**GUIDELINES**

|  |  |  |  |
| --- | --- | --- | --- |
| **Solution** | Big, Project – Application of Knowledge | | |
| **Description** | Work count project for Apache Spark | | |
| **Author** | Viet Hung Duong | **Drafted Date** | 10/20/2016 |
| **Reviewed** |  | **Reviewed Date** |  |
| **Version** | 0.1 | **Last Modified Date** |  |

Table of Contents

Purpose 3

Apache Spark setup 3

Configuration Prequiste and cloudera setup 3

Word Count project 3

Create Apache Spark project with Maven 3

Run batch file 4

How batch file work 4

References 4

# Purpose

This document describes how to set up and configure Apache Spark installation.

# Apache Spark setup

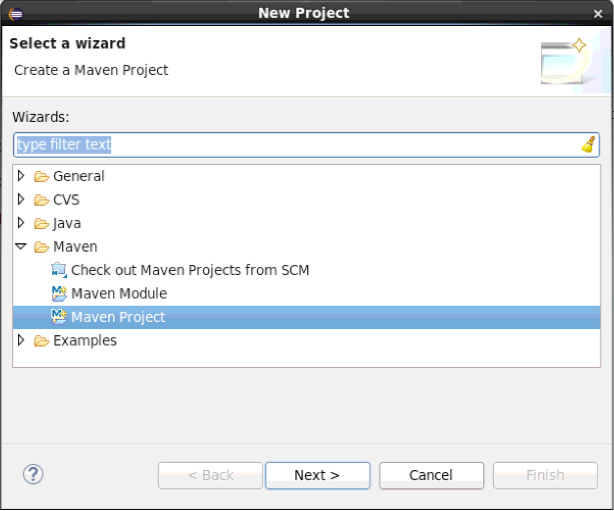
## Configuration Prequiste and cloudera setup

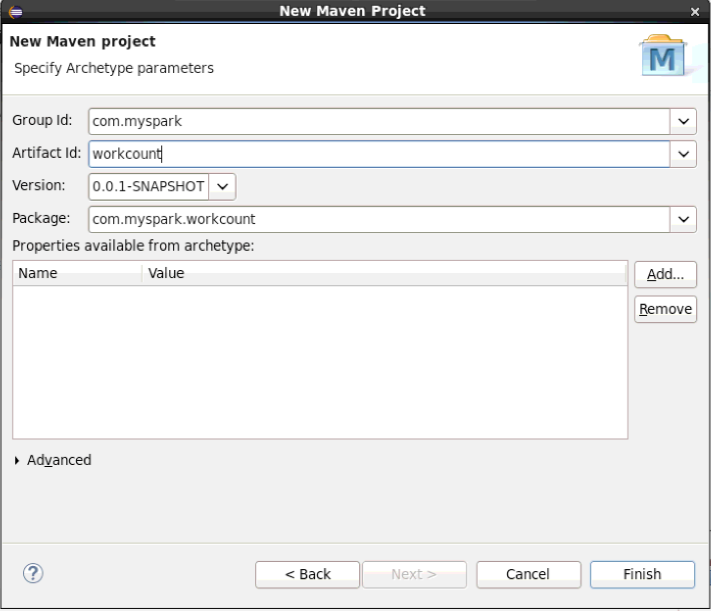
Please refer to document in Hadoop project.

## Word Count project

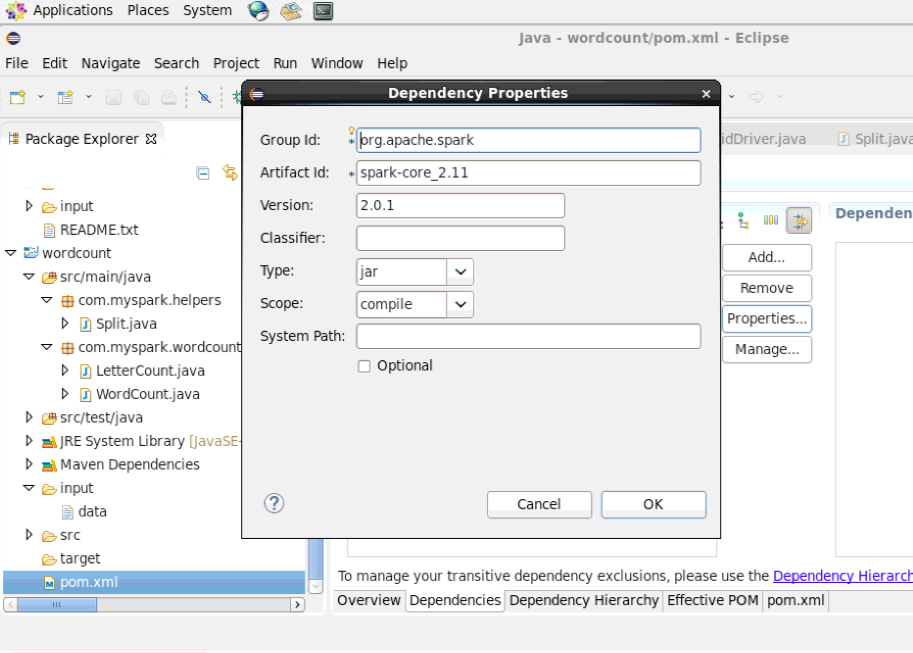
### Create Apache Spark project with Maven

1. Create new Maven project





1. Add spark-core dependency



### Run batch file

1. Copy the Spark project folder to local
2. Open the terminal
3. Cd to Spark project
4. Complile the batch file by command: **chmod +x run\_spark\_job.sh**
5. Execcute command: **sh ./run\_spark\_job.sh type** with type {1, 2, 3}

### How batch file work

Below the steps that batch file will work

Step 0: Create Jar filename – If jar file is existed, delete it and create new.

Step 1: Remove input and output in Hadoop

Step 2: Create input folder to store input data

Step 3: Copy data to input folder

Step 4: Run Spark on Hadoop cluster

Step 5: Copy to results to local

Step 6: View all results

# References

<http://spark.apache.org/docs/latest/index.html>

<https://www.tutorialspoint.com/apache_spark/apache_spark_rdd.htm>