Ephrem Glushchenko
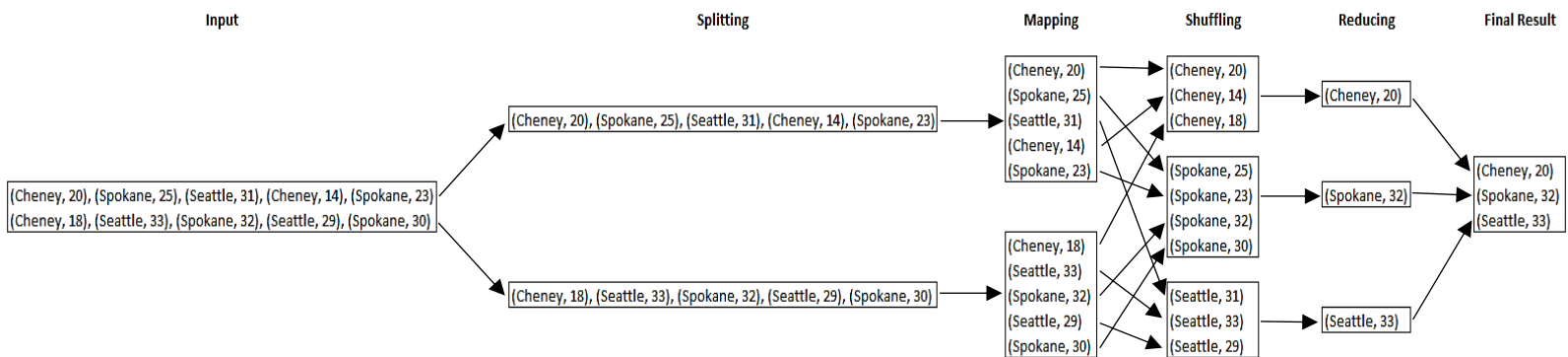Assignment 11
5/29/2019
MISC 485

Big Data and NoSQL

Question 1:

Assume there are two files each of which contains city names and their recorded temperatures (note that each file might have the same city represented multiple times) as follows:

File 01: (Cheney, 20), (Spokane, 25), (Seattle, 31), (Cheney, 14), (Spokane, 23)

File 02: (Cheney, 18), (Seattle, 33), (Spokane, 32), (Seattle, 29), (Spokane, 30)

Given these two files as the input, illustrate the Splitting, Mapping, Shuffling, Reducing, and Final Result steps (in this order) in the Map Reduce framework to return the maximum temperature for each city.

Input	Splitting	Mapping	Shuffling	Reducing	Final Result

(Cheney, 20), (Spokane, 25), (Seattle, 31), (Cheney, 14), (Spokane, 23)
(Cheney, 18), (Seattle, 33), (Spokane, 32), (Seattle, 29), (Spokane, 30)

(Cheney, 20), (Spokane, 25), (Seattle, 31), (Cheney, 14), (Spokane, 23)

(Cheney, 18), (Seattle, 33), (Spokane, 32), (Seattle, 29), (Spokane, 30)

(Cheney, 20)
(Spokane, 25)
(Seattle, 31)
(Cheney, 14)
(Spokane, 23)

(Cheney, 18)
(Seattle, 33)
(Spokane, 32)
(Seattle, 29)
(Spokane, 30)

(Cheney, 20)
(Cheney, 14)
(Cheney, 18)

(Spokane, 25)
(Spokane, 23)
(Spokane, 32)
(Spokane, 30)

(Seattle, 31)
(Seattle, 33)
(Seattle, 29)

(Cheney, 20)

(Spokane, 32)

(Seattle, 33)

(Cheney, 20)
(Spokane, 32)
(Seattle, 33)

Question 2:

Why in the CP model may some data not be accessible?

> CP is when availability is sacrificed only in the case of a network partition. Clients will be separate and won't be able to access the data until they are joined and updated.

Ephrem Glushchenko
Assignment 11
5/29/2019
MISC 485
Question 3:

Represent the information in the following relational tables as key-value pairs.

Building ID	NoOfFloors
A	3
B	2
C	2

BuildingID	AptNo	NoOfBedrooms
A	101	4
A	201	4
A	301	5
B	101	2
B	201	2
C	101	3
C	102	3
C	201	4

A: NoOfFloors	3
B: NoOfFloors	2
C: NoOfFloors	2

A: 101: NoOfBedrooms	4
A: 201: NoOfBedrooms	4
A: 301: NoOfBedrooms	5
B: 101: NoOfBedrooms	2
B: 201: NoOfBedrooms	2
C: 101: NoOfBedrooms	3
C: 102: NoOfBedrooms	3
C: 201: NoOfBedrooms	4

Question 4:

Why does column-oriented data storage make data access on disks faster than row-oriented data storage?

> It is faster because the disk can quickly loop over all of the column values and loading them all in a single memory reference while row-oriented data has to be skipped through since it may not me continuous making reading and writing speeds slower than column-oriented data storage.