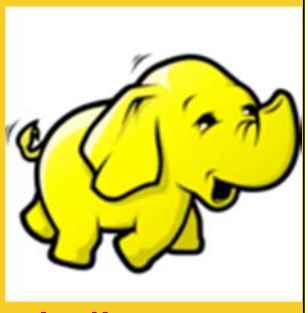
Virtual Machine (VM) For Hadoop Training

Originals of slides and source code for examples: http://www.coreservlets.com/hadoop-tutorial/
Also see the customized Hadoop training courses (onsite or at public venues) – http://courses.coreservlets.com/hadoop-training.html

Customized Java EE Training: http://courses.coreservlets.com/
Hadoop, Java, JSF 2, PrimeFaces, Servlets, JSP, Ajax, jQuery, Spring, Hibernate, RESTful Web Services, Android.
Developed and taught by well-known author and developer. At public venues or onsite at *your* location.

© 2012 coreservlets.com and Dima May



For live customized Hadoop training (including prep for the Cloudera certification exam), please email info@coreservlets.com

Taught by recognized Hadoop expert who spoke on Hadoop several times at JavaOne, and who uses Hadoop daily in real-world apps. Available at public venues, or customized versions can be held on-site at your organization.

- Courses developed and taught by Marty Hall
  - JSF 2.2, PrimeFaces, servlets/JSP, Ajax, jQuery, Android development, Java 7 or 8 programming, custom mix of topics
  - Courses available in any state or country. Maryland/DC area companies can also choose afternoon/evening courses.
- Courses developed and taught by coreservlets.com experts (edited by Marty)
  - Spring, Hibernate/JPA, GWT, Hadoop, HTML5, RESTful Web Services

Contact info@coreservlets.com for details

Agenda

- Overview of Virtual Machine for Hadoop Training
- Eclipse installation
- Environment Variables
- Firefox bookmarks
- Scripts
- Developing Exercises
- Well-Known Issues

Virtual Machine

- In this class we will be using Virtual Box , a desktop virtualization product, to run Ubuntu
  - https://www.virtualbox.org
- Ubuntu image is provided with Hadoop products pre-installed and configured for development
  - Cloudera Distribution for Hadoop (CDH) 4 is used; installed products are:
    - Hadoop (HDFS and YARN/MapReduce)
    - HBase
    - Oozie
    - Pig & Hive

Installing Virtual Box

- **Download the latest release for your specific OS**
  - https://www.virtualbox.org/wiki/Downloads
- **After download is complete, run Virtual Box installer**
- **Start Virtual Box and import provided Ubuntu image/appliance**
  - File → Import Appliance
- **Now that new image is imported, select it and click 'Start'**

VM Resource

- **VM is set up with**
  - 3G of RAM and 2CPUs and 13G of Storage
- **If you can spare more RAM and CPU adjust VM Settings**
  - Virtual Box Manager → right click on VM → Settings → System → adjust under Motherboard and Processor tabs

