Siddharth Chaphekar

CSC 555 – Mining Big Data

Summer 2019

Assignment 1

**PART 1 – Computation Problems**

1. Discrete Computations

* 210 = **1024**
* 45 = **1024**
* 85 = **32768**
* 837 MOD 100 = **37**
* 842 MOD 20 = **2**
* 16 MOD 37 = **16**
* 37 MOD 16 = **5**

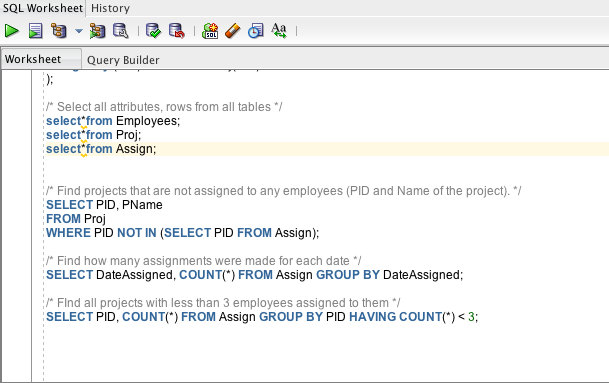
1. Vector Calculations

* V2 – V1 = (2-1, 1-2, 2-3) = **(1, -1, -1)**
* V1 + V1 = (1+1, 2+2, 3+3) = **(2, 4, 6)**
* |V1| = SQRT(12 + 22 + 32) = SQRT(14) =
* |V2| = SQRT(22 + 12 + 22) = SQRT(9) = **3**
* M \* V2 = **(11, 8, 6)**
* M2 = **[(8, 4, 14), (6, 5, 11), (4, 1, 7)]**
* M3 = **[(34, 16, 60), (28, 16, 50), (16, 6, 28)]**

1. Coin Flip Probabilities

* HTHT = 0.6 x 0.4 x 0.6 x 0.4 = **0.0576**
* THTT = 0.4 x 0.6 x 0.4 x 0.4 = **0.0384**
* Exactly 1 Head = {HTTT, THTT, TTHT, TTTH} = **0.1536**
* Exactly 1 Tail = {THHH, HTHH, HHTH, HHHT} = 4 x 0.0864 = **0.3456**

1. SQL Queries



1. Mining of Massive Datasets, Ex. 1.3.3

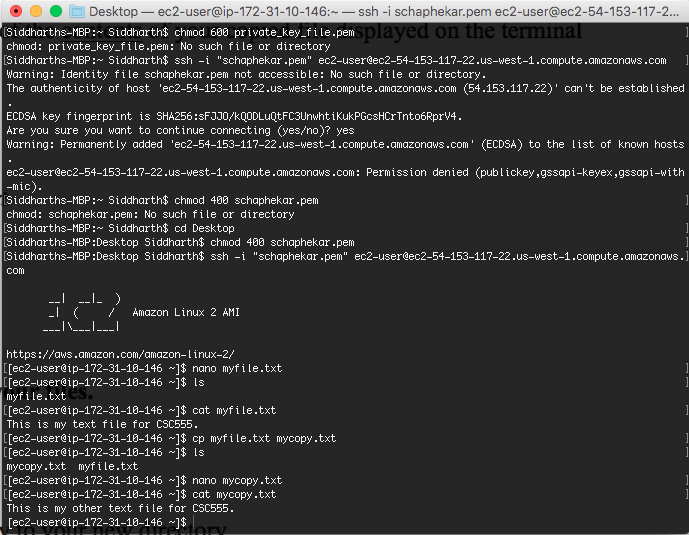
The hash function h(x) will only give values between 0 and 15 since it is x MOD 15. If we choose c = 1, then we can get hash keys that are divided equally into all of the buckets, but we choose c = 2 then we only get hash key distributions across buckets 0,2,4,6,8, … etc. Similar for c = 3, c = 4, etc. But for c = 16, we again get equal distribution, so **c should be one more than multiple of bucket size**. The number should be co-prime. So 2 and 4 and 16 are indeed good, but 3 is actually bad because 15 is 3x5.

