# Create Spark Scala project using sbt and Eclipse

*Run in linux, with windows you need to install winutils*

* Create working directory for new project

/home/hadoop/Projects/SparkProjects/WordCount

* Under working directory create src/main/scala directory structure

mkdir -p src/main/scala

* Create build.sbt

name := "Word Count"

version := "1.0"

organization := "org.dikshit"

scalaVersion := "2.11.8"

libraryDependencies += "org.apache.spark" %% "spark-core" % "2.3.1"

resolvers += Resolver.mavenLocal

Run sbt package – it will take some time to download all the dependencies

* Run sbt eclipse
  + First create a file “plugins.sbt” inside project folder, which contains the following

addSbtPlugin("com.typesafe.sbteclipse" % "sbteclipse-plugin" % "5.2.4")

(For sbt 0.13 and up)

<https://github.com/sbt/sbteclipse>

* Import into Scala IDE for Eclipse

File → Import → General/Existing Project into Workspace.

* Develop simple spark program

Go to “File -> New -> Scala Object”

Name it as “SimpleApp”

package org.dikshit.spark

import org.apache.spark.SparkContext, org.apache.spark.SparkConf

object SimpleApp {

def main(args: Array[String]) {

val conf = new SparkConf().setAppName("scala spark").setMaster(args(0))

val sc = new SparkContext(conf)

val i = List(1, 2, 3, 4, 5)

val dataRDD = sc.parallelize(i)

dataRDD.saveAsTextFile(args(1))

}

}

### Run the Spark Application using Scala IDE

* Right Click on the program, go to “Run As” -> “Run Configurations”
* Choose Project and Main Class
* Click on Arguments and update program arguments
* SimpleApp main function expects 2 parameters
  + Cluster mode (pass local)
  + Output path (pass output directory) – don’t create the output directory. It will be created by itself
* Click on Run
* Make sure there are no logs and validate the path
* It should have data from the collection that is created as part of th

Running on YARN

spark-submit --class "org.dikshit.spark.SimpleApp" --master yarn --executor-memory 512m --total-executor-cores 1 target/scala-2.11/word-count\_2.11-1.0.jar yarn-client /dikshit/output

Read Output

hdfs dfs -cat /dikshit/output/part-00000

Common troubleshooting

Setup Hadoop\_Conf\_Dir to HADOOP\_HOME/etc/hadoop

spark-submit --class org.dikshit.etl.Etl1 \

--master yarn \

--deploy-mode cluster \

--driver-memory 4g \

--executor-memory 1g \

--executor-cores 1 \

etl.jar \

/dataset1/booking

name := "example"

organization := "com.databricks"

version := "0.1-SNAPSHOT"

scalaVersion := "2.11.8"

// Spark Information

val sparkVersion = "2.2.0"

// allows us to include spark packages

resolvers += "bintray-spark-packages" at

"https://dl.bintray.com/spark-packages/maven/"

resolvers += "Typesafe Simple Repository" at

"http://repo.typesafe.com/typesafe/simple/maven-releases/"

resolvers += "MavenRepository" at

"https://mvnrepository.com/"

libraryDependencies ++= Seq(

// spark core

"org.apache.spark" %% "spark-core" % sparkVersion,

"org.apache.spark" %% "spark-sql" % sparkVersion,

)