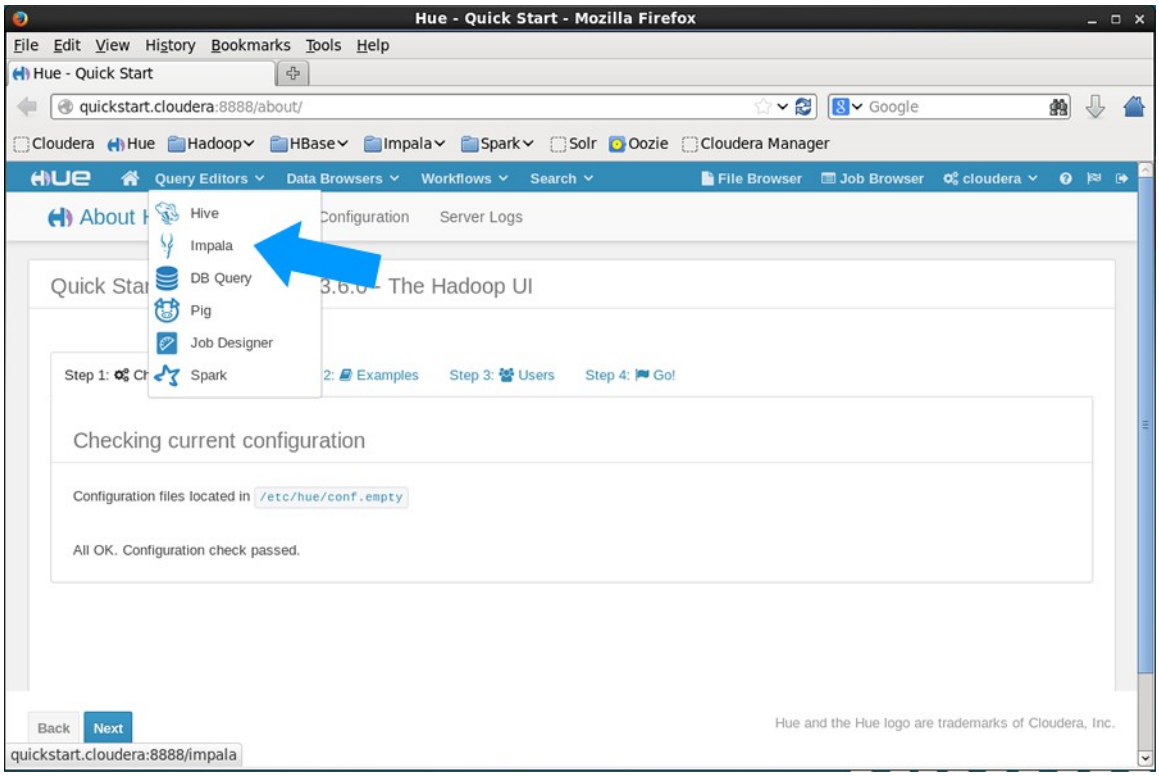
and same machines. Another notable advantage of Impala is that if you're familiar with using Hive, you're also familiar with using Impala, as the same query language is supported, with only some minor differences.

As with Hive, Impala can be used from the Impala shell or the Hue web UI. You will use the Hue UI in this exercise as it gives a richer experience. To launch the Impala shell instead, run the `impala-shell` command in the terminal window.

Querying

Since Impala supports the same syntax as Hive, in the following steps you'll rerun the Hive queries from the previous exercise to compare the performance and results.

1. In the Hue browser window, click on *Query Editors* → *Impala* link on the Hue menu bar.



2. In the *Query Editor* box, enter:

   ```
   invalidate metadata;
   ```

   This command ensures that Impala is operating on the same metadata as Hive.

3. Click on the *Execute* button.

   The output will be empty as this SQL statement does not return any results.

4. In the *Query Editor* box, enter:

   ```
   select id, fname, lname from users where state='CA';
   ```

   The output will show all of the users located in California. Compare the processing time with the processing time for the same query in the previous exercise.

5. Click on the *Query Editor* tab in the menu bar to return to the query editor page.

6. In the *Query Editor* box, enter:

   ```
   select state, count(*) as count from users group by state;
   ```

7. Click on the *Execute* button.

   The output will show each state along with the number of users in that state.

You may have noticed that you only ran queries against the users table. The reason is that the logins table uses a Hive SerDe to read the data, and Impala does not support Hive SerDes. This difference is one of the main cases where portability across Hive and Impala is not seamless. To get around the issue, you should use Hive to read the log data and write it into a more traditionally delimited table. This approach is, in fact, a best practice as the performance of queries that have to parse text files using UDFs or SerDes is generally poor.

Summary

In this exercise you reran the same queries from the previous exercise. You saw that Impala was able to use the same tables and data as Hive in many cases, and that the query times in Impala are considerably shorter, even on a single-node cluster running in a VM.