1. MapReduce Implementation

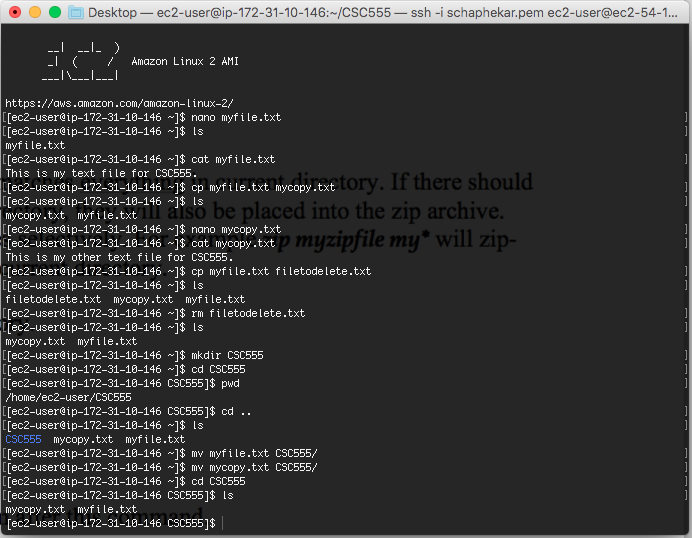
The Map function transforms input into meaningful key-value pairs. For example, if you were dealing with a dataset containing cities and states, it would combine those two attributes to form a key; the value would be a combination of other attributes related to that city. So, a MapReduce job generally divides the input data into smaller blocks that are processed in parallel by the map tasks.

**PART 2 – Linux Problems**

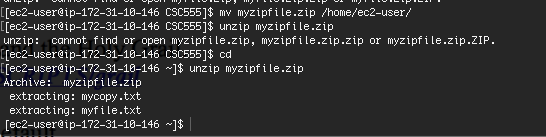
1) Contents of copied file displayed on terminal screen



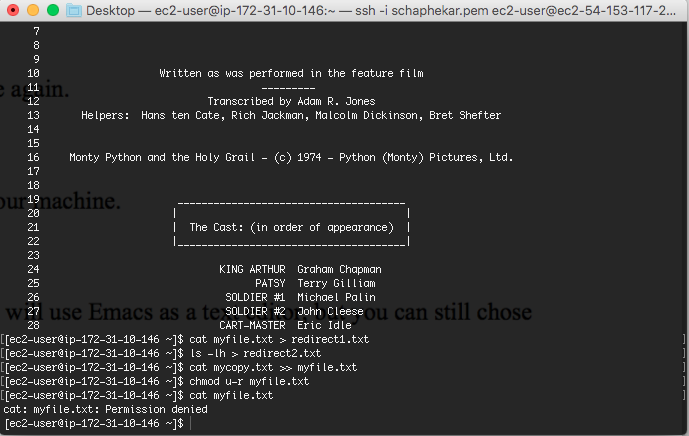
2) Files in CSC555 directory



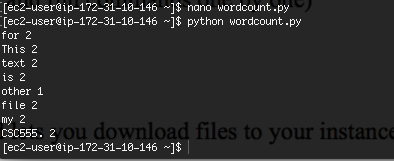
3) Unzipping myzipfile.zip



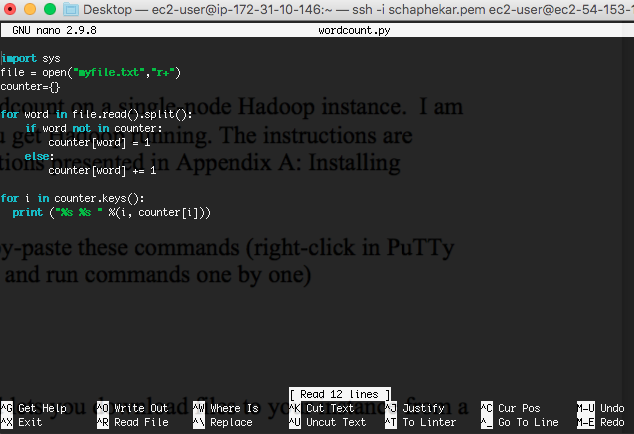
3) Size of grail file = 9 bytes + Permission denied



