**Map Reduce Job:**

We have created a Map reduce job that takes Wikipedia files uploaded on S3 bucket as input and generates key value pairs, which contain page names and their hit counts over a day. This map reduce job consists of 2 jobs chained together. The first job, collects all the page names from the wiki files and associate their hit counts with them. But the output we get is sorted according to the page names instead of the count. So the output from the first job is fed to the second job. The second Map Reduce job sorts all the records in descending order of page hits. Hence it is now easier to get top ‘n’ trending pages by reading first n records from the output file generated from second job.

**Job 1**

**Input: Wiki Files for page counts over a day.**

|  |  |
| --- | --- |
| **File 1: pagecounts-20151102-000000** | **File 2: pagecounts-20151102-010000** |
| en Christmas 11 1478  en Black\_Hole 121 1347  en Barack\_Obama 21 500 | en Christmas 7 1478  en Black\_Hole 12 1347  en Barack\_Obama 36 500 |

**Mapper**

Christmas 11

Black\_Hole 121

Barack\_Obama 21

Christmas 7

Black\_Hole 12

Barack\_Obama 36

**Reducer**

**Barack\_Obama 57**

**Black\_Hole 133**

**Christmas 17**

**Job 2**

This Map Reduce job sorts the results we get from first job, and gives us the sorted output. The default sort provided by Hadoop is for increasing order. But that can be changed by setting SortComparatorClass to DecreasingComparator.class

**Input: Output files from job 1**

**Barack\_Obama 57**

**Black\_Hole 133**

**Christmas 17**

**Mapper**

**57 Barack\_Obama**

**133 Black\_Hole**

**17 Christmas**

**Reducer**

**133 Black\_Hole**

**57 Barack\_Obama**

**17 Christmas**