cloudera®

Ask Bigger Questions

CAP - Developing with Spark and Hadoop:

Homework Assignment Guide for Students

Homework: Setup and General Notes

Cloudera's training courses use a Virtual Machine running the CentOS Linux distribution. This VM has CDH (Cloudera's Distribution, including Apache Hadoop) installed in Pseudo-Distributed mode. Pseudo-Distributed mode is a method of running Hadoop whereby all Hadoop daemons run on the same machine. It is, essentially, a cluster consisting of a single machine. It works just like a larger Hadoop cluster, the only difference (apart from speed, of course!) being that the block replication factor is set to 1, since there is only a single DataNode available.

Getting Started

1. Before starting the homework assignments, run the course setup script in a terminal window:

```
$ $DEV1/scripts/training_setup_dev1.sh
```

   This script will enable services and set up any data required for the course. You must run this script before starting the Homework.

Working with the Virtual Machine

1. The VM is set to automatically log in as the user `training`. Should you log out at any time, you can log back in as the user `training` with the password `training`.

2. Should you need it, the root password is `training`. You may be prompted for this if, for example, you want to change the keyboard layout. In general, you should not need this password since the `training` user has unlimited sudo privileges.

3. In some command-line steps in the homework, you will see lines like this:

```
$ hdfs dfs -put shakespeare \
/user/training/shakespeare
```

The dollar sign (**$**) at the beginning of each line indicates the Linux shell prompt. The actual prompt will include additional information (e.g., `[training@localhost workspace]$`) but this is omitted from these instructions for brevity.

The backslash (`\`) at the end of the first line signifies that the command is not completed, and continues on the next line. You can enter the code exactly as shown (on two lines), or you can enter it on a single line. If you do the latter, you should *not* type in the backslash.

Points to note during the homework

1. The main directory for the homework for this course is `$DEV1/exercises` (`~/training_materials/dev1/exercises`). Professors may choose to demonstrate solutions as part of a lecture. Please note, we use the terms homework and exercise interchangeably.

2. Each directory under the main directory corresponds to a homework assignment or set of assignments – this is referred to in the instructions as "the exercise directory". Any scripts or files required for the exercise (other than data) are in the exercise directory.

3. Within each exercise directory you may find the following subdirectories (on the professor VM only):

    a. `solution` – Solution code for each exercise.

    b. `stubs` – A few of the exercises depend on provided starter files containing skeleton code.

    c. Maven project directories – For exercises for which you must write Scala classes, you have been provided with preconfigured Maven project directories. Within these projects are two packages: `stubs`,

where you will do your work using starter skeleton classes; and `solution`, containing the solution class.

4. Data files used in the exercises are in `$DEV1DATA` (`~/training_materials/data`). Usually you will upload the files to HDFS before working with them.

5. As the exercises progress, and you gain more familiarity with Hadoop and Spark, we provide fewer step-by-step instructions; as in the real world, we merely give you a requirement and it's up to you to solve the problem!

6. There are additional optional homework assignments for some of the chapters.

This is the end of the Homework