In general, the rule of thumb for deciding between Hive and Impala is to use Impala for exploratory or interactive queries against tabular data, and to use Hive for very large scale batch queries or in cases when a particular Hive SerDe or UDF is needed.

Exercise 4: Working With Pig [30 minutes]

In this lab, you will be interacting with Hadoop through the command line in a terminal and through the Hue web user interface using the Firefox browser. If you do not already have a terminal window and a Firefox browser window open, open them now by following steps 4 and 5 in Exercise 0 above.

Introduction

Apache Pig is an abstraction layer on top of Hadoop to offer high-level data processing as an alternative to writing low-level MapReduce code. Pig is commonly used for extracting, aggregating and transforming data. The Pig interpreter runs on the client machine and translates PigLatin scripts into MapReduce jobs, which are then submitted to the cluster. Pig itself is written in Java, and the Pig data types correspond to the Java data types.

In many ways, Pig is very similar to Hive. Both are abstractions over MapReduce that simplify the process of writing queries and transformations. Both operate by running MapReduce jobs under the covers. Both have a client side “shell”. The main difference is the language each supports. Hive supports HiveQL, and Pig supports PigLatin. While there are some tasks that are easier in Hive or easier in Pig, the primary deciding factor on which to use is typically the skill set of the users. Users with SQL experience will tend to prefer Hive, and users with a preference for scripting with tend to use Pig.

Basic Data Operations

In this exercise, you'll be using the Grunt shell. Pig scripts can also be run via the Hue UI, but because of the constructive nature of Pig, Grunt is often easier to use. Once you have a complete and tested Pig script, running it through the Hue UI may be a better alternative.

You can start the Grunt shell in local or cluster mode. In local mode, all commands are run locally in the Grunt shell's JVM using data on the local filesystem. In cluster mode, all commands are translated into MapReduce jobs that operate on data from HDFS and are submitted to the cluster. Local mode is useful when debugging scripts using small amounts of data. In this lab, you'll work with Pig in clustered mode.

1. In the terminal window, run the following command:

   ```
   pig
   ```

This command will enter you into the Grunt shell in clustered mode. Your prompt will change to the Grunt prompt: `grunt>`. To use local mode, you'd run `pig -x local`.

In the Grunt shell, you build scripts line by line. As long as you don't enter a command that requires Pig to materialize the data in order to proceed (such as printing the results or storing the results to a file), the commands are not actually executed. When a command is entered that requires the data to be materialized, then all preceding commands that are required to produce the data are executed as a single script.

---

2. First, load the user data that you've already stored in HDFS:

   ```
   users = LOAD 'data/users' using PigStorage(',')
   AS (id, fname, lname, address, city, state, zip);
   ```

   In Pig terminology, users is called a relation. Notice how quickly the Grunt shell returns after entering the command. It's fast because the shell didn't actually do anything yet. The data will not be loaded until it is needed.

   When loading data in Pig, you can specify the format of the data through the SerDe, similar to Hive. The PigStorage SerDe assumes tabular data delimited by the given character.

3. Next, look at the structure (schema) of users:

   ```
   DESCRIBE users;
   ```

   You can see that `users` contains 6 fields, named and ordered as specified in the `LOAD` command. Notice, though, that they're all of type `bytearray`. For your purposes that's actually OK, but you could do better. You know, for example, that the ID and zip code are numbers, and the other fields are strings. If you give Pig that information, it will be able to catch bad records earlier, saving you debugging time.

4. To load the data with field types, run the following command:

   ```
   users = LOAD 'data/users' using PigStorage(',')
   AS (id:int, fname:chararray, lname:chararray,
   address:chararray, city:chararray, state:chararray, zip:int);
   ```

   Look at the schema for `users` again:

   ```
   DESCRIBE users;
   ```

   You can see that the ID and zip code fields are now of type int, which means that non-numeric Ids and zip codes will be caught earlier by Pig. If you had left off the type for any of the fields, Pig would have assumed they were of type `bytearray`.

5. To see the data that Pig has loaded, you use the `DUMP` command. Because there is quite a bit of data, you don't want to dump it all to the screen. You can reduce the data set down to 10 lines with the following command:

   ```
   users10 = LIMIT users 10;
   ```

   In Pig, relations are immutable. When you wish to alter a relation, you have to store the results in a new relation.

6. To see the data, run the following command:

   ```
   DUMP users10;
   ```

Pig will now run a MapReduce job to gather the data and reduce it to the first 10 lines. When the job completes, you will see the first 10 lines of the data.

7. To save the data in a relation to HDFS, you use the STORE command. When storing the data, you have the option of setting the serialization format and the delimiter, just as you did when loading the data. To save the data to HDFS as a tab-delimited file, run the following command:

   ```
   STORE users INTO 'data/users.tsv' USING PigStorage('\t');
   ```

   Because the tab character is the default delimiter used by Pig, you could have left off the `USING PigStorage('\t')` part of the command and gotten the same results.

