



Enterprise Computing: Exercise 3 – Elastic Map Reduce

Markus Klems, Stefan Tai

Task 1 – Word Count example

1. Word Count reads text files and counts how often words occur.
 - The input are text files
 - The output are text files, each line of which contains a word and the count of how often it occurred, separated by a tab.
2. Mapper
 - Each mapper takes a line as input and breaks it into words.
 - It then emits a key/value pair of the word and 1.
3. Reducer
 - Each reducer sums the counts for each word and
 - emits a single key/value with the word and sum.

Source: <http://wiki.apache.org/hadoop/WordCount>

Task 1 – Launch EMR cluster

1. Sign in at <https://336380577901.signin.aws.amazon.com/console>
2. Open the Amazon Elastic MapReduce console at <https://console.aws.amazon.com/elasticmapreduce/>.
3. Click “Create cluster”.
4. In the Create Cluster page, click “Configure sample application”.

Configure Sample Application

Select a sample application to auto-populate the Create Cluster page

Select sample application: Word count

Output location: s3://<bucket-name>/wordcount/output/2014-04-10/15-1

Logging: ☒ Enabled
s3n://<bucket-name>/wordcount/logging/
s3://<bucket-name>/<folder>/

Debugging: ☒ Enabled

Cancel Ok

See: <http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-get-started-count-words-step-5.html>

Task 1 – Launch EMR cluster

5. Configure the sample application "Word count" as follows

Master m1.medium -> 1 instance

Core m1.medium -> 2 instances

Task m1.medium -> 0 instances

Hardware Configuration

i Specify the [networking](#) and [hardware](#) configuration for your cluster. If you need more than 20 EC2 instances, [complete this form](#).
[Request Spot instances](#) (unused EC2 capacity) to save money.

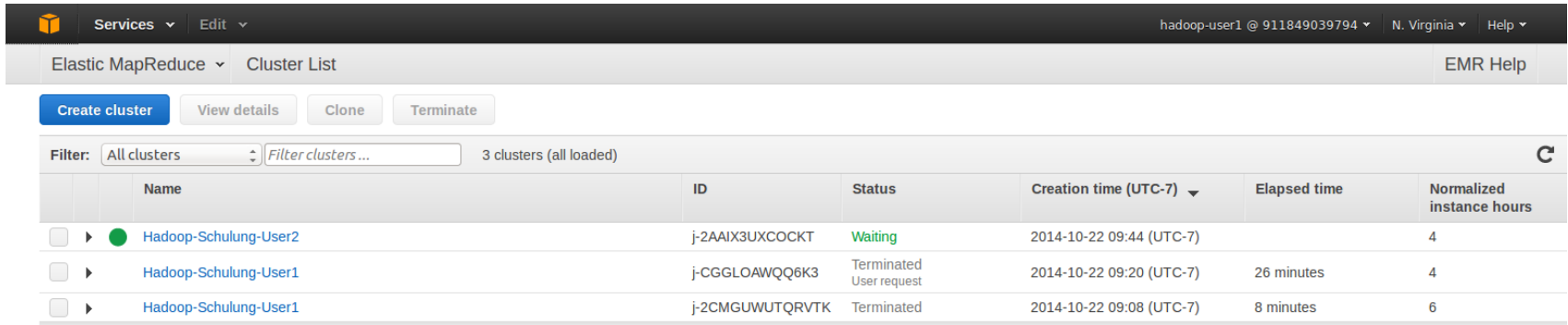
Network Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. [Create a VPC](#)

EC2 Subnet [Create a Subnet](#)

	EC2 instance type	Count	Request spot	
Master	<input type="text" value="m1.medium"/>	<input type="text" value="1"/>	<input type="checkbox"/>	The Master instance assigns Hadoop tasks to core and task nodes, and monitors their status.
Core	<input type="text" value="m1.medium"/>	<input type="text" value="2"/>	<input type="checkbox"/>	Core instances run Hadoop tasks and store data using the Hadoop Distributed File System (HDFS).
Task	<input type="text" value="m1.medium"/>	<input type="text" value="0"/>	<input type="checkbox"/>	Task instances run Hadoop tasks.

See: <http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-get-started-count-words-step-5.html>

Task 1 – Monitor cluster status and progress



The screenshot shows the AWS EMR console interface. At the top, there's a navigation bar with 'Services' and 'Edit' dropdowns, and user information 'hadoop-user1 @ 911849039794' and 'N. Virginia' region. Below this is a breadcrumb 'Elastic MapReduce > Cluster List' and an 'EMR Help' link. A toolbar contains buttons for 'Create cluster', 'View details', 'Clone', and 'Terminate'. A filter section shows 'All clusters' selected, with a search box and '3 clusters (all loaded)'. The main table lists three clusters with columns for Name, ID, Status, Creation time, Elapsed time, and Normalized instance hours.

	Name	ID	Status	Creation time (UTC-7) ▾	Elapsed time	Normalized instance hours
<input type="checkbox"/>	Hadoop-Schulung-User2	j-2AAIX3UXCOCKT	Waiting	2014-10-22 09:44 (UTC-7)		4
<input type="checkbox"/>	Hadoop-Schulung-User1	j-CGGLOAWQQ6K3	Terminated User request	2014-10-22 09:20 (UTC-7)	26 minutes	4
<input type="checkbox"/>	Hadoop-Schulung-User1	j-2CMGUWUTQRVTK	Terminated	2014-10-22 09:08 (UTC-7)	8 minutes	6

Task 1 - Questions

- a) What is the word count of the word "accepted"?
- b) What is the word count of the word "gibraltar"?

