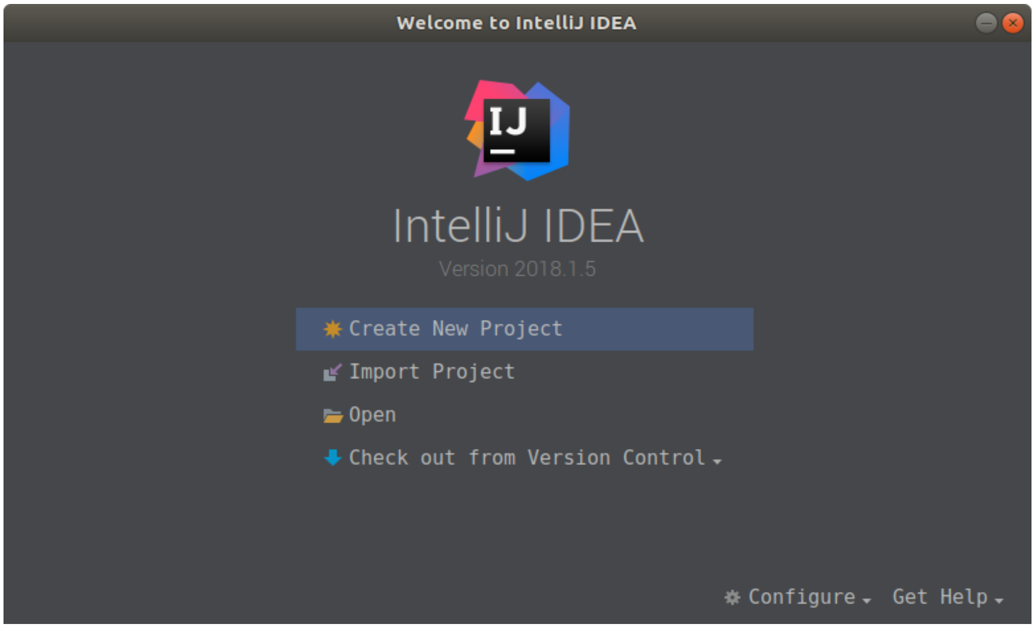
Spark Project in Scala documentation

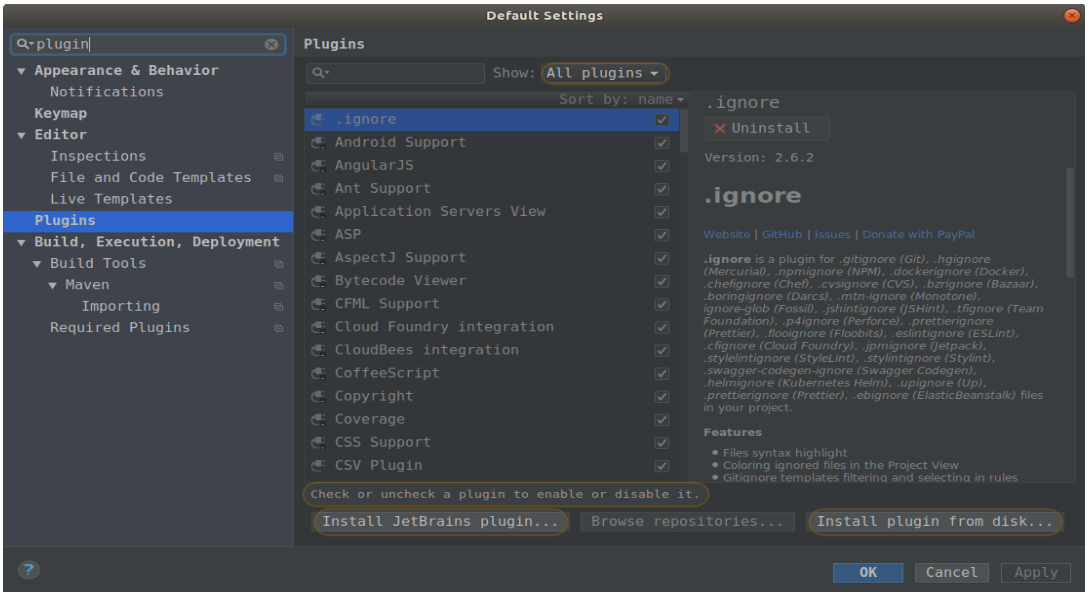
I have chosen to use **IntelliJ IDEA** environment and reason is that I have been using this tools for 5 years. It is very easy to use and supports many languages such as Java, Scala, Kotlin, Node.js, Spring framework project and so on. So, let's get started.
First, let's open IntelliJ IDEA environment.

Licensed to Bakhodir Boydedaev
Subscription is active until November 10, 2018

After opening the the tool, we need to configure IntelliJ IDEA in order to supports Scala project. So, click configure menu from the view.



Next step is to select plugins menu and type **Scala** keyword



.ignore

✕ Uninstall

Version: 2.6.2

.ignore

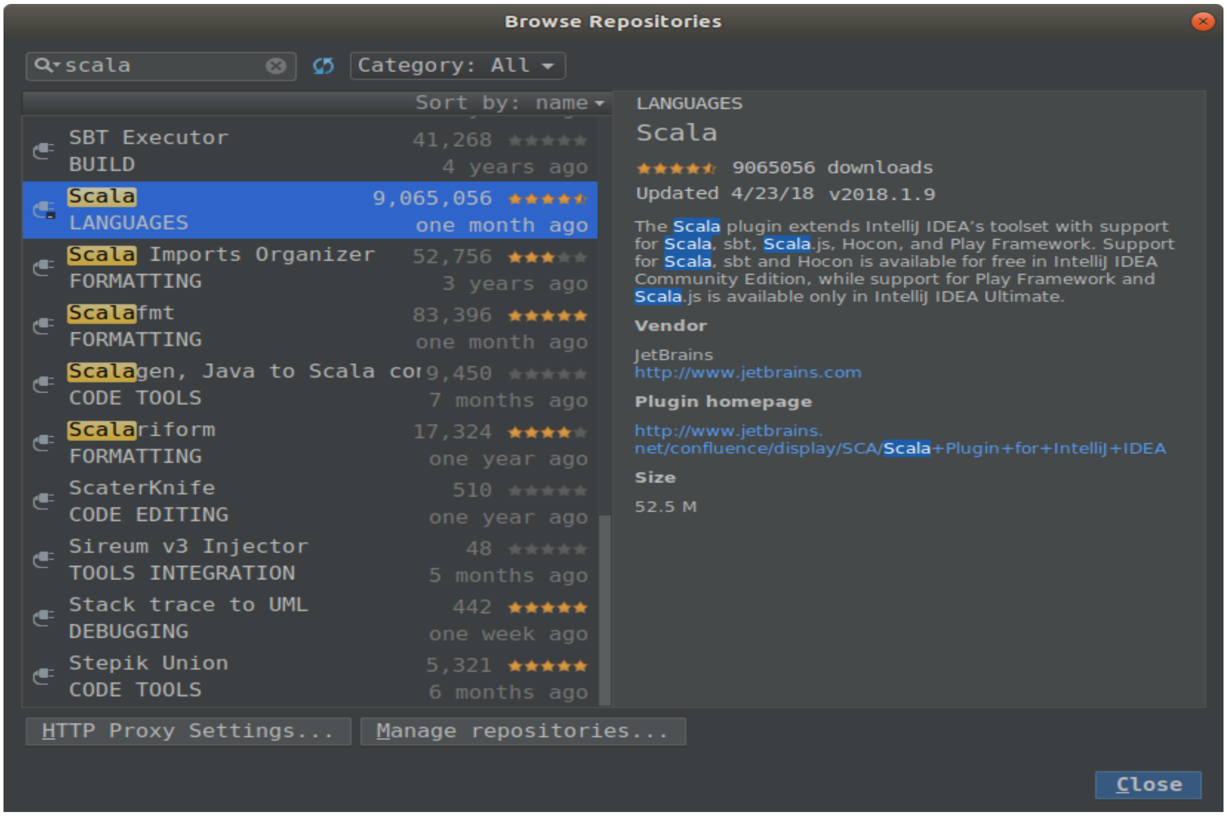
Website | GitHub | Issues | Donate with PayPal