8. To see the results, either open another terminal window or tab and run:

   ```
   hadoop fs -cat 'data/users.tsv/*' | head
   ```

   or use the File Browser in the Hue UI to view the data.

Queries and Transformations

Pig can is useful for both queries and data transformations. It is particularly good at the latter.

1. To find the users from Washington DC, enter the following commands:

   ```
   dc_users = FILTER users BY state == 'DC';

   DUMP dc_users;
   ```

2. To find the list of states and how many users are from each state, enter the following commands:

   ```
   by_state = GROUP users BY state;
   ```

   The `GROUP` command recombines the relation into a new relation where every entry is a tuple with two members. The first is the element on which the data is grouped, the state in this case. This tuple member is always named `group`. The second member is the original data record. This member is named for the grouped relation, `users` in this case.

3. To see the structure of the grouped relation, enter:

   ```
   DESCRIBE by_state;
   ```

4. Next, find the counts:

   ```
   count_by_state = FOREACH by_state GENERATE group, COUNT(users);
   ```

   The `FOREACH-GENERATE` command lets you create new data structures from existing ones. In this case, you're changing the previous tuple into a simpler one that contains only the state and the number of users in each state.

5. Finally, show the results:

   ```
   DUMP count_by_state;
   ```

6. Pig includes a variety of built-in functions and tools for processing. Among the built-in tools are functions for processing text using regular expressions. Using those tools you can load and process the login data as well. First, though, you have to load the data from HDFS. Run the following command:

```
raw_logins = LOAD 'data/logins' USING TextLoader() AS (text);
```

Because the data is not tabular, it makes more sense to load it with the `TextLoader`.

7. Next, to extract the data, use the `REGEX_EXTRACT_ALL` function:

```
unlabeled_logins = FOREACH raw_logins
GENERATE flatten(REGEX_EXTRACT_ALL(text,
'(\\d+) in ([A-Z]{2}) at (\\d{2}:\\d{2}:\\d{2}) on (\\d{2}/\\d{2}/\\d{2})'
));
```

You may see WARNING output about implicit type casts. You can ignore that for this step and the rest of this exercise.

The `REGEX_EXTRACT_ALL` function uses the regular expression to parse the text and returns a tuple containing all matching capture groups. Because `REGEX_EXTRACT_ALL` returns a tuple, and each data row is already a tuple, the result is a tuple that contains a tuple that contains the data. To eliminate the redundant tuple, you have to use the `flatten` function, which tells Pig to remove the outer layer of grouping.

Notice that the regular expression is the same regular expression you used when creating the `logins` table in Hive.

8. The relation produced above does not have the fields labeled. To make the following steps easier, label the fields:

```
logins = FOREACH unlabeled_logins GENERATE $0 as id,
$1 as state, $2 as time, $3 as date;
```

Regardless of the field labels, you can always refer to fields using ordinals. `$0` is the first first, `$1` is the second, etc. Note that the types specified in the above command are optional. The default assumption is `bytearray`.

9. To make sure the data is loading correctly, down-sample and print it.

```
logins10 = LIMIT logins 10;

DUMP logins10;
```

10. Now, you can join the login data to the user data and produce a list of users and from how many states they've logged in. Start with the join:

```
joined = JOIN users BY id, logins BY id;
```

The join specifies the field to be used for the join in all tables. One interesting thing about Pig is that it supports joining more than two data sets. Notice that the join field for the login data is `$0` and not a name. The logins relation produced in the previous steps does not have labels for the fields.

11. Now group the resulting join by the user IDs:

```
joined_by_user = GROUP joined BY users::id;
```

The fields in the joined relation are scoped by their source relations, which is why you much refer to the grouping field as `users::id`.

12. Next, reduce the data set down to the ID and count:

```
user_count = FOREACH joined_by_user GENERATE group, COUNT(joined);
```

13. Finally, down-sample the data and print it:

```
user_count10 = LIMIT user_count 10;

DUMP user_count10;
```

Summary

What you've seen in this exercise is how to use Pig to do some of the same kinds of operations that you perform in the other exercises. You have also gotten a feel for what it's like to write scripts in PigLatin. Because Hive and Pig have very similar capabilities, the choice between them comes down to comfort level with HiveQL versus with PigLatin.

For more information about Hive, these resources are recommended:

- Hadoop: the Definitive Guide: http://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1449311520
- Getting Started: http://pig.apache.org/docs/r0.13.0/start.html
- Pig Documentation: http://pig.apache.org/docs/r0.13.0/

Additional Exercises

1. Assemble the commands from all the steps in this exercise into a single script, and run that script through the Pig Query Editor in Hue. You will have to remove all of the intermediate DUMP commands. You may also wish to print the full result set at the end without down-sampling it.
2. Write a PigLatin script to count the number of unique users who have logged in in the data in the data/`logins` directory in HDFS.
3. Write a PigLatin script to count the number of unique logins in the data in the data/`logins` directory in HDFS.
4. Write a PigLatin script to count the number of users who have logged in from each state.
5. Write a PigLatin script to list the states from which each user has logged in other than the user's home state.

