Assignment_wk14

Prashant B. Bhuyan

December 17, 2014

Problem:

Your job here is to design (but not implement) a Map Reduce job to process this log data. You should start by describing what you want to accomplish, e.g. "What is the question that you will answer?" Be as specific as possible about what each stage (Map, Reduce, and if applicable, Sort and Shuffle) will be responsible for. Show mockups of what the data looks like as inputs and outputs to each stage.

Solution:

Below is a sampling of a file from http://dumps.wikimedia.org/other/pagecounts-raw/:

Here are a few sample lines from one file:

```
fr.b Special:Recherche/Achille_Baraguey_d%5C%27Hilliers 1 624
fr.b Special:Recherche/Acteurs_et_actrices_N 1 739
fr.b Special:Recherche/Agrippa_d/%27Aubign%C3%A9 1 743
fr.b Special:Recherche/All_Mixed_Up 1 730
fr.b Special:Recherche/Andr%C3%A9_Gazut.html 1 737
```

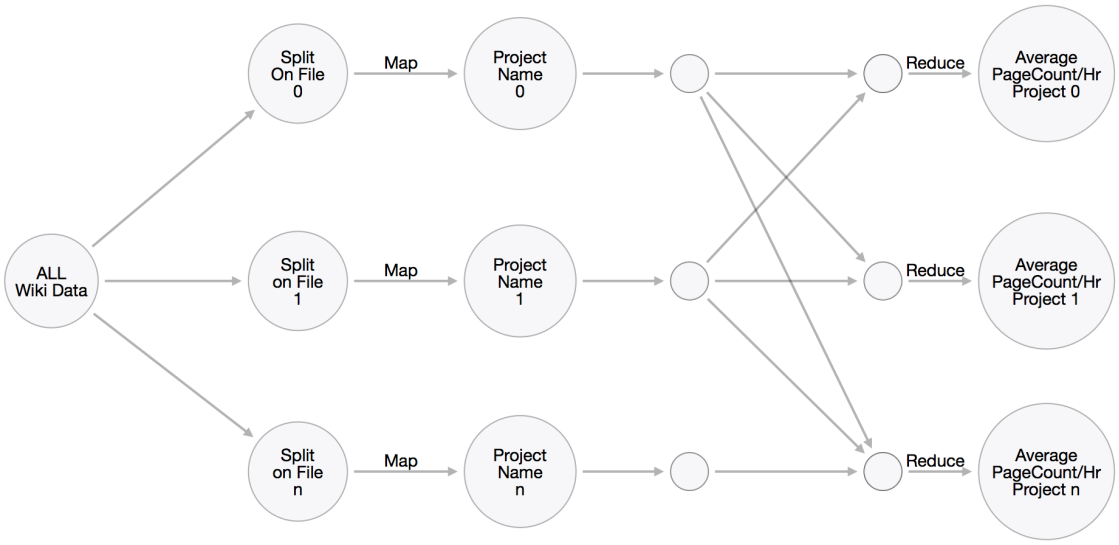
In the above, the first column "fr.b" is the project name. The following abbreviations are used:

wikibooks: ".b" wiktionary: ".d" wikimedia: ".m" wikipedia mobile: ".mw" wikinews: ".n" wikiquote: ".q" wikisource: ".s" wikiversity: ".v" mediawiki: ".w"

Goal: The map reduce job that I am designing will return the average number of page requests per hour by project. In other words, I want to know the average page requests per hour for wikipedia mobile across the entire data set from 2007 to 2011. I want to know the average page request per hour for te wikiversity project across the entire data set from 2007 to 2011, etc.

Design:

```
library(grid)
library(png)
img <- readPNG("/Users/MicrostrRes/Desktop/mapreduce_job_design.png")
grid.raster(img)
```


ALL
Wiki Data
Split
On File
0
Split
on File
1
Split
on File
n
Map
Map
Map
Project
Name
0
Project
Name
1
Project
Name
n
Reduce
Reduce
Reduce
Average
PageCount/Hr
Project 0
Average
PageCount/Hr
Project 1
Average
PageCount/Hr
Project n