

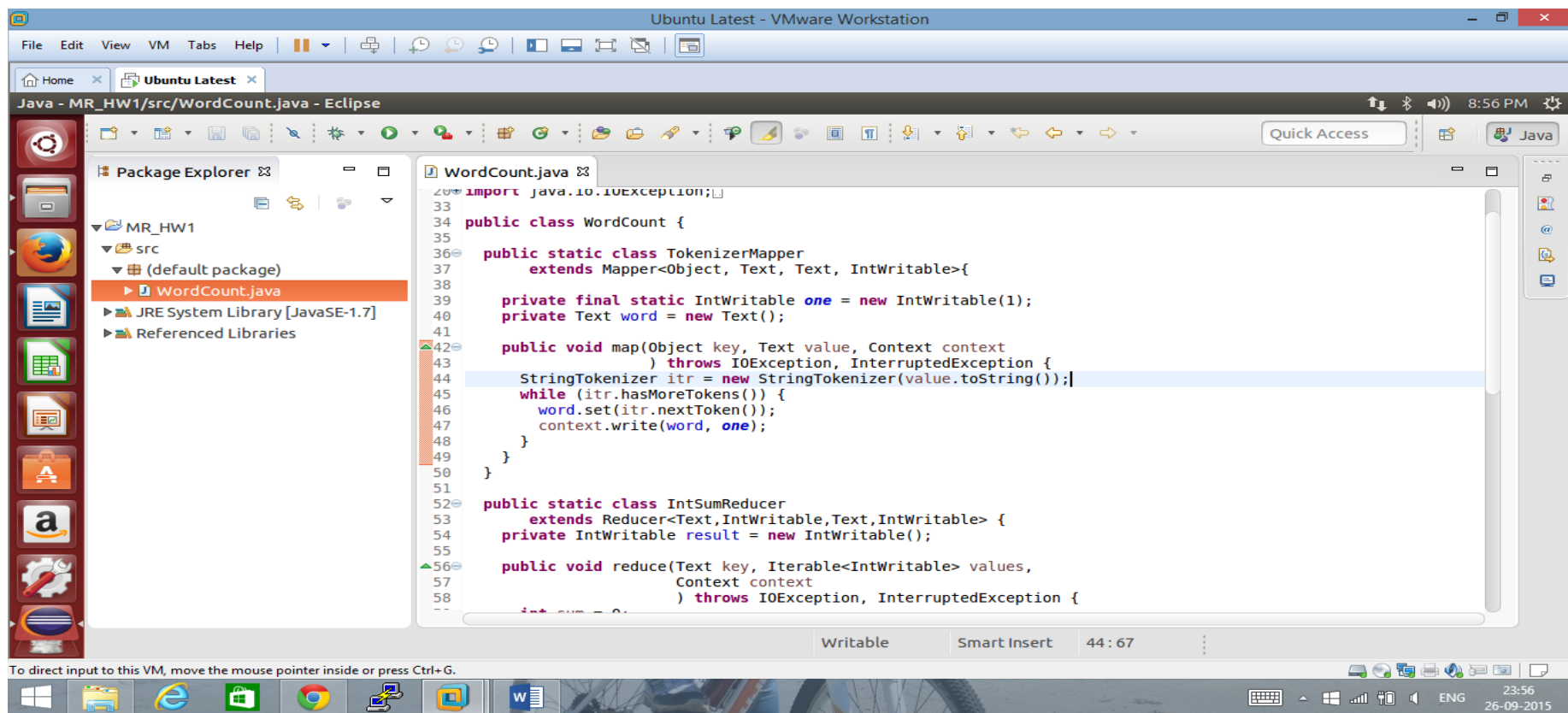
Class Number: CS6240 [Tuesday Evening Section] – Section 01

