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**, zvith **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

$$part\_files = 4 \times cores - 1 \tag{1}$$

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:

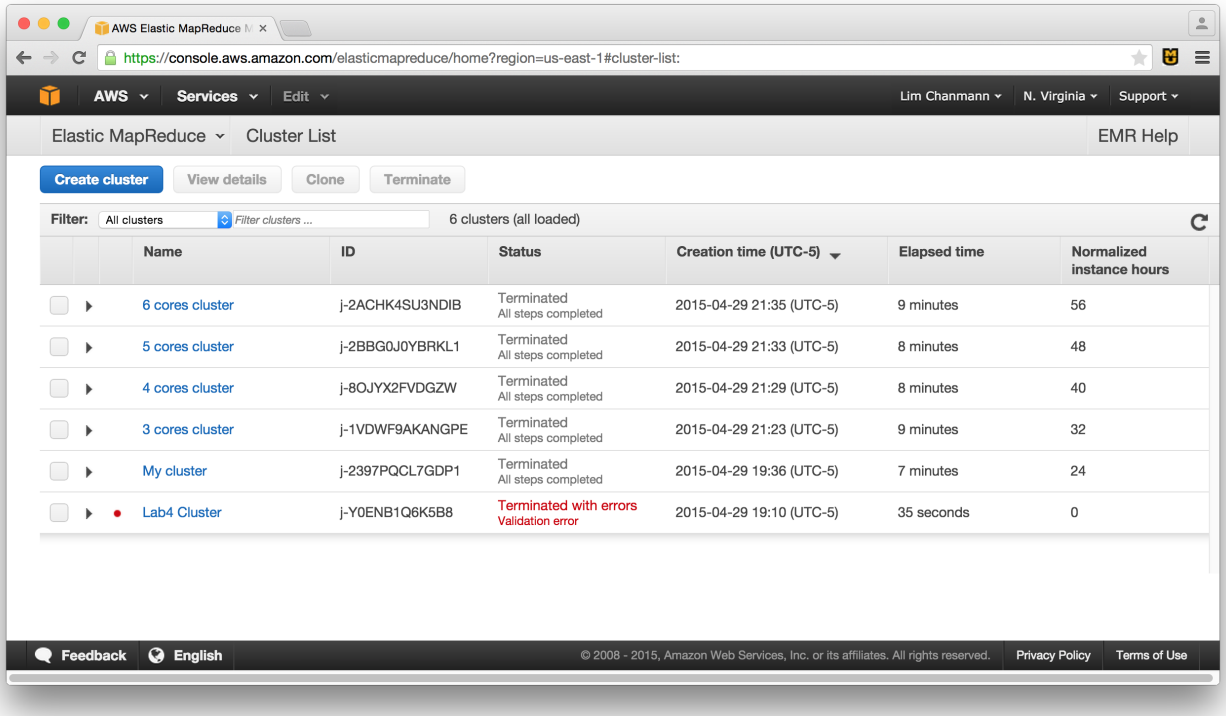

Figure 1: Amazon EMR clusters