

# Introduction to Hadoop & MapReduce

## Lab Instructions

# Agenda

- HDFS command line tools
- Alternatives to HDFS command line tools
- Hadoop MapReduce jobs
- Lab Scenarios

# Accessing HDFS Via The Command Line

- HDFS is not a general purpose file system
  - HDFS files cannot be accessed through the host OS
  - End users typically access HDFS via the `hadoop fs` command
- Example:
  - Display the contents of the `/user/fred/sales.txt` file

**hdfs dfs commands perform similar functions with minute differences**

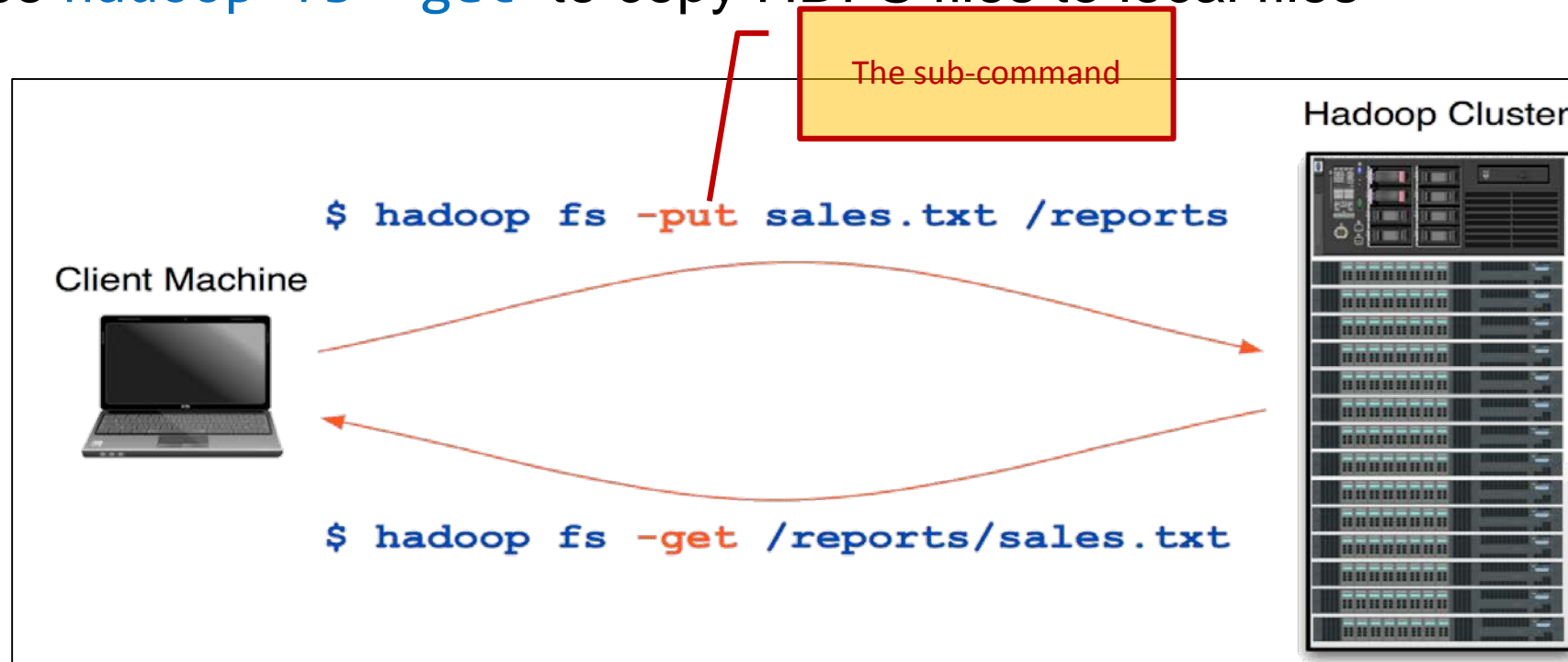
```
hadoop fs -cat /user/fred/sales.txt
```

- Create a directory (below the root) called reports

```
hadoop fs -mkdir /reports
```

# Copying Local Data To And From HDFS

- Remember that HDFS is separated from your local filesystem
  - Use `hadoop fs -put` to copy local files to HDFS
  - Use `hadoop fs -get` to copy HDFS files to local files



# More Command Examples

- Copy file input.txt from the local disk to the user's home directory in HDFS

```
hadoop fs -put input.txt /user/cloudera/input.txt
```

- This is equivalent to `hadoop fs -put input.txt input.txt`
- Get a directory listing of the HDFS root directory

```
hadoop fs -ls /
```

- Delete the **file** /reports/sales.txt

```
hadoop fs -rm /reports/sales.txt
```

- Delete an entire directory and all of its subdirs (be careful not to do this at the root level)

```
hadoop fs -rm -r /dualcore/example/
```

# Are There Alternatives to the Command Line Interface?

- Web UI:
  - On the Hadoop NameNode, access the following URL:
  - <http://localhost:50070> (suppose the localhost is a NameNode)
- Via **Hue** (running over http on an edge node)
  - Hue (Hadoop User Experience) is a Cloudera product that is supported in a variety of Hadoop distributions.
  - On our VMs, username = **cloudera** and Password = **cloudera**
  - <http://localhost:8888>

## Review Question

- HDFS command line tool



# Hue HDFS File Manager

The screenshot shows the Hue HDFS File Manager interface. The left sidebar contains a 'Toggle Menu' button (a left arrow icon) and a 'Files' button. The main area displays the 'File Browser' for the '/dualcore' directory. A table lists files and directories with columns for Name, Size, User, Group, Permissions, and Date. Annotations with callout boxes identify key features: 'Toggle Menu', 'Open Root Directory' (pointing to the '/' icon), 'Manage Files' (pointing to the 'Actions' dropdown), 'Upload Files' (pointing to the 'Upload' button), 'Open File Browser' (pointing to the 'Files' button), and 'Browse Files' (pointing to the file list).