Logging In

- **Username: hadoop**
- **Password: hadoop**

Desktop Screen

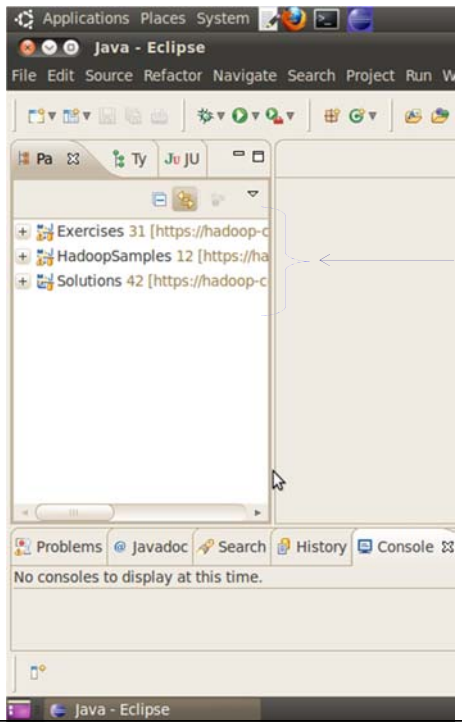
Command line terminal

Eclipse is installed to assist in developing Java code and scripts

Directory Locations

Training	All the training artifacts; located in the user's home directory
CDH4	Installation directory for Hadoop products
eclipse	Eclipse installation
eclipse_workspace	Code, resources and scripts managed via Eclipse
exercises	Data for exercises
hadoop_work	Hadoop is configured to store its data here
jdk1.6.0_29	Java Development Kit (JDK) installation
logs	Logs are configured to be saved in this directory
PigPen	Eclipse Plugin to enable highlighting of Pig Scripts
play_area	Execute Java code, MapReduce Jobs and scripts from here
scripts	Well known shell scripts

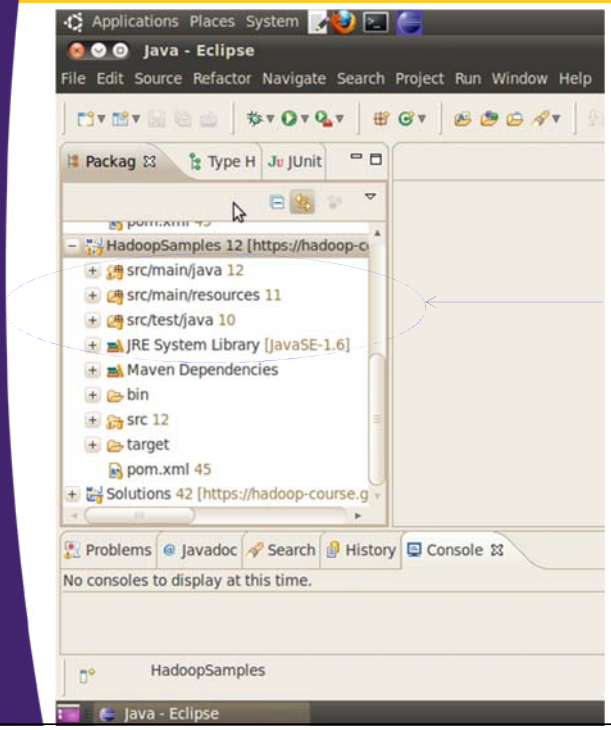
Eclipse



Eclipse workspace will contain three projects:

- **Exercises** – you will implement hands-on exercises in this project
- **Solutions** – the solutions to the exercises can be found here
- **HadoopSamples** – code samples used throughout the slides

Eclipse Project



Projects follow maven directory structure

- **/src/main/java** – Java packages and classes reside here
- **/src/main/resources** – non-Java artifacts
- **/src/main/test/java** – Java unit tests go here

To further learn about maven please visit http://maven.apache.org

Environment Variables

- **VM is set up with various environment variables to assist you with referencing well-known directories**
- **Environment variables are sourced from**
  - /home/hadoop/Training/scripts/hadoop-env.sh
- **For example:**
  - $ echo $PLAY_AREA
  - $ yarn jar $PLAY_AREA/Solutions.jar ....

