ISEngineering
Wirtschaftsinformatik –
Information Systems Engineering

Technische Universität Berlin

Prof. Dr. Stefan Tai
Markus Klems

Your name: Budi Yanto (308819)

Enterprise Computing (WS 2014)

Exercise 3 (3 Portfoliopunkte)

Info:

- The solution to this exercise must be handed in by Tuesday, Nov 18th 2014, 2PM to Markus Klems.
- The solution must be printed out. Please write your name on the solution sheet.

Task 1 – Simple Word Count (20%)

- Launch an AWS EMR cluster
- Configure the sample application "Word count" as follows

Master m1.medium -> 1 instance

Core m1.medium -> 2 instances

Task m1.medium -> 0 instances

a) What is the word count of the word "accepted"? (10%)

Solution: 402

b) What is the word count of the word "gibraltar"? (10%)

Solution: 931

Don't forget to terminate the EMR cluster. There should be no running EC2 instances after you have completed the exercise.

Next, we want to program the MapReduce program by ourselves in Java.

Prerequisites:

Install the following plugins in your Eclipse IDE (if not already installed)

- http://download.eclipse.org/releases/kepler -> "Database Development"
- http://aws.amazon.com/eclipse

Instructions:

- Clone Markus's git repository: $ git clone git@gitlab.tubit.tu-berlin.de:klems/ecmapreduce.git
- Open the project in Eclipse as AWS Java project
- Set up your AWS credentials file like this:

#/Users/markus/.aws/credentials

[default]

aws_access_key_id = your_access_key_id

aws_secret_access_key = your_secret_access_key

- Run Maven install to build the project
- Complete the missing code marked with TODO
- Generate a jar file with Maven
- Upload the jar file to S3
- Create a new EMR cluster and select your custom jar file while creating the EMR cluster
- Enter the CLUSTER_ID in the MapReduceClient code
- Run the MapReduceClient as local Java application in Eclipse

Task 2 – Modified Word Count (40%)

Complete the missing code in WordCountABA and answer the following question: which words end with 'aba' and what is their word count?

Solution:

Ababa	10
Aqaba	3
Banaba	16
Jaraba	3
Kairaba	3
Maneaba	3
Shaba	1
Ayribaba	3
Madaba	3
Orizaba	3
Saba	10
Asaba	3
Baba	3
Ggaba	3
[Hajibaba	3

Task 2 – Palindrome Count (40%)

A palindrome is a word, phrase, number, or other sequence of symbols or elements that reads the same forward or reversed, with general allowances for adjustments to punctuation and word dividers. Complete the missing code in PalindromeCount and answer the questions: which words with at least 5 letters are palindromes and what is their word count? Do not include words that only consist of numbers in your answer.

Solution:

TALAT	6
level	161
INONI	3
OGEGO	3
SALAS	3
sexes	9
civic	12
refer	11

References:

- Launch an EMR cluster
  http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-get-started-count-words.html
- Java code example
  http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-common-programming-sample.html