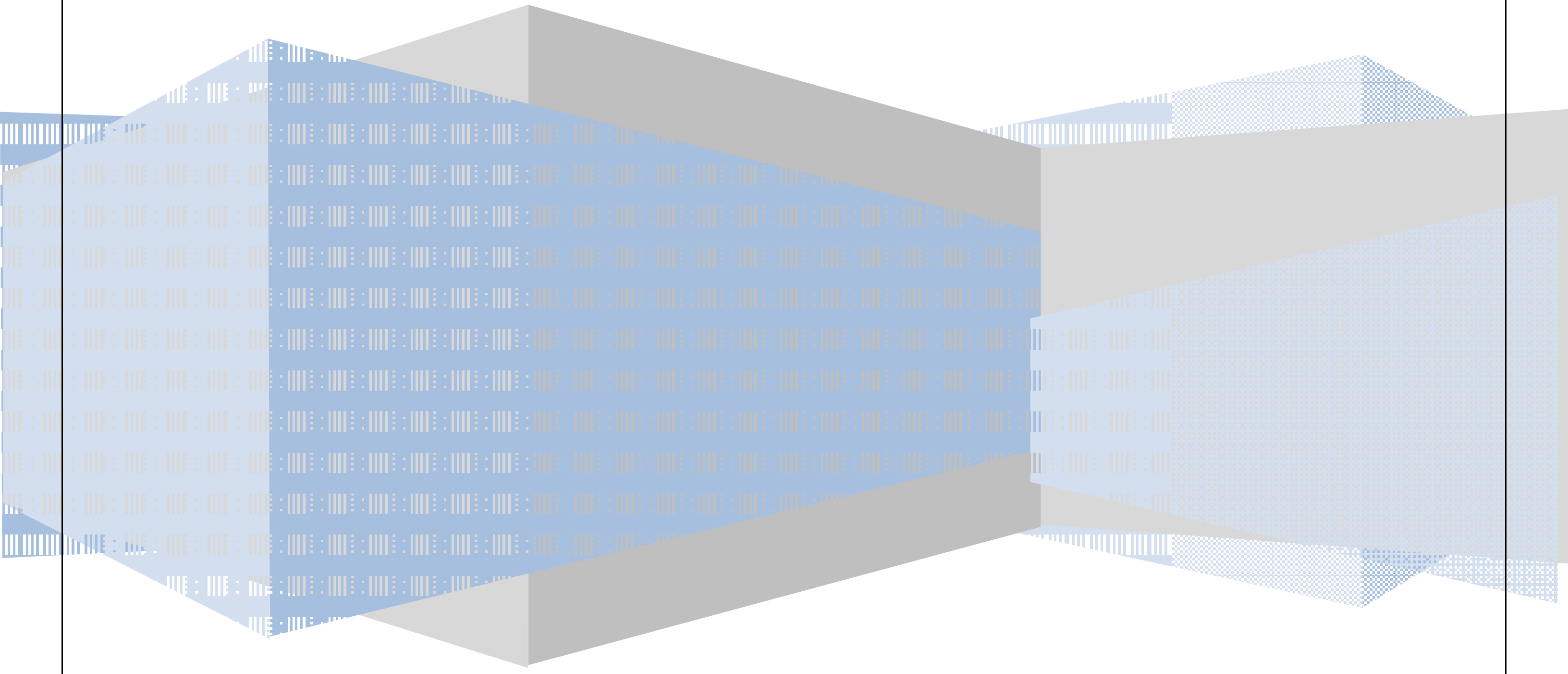
Collabera
Value. Accelerated.

Collabera
TACT
360° Training
Technology Academy for Competency Training

IT TOOK YOU YEARS OF EXPERIENCE
TO REACH WHERE YOU ARE TODAY.
ANOTHER 45 HOURS OF TRAINING
WITH TACT WILL TAKE YOU HIGHER.
TACT is an online technology academy for competency training to train the IT professionals and uplift their career to discover more opportunities in the space of emerging technologies.

Big Data/Hadoop Training

MapReduce Assignment



MapReduce Assignment

Exercise

1) Run the WordCount program and make sure the output is correct.

2) Put a break point the in the WordCount program within Eclipse and check the input K and the V pairs for the map and the reducers and also check how many times the map and the reduce function is called.

3) Use a combiner for the above WordCount program and the input K and the V pairs for the map and the reducers and also check how many times the map and the reduce function is called using a Break Point.

4) Run the WordCount program in Python and Ruby.

5) Modify the WordCount program to calculate index.

The program input would be
file1.txt
Hadoop is easy
file2.txt
Hadoop is fast

and the expected output is

Hadoop,{file1.txt,file2.txt}
is,{file1.txt,file2.txt}
easy,{file1.txt }
fast, {file2.txt}

6) Modify the WordCount program to get the word count per file/book.

The program input would be

file1.txt

Hadoopiseasy
Hadoop is cool

file2.txt

Hadoopisfast
Pig is fast

and the expected output is

file1.txt+Hadoop,2
file1.txt+is,2
file1.txt+easy,1
file1.txt+cool,1
file2.txt+Hadoop,1

```
file2.txt+Pig,1
file2.txt+is,2

file2.txt+fast,2
```

7) For the above program, set the number of reducers to 4 and run the MR program again. In Apache Hadoop the default number of reducers is 1 and Cloudera CDH is 2.

8) For the above program, change the job name to 'Problem8' and execute it.

9) Given a school database in the following format in a file.

Roll Number | School Name | Name | Age | Gender | Class | Subject | Marks

Develop the algorithm and the MapReduce code for the following questions and use the students-db.txt for the corresponding data.

a) Who got the highest for each class?
b) Who got the highest across all the schools for each class?
c) Sort the students according to the total marks for each school?
d) Did boys fare better or girls for each class?

10) Give the movie and a movie rating data in the following format

movie
id(pk) | name | year

movierating
userid | movieid (fk) | rating

Develop the algorithm and the MapReduce code for the following questions and use the movie.txt and movierating.txt for the corresponding data

a) What is the oldest known movie in the database? Note that movies with unknown years have a value of 0 in the year field - these do not belong in your answer.

b) List the name and year of all unrated movies (movies where the movie data has no related movierating data).

c) Produce an updated copy of the movie data with two new fields:
- numrating (the number of ratings for the movie)
- avgrating (the average rating for the movie)

d) What are the 3 highest-rated movies?