4) Python word counter, executed on myfile.txt

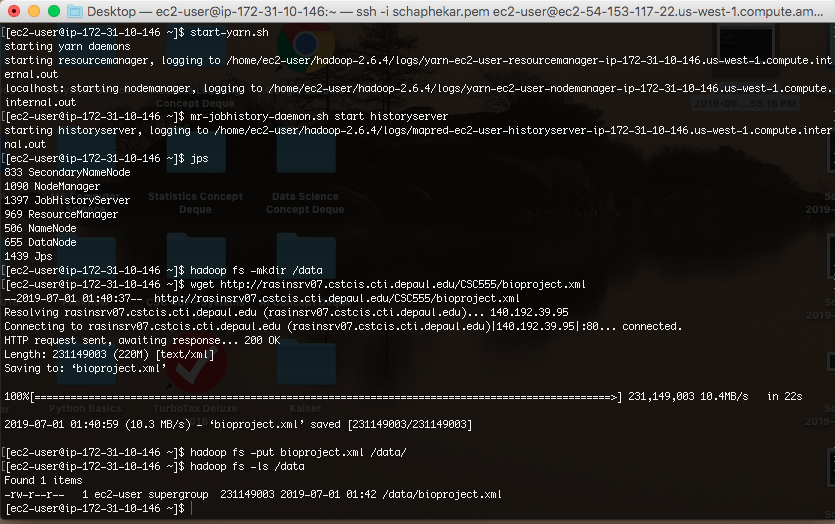


5) Python word counter code

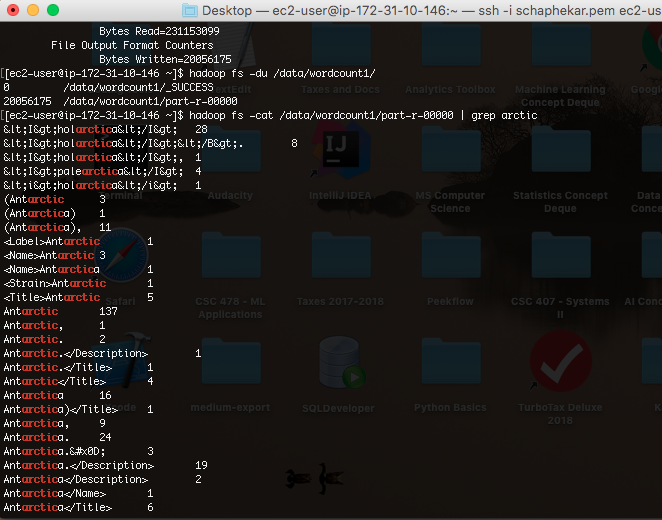


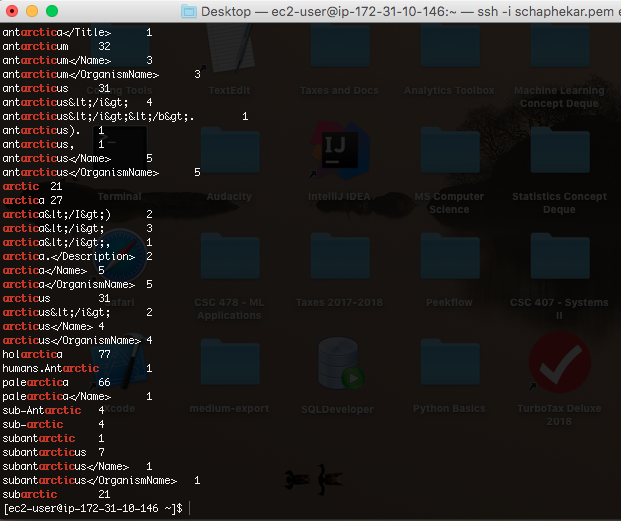
**PART 3 – Word Count Lab**

1) Verifying that the file was uploaded to HDFS



2) Occurrences of “arctic”





3) Job completion time