Exercise 5: Working With Hadoop Streaming [20 minutes]

In this lab, you will be interacting with Hadoop through the command line in a terminal and through the Hue web user interface using the Firefox browser. If you do not already have a terminal window and a Firefox browser window open, open them now by following steps 4 and 5 in Exercise 0 above.

Introduction

As explained in the presentation available here:

http://www.slideshare.net/templedf/java-one14-handsonhadoop

Hadoop is composed of two core components: HDFS and MapReduce. As you saw in exercise 1, HDFS is a distributed file system for storing data. MapReduce is the execution engine that underlies Hive and Pig and is the component that did all the actual work in exercises 2 and 4. (Impala does not rely on MapReduce.) In addition to using an abstraction layer like Hive or Pig, you can also use MapReduce directly.

MapReduce jobs have two main components: map tasks and a reduce tasks. To understand these components, imagine you have three friends, and you'd like the to sort a shuffled deck of cards for you so that all the cards of each suit are together, and all the cards within a suit are sorted from 2 to 10, jack, queen, king, ace. How would you do it most efficiently? One approach you might take is to divide the deck into three chunks and hand one chunk to each friend. This is analogous to storing a data set in HDFS where it is broken into chunks that are distributed across the cluster. Each of your friends would then take her cards and sort them into four piles, one for each suit. This is analogous to the map phase. When everyone had finished sorting, you'd then take all of the hearts and give them to one friend. You give the clubs to another, and so on. (Note that this means one friend will have the cards for two suits.) Each of your friends would then sort the cards in each of her suits. This is analogous to the reduce phase. Finally, when all the sorting is done, you'd collect the cards and stack them up to produce a sorted deck of cards. This is analogous to a getmerge from HDFS.

In more concrete terms, MapReduce will read an input data set from HDFS and feed it, record by record, to a set of independent map tasks. Each map task will produce from each record some amount of intermediate data that has a key associated with it. When all of the input data has been processed by the map tasks, MapReduce will gather up all the intermediate data, group it by keys, and pass it to a set of independent reduce tasks. Each reduce task is guaranteed to get all of the data for each key that it receives. Each reduce task will read the intermediate data, key by key, and produces some amount of final data that is output back into HDFS. Can you see now how sorting cards fits into this model?

There are a couple of methods for using MapReduce, primarily MapReduce jobs written in Java and MapReduce jobs executed through the Hadoop streaming tool. In this exercise we'll focus on Hadoop streaming. Hadoop streaming actually has nothing to do with data streaming. It is instead a tool for running MapReduce jobs not written in Java. (There is also a separate tool for writing jobs in C++ called Hadoop pipes.) Hadoop streaming can be used to run jobs written in any language, including Python, Perl, C, Java, and even Linux command-line utilities.

In this exercise, you'll use Hadoop streaming to operate on the data you uploaded into HDFS in exercise 1.

Hadoop Streaming

1. In the terminal window, run the following command:

   ```
   export STREAMING=/usr/lib/hadoop-mapreduce/hadoop-streaming.jar
   ```

   This command will add an environment variable that you'll use in the following steps. [Should you open a new terminal window or new tab, be sure to run this command there as well.]

2. To get familiar with Hadoop streaming, you'll start with the simplest job that does something useful: translate the `users.csv` file from comma-delimited to tab-delimited.

   In the terminal window, run the following command:

   ```
   hadoop fs -rm -R data/users.tsv

   hadoop jar $STREAMING -mapper "tr , '\t'" -numReduceTasks 0 \
   -input data/users -output data/users.tsv
   ```

   The `-mapper` argument tells Hadoop streaming what command to run as the map task. The `-numReducerTasks` sets the number of reducers, in this case none. The `-input` and `-output` arguments tell Hadoop streaming where to get the data and put the results.

   When the job has finished, the output will show you many lines of data about the job execution. You can see, for example, the number of records and the number of bytes of data that were processed in each phase. Notice that there is no information about the reduce phase because this was a “map-only” job.

NOTE: When you run a Hadoop job, the output directory must not already exist. If you are attempting to rerun a job, even if the job failed on the previous attempt, you must always remove the output directory first. The command to remove the output directory for the above job is:

```
hadoop fs -rm -R data/users.tsv
```

The `-R` works just like it does in the Linux `rm` command – it tells Hadoop to delete the directory and everything in it.

---

3. To see the results produced by the job, run the following commands:

   ```
   hadoop fs -ls data/users.tsv
   ```

   Notice that the file listing shows that the `users.tsv` "file" is really a directory that contains a number of "part" files, each produced by one map task. Typically the part files are produced by the reduce tasks, but since this was a map-only job, the map tasks produced the output directly.