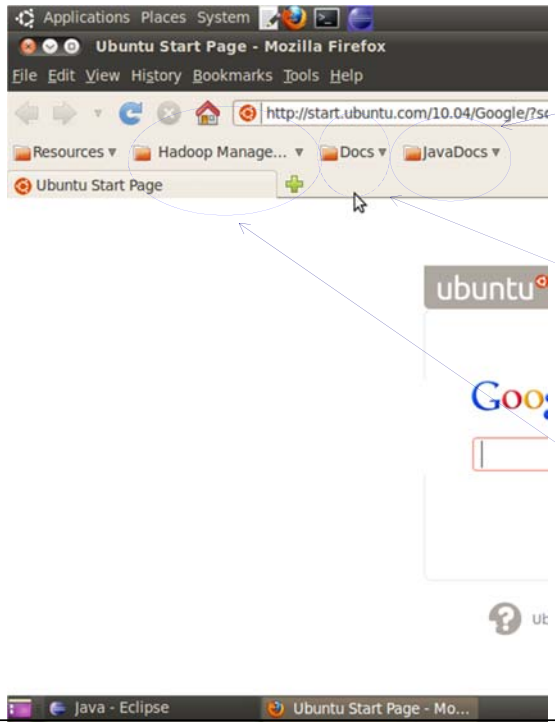
Environment Variables

- **PLAY_AREA=/home/hadoop/Training/play_area**
  - Run examples, exercises, and solutions from this directory
  - Jar files are copied here (by maven)
- **TRAINING_HOME=/home/hadoop/Training**
  - Root directory for all of the artifacts for this class
- **HADOOP_LOGS=$TRAINING_HOME/logs**
  - Directory for logs; logs for each product are stored under
  - $ ls $HADOOP_LOGS/
    - hbase hdfs oozie pig yarn
- **HADOOP_CONF_DIR=$HADOOP_HOME/conf**
  - Hadoop configuration files are stored here

Environment Variables

- **There is a variable per product referencing it's home directory**
  - CDH_HOME=$TRAINING_HOME/CDH4
  - HADOOP_HOME=$CDH_HOME/hadoop-2.0.0-cdh4.0.0
  - HBASE_HOME=$CDH_HOME/hbase-0.92.1-cdh4.0.0
  - OOZIE_HOME=$CDH_HOME/oozie-3.1.3-cdh4.0.0
  - PIG_HOME=$CDH_HOME/pig-0.9.2-cdh4.0.0
  - HIVE_HOME=$CDH_HOME/hive-0.8.1-cdh4.0.0

Firefox Bookmarks



Folder with bookmarks to Javadocs for each product used in this class

Folder with bookmarks to documentation packaged with each product used in this class

Folders with bookmarks to management web applications for each product; of course the Hadoop product has to be running for those links to work

Scripts

- **Scripts to start/stop ALL installed Hadoop products**
  - startCDH.sh - start ALL of the products
  - stopCDH.sh - stop ALL of the products
  - These scripts are located in ~/Training/scripts/
  - Scripts are on the PATH, you can execute from anywhere

```
$ startCDH.sh
...
...
$ stopCDH.sh
...
$ ps -ef | grep java
...
$ kill XXXX
```