HW Number: 1

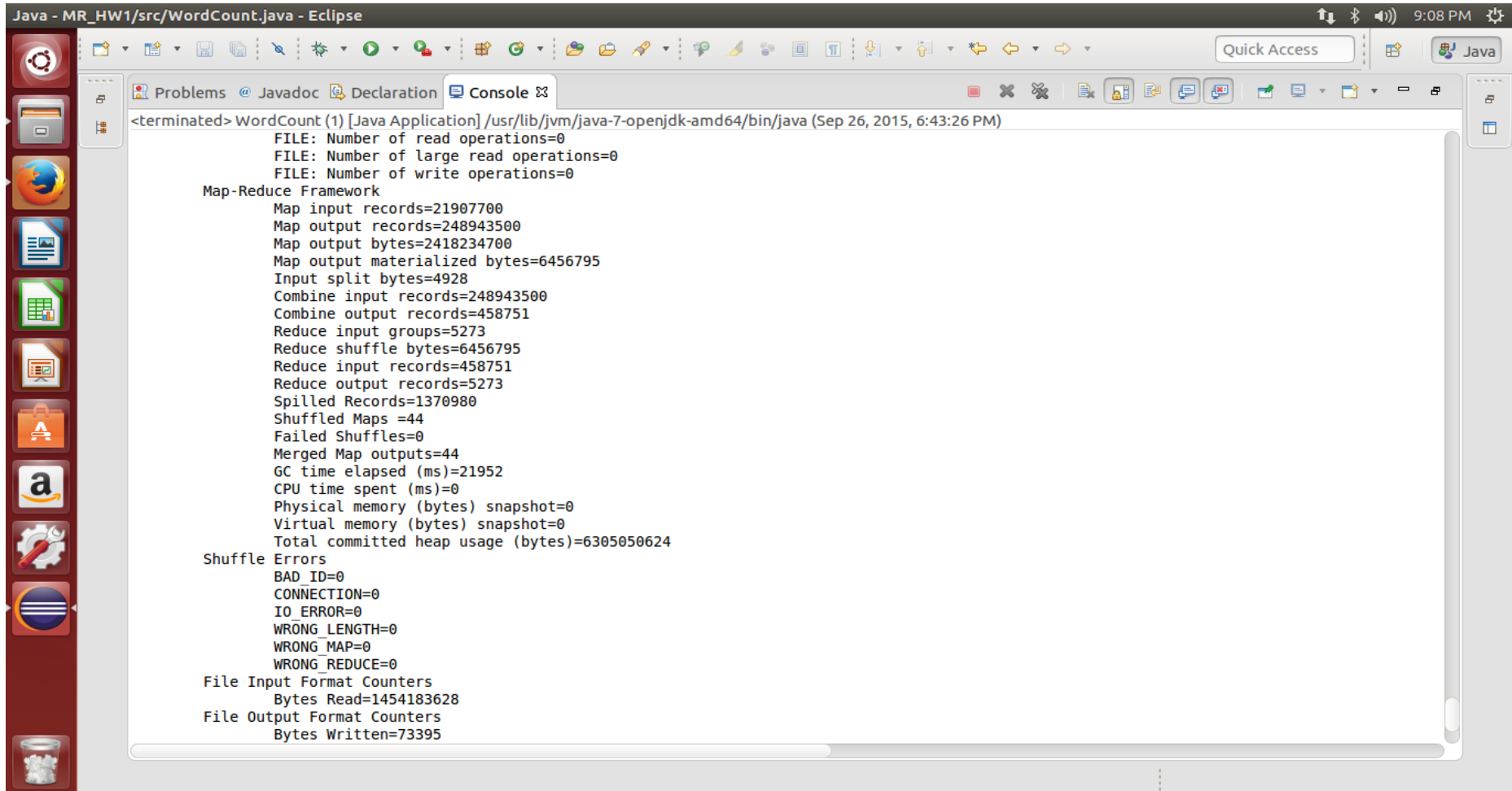
Name: Pavitra Srinivasan

## Local Execution

Project directory structure, showing that the WordCount.java file is somewhere in the src directory.