4. To view the output, run:

   ```
   hadoop fs -cat 'data/users.tsv/part-00000' | head
   ```

   You can see that all of the commas in the file have now been replaced by tabs.

5. To try something a little more complicated, count the number of users in Washington DC. Run the following command:

   ```
   hadoop jar $STREAMING -mapper "grep ,DC," -reducer "wc -l" \
   -numReduceTasks 1 -input data/users -output data/dccount
   ```

   Here you've set `grep` as the map task, looking for records that have "DC" as a field, and you've set `wc -l` as the reduce task. Note that you've set the number of reduce tasks to 1 explicitly. Why only one?

6. To view the output, run:

   ```
   hadoop fs -cat 'data/dccount/*'
   ```

   The results should agree with the count you found earlier.

7. To get the counts for all the states, run the following commands:

   ```
   hadoop jar $STREAMING -mapper "awk -F, '{print $6}'" -reducer "uniq -c" \
   -input data/users -output data/count
   ```

   ```
   hadoop fs -cat 'data/count/*' | head
   ```

   Notice that you did not need to explicitly set the number of reducers to 1 this time. Why not?

   The results should again agree with the counts you found earlier.

   Recall that `uniq -c` requires that the data be in sorted order. You can see from this command, then, that between the map and reduce phases, MapReduce is collecting and sorting the intermediate data, a phase know as the shuffle and sort phase.

8. Now, to understand Hadoop streaming a little deeper, you'll write a short MapReduce job in Python. Create a file in your home directory called `map.py` with the following contents:

   ```
   import sys

   # Read records from stdin
   for line in sys.stdin:
       # Split them on commas
       fields = line.strip().split(',')
   ```

```
    # Output the state followed by the full name
    print "%s\t%s %s" % (fields[6], fields[1], fields[2])
```

(You can either use `vi` or `emacs` from the command line, or you can open a text editor by selecting *Applications* → *Accessories* → *gEdit Text Editor* from the desktop menu bar.)

The most important thing to note that this script is that it is reading input records from STDIN and writing the intermediate data out to STDOUT. This input/output contract is the only requirement placed on the commands executed by Hadoop streaming. By default, the tab character (`\t`) is taken as the delimiter between the key and the value in the output from a task, but as long as both the map and reduce phases are commands begin run through Hadoop streaming, the key/value delimiter makes very little difference. (It is possible to include Java MapReduce tasks in a Hadoop streaming job, in which case use of the key/value delimiter matters.)

Create a second file in your home directory called `reduce.py` with the following contents:

```
import sys

# Track the last key seen
last = None

# Read records from stdin
for line in sys.stdin:
    # Split the records on tabs
    fields = line.strip().split('\t')

    # If this is a new key...
    if last != fields[0]:
        # If there was a previous key...
        if last != None:
            # Print out the cached data
            print "%s,%s" % (last, str(users))

        # Store the new key and reset the cache
        last = fields[0]
        users = list()

    # Add this name to the cache
    users.append(fields[1])

# If there was a last key, print out the cached data
if last != None:
    print "%s,%s" % (last, str(users))
```

Notice that in Hadoop streaming, it's up to you to manage the keys. In this script, you explicitly track the keys so that when the next key appears you can output the data for the previous key. (Remember that the intermediate data comes to the reduce tasks already sorted by key.)

To run this job, run the following commands:

```
hadoop jar $STREAMING --files ~/map.py,~/reduce.py \
-mapper "python map.py" -reducer "python reduce.py" \
-input data/users -output data/stateusers
```

This time you have given your scripts as the map and reduce tasks.

Note that you did not specify the number of reducer tasks, which means that you will get the

default number of reducers as specified by the cluster config. How many is that?

Note also the `--files` argument. In order to run your scripts, a copy must be available from each of the nodes where the tasks will run. The `--files` argument tells Hadoop streaming to place a copies of the specified files into the working directory of every task on every node where a task will run. This process is called "file staging."

9. To view the output, run:

   ```
   hadoop fs -cat 'data/stateusers/*' | head
   ```

   The output should list each state followed by the names of the users in that state.

Summary

In this quick tour of Hadoop streaming, you've seen how to use Hadoop streaming to parallelize Linux command-line tools to operate on large data sets. You've also seen how Hadoop streaming let's you write Hadoop jobs using scripts and programs that adhere to a simple input/output contract. Because the input/output contract is the only requirement, Hadoop streaming is a very flexible tool, accommodating jobs written in any language that the worker nodes can run.

To read more about Hadoop streaming, these resources are recommended:

- Hadoop: the Definitive Guide: http://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1449311520
- Hadoop Streaming Manual: http://hadoop.apache.org/docs/r2.5.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/HadoopStreaming.html

Additional Exercises

1. Write a Hadoop streaming job that counts the number of unique users who have logged in in the data in the data`/logins` directory in HDFS.
2. Write a Hadoop streaming job that counts the number of unique logins in the data in the data`/logins` directory in HDFS.
3. Write a Hadoop streaming job that parses the data in the data`/logins` directory in HDFS and outputs the user id, state, time, and date, all separated by commas, for each login.
4. Write a Hadoop streaming job that parses the data in the data`/logins` directory in HDFS and outputs the each user id and a list of the states from which that user has logged in.