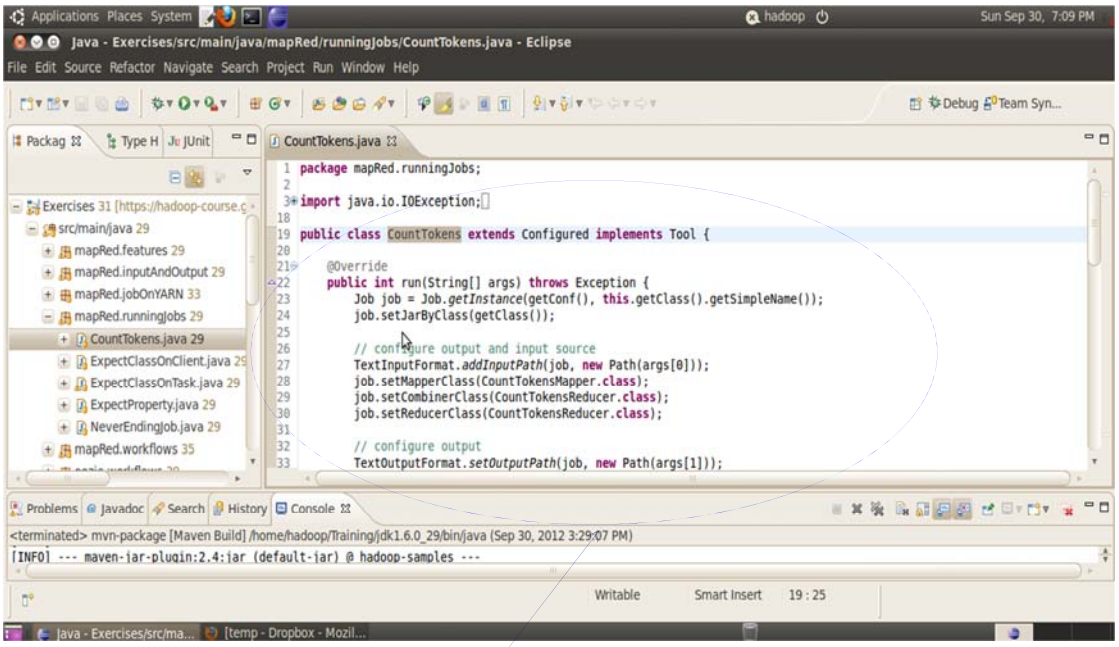
Start then stop all of the products

Check if any processes failed to shut down, if so kill them by PID

Developing Exercises

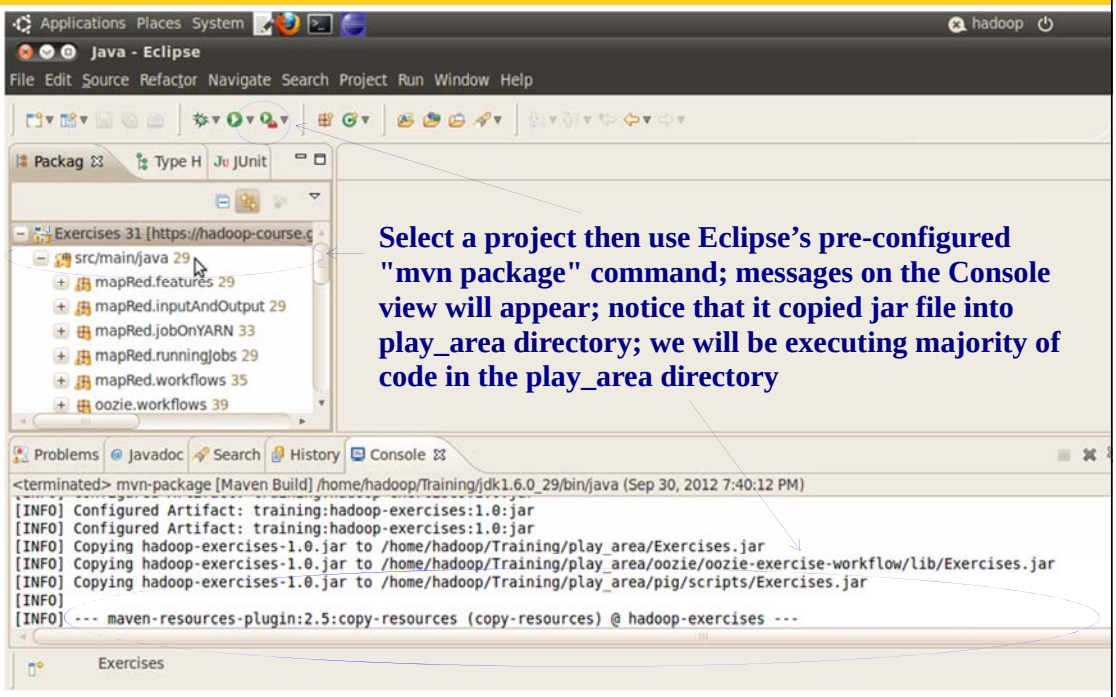
- **Proposed steps to develop code for training exercises**
  1. Add code, configurations and/or scripts to the Exercises project
     - Utilize Eclipse
  2. Run mvn package
     - Generates JAR file with all of the Java classes and resources
     - For your convenience copies JAR file to a set of well-known locations
     - Copies scripts to a well-known location
  3. Execute your code (MapReduce Job, Oozie job or a script)