.ignore is a plugin for *.gitignore (Git), .hgignore (Mercurial), .npmignore (NPM), .dockerignore (Docker), .chefignore (Chef), .cvsignore (CVS), .bzrignore (Bazaar), .boringignore (Darcs), .mtn-ignore (Monotone), ignore-glob (Fossil), .jshintignore (JSHint), .tfignore (Team Foundation), .p4ignore (Perforce), .prettierignore (Prettier), .flooignore (Floobits), .eslintignore (ESLint), .cfignore (Cloud Foundry), .jpmignore (Jetpack), .stylelintignore (StyleLint), .stylintignore (Stylint), .swagger-codegen-ignore (Swagger Codegen), .helmignore (Kubernetes Helm), .upignore (Up), .prettierignore (Prettier), .ebignore (ElasticBeanstalk)* files in your project.

Features

- Files syntax highlight
- Coloring ignored files in the Project View
- Gitignore templates filtering and selecting in rules

In my machine it is already installed, if you haven't installed yet there is **install** to instal it. You just need to click it.



SBT Executor 41,268
BUILD 4 years ago

Scala 9,065,056
LANGUAGES one month ago

Scala Imports Organizer 52,756
FORMATTING 3 years ago

Scalafmt 83,396
FORMATTING one month ago

Scalagen, Java to Scala cor 9,450
CODE TOOLS 7 months ago

Scalariform 17,324
FORMATTING one year ago

ScaterKnife 510
CODE EDITING one year ago

Sireum v3 Injector 48
TOOLS INTEGRATION 5 months ago

Stack trace to UML 442
DEBUGGING one week ago

Stepik Union 5,321
CODE TOOLS 6 months ago

LANGUAGES

Scala

★★★★☆ 9065056 downloads

Updated 4/23/18 v2018.1.9

The Scala plugin extends IntelliJ IDEA's toolset with support for Scala, sbt, Scala.js, Hocon, and Play Framework. Support for Scala, sbt and Hocon is available for free in IntelliJ IDEA Community Edition, while support for Play Framework and Scala.js is available only in IntelliJ IDEA Ultimate.