Exercise 6: Java MapReduce [Bonus]

In this lab, you will be interacting with Hadoop through the command line in a terminal and through the Hue web user interface using the Firefox browser. If you do not already have a terminal window and a Firefox browser window open, open them now by following steps 4 and 5 in Exercise 0 above.

Introduction

As discussed in the previous exercise, Hadoop offers different ways to run jobs. In the previous exercise you looked at Hadoop streaming, which is convenient in many cases, but is not the most common way to run jobs. The most common approach is to write Java MapReduce jobs. In this bonus exercise, you'll examine the anatomy of a Java MapReduce job and then run it.

Job Code

The following code is a complete Java MapReduce job that will produce for each user the list of states from which that user has logged in that are not the user's home state. In the sections following the code you will find a full description of how the code works.

```
 1 package handsonhadoop;
 2
 3 import java.io.IOException;
 4 import java.util.HashSet;
 5 import java.util.Set;
 6 import java.util.regex.Matcher;
 7 import java.util.regex.Pattern;
 8 import org.apache.hadoop.conf.Configuration;
 9 import org.apache.hadoop.conf.Configured;
10 import org.apache.hadoop.fs.Path;
11 import org.apache.hadoop.io.IntWritable;
12 import org.apache.hadoop.io.LongWritable;
13 import org.apache.hadoop.io.Text;
14 import org.apache.hadoop.mapreduce.Job;
15 import org.apache.hadoop.mapreduce.Mapper;
16 import org.apache.hadoop.mapreduce.Reducer;
17 import org.apache.hadoop.mapreduce.lib.input.MultipleInputs;
18 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
19 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
20 import org.apache.hadoop.util.Tool;
21 import org.apache.hadoop.util.ToolRunner;
22
23 public class HandsOnHadoop extends Configured implements Tool {
24   private static final String LOGIN =
25           "(\\d+) in ([A-Z]{2}) at (\\d{2}:\\d{2}:\\d{2}) "
26           + "on (\\d{2}/\\d{2}/\\d{2})";
27
28   public static void main(String[] args) throws Exception {
29     ToolRunner.run(new Configuration(), new HandsOnHadoop(), args);
30   }
31
32   @Override
33   public int run(String[] args) throws Exception {
34     Job job = Job.getInstance(getConf());
35
36     job.setJarByClass(HandsOnHadoop.class);
37
38     MultipleInputs.addInputPath(job, new Path(args[0]),
39             TextInputFormat.class, UsersMap.class);
40     MultipleInputs.addInputPath(job, new Path(args[1]),
41             TextInputFormat.class, LoginsMap.class);
42     FileOutputFormat.setOutputPath(job, new Path(args[2]));
43
44     job.setReducerClass(Reduce.class);
45     job.setOutputKeyClass(IntWritable.class);
46     job.setOutputValueClass(Text.class);
47
48     return job.waitForCompletion(true) ? 0 : 1;
49   }
50
51   public static class UsersMap extends Mapper<LongWritable, Text, IntWritable, Text> {
52     private IntWritable k = new IntWritable();
53     private Text v = new Text();
54
55     @Override
56     protected void map(LongWritable key, Text value, Context context)
57             throws IOException, InterruptedException {
58       String[] parts = value.toString().split(",");
59
60       k.set(Integer.parseInt(parts[0]));
61       v.set("-" + parts[5]);
62
63       context.write(k, v);
64     }
65   }
66
67   public static class LoginsMap extends Mapper<LongWritable, Text, IntWritable, Text> {
68     private final Pattern p = Pattern.compile(LOGIN);
```

```
 69     private IntWritable k = new IntWritable();
 70     private Text v = new Text();
 71
 72     @Override
 73     protected void map(LongWritable key, Text value, Context context)
 74             throws IOException, InterruptedException {
 75       Matcher m = p.matcher(value.toString());
 76
 77       if (m.matches()) {
 78         k.set(Integer.parseInt(m.group(1)));
 79         v.set(m.group(2));
 80
 81         context.write(k, v);
 82       }
 83     }
 84   }
 85
 86   public static class Reduce extends Reducer<IntWritable, Text, IntWritable, Text> {
 87     private Text v = new Text();
 88
 89     @Override
 90     protected void reduce(IntWritable key, Iterable<Text> values, Context context)
 91             throws IOException, InterruptedException {
 92       Set<String> logins = new HashSet<String>();
 93       String home = null;
 94
 95       for (Text state: values) {
 96         if (state.charAt(0) == '-') {
 97           home = state.toString().substring(1);
 98         } else {
 99           logins.add(state.toString());
100         }
101       }
102
103       logins.remove(home);
104       v.set(logins.toString());
105
106       context.write(key, v);
107     }
108   }
109 }
```

Code Explanation

Line 1 is the package statement. This class is in the `handsonhadoop` package.