**Toggle Menu**

**Open Root Directory**

**Manage Files**

**Upload Files**

**Open File Browser**

**Browse Files**

Name	Size	User	Group	Permissions	Date
hdfs		hdfs	supergroup	drwxr-xr-x	November 13, 2017 12:45 PM
training		training	supergroup	drwxrwxrwx	November 13, 2017 02:50 PM
ad_data1				drwxrwxrwx	November 13, 2017 02:50 PM
ad_data1.txt	30.1 MB			-rw-rw-rw-	November 13, 2017 02:43 PM
customers		training	supergroup	drwxrwxrwx	November 13, 2017 01:35 PM



# Run MapReduce Jobs

- MapReduce code for Hadoop is typically written in Java
- Minimally, MapReduce applications specify the input/output locations and supply *map* and *reduce* functions.
  - The rest is handled by the Hadoop MapReduce framework and the user can monitor the progress by a JobTracker
  - E.g.
  - `hadoop jar /usr/joe/wordcount.jar org.myorg.WordCount /usr/joe/wordcount/input /usr/joe/wordcount/output`

Java class defines map and reduce functions

Input and output paths

Source: [http://hadoop.apache.org/docs/stable1/mapred\\_tutorial.html#Usage](http://hadoop.apache.org/docs/stable1/mapred_tutorial.html#Usage)

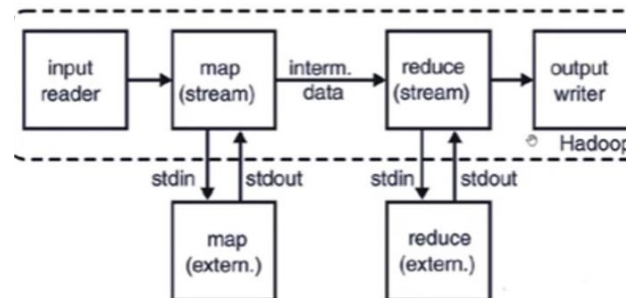
# Hadoop Streaming

- Alternatively, Hadoop MapReduce can leverage Hadoop streaming.
  - Example (if written in Python)

```
hadoop jar /path/to/Hadoop-streaming.jar \  
-file /path/to/mapper.py -mapper /path/to/mapper.py \  
-file /path/to/reducer.py -reducer /path/to/reducer.py \  
-input /path/to/input \  
-output /path/to/output
```

Java jar file that streams data to/from external scripts

Allow the command to break into multiple lines



[http://hadoop.apache.org/docs/stable1/mapred\\_tutorial.html#Usage](http://hadoop.apache.org/docs/stable1/mapred_tutorial.html#Usage)

# Labs: Scenario Explanation

- Hands-On exercises throughout the course will reinforce the topics being discussed
  - Exercises simulate the kind of tasks often performed using the tools you will learn about in class
  - Most exercises depend on data generated in earlier exercises
- Scenario: Dualcore Inc. is a leading electronics retailer
  - More than 1,000 brick and mortar stores
  - Dualcore also has a thriving e-commerce Web site
- Dualcore has hired you to help find value in their data
  - You will process and analyze data from internal and external sources
  - Identify opportunities to increase revenue
  - Find new ways to reduce costs
  - Help other departments achieve their goals

# Labs: Fact Sheet

- Name: Dualcore
- Locations: 1,057 (all in the continental United States)
- Employees: 60,000
  - Concentrated near headquarters in the Bay Area, with many more elsewhere in California and on the West coast
  - Headquartered in Palo Alto, Ca. 94305
  - Numbers generally decrease eastward throughout US
- History
  - Company founded in 1985
  - 500 brick-and-mortar locations closed recently
  - E-commerce division has grown significantly and now accounts for a majority of revenue
- Products
  - PCs, servers, tablet devices, electronics (like DVD players and televisions), and related accessories (like printers, disks, batteries, cables and adapters)

## Labs: Important!

- Most lab exercises depend on data generated in earlier exercises
- Therefore, you cannot proceed with an exercise unless you've completed the one that precedes it
- If you are unable to complete one of the exercises (or suspect you may have made a mistake), there is a procedure for automatically “catching up”, described in the lab manual

# Lab 01 – Using HDFS

- During this course, you will perform numerous Hands-On exercises using the Cloudera Training Virtual Machine (VM)
- The VM has Hadoop installed in pseudo-distributed mode
  - Simply a cluster comprised of a single node
  - Typically used for testing code before deploying to a large cluster
- Read the general notes first in the lab handout
- In the lab titled “Using HDFS” you will begin to get acquainted with the Hadoop tools. You will manipulate files in HDFS, the Hadoop Distributed File System.