Vendor

JetBrains
http://www.jetbrains.com

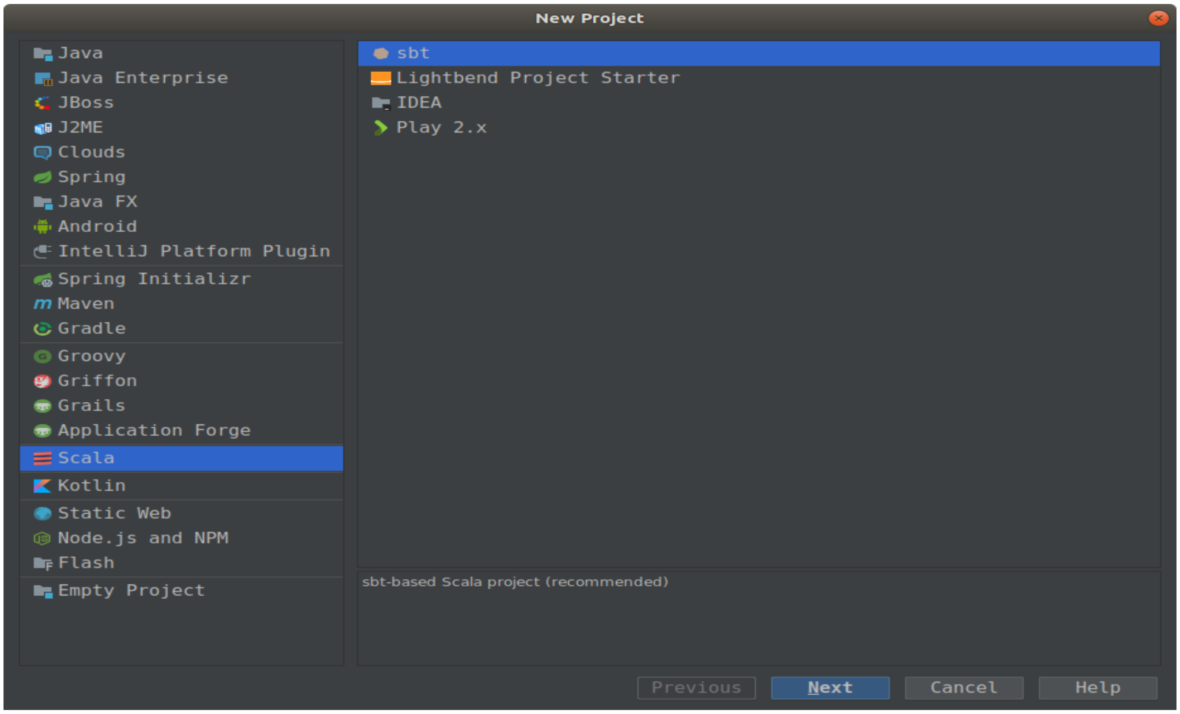
Plugin homepage

http://www.jetbrains.net/confluence/display/SCA/Scala+Plugin+for+IntelliJ+IDEA

Size

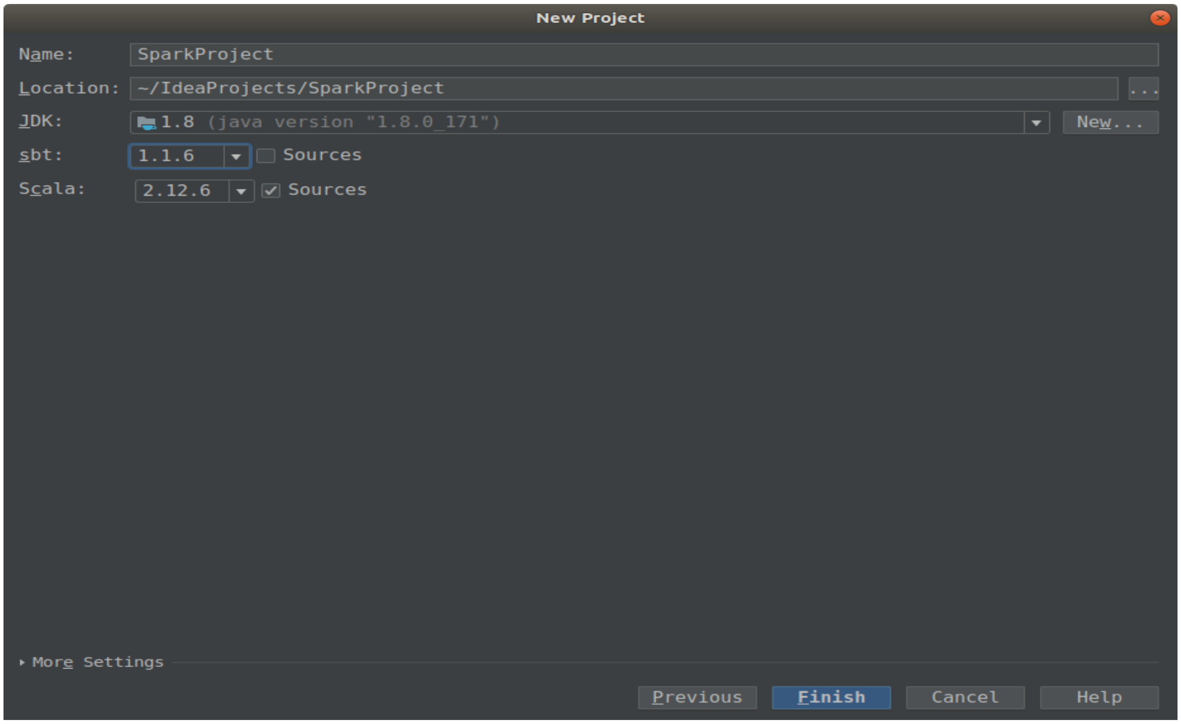
52.5 M

Now, everything is ready to start creating a project. Select Scala on the left menu and sbt on right menu as it is shown here.



sbt-based Scala project (recommended)

Select as it is shown here.



Name: SparkProject

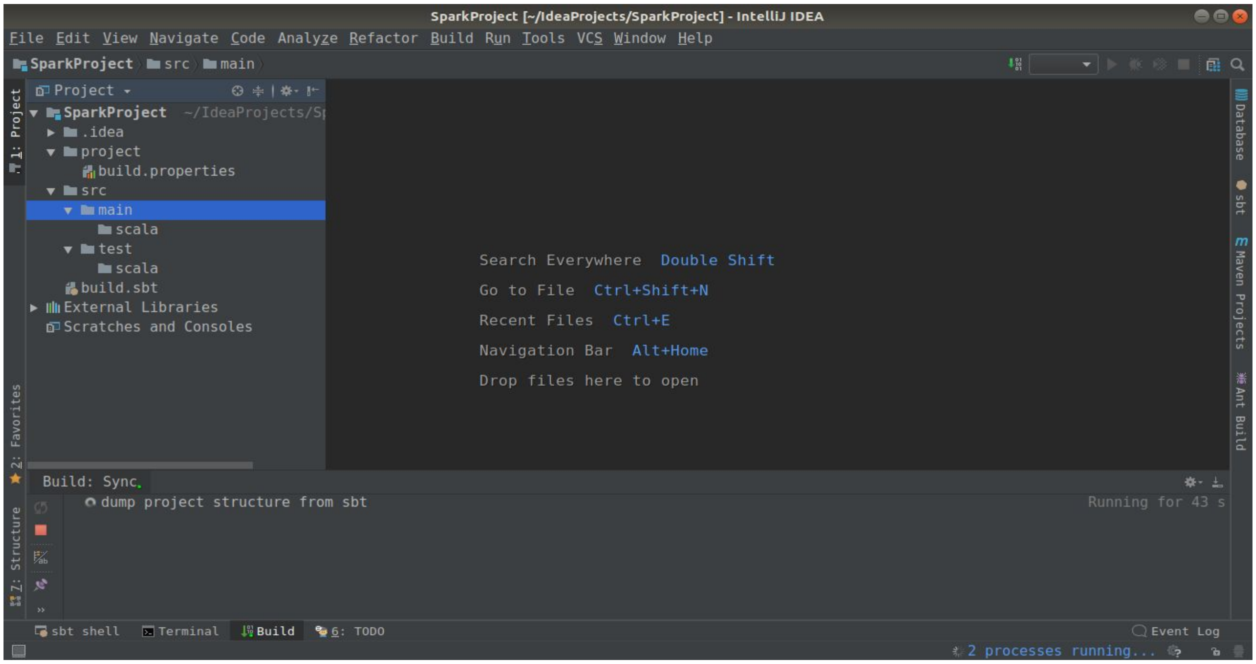
Location: ~/IdeaProjects/SparkProject

JDK: 1.8 (java version "1.8.0_171")

sbt: 1.1.6 □ Sources

Scala: 2.12.6 ☑ Sources

After then, this view shows up. Now, let's create test scala object and run it.

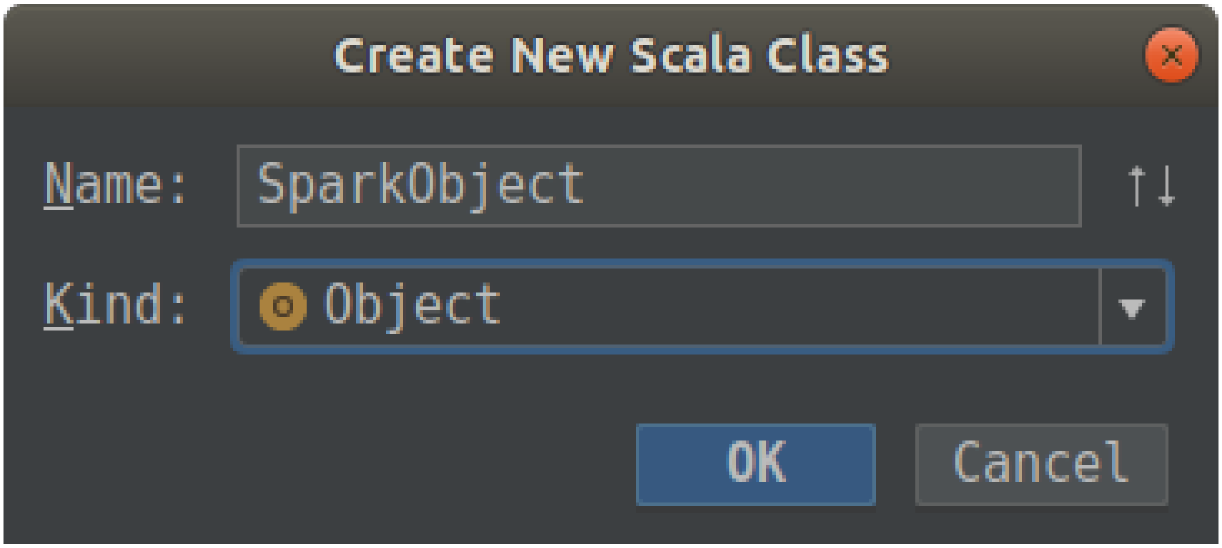


SparkProject [~/IdeaProjects/SparkProject] - IntelliJ IDEA

```
dump project structure from sbt                                   Running for 43 s
```

2 processes running...