Lines 3-21 are the imports for this class. You can see that MapReduce jobs need to make use of many different classes from the Hadoop API.

Line 23 is the class declaration. Notice that the class extends `Configured` and implements `Tool`, which is a common idiom in MapReduce jobs. It helps with parsing Hadoop command-line arguments and gives a common structure to your job.

Line 24 is the same regular expression string that you used in the Hive exercise when creating the `logins` table. You use it later in the reduce task.

Lines 28-30 are the `main` method. For classes that extend `Configured`, the `main` method usually consists only of a call to `ToolRunner.run()` to execute the job. The `ToolRunner.run()` method will look for leading Hadoop arguments in the `args` array, process them, and remove them before passing the `args` array to the job. The `Configuration` object that is created and passed to the job holds all of the information about the job and how it should be executed.

Lines 32-48 are known as the “driver code”. The driver code sets up the job to be executed.

Line 34 creates a `Job` object, which is a helper class that makes it easier to set the required properties in the `Configuration` object.

Line 36 sets up the classpath that will be used by the workers running the job tasks to include the JAR

file that contains that job's classes.

Lines 38-41 set up the input paths. Typically there will be a single call to `FileInputFormat.addInputPath()` and `job.setMapperClass()`, but in this case, there are two separate input sources: the user data and the login data, and both need to be processed differently. Instead of the typical approach, this job calls `MulitpleInputs.addInputPath()` twice, each time passing it the path to an input data set and the name of the class that will process it.

Line 42 sets up the output path. There can only be one output path.

Line 44 names the class that will be used as the reducer task.

Lines 45-46 declare the types of the keys and values produced by **both** the map and reduce tasks. The map and reduce tasks will produce integer keys and string values.

In Hadoop, the usual Java types are wrapped in instances of `Writable.` `Writable` is Hadoop's equivalent to Java's `Externalizable`, but it is designed to perform well at scale.

In the case that the map and reduce tasks do not have the same key and value types, then the types for the reduce tasks are declared as they are in lines 45-46, and the types for the map tasks are declared through calls to `job.setMapOutputKeyClass()` and `job.MapOutputValueClass()`. (The reason the types must be declared at all is that type erasure prevents the types declared using generics from being available at run time.)

Line 48 executes the job and waits for it to complete. If it completes successfully, it returns a 0. If it fails for any reason, it returns a 1.

Lines 51-65 are the first map task class. This class is built to parse the user data.

Line 51 defines the class as a subclass of `Mapper` and declares the types of the input and intermediate keys and values. The input key type for a map task is usually `LongWritable`, and the input value type is usually `Text.` In most cases, the input value will a full line of the input data, and the input key will be the byte offset of that data into the source file. In most cases the input key is ignored. The output key and value types must match the types declared in the driver code.

Lines 52-53 declare `Writable` objects to be used as the key and value for the intermediate data. In Hadoop, you have to assume that everything you do will be done at massive scale. On a smaller scale, creating a new `Writable` object for every intermediate data record might not be a bad thing. At Hadoop scales, it can create a significant garbage collection overhead. Because the `Writable` class are designed to be reused, it's common practice to reuse the key and value classes.

Lines 55-64 are the `map()` method, which is where the work of the map task is done.

Line 58 retrieves the text data from the input value object and splits it on commas.

Line 60 sets the intermediate key object to be the first field of the input parsed as an integer.

Line 61 sets the intermediate value object to be the second field of the input with a '`-`' prepended. When there is more than one input data source, it is often useful to flag the data from one or more of the sources so that the reducer can tell them apart.

Line 63 emits the intermediate key and value.

Lines 67-84 are the second map task class. This class is built to parse login data.

Line 67 defines the map class as a subclass of `Mapper` and declares the types for the input and intermediate keys and values, the same as was done in the first map task class.

Line 68 defines regex pattern to use for parsing the login data.

Lines 69-70 define key and value objects to use when emitting intermediate data.

Lines 72-83 are the `map()` method.

Lines 75 and 77 test whether the line of input data matches the regular expression.

If the regular expression matches, then line 78 sets the key to the first matching group parsed as an integer; line 79 sets the value to the second matching group; and line 81 emits the intermediate key and value.

Lines 86-108 are the reduce task class.

Line 86 defines the reduce class as a subclass of `Reducer` and declares the types for the intermediate and output keys and values. The types for the intermediate keys and values must match the types declared by the map task class(es). In this case, because there are two map task classes, both map task classes must have the same types for the intermediate keys and values. In the case of a single job, the output data will be converted to text and written to the output path, so the types of the output keys and values are less important. In the case of "chained" jobs, however, the output from one job will be fed as input to the next job, so the output types may be important.

