# Big Data Analytics

**Homework 4 (MapReduce)**

**Instructor: Ruoming Jin**

**Shiva Kumar Peddapuram (**[**speddapu@kent.edu**](mailto:speddapu@kent.edu)**)**

In this homework, there are 4 questions + 1 bonus question, covering the topic of MapReduce. If you can answer the bonus question correctly, you can obtain 20 extra points. The maximum mark for this homework is **120 points**, which will be later scaled.

1. **Please use your own words to describe the Map and Reduce functions. [20 points]**

**Answer:**

Map reduce is the one of programming models that is basically used for the processing of large scale of datasets.

This shows lot of impact on the distributing computing’s.

In the map function the required business logic will be written to processing the data.

1. **Map function:**

* This is the initial step in the map reducing function.
* The map function usually gets the data from the user and then stores it in the form of key and the value as pair.
* We get the output in the form of intermediate key pair values.

1. **Reduce Function:**

* This is the second step in the map reduce function
* In this step it gets the data from the mapper
* The best part of reduce function is it can sort, shuffle and aggregate imtermediate results and then it generates the final output.
* It combines all final intermediate keys and makes a final key.

2. Assume that we have a relational database *D* containing tuples *t* with attributes *t*[*A*]. Write the pseudo code of Map and Reduce functions for retrieving all tuples with attribute *t*[*A*] > 100. [20 points]

**Answer:**

From the above given question the Re;ational Database D it contains tuples ‘t’ with Attributes t[A]

First let me briefly describe about mapper and reducer function

**Mapper Function:** it takes input as key/value and emits intermediate key/value pairs.

**Reduction Function:** take intermediate key/value pairs. Produce/emit output key/value pairs.

Text, letter

Description automatically generated

Pseudo code:

Input key=t

Input value=t(A)

Map(input key,input value)

{

If (input key > 100)

{

Emitintermediate(input\_key, input\_value)

}

}

Reduce(intermediate\_key, intermediate\_values[])

{

For each tuple in intermediate\_values[]

{

EmitOutput (t)

}

}

**3. Assume that we have two relational databases *R* and *S* containing tuples *r* and *s*, respectively. Write the pseudo code of Map and Reduce functions for joining databases *R* and *S* on attribute *A* (i.e., joining tuples *r* and *s* satisfying *r*[*A*] = *s*[*A*]). [20 points]**

Answer:

Given,

Relational databases

R – containing r tuples

S – containing s tuples

Given, condition to join both the databases r[A] = s[A]

Pseudo Code-

**Mapper logic :**

Map(input key, input value)

{

If(input key==D[A])

{

Emit(intermediate\_key, intermediate\_value)

}

}

//from question 1 we know that Reduce function is used to looping over intermediate keys over each group.

Reduce(output key, output values[])

For each tuple in R

{

For each tuple in S

{

If (R[A]==S[A])

Emit(current value)

}

}

**4. Please write a tutorial for installing Apache Hadoop (on either windows or MacOS) or configuring Amazon Web Services. *Please use screen captures and record the entire installation process step by step.* [40 points]**

Answer:

**Installing Hadoop on windows 11**

Initially I have checked whether java is installed or not in my pc but I came to know that java is not installed so initially I am installing java from the official oracle site

Reference: <https://www.oracle.com/java/technologies/downloads/>

**So here is the first step of the java installation.**

Graphical user interface, text, application

Description automatically generated

Here I am attaching the site link from where I have downloaded the java

Graphical user interface, text, application, email

Description automatically generated

Here are the series of the steps of installation process.

Graphical user interface, text, application

Description automatically generated

**Java installed successfully**

Graphical user interface, application

Description automatically generated

**Setting environment variables for java**

Graphical user interface, application

Description automatically generated

**Checking whether java installed or not for our information java got installed succesfuly**

Text

Description automatically generated

**Downloading apache from the official site**

Graphical user interface, text, website

Description automatically generated

**Succefully installed hadoop**

Graphical user interface, text, application

Description automatically generated

**Extracting the files**

Graphical user interface, application, Teams

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Showing all the contents of the extraction process

Graphical user interface, application, table

Description automatically generated

**Setting Hadoop variable in the environments**

Graphical user interface, application

Description automatically generated

**Java home setting variable**

A picture containing text, indoor, screenshot, several

Description automatically generated

**Hadoop system variable**

Graphical user interface, application

Description automatically generated

**Next I configured core-site.xml file**

Graphical user interface, text, application

Description automatically generated

**Next I have changed the file name from mapred-site.xml template to mapred-site.xml**

Graphical user interface, text, application

Description automatically generated

**In the next step I have create a main folder named “data ” under C:\Program Files\hadoop-dependencies-3.2.2\**

**Then further I have created two folders named as data node and namenode under the C:\Program Files\hadoop-dependencies-3.2.2\data**

**In the next step I have edited “hdfs-site.xml” and “yarn-site.xml”**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

In the next step I have edited command line in “Hadoop-env.cmd”

Setting java home location

Text, application

Description automatically generated

Now Hadoop installation process is done successfully.

**References:**

1. <https://www.oracle.com/java/technologies/downloads/>
2. <https://www.datasciencecentraal.com/profiles/blogs/how-to-install-and-run-hadoop-on-windows-for-beginners>
3. [Apache Hadoop](https://hadoop.apache.org/releases.html)
4. <https://hadoop.apache.org/releases.html>

**Bonus Question [20 extra points]**

5. Implement the Word Count example (as mentioned in the lecture slides) on Hadoop or AWS. *Please include the source code and add screen captures to illustrate the execution process.*

***Answer:***