Here we need to select to **Object** and click **OK**.

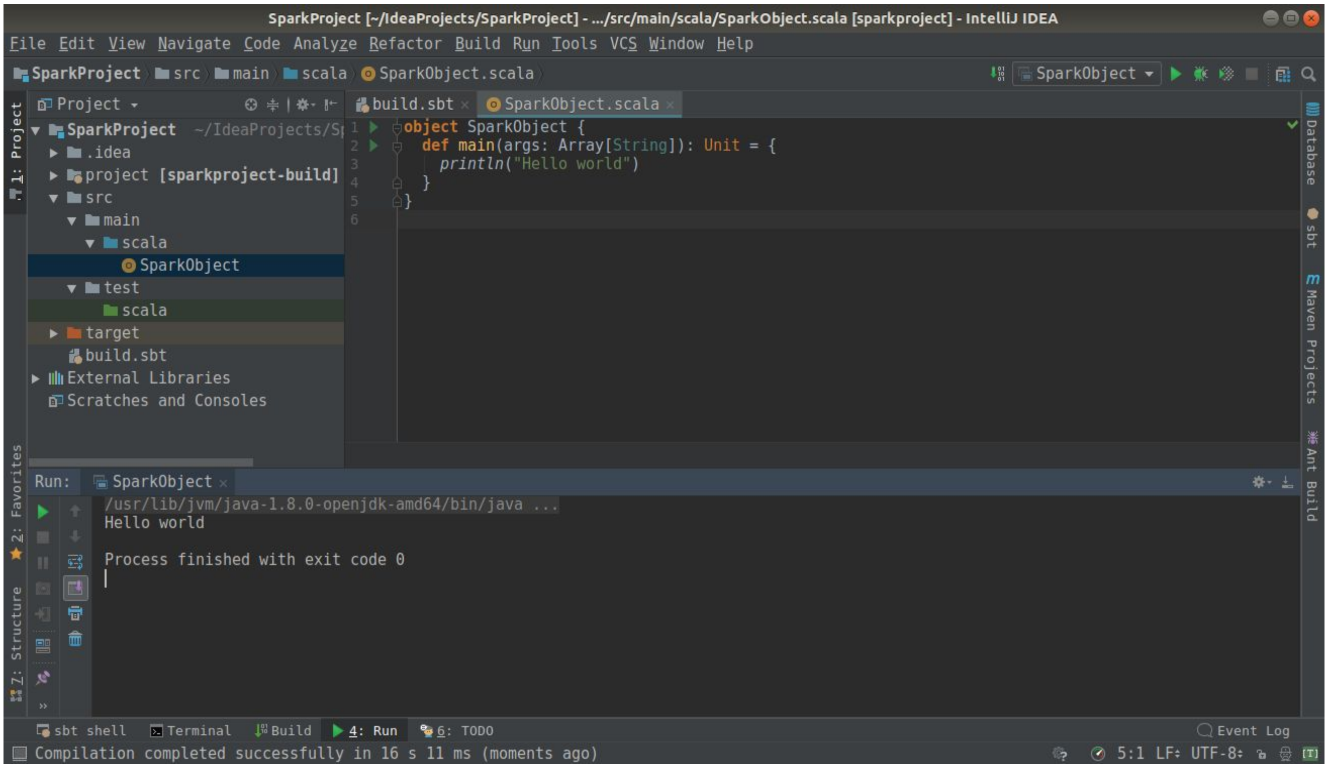


Create New Scala Class

Name: SparkObject

Kind: Object

Here is sample example of Scala project. Work of the **SparkObject** is just printing **Hello World** as it is good example to begin learning a new language -:)). We can see the result from bottom view and it means scala project has successfully worked. The next step is to add **Spark** library to the project and use it.

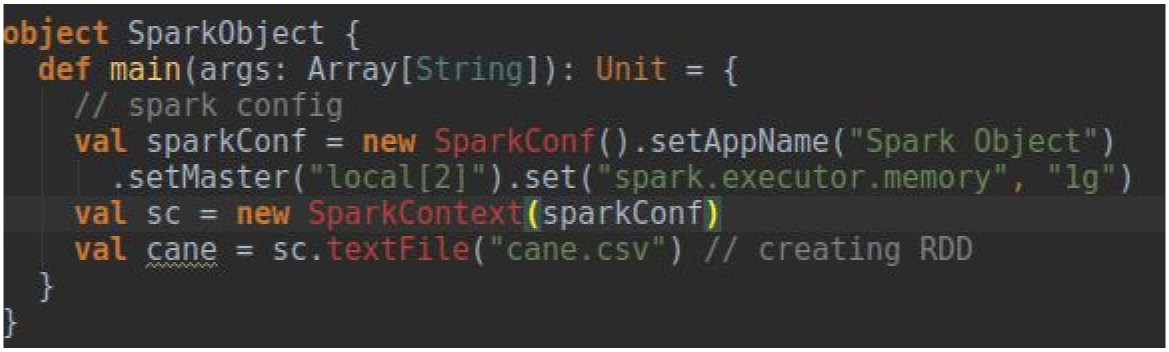

SparkProject [~/IdeaProjects/SparkProject] - .../src/main/scala/SparkObject.scala [sparkproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SparkProject src main scala SparkObject.scala
SparkObject
Project
build.sbt
SparkObject.scala
SparkProject ~/IdeaProjects/Sp
.idea
project [sparkproject-build]
src
main
scala
SparkObject
test
scala
target
build.sbt
External Libraries
Scratches and Consoles
1: Project
2: Favorites
Z: Structure
Database
sbt
Maven Projects
Ant Build
Run:
SparkObject
sbt shell
Terminal
Build
4: Run
6: TODO
Event Log
5:1 LF: UTF-8:

```
object SparkObject {
  def main(args: Array[String]): Unit = {
    println("Hello world")
  }
}
```

```
/usr/lib/jvm/java-1.8.0-openjdk-amd64/bin/java ...
Hello world

Process finished with exit code 0
```