1: Add Code to the Exercises Project



Write and edit code

2: Run mvn package



Select a project then use Eclipse's pre-configured "mvn package" command; messages on the Console view will appear; notice that it copied jar file into play_area directory; we will be executing majority of code in the play_area directory

3: Execute your code

- **Utilize the jar produced by step #2**
- **Run your code in $PLAY_AREA directory**

```
$ cd $PLAY_AREA
```

Produced by previous step Exercises.jar will reside in $PLAY_AREA directory

```
$ yarn jar $PLAY_AREA/Exercises.jar \
      mapRed.workflows.CountDistinctTokens \
      /training/data/hamlet.txt \
      /training/playArea/firstJob
```

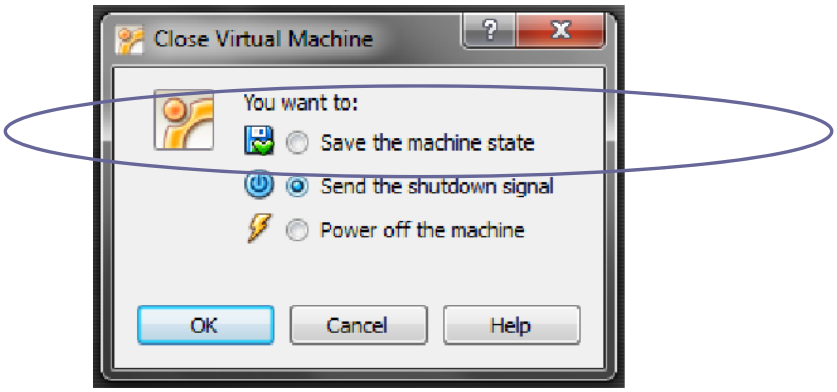
This is a MapReduce job implemented in the Exercises project and then package into a JAR file

Clean up after yourself; Delete output directory

```
$ hdfs dfs -rm -r /training/playArea/firstJob
```

Save VM Option

- **Instead of Shutting down OS you can save current OS State**
  - When you load it again the saved state will be restored



Well-Known Issues

- **If you "save the machine state", instead of restarting VM, HBase will not properly reconnect to HDFS**
  - Solution: shutdown all of the Hadoop products prior closing VM (run stopCDH.sh script)
- **Current VM allocates 3G of RAM; it is really not much given all of the Hadoop and MapReduce daemons**
  - Solution: If your machine has more RAM to spare, increase it. When the VM is down go to Settings → System → Base Memory

Wrap-Up

Summary

- We now know more about Ubuntu VM
- There are useful environment variables
- There are helpful Firefox bookmarks
- Use management scripts to start/stop Hadoop products
- Develop exercises utilizing Eclipse and Maven
- Look out for well-known issues with running Hadoop on top of Virtual Box VM

Questions?

More info:

http://www.coreservlets.com/hadoop-tutorial/ – Hadoop programming tutorial

http://courses.coreservlets.com/hadoop-training.html – Customized Hadoop training courses, at public venues or onsite at *your* organization

http://courses.coreservlets.com/Course-Materials/java.html – General Java programming tutorial

http://www.coreservlets.com/java-8-tutorial/ – Java 8 tutorial

http://www.coreservlets.com/JSF-Tutorial/jsf2/ – JSF 2.2 tutorial

http://www.coreservlets.com/JSF-Tutorial/primefaces/ – PrimeFaces tutorial

http://coreservlets.com/ – JSF 2, PrimeFaces, Java 7 or 8, Ajax, jQuery, Hadoop, RESTful Web Services, Android, HTML5, Spring, Hibernate, Servlets, JSP, GWT, and other Java EE training