## 15-440 Distributed Systems Recitation 11

Laila Elbeheiry Slides by: Zeinab Khalifa

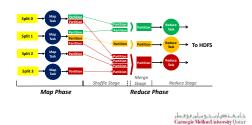
> جامعہ دارنیجی میلور ہی قطر Carnegie Mellon University Qatar

## **Project 4**

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

> جامعے ہار نیجی میلوں ہی قطر Carnegie Mellon University Qatar

### **MapReduce: A Systems View**



### **Data Structure: Keys and Values**

- In a MapReduce program, the programmer has to specify <u>two functions</u>: the Map function and the <u>Reduce function</u> that implement the <u>Mapper</u> and the <u>Reducer</u>, 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



# How are the keys and loss

#### How are the keys and values determine? A Chunk of File The course s Distributed Systems 1 How are the keys and values determine? A Chunk of File The course determined 15 -440 is Distributed Systems 1 How are the determined to the course to main the determined to the course determined to the determined tothe determined to the determined to the determined t

# 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

جامعے کا ریے جی میلوں ہی ہطر Carnegie Mellon University Qatar

# What about Multiple Iterations?

جامعہ کاربیجی میلوں ہی قطر Carnegie Mellon University Qatar