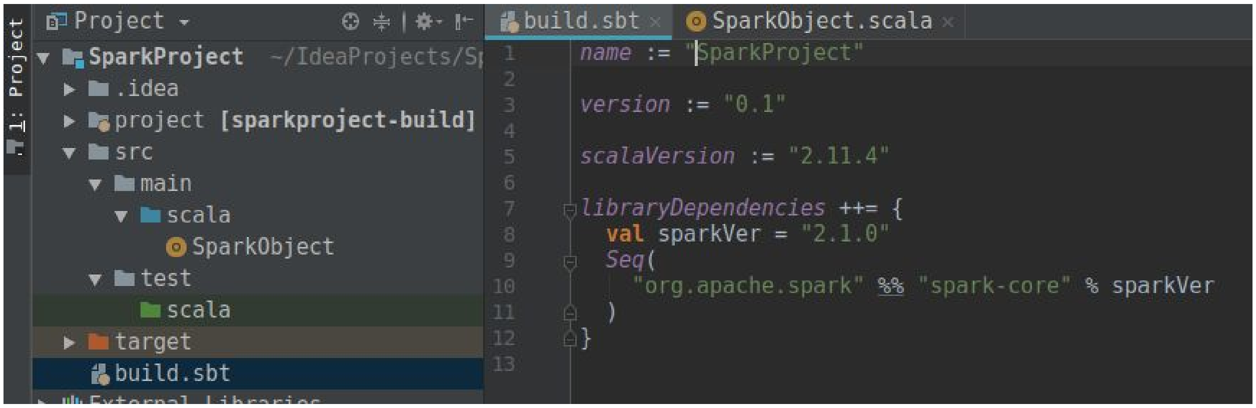
Compilation completed successfully in 16 s 11 ms (moments ago)

Here is sample code of the code of Spark. Spark's keywords are red as you can see. It means the project doesn't support **Spark** now. So let's add Spark library into the project.



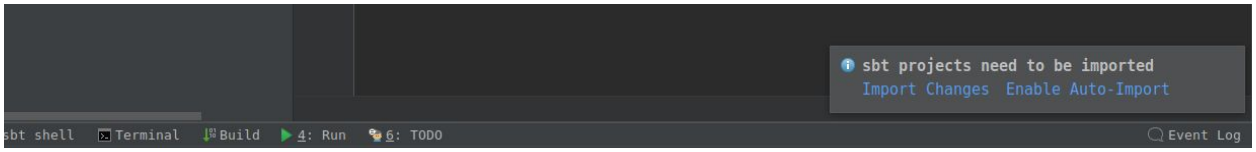
```
object SparkObject {
  def main(args: Array[String]): Unit = {
    // spark config
    val sparkConf = new SparkConf().setAppName("Spark Object")
      .setMaster("local[2]").set("spark.executor.memory", "1g")
    val sc = new SparkContext(sparkConf)
    val cane = sc.textFile("cane.csv") // creating RDD
  }
}
```

We need to open **build.sbt** file to add Spark library. **Build.sbt** stores all the dependencies that are used in the project such as maven or gradle do.



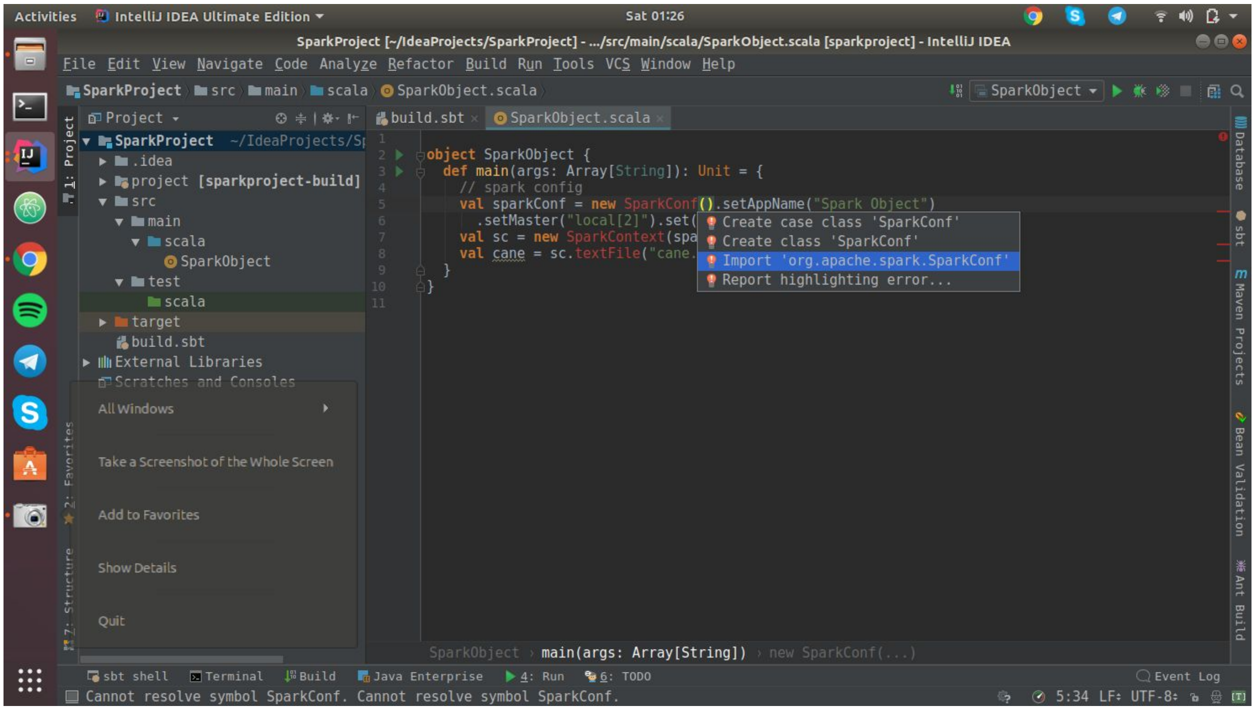
```
name := "SparkProject"

version := "0.1"

scalaVersion := "2.11.4"

libraryDependencies ++= {
  val sparkVer = "2.1.0"
  Seq(
    "org.apache.spark" %% "spark-core" % sparkVer
  )
}
```

Once we add the code of the library, it asks us if we want to import the library. We will choose yes.



sbt projects need to be imported

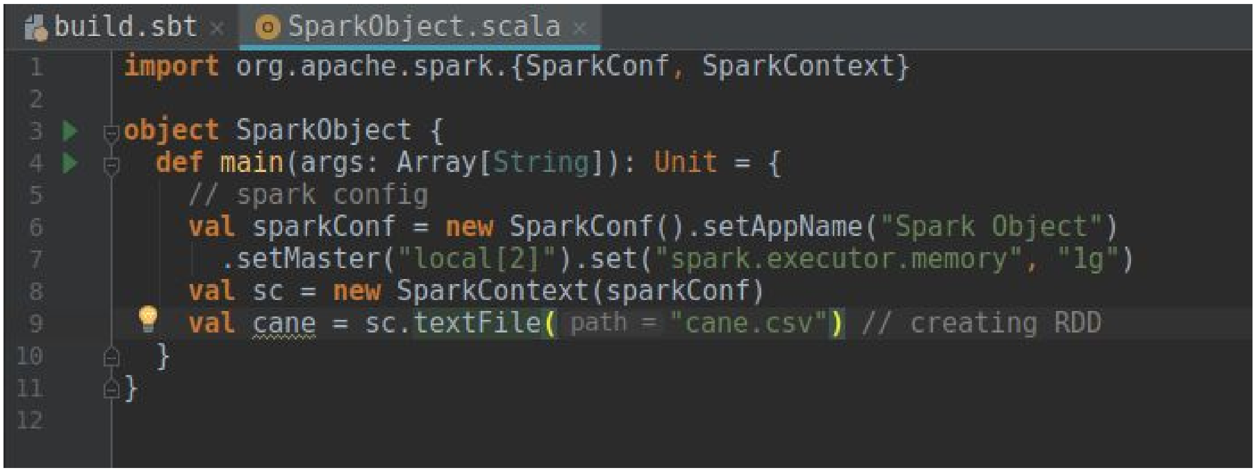
After that we can import Spark codes.


