CLOUD COMPUTING CONCEPTS

with Indranil Gupta (Indy)

MAPREDUCE

Lecture B

MAPREDUCE EXAMPLES

Some Applications of MapReduce

Distributed Grep:

- Input: large set of files
- Output: lines that match pattern

- Map – *Emits a line if it matches the supplied pattern*
- Reduce – *Copies the intermediate data to output*

Some Applications of MapReduce (2)

Reverse Web-Link Graph

- Input: Web graph: tuples (a, b) where (page a → page b)
- Output: For each page, list of pages that link *to* it

- Map – *process web log and for each input <source, target>, it outputs <target, source>*
- Reduce – *emits <target, list(source)>*

Some Applications of MapReduce (3)

Count of URL access frequency

- Input: Log of accessed URLs, e.g., from proxy server
- Output: For each URL, % of total accesses for that URL

- Map – *Process web log and outputs <URL, 1>*
- Multiple reducers – *Emits <URL, URL_count>*

(So far, like WordCount. But still need %)

- Chain another MapReduce job after above one
- Map – *Processes <URL, URL_count>* and outputs *<1, (<URL, URL_count>)>*
- 1 Reducer – Sums up *URL_count's* to calculate overall_count.

Emits multiple <URL, URL_count/overall_count>

Some Applications of MapReduce (4)

Map task's output is sorted (e.g., quicksort)

Reduce task's input is sorted (e.g., mergesort)

Sort

- Input: Series of (key, value) pairs
- Output: Sorted <value>s

- Map – *<key, value> -> <value, _> (identity)*
- Reducer – *<key, value> -> <key, value> (identity)*
- Partitioning function – partition keys across reducers based on ranges
  - Take data distribution into account to balance reducer tasks