Line 87 defines an output value object. In this reduce task class the intermediate key output is reused as the output key object.

Lines 89-107 are the `reduce()` method, where the work of the reduce task is done.

Line 92 defines a `Set` in which to keep the login states for the current user.

Line 93 defines a `String` to hold the current user's home state.

Lines 95-101 iterate through the values (states) that are associated with the current key (user ID). There are two important things to know about iterating through the values array in the `reduce()` method. First, you can only iterate through the values once. Any subsequent iterations will see no data. This limitation is why this code keeps a `Set` of states rather than iterating through the values twice. Second, in each iteration through the loop the **same** key and value object will be reused. The contents of the objects will change with every iteration, but the objects themselves will always be the same objects. This behavior is why the object that stores the home state is a `String` and not a `Text`. Were you to store the `Text` object itself, its contents would change with any subsequent iterations through the values. Both of these sometimes surprising behaviors are done to optimize for scalability.

Line 96 tests for the leading '-' that marks the state as having come from the user data. If the value has a leading '-', then line 97 sets it, minus the leading '-', as the home state. If not, then line 99 adds the value to the set of login states.

Line 103 removes the home state from the login states.

Line 104 sets the output value to list of login states.

Finally, line 106 emits the output key and value, reusing the intermediate key and the output key.

Execution

1. To launch this job, run the following command:

   ```
   hadoop jar ~/handsonhadoop/mapreduce/target/mapreduce-1.0.jar \
   handsonhadoop.HandsOnHadoop data/users data/logins data/nohome
   ```

   The three arguments are the path to the user data, the path to the login data, and the output path, in that order.

2. To view the results, run:

   ```
   hadoop fs -cat 'data/nohome/*' | head
   ```

The output should show you a user ID followed by a list of the states from which the user has logged in, excluding his or her home state.

Summary

In this exercise you have walked through a non-trivial Java MapReduce job. You have seen that there are many more moving parts to a Java MapReduce job than with Hadoop streaming or Pig or Hive, but that there is also much power and flexibility. This example only scratched the surface of what's possible from a Java MapReduce job. It did not cover use of the distributed cache, job chaining, secondary sort, custom grouping comparators, custom partitioners, etc.

You can find the source code for the class in the `~/handsonhadoop/mapreduce/src/java` directory. The `~/handsonhadoop/mapreduce` directory is a Maven project, but if you are doing the lab at a location with limited network bandwidth, it is recommended that you do not build the project as it will download dependencies, which may be difficult given the conference network bandwidth constraints.

For a more information about Java MapReduce jobs, see:

- Hadoop: the Definitive Guide: http://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1449311520
- Udacity's Intro to Hadoop and MapReduce: https://www.udacity.com/course/ud617
- Yahoo!'s Hadoop Tutorial: https://developer.yahoo.com/hadoop/tutorial/

Additional Exercises

1. Write a MapReduce job to return the list of unique states where users live.
2. Write a MapReduce job to return the Ids of all users who logged in between 00:00 (midnight) and 01:00.
3. Write a MapReduce job to calculate the mean number of states from which users have logged in. For example, if Bob logged in from 1 state, and Sally logged in from 2 states, and there are no other users, the mean is (1+2)/2 = 1.5.

Summary

In this lab, you have seen the core building blocks of Hadoop: HDFS, MapReduce, Hive, Pig, and Impala. Using these tools, you can tackle most big data problems. These tools, however, are only a few of the tools that are available to you as part of the Hadoop ecosystem.

The Hadoop ecosystem is broad and growing quickly. Cloudera offers certifications in Hadoop administration and development, and in data science. These certifications can help you identify prospects with the skills needed to be successful with big data or self-identify as someone with those skills. For more information about Cloudera's certifications, see http://www.cloudera.com/content/cloudera/en/training/certification.html.

What follows is an incomplete list of other tools and technologies you may wish to explore (items with an asterisk do not come pre-installed on the Quickstart VM):

- Avro – binary storage format that encapsulates metadata
- Crunch – framework to simplify writing MapReduce code
- DataFu – additional utilities for use with Pig
- Flume – tool for streaming data into HDFS
- HBase – distributed key-value store built on HDFS

- HCatalog – shared metadata service for Hive, Pig, Impala, and MapReduce
- Kafka* – tool for handling streaming events
- Mahout – machine learning library partially built on Hadoop
- MRJob* – Hadoop streaming library for Python
- MRUnit – unit testing for MapReduce jobs
- Oozie – orchestrates the execution of Hadoop workflows
- Oracle NoSQL Database* – distributed key-value store
- Spark – alternative to MapReduce that takes advantage of in-memory caching
    - Spark Streaming – using Spark to processing streaming data
    - Spark ML/MLLib – machine learning library built on Spark
    - GraphX – graph processing library built on Spark
- Sqoop – tool to automate the transfer of data between structured data sources (e.g. RDBMS) and HDFS and Hive
- Cloudera Manager – cluster installation and management tool