Activities
IntelliJ IDEA Ultimate Edition
Sat 01:26
SparkProject [~/IdeaProjects/SparkProject] - .../src/main/scala/SparkObject.scala [sparkproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SparkProject src main scala SparkObject.scala
SparkObject
Project
SparkProject ~/IdeaProjects/Sp
.idea
project [sparkproject-build]
src
main
scala
SparkObject
test
scala
target
build.sbt
External Libraries
Scratches and Consoles
All Windows
Take a Screenshot of the Whole Screen
Add to Favorites
Show Details
Quit
build.sbt
SparkObject.scala
Create case class 'SparkConf'
Create class 'SparkConf'
Import 'org.apache.spark.SparkConf'
Report highlighting error...
SparkObject › main(args: Array[String]) › new SparkConf(...)
sbt shell
Terminal
Build
Java Enterprise
4: Run
6: TODO
Event Log
5:34 LF: UTF-8:
Database
sbt
Maven Projects
Bean Validation
Ant Build
1: Project
2: Favorites
7: Structure

```scala

object SparkObject {
  def main(args: Array[String]): Unit = {
    // spark config
    val sparkConf = new SparkConf().setAppName("Spark Object")
      .setMaster("local[2]").set(
    val sc = new SparkContext(spa
    val cane = sc.textFile("cane.
  }
}
```

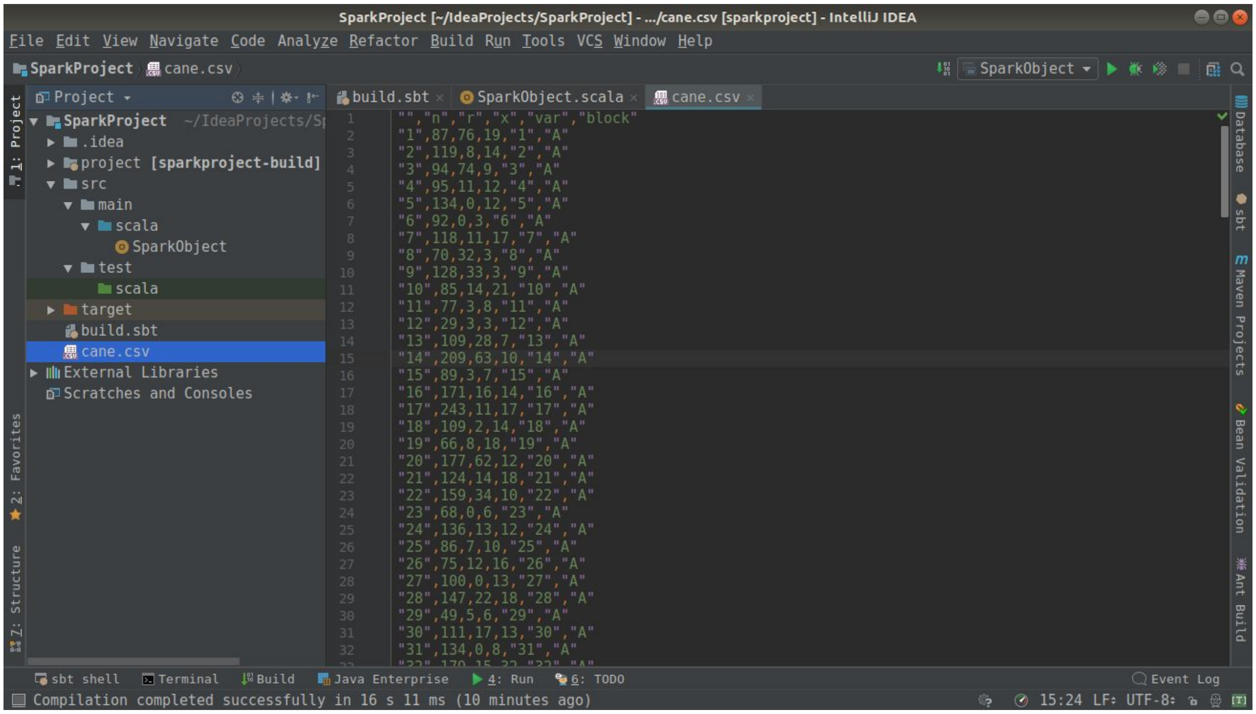
Cannot resolve symbol SparkConf. Cannot resolve symbol SparkConf.

It is the view when we import all the keywords and now we can test the code by running it.



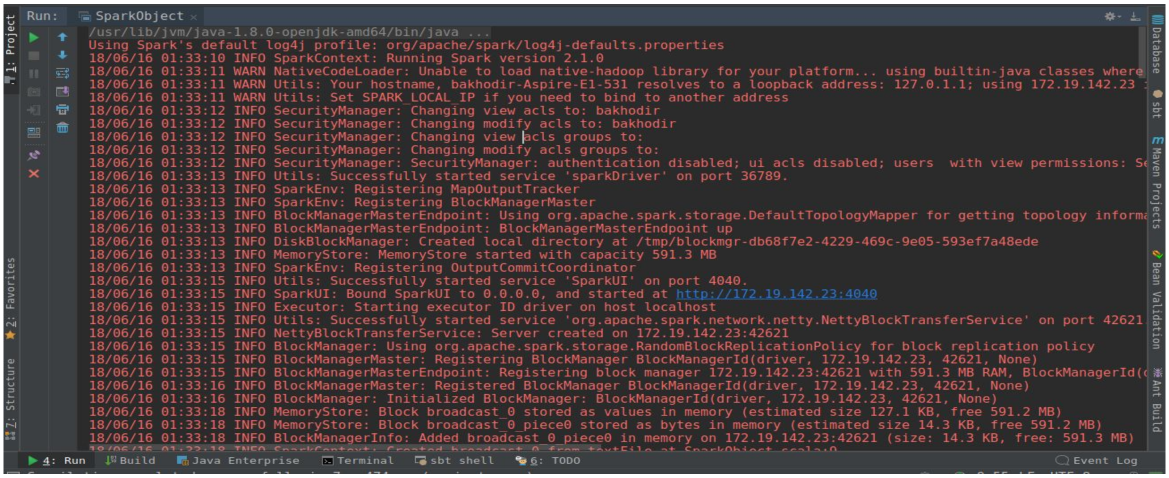
```scala
import org.apache.spark.{SparkConf, SparkContext}

object SparkObject {
  def main(args: Array[String]): Unit = {
    // spark config
    val sparkConf = new SparkConf().setAppName("Spark Object")
      .setMaster("local[2]").set("spark.executor.memory", "1g")
    val sc = new SparkContext(sparkConf)
    val cane = sc.textFile( path = "cane.csv") // creating RDD
  }
}
```

Here is cvs sample file to check the project. We read the file, create a RDD, perform the output result. Let's run it.

