

CS 7001-03: Homework 4: AWS-EMR

Chanmann Lim
cl9p8@mail.mail.missouri.edu

April 30, 2015

1. By default there is one Master instance with "m3.xlarge" EC2 instance type created in the EMR cluster and the main jobs of the Master node is to assign Hadoop tasks to the **core** and **task** nodes and monitor their status and there are two Core instances are also created in the EMR cluster for processing Hadoop tasks and storing the data using the Hadoop Distributed File System (HDFS) and to S3 using EMR File System (EMRFS).

2. List of the first and last word with the associated number of occurrences from each output part file

part-00000: abbate **1**, zoology **3**

part-00001: abacus **3**, zwith **1**

part-00002: a **4215**, zenith **2**

part-00003: abandon **6**, zweite **1**

part-00004: abbreviated **1**, yolk **1**

part-00005: abandons **2**, zeitschrift **1**

part-00006: ab **1**, zwanzig **1**

3. AWS-EMR execution chart:

Master	Cores	Output parts	Cluster ID
1	2	7	j-2397PQCL7GDP1
1	3	11	j-1VDWF9AKANGPE
1	4	15	j-8OJYX2FVDGZW
1	5	19	j-2BBG0J0YBRKL1
1	6	23	j-2ACHK4SU3NDIB

The number of output part files increased linearly as the number core instances get large. The observations above show that

$$part_files = 4 \times cores - 1 \quad (1)$$

The more core instances initiated to handle the MapReduce tasks the more chunks can be split from the original input file for parallel processing.

4. Screenshot of the clusters created with EMR:

Figure 1 shows the AWS Elastic MapReduce console interface. At the top, there's a navigation bar with 'AWS', 'Services', and 'Edit' menus. The main header indicates 'Elastic MapReduce' and 'Cluster List'. Below this, there are buttons for 'Create cluster', 'View details', 'Clone', and 'Terminate'. A filter section shows 'All clusters' selected, with a search bar and a count of '6 clusters (all loaded)'. The main content is a table listing the clusters.

	Name	ID	Status	Creation time (UTC-5)	Elapsed time	Normalized instance hours
<input type="checkbox"/>	6 cores cluster	j-2ACHK4SU3NDIB	Terminated All steps completed	2015-04-29 21:35 (UTC-5)	9 minutes	56
<input type="checkbox"/>	5 cores cluster	j-2BBG0J0YBRKL1	Terminated All steps completed	2015-04-29 21:33 (UTC-5)	8 minutes	48
<input type="checkbox"/>	4 cores cluster	j-8OJYX2FVDGZW	Terminated All steps completed	2015-04-29 21:29 (UTC-5)	8 minutes	40
<input type="checkbox"/>	3 cores cluster	j-1VDWF9AKANGPE	Terminated All steps completed	2015-04-29 21:23 (UTC-5)	9 minutes	32
<input type="checkbox"/>	My cluster	j-2397PQCL7GDP1	Terminated All steps completed	2015-04-29 19:36 (UTC-5)	7 minutes	24
<input type="checkbox"/>	Lab4 Cluster	j-Y0ENB1Q6K5B8	Terminated with errors Validation error	2015-04-29 19:10 (UTC-5)	35 seconds	0

The footer contains 'Feedback', 'English', copyright information '© 2008 - 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.', and links to 'Privacy Policy' and 'Terms of Use'.

Figure 1: Amazon EMR clusters