Task 2 – 3 Preparation

Next, we want to program the MapReduce program by ourselves in Java.

Prerequisites:

Install the following plugins in your Eclipse IDE (if not already installed)

- <http://download.eclipse.org/releases/kepler> -> "Database Development"
- <http://aws.amazon.com/eclipse>

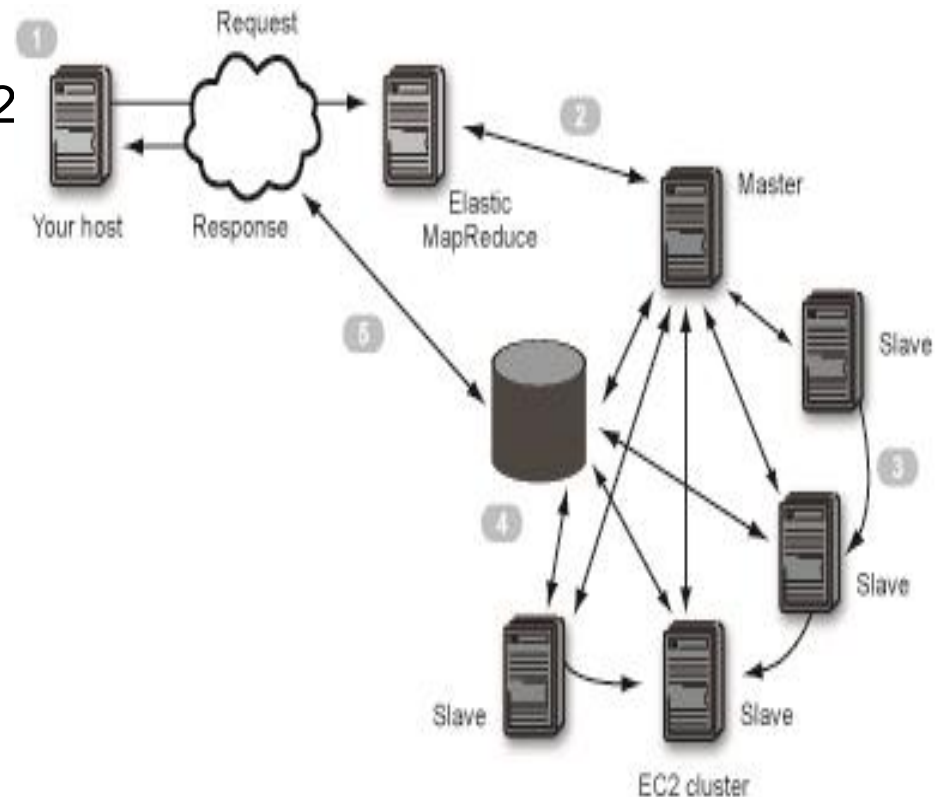
Task 2 – 3 Preparation

Instructions:

- Clone Markus's git repository: `$ git clone git@gitlab.tubit.tu-berlin.de:klems/ecmapreduce.git`
- Open the project in Eclipse as AWS Java project
- Run Maven install to build the project
- Complete the missing code marked with TODO
- Generate a jar file with Maven
- Upload the jar file to S3
- Create a new EMR cluster and select your custom jar file while creating the EMR cluster
- Enter the CLUSTER_ID in the MapReduceClient code
- Run the MapReduceClient as local Java application in Eclipse

AWS EMR Workflow

1. Load data, Map and Reduce executables to S3
2. Elastic MapReduce starts EC2 Hadoop-Cluster (Master + Slaves)
3. Hadoop generates Jobflow to distribute S3 data to cluster and to process it
4. Results are copied over to S3
5. Message is sent at the end: Retrieve the results from S3 (Browser, wget,...)



Task 2

Complete the missing code in WordCountABA and answer the following question: which words end with 'aba' and what is their word count?

Task 3

A palindrome is a word, phrase, number, or other sequence of symbols or elements that reads the same forward or reversed, with general allowances for adjustments to punctuation and word dividers. Complete the missing code in `PalindromeCount` and answer the questions: which words with at least 5 letters are palindromes and what is their word count? Do not include words that only consist of numbers in your answer.

224-word palindrome poem by Demetri Martin

Dammit I'm mad.
Evil is a deed as I live.
God, am I reviled? I rise, my bed on a sun, I melt.
To be not one man emanating is sad. I piss.
Alas, it is so late. Who stops to help?
Man, it is hot. I'm in it. I tell.
I am not a devil. I level "Mad Dog".
Ah, say burning is, as a deified gulp,
In my halo of a mired rum tin.
I erase many men. Oh, to be man, a sin.
Is evil in a clam? In a trap?
No. It is open. On it I was stuck.
Rats peed on hope. Elsewhere dips a web.
Be still if I fill its ebb.

Ew, a spider... eh?
We sleep. Oh no!
Deep, stark cuts saw it in one position.
Part animal, can I live? Sin is a name.
Both, one... my names are in it.
Murder? I'm a fool.
A hymn I plug, deified as a sign in ruby ash,
A Goddam level I lived at.
On mail let it in. I'm it.
Oh, sit in ample hot spots. Oh wet!
A loss it is alas (sip). I'd assign it a name.
Name not one bottle minus an ode by me:
"Sir, I deliver. I'm a dog"
Evil is a deed as I live.
Dammit I'm mad.