The console output for a successful run of the WordCount program inside the IDE. Show at least the last 20 lines of the console output.



```
Java - MR_HW1/src/WordCount.java - Eclipse
<terminated> WordCount (1) [Java Application] /usr/lib/jvm/java-7-openjdk-amd64/bin/java (Sep 26, 2015, 6:43:26 PM)
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
  Map input records=21907700
  Map output records=248943500
  Map output bytes=2418234700
  Map output materialized bytes=6456795
  Input split bytes=4928
  Combine input records=248943500
  Combine output records=458751
  Reduce input groups=5273
  Reduce shuffle bytes=6456795
  Reduce input records=458751
  Reduce output records=5273
  Spilled Records=1370980
  Shuffled Maps =44
  Failed Shuffles=0
  Merged Map outputs=44
  GC time elapsed (ms)=21952
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=6305050624
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=1454183628
File Output Format Counters
  Bytes Written=73395
```

## AWS Execution:

Evidence of a successful run of the WordCount program on AWS.

The screenshot displays the AWS Elastic MapReduce console interface. The browser address bar shows the URL: <https://us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#cluster-details:j-O7ED4JSUHRTF>. The navigation bar includes 'AWS', 'Services', 'EMR', 'S3', and 'Edit'. The main header shows 'Elastic MapReduce' and 'Cluster List > Cluster Details'. Below the header are buttons for 'Add step', 'Resize', 'Clone', and 'Terminate'. The cluster name is 'Assignment1\_HW1' with a status of 'Waiting' and a note 'Waiting after step completed'. The 'Connections' section includes links for 'Enable Web Connection', 'Master public DNS', and 'Tags'. The main content area is divided into four tabs: 'Summary', 'Configuration Details', 'Network and Hardware', and 'Security and Access'. The 'Summary' tab is active, showing details such as ID, creation date, elapsed time, and auto-terminate settings. The 'Configuration Details' tab shows release label, Hadoop distribution, applications, log URI, and EMRFS status. The 'Network and Hardware' tab shows availability zone, subnet ID, and instance counts. The 'Security and Access' tab shows key name, EC2 instance profile, EMR role, and security groups. The bottom of the screen shows a Windows taskbar with various application icons and a system clock indicating 12:59 on 27-09-2015.

Cluster: Assignment1\_HW1 **Waiting** Waiting after step completed

**Connections:** [Enable Web Connection](#) – Resource Manager ... (View All)  
**Master public DNS:** [ec2-52-89-156-137.us-west-2.compute.amazonaws.com](#) [SSH](#)  
**Tags:** -- [View All / Edit](#)

Summary	Configuration Details	Network and Hardware	Security and Access
<b>ID:</b> j-O7ED4JSUHRTF <b>Creation date:</b> 2015-09-27 12:28 (UTC-4) <b>Elapsed time:</b> 28 minutes <b>Auto-terminate:</b> No <b>Termination protection:</b> Off <a href="#">Change</a>	<b>Release label:</b> emr-4.0.0 <b>Hadoop distribution:</b> Amazon 2.6.0 <b>Applications:</b> -- <b>Log URI:</b> <a href="#">s3://aws-logs-513576545204-us-west-2/elasticmapreduce/</a> <b>EMRFS consistent view:</b> Disabled	<b>Availability zone:</b> us-west-2c <b>Subnet ID:</b> subnet-79b83c20 <b>Master:</b> <b>Running</b> 1 m1.medium <b>Core:</b> <b>Running</b> 2 m1.medium <b>Task:</b> --	<b>Key name:</b> -- <b>EC2 instance profile:</b> EMR_EC2_DefaultRole <b>EMR role:</b> EMR_DefaultRole <b>Visible to all users:</b> <a href="#">Change</a> <b>Security groups for Master:</b> <a href="#">sg-1c20aa78</a> (ElasticMapReduce-master) <b>Security groups for Core &amp; Task:</b> <a href="#">sg-1d20aa79</a> (ElasticMapReduce-slave)

► Monitoring  
▼ Hardware

## Step Details:

The screenshot shows the AWS Elastic MapReduce console interface. The browser address bar displays the URL: `https://us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#cluster-details:j-O7ED4JSUHRTF`. The page title is "Steps".