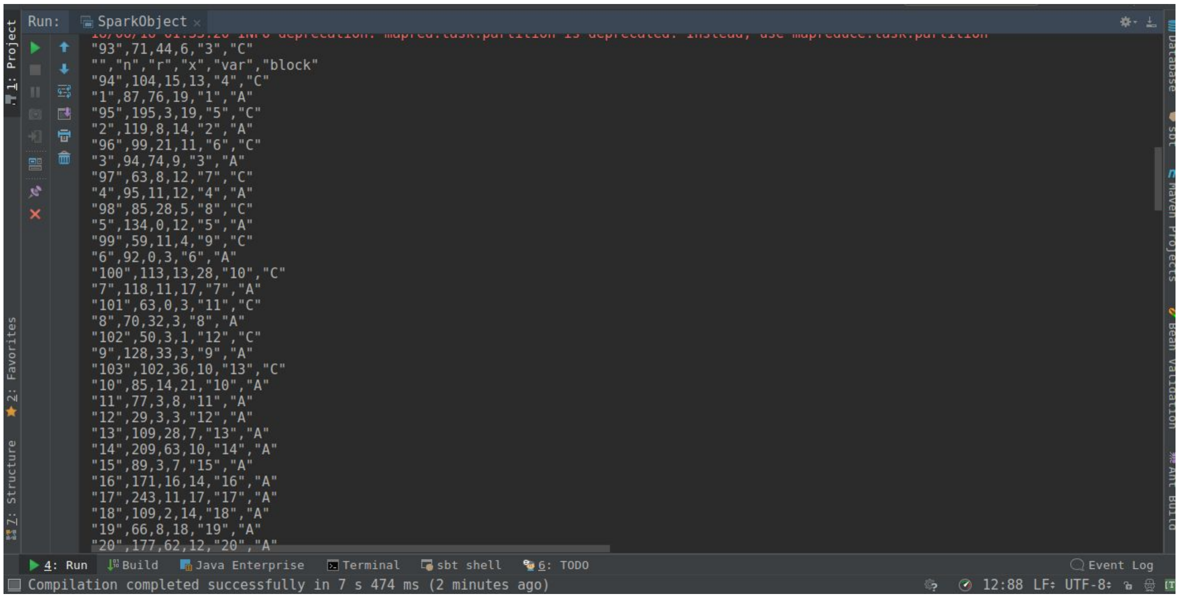


```
"","n","r","x","var","block"
"1",87,76,19,"1","A"
"2",119,8,14,"2","A"
"3",94,74,9,"3","A"
"4",95,11,12,"4","A"
"5",134,0,12,"5","A"
"6",92,0,3,"6","A"
"7",118,11,17,"7","A"
"8",70,32,3,"8","A"
"9",128,33,3,"9","A"
"10",85,14,21,"10","A"
"11",77,3,8,"11","A"
"12",29,3,3,"12","A"
"13",109,28,7,"13","A"
"14",209,63,10,"14","A"
"15",89,3,7,"15","A"
"16",171,16,14,"16","A"
"17",243,11,17,"17","A"
"18",109,2,14,"18","A"
"19",66,8,18,"19","A"
"20",177,62,12,"20","A"
"21",124,14,18,"21","A"
"22",159,34,10,"22","A"
"23",68,0,6,"23","A"
"24",136,13,12,"24","A"
"25",86,7,10,"25","A"
"26",75,12,16,"26","A"
"27",100,0,13,"27","A"
"28",147,22,18,"28","A"
"29",49,5,6,"29","A"
"30",111,17,13,"30","A"
"31",134,0,8,"31","A"
"32",170,15,32,"32","A"
```

Compilation completed successfully in 16 s 11 ms (10 minutes ago)

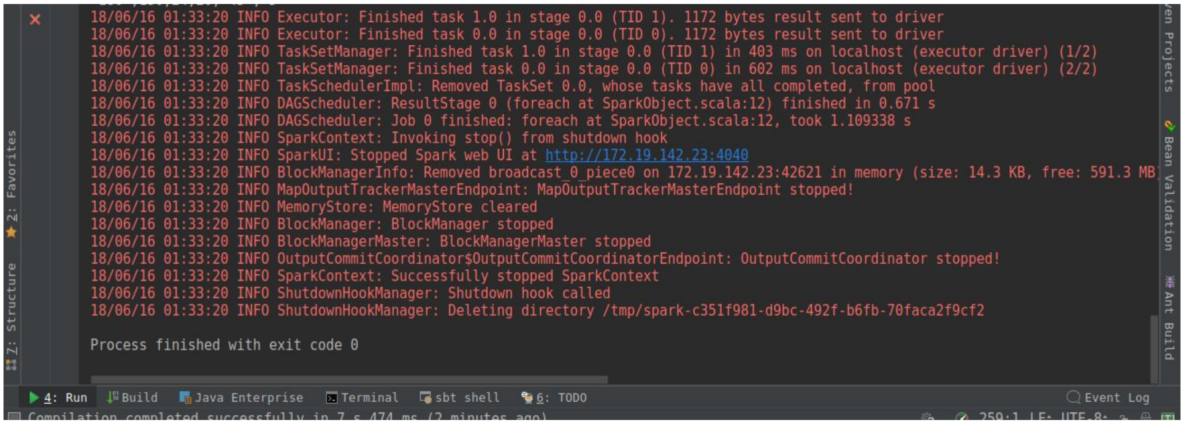


```
/usr/lib/jvm/java-1.8.0-openjdk-amd64/bin/java ...
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
18/06/16 01:33:10 INFO SparkContext: Running Spark version 2.1.0
18/06/16 01:33:11 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
18/06/16 01:33:11 WARN Utils: Your hostname, bakhodir-Aspire-E1-531 resolves to a loopback address: 127.0.1.1; using 172.19.142.23 i
18/06/16 01:33:11 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/06/16 01:33:12 INFO SecurityManager: Changing view acls to: bakhodir
18/06/16 01:33:12 INFO SecurityManager: Changing modify acls to: bakhodir
18/06/16 01:33:12 INFO SecurityManager: Changing view acls groups to:
18/06/16 01:33:12 INFO SecurityManager: Changing modify acls groups to:
18/06/16 01:33:12 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Se
18/06/16 01:33:13 INFO Utils: Successfully started service 'sparkDriver' on port 36789.
18/06/16 01:33:13 INFO SparkEnv: Registering MapOutputTracker
18/06/16 01:33:13 INFO SparkEnv: Registering BlockManagerMaster
18/06/16 01:33:13 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology informa
18/06/16 01:33:13 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
18/06/16 01:33:13 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-db68f7e2-4229-469c-9e05-593ef7a48ede
18/06/16 01:33:13 INFO MemoryStore: MemoryStore started with capacity 591.3 MB
18/06/16 01:33:13 INFO SparkEnv: Registering OutputCommitCoordinator
18/06/16 01:33:15 INFO Utils: Successfully started service 'SparkUI' on port 4040.
18/06/16 01:33:15 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://172.19.142.23:4040
18/06/16 01:33:15 INFO Executor: Starting executor ID driver on host localhost
18/06/16 01:33:15 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 42621.
18/06/16 01:33:15 INFO NettyBlockTransferService: Server created on 172.19.142.23:42621
18/06/16 01:33:15 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
18/06/16 01:33:15 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 172.19.142.23, 42621, None)
18/06/16 01:33:15 INFO BlockManagerMasterEndpoint: Registering block manager 172.19.142.23:42621 with 591.3 MB RAM, BlockManagerId(d
18/06/16 01:33:16 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 172.19.142.23, 42621, None)
18/06/16 01:33:16 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 172.19.142.23, 42621, None)
18/06/16 01:33:18 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 127.1 KB, free 591.2 MB)
18/06/16 01:33:18 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 14.3 KB, free 591.2 MB)
18/06/16 01:33:18 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 172.19.142.23:42621 (size: 14.3 KB, free: 591.3 MB)
18/06/16 01:33:18 INFO SparkContext: Created broadcast 0 from textFile at SparkObject.scala:9
```



```
18/06/16 01:33:20 INFO deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
"93",71,44,6,"3","C"
"","n","r","x","var","block"
"94",104,15,13,"4","C"
"1",87,76,19,"1","A"
"95",195,3,19,"5","C"
"2",119,8,14,"2","A"
"96",99,21,11,"6","C"
"3",94,74,9,"3","A"
"97",63,8,12,"7","C"
"4",95,11,12,"4","A"
"98",85,28,5,"8","C"
"5",134,0,12,"5","A"
"99",59,11,4,"9","C"
"6",92,0,3,"6","A"
"100",113,13,28,"10","C"
"7",118,11,17,"7","A"
"101",63,0,3,"11","C"
"8",70,32,3,"8","A"
"102",50,3,1,"12","C"
"9",128,33,3,"9","A"
"103",102,36,10,"13","C"
"10",85,14,21,"10","A"
"11",77,3,8,"11","A"
"12",29,3,3,"12","A"
"13",109,28,7,"13","A"
"14",209,63,10,"14","A"
"15",89,3,7,"15","A"
"16",171,16,14,"16","A"
"17",243,11,17,"17","A"
"18",109,2,14,"18","A"
"19",66,8,18,"19","A"
"20",177,62,12,"20","A"
```

