EE 382N: Distributed Systems **Fall 2017**

Lecture 7: October 14

Lecturer: Vijay Garg *Scribe: Kyle Sung*

7.1 Introduction

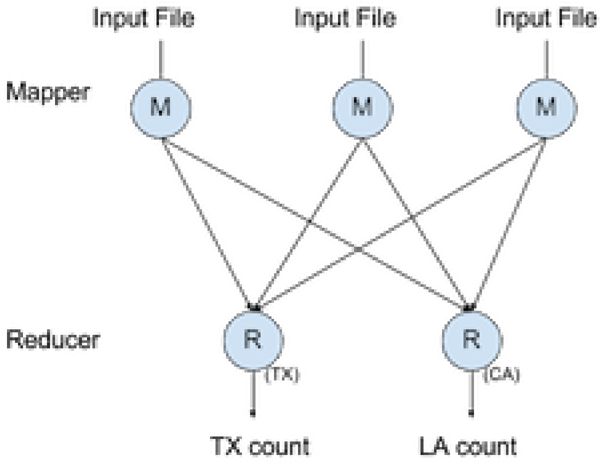
Hadoop is a software produeced by Yahoo based on the paper originally published by Google. It's goal is to help parallelism easier in software development by making the development process easier and taking care of fault tolerance.

There are two approaches to parallelism: Control flow: partition based on activities. (e.g. divide into threads) Data flow: partition based on data (e.g. hadoop, which divide data into key, value pairs)

7.2 MapReduce components

Mapper: divides input into (key,value)→ $setof(key, value) Reducer : setof(key, value) \rightarrow (key, value)$

Example: Word Count



Mapper: Input: documents containing words Output: pairs with intermediate keys Mapper → $(Texas, 2), (California, 1)$

Reducer: Performs operations that are associative and communtative. Such as sum, max, min. Graph here.

7.3 Fault Tolerance

Input from GFS: Files are divided up into chunks, each chunk replicated in multiple places. A master node tracks nodes of mapper activity and redistribute work to other mapper.

7.4 Modern development

Apache Spark: Spark's data are stored in a collection of servers that are resilient distributed dataset (RDD). Generally faster by voiding queueing on hard disk. MapReduce handles input in batched files.

Students in EE 382V are required to scribe lecture notes for one lecture. These lecture notes will be done using the document processing system called Latex. We have posted the file *scribe.tex* on the Canvas system. You can run `pdflatex` on that file to generate *scribe.pdf*. The remaining document shows usage of some of the commands in Latex.

References

[URL] , https://developer.yahoo.com/hadoop/tutorial/