Session 21

Assignment 1

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 21 – Assignment 1 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 21 – Assignment 1** |
|  |  |
| **Start Date:** | 22/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc496417441)

[Change History 3](#_Toc496417442)

[1. Problem Statement 4](#_Toc496417443)

[2. Datasets 4](#_Toc496417444)

[3. Solution 5](#_Toc496417445)

[4. Results 6](#_Toc496417446)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 22/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

* Count the numbers of blank lines in a text file using accumulators

# Datasets

For this I used the readme.MD file that is in the Spark installation directory exert of the file below showing populated and blank lines:

# Apache Spark

Spark is a fast and general cluster computing system for Big Data. It provides

high-level APIs in Scala, Java, Python, and R, and an optimized engine that

supports general computation graphs for data analysis. It also supports a

rich set of higher-level tools including Spark SQL for SQL and DataFrames,

MLlib for machine learning, GraphX for graph processing,

and Spark Streaming for stream processing.

<http://spark.apache.org/>

## Online Documentation

You can find the latest Spark documentation, including a programming

# Solution

**Code written**

**package** com.duncb.spark

**object** aggregation {

**import** org.apache.spark.\_

**import** org.apache.spark.SparkContext.\_

**import** org.apache.log4j.\_

**import** org.apache.spark.sql.\_

**import** org.apache.spark.sql.functions.\_

**def** main(args: Array[*String*]) {

// Create a SparkContext using every core of the local machine

**val** sc = **new** SparkContext("local[\*]", "aggregation")

**val** sqlContext = **new** org.apache.spark.sql.SQLContext(sc)

**import** sqlContext.implicits.\_

**val** spark = SparkSession

.builder

.appName("SparkSQL")

.master("local[1]")

.config("spark.sql.warehouse.dir", "file:///C:/temp") // Necessary to work around a Windows bug in Spark 2.0.0; omit if you're not on Windows.

.getOrCreate()

Logger.getLogger("org").setLevel(Level.ERROR)

**val** readme = sc.textFile("file:///N:/Datasets/README.md")

**val** blLines = sc.~~accumulator~~(0)

**val** popLines = sc.~~accumulator~~(0)

**val** count = readme.foreach {line => **if** (line.length() ==0) blLines +=1 **else** popLines +=1}

println( "Number of blank lines in file is " + blLines.value)

println( "Number of populated lines in file is " + popLines.value)

}

}

# Results

The results are checked and correct after manually counting 39 blank and 64 populated.

Number of blank lines in file is 39

Number of populated lines in file is 64