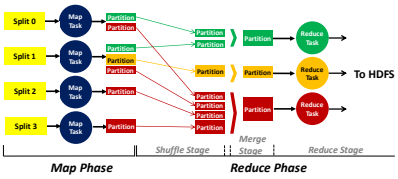
15-440
Distributed Systems
Recitation 11

Laila Elbeheiry
Slides by: Zeinab Khalifa

Project 4

Apply MapReduce to cluster analysis, using the **K-Means** algorithm

MapReduce: A Systems View

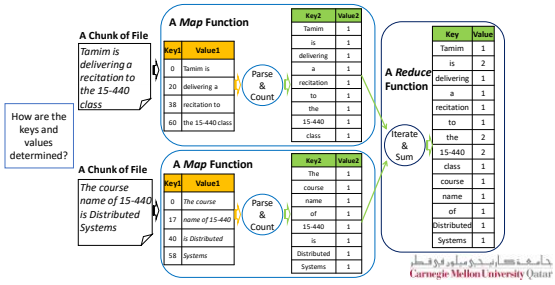


Data Structure: Keys and Values

- In a MapReduce program, the programmer has to specify **two functions**: the Map function and the Reduce function that implement the ***Mapper*** and the ***Reducer***, respectively
- In MapReduce, data elements are always structured as **key-value (i.e., (K, V)) pairs**
- Therefore, the Map and Reduce functions *receive* and *emit* (K, V) pairs

More accurately:
•**map:** $(K_1, V_1) \rightarrow \text{list}(K_2, V_2)$
•**reduce:** $(K_2, \text{list}(V_2)) \rightarrow \text{list}(K_3, V_3)$

MapReduce: An Application View



WordCount.java (Helpers)

- **Scanner Object:**

- A Scanner breaks its input into tokens using a delimiter pattern, which matches whitespace by default.
- hasNext(): checks if the Scanner has another token in its input.
- next(): gets the next token

- **MR Text object:**

- .set(token): sets a token to a Hadoop Text object

- **OutputCollector<Text, IntWritable> object:**

- .collect(x, y) sets a text x and Int y (k,v) paris output to the reduce function

What about Multiple Iterations?