At the top, there are buttons for "Add step" and "Clone step". Below these, the "Steps" section is visible, with links for "View all interactive jobs" and "View all jobs".

The "Steps" table shows two steps:

ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files	Actions
s-2YQ9G9OQDRNKV	Custom JAR	Completed	2015-09-27 12:36 (UTC-4)	10 minutes	<a href="#">controller</a>   <a href="#">syslog</a>   <a href="#">stderr</a>   <a href="#">stdout</a>	<a href="#">View jobs</a>
s-3RB2VANJO3S8J	Setup hadoop debugging	Completed	2015-09-27 12:36 (UTC-4)	7 seconds	<a href="#">controller</a>   <a href="#">syslog</a>   <a href="#">stderr</a>   <a href="#">stdout</a>	<a href="#">View jobs</a>

Below the table, the details for the selected step (s-2YQ9G9OQDRNKV) are shown:

- JAR location:** s3://cs6240assignment/1/code/hw1.jar
- Main class:** None
- Arguments:** s3://cs6240assignment/1/input/hw1 s3://cs6240assignment/1/output
- Action on failure:** Terminate cluster

The bottom of the screenshot shows the Windows taskbar with various application icons and the system clock displaying 13:03 on 27-09-2015.

Using at least three small machines, i.e., one master node and two core nodes

Browser tabs: Pavitra Srinivasan, AWS Elastic MapReduce, https://aws-logs-5135765...

Address bar: https://us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#cluster-details:j-O7ED4JSUHRTF

Bookmarks: Apps, INTERNSHIP, MSD, MR, IR, App, Jobs, myNEU, Splitwise, Zimbra, Piazza, RSO, Blackboard, Outlook, Gmail, Husky Id, Fall Career Fair, Other bookmarks

### ▼ Hardware

**Add task instance group**

**Instance Groups**

Filter:  2 instance groups (all loaded)

ID	Name	Status	Type	Instance Type	Count	Bid Price	Actions
▶ ig-XUGEIWOE4575	Master instance group - 1	Running	MASTER	m1.medium	1		<a href="#">View EC2 ins</a>
▶ ig-LOZ2EUQW25C5	Core instance group - 2	Running	CORE	m1.medium	2 <a href="#">Resize</a>		<a href="#">View EC2 ins</a>

### ▼ Steps

**Add step** **Clone step**

**Steps** [View all interactive jobs](#) | [View all jobs](#)

Filter:   2 steps (all loaded)

ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files	Actions
▶ s-2YQ9G9OQDRNKV	Custom JAR	Completed	2015-09-27 12:36 (UTC-4)	10 minutes	<a href="#">controller</a>   <a href="#">syslog</a>   <a href="#">stderr</a>   <a href="#">stdout</a>	<a href="#">View jobs</a>
▶ s-3RB2VANJO3S8J	Setup hadoop debugging	Completed	2015-09-27 12:36 (UTC-4)	7 seconds	<a href="#">controller</a>   <a href="#">syslog</a>   <a href="#">stderr</a>   <a href="#">stdout</a>	<a href="#">View jobs</a>

Taskbar: IMG\_2674.JPG, data01.zip, Show all downloads...

System tray: 13:00, 27-09-2015

Controller, syslog log files and final result are saved in the folder.

### **Analysis of Syslog:**

The job client submits the job to the resource manager. Then the job client continues to monitor the execution of the job and report back to the console with the progress of the map and reduce containers. That is why we see the map 0% reduce 0%, map 1% reduce 0% and so on. Then the reduce phases i.e. shuffle and sort start to copy the data and group the keys together. Hence you start seeing map 57% reduce 6% and so on. However the actual reduce starts only after mapper finishes. After both mapper and reducer finish, the counter values are written indicating the record details.