Session 4

Assignment 3

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 4 – Assignment 3 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 4 – Assignment 3** |
|  |  |
| **Start Date:** | 15/09/2017 |
|  |  |
|  |  |

# 

# Contents

[Contents 2](#_Toc493255589)

[Change History 3](#_Toc493255590)

[1. Problem Statement 4](#_Toc493255591)

[2. Solutions 5](#_Toc493255592)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 07/09/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

Write a program to implement wordcount using Pig.

Share the screenshots of the commands used with its associated output.

# Solutions

The following pig script finds the number of times a word repeated in a file:

So we count all the words as one, and **The** is not different from **the** a UDF is created in java to change all words to lowercase.

Java class created in eclipse.

*package pig;*

*import java.io.IOException;*

*import org.apache.pig.EvalFunc;*

*import org.apache.pig.data.Tuple;*

*public class pigudf extends EvalFunc<String> {*

*public String exec(Tuple input) throws IOException {*

*if (input == null || input.size() == 0)*

*return null;*

*String str = (String) input.get(0);*

*return str.toLowerCase();*

*}*

*}*

**The class is the exported to create a jar file.

The jar file is then copied to the hdfs file system

$ hadoop fs -copyFromLocal pigudf.jar

Pig is run in mapreduce mode.

Register the jar file just created using REGISTER.

grunt> REGISTER pigudf.jar

Define the jar file stating package.classfile

grunt> DEFINE mypigudf pig.pigudf;

File created and previously copied to HDFS.

**File words**

*the*

*THe*

*The*

*What*

*when*

*wherE*

*were*

*DOG*

Run the words file using the UDF

*grunt> lower\_case = FOREACH pigwords GENERATE mypigudf(word);*

(the)

(the)

(the)

(what)

(when)

(where)

(were)

(dog)

All word converted to lower case.

Now to perform a word count of the newly created lower\_case file.

*words = FOREACH lower\_case GENERATE FLATTEN(TOKENIZE(word)) as word1;*

*grouped = GROUP words BY word1;*

*wordcount = FOREACH grouped GENERATE group, COUNT(words);*

*DUMP wordcount;*

**The results**

(dog,1)

(the,3)

(were,1)

(what,1)

(when,1)

(where,1)