**CS5543 Real-Time Big Data Analytics**

MapReduce & Spark Programing

InClassEx-4

9/13/2015

Name:

Class ID:

**MapReduce & Spark Programming – Joining two data sets**

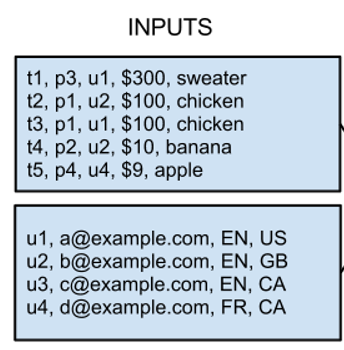
Two datasets are given as follows:

1. User information (id, email, language, location)
2. Transaction information (transaction-id, product-id, user-id, purchase-amount, item-description)

Given these datasets, find the number of unique locations in which each product has been sold.

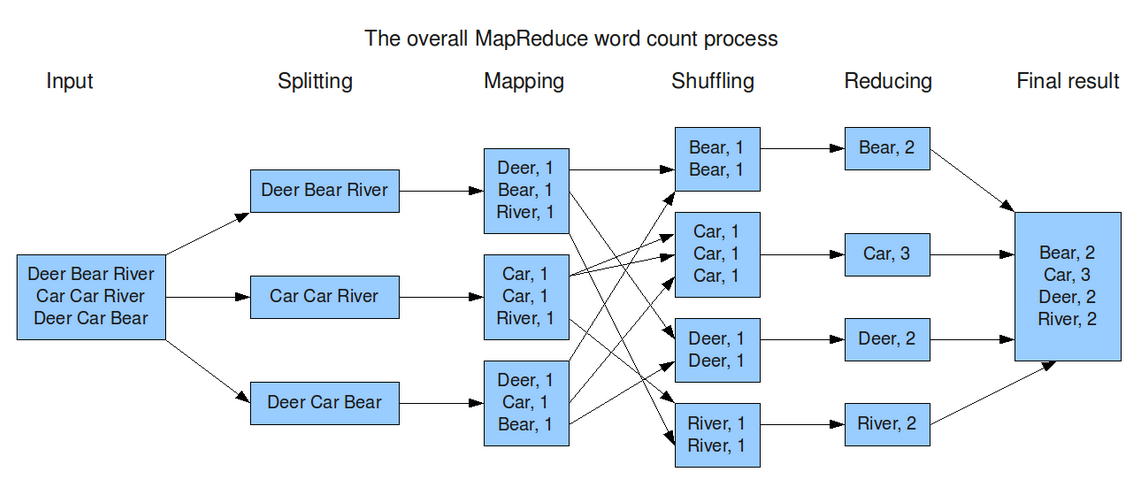
1. Draw the MapReduce Diagram.
2. Sketch the MapReduce Algorithm.
3. Sketch the Spark Scala implementation

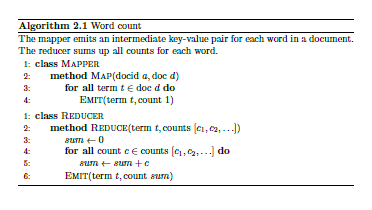
**Example**



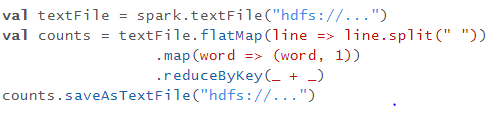
The result is:







**Spark Scala Code for WordCount**



|  |  |
| --- | --- |
| **flatMap(func)** | Similar to map, but each input item can be mapped to 0 or more output items (so func should return a Seq rather than a single item). |
| **reduceByKey**(*func*, [*numTasks*]) | When called on a dataset of (K, V) pairs, returns a dataset of (K, V) pairs where the values for each key are aggregated using the given reduce function *func*, which must be of type (V,V) => V. Like in groupByKey, the number of reduce tasks is configurable through an optional second argument. |