Session 19

Assignment 3

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 19 – Assignment 3 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 19 – Assignment 3** |
|  |  |
| **Start Date:** | 19/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc496319310)

[Change History 3](#_Toc496319311)

[1. Problem Statement 4](#_Toc496319312)

[2. Solution 4](#_Toc496319313)

[3. Results 4](#_Toc496319314)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 20/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

* Create a dataframe with 1 to 100 and save as parquet file.

# Solution

I have deliberately commented and printed out all steps for enhanced learning

**Code written**

**package** com.duncb.spark

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

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

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

**object** parquet {

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

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

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

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

**import** sqlContext.implicits.\_

**val** kk = sc.parallelize(1 to 100)

**val** pdata = kk.toDF()

pdata.write.format("parquet").save("file:///N:/Datasets/pdata.parquet")

}

}

# Results

A directory was created called pdata.parquet

The resultant files are in the following screen shot