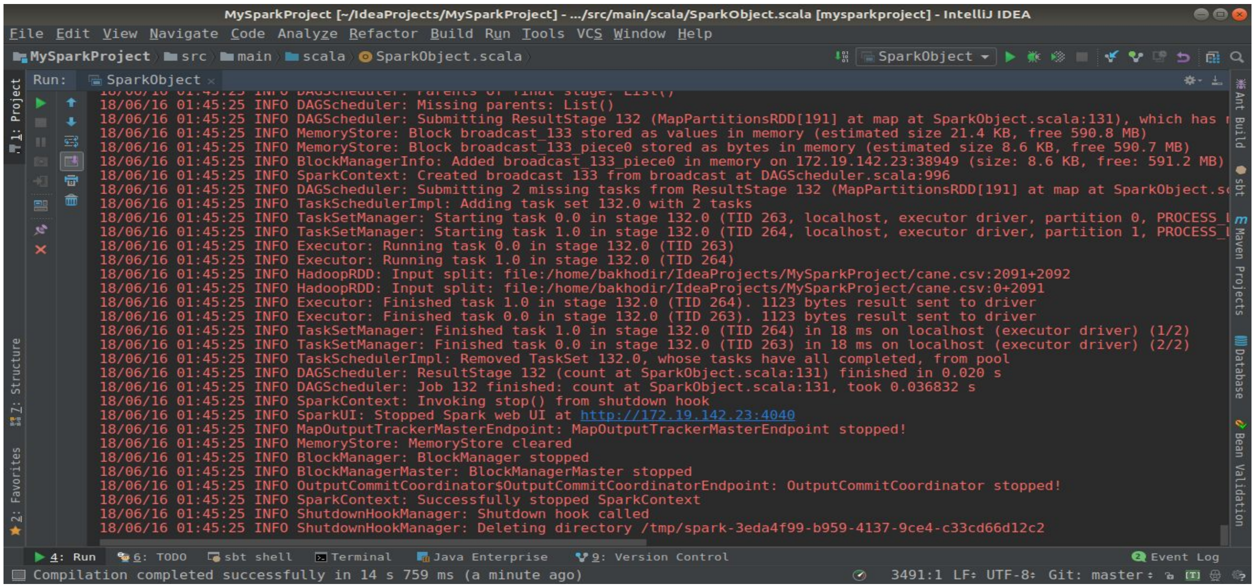
Compilation completed successfully in 7 s 474 ms (2 minutes ago)



```
18/06/16 01:33:20 INFO Executor: Finished task 1.0 in stage 0.0 (TID 1). 1172 bytes result sent to driver
18/06/16 01:33:20 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1172 bytes result sent to driver
18/06/16 01:33:20 INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 403 ms on localhost (executor driver) (1/2)
18/06/16 01:33:20 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 602 ms on localhost (executor driver) (2/2)
18/06/16 01:33:20 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
18/06/16 01:33:20 INFO DAGScheduler: ResultStage 0 (foreach at SparkObject.scala:12) finished in 0.671 s
18/06/16 01:33:20 INFO DAGScheduler: Job 0 finished: foreach at SparkObject.scala:12, took 1.109338 s
18/06/16 01:33:20 INFO SparkContext: Invoking stop() from shutdown hook
18/06/16 01:33:20 INFO SparkUI: Stopped Spark web UI at http://172.19.142.23:4040
18/06/16 01:33:20 INFO BlockManagerInfo: Removed broadcast_0_piece0 on 172.19.142.23:42621 in memory (size: 14.3 KB, free: 591.3 MB
18/06/16 01:33:20 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/06/16 01:33:20 INFO MemoryStore: MemoryStore cleared
18/06/16 01:33:20 INFO BlockManager: BlockManager stopped
18/06/16 01:33:20 INFO BlockManagerMaster: BlockManagerMaster stopped
18/06/16 01:33:20 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/06/16 01:33:20 INFO SparkContext: Successfully stopped SparkContext
18/06/16 01:33:20 INFO ShutdownHookManager: Shutdown hook called
18/06/16 01:33:20 INFO ShutdownHookManager: Deleting directory /tmp/spark-c351f981-d9bc-492f-b6fb-70faca2f9cf2

Process finished with exit code 0
```

Compilation completed successfully in 7 s 474 ms (2 minutes ago)

Everything is fine. We can do all the project requirements that is given. After being done, let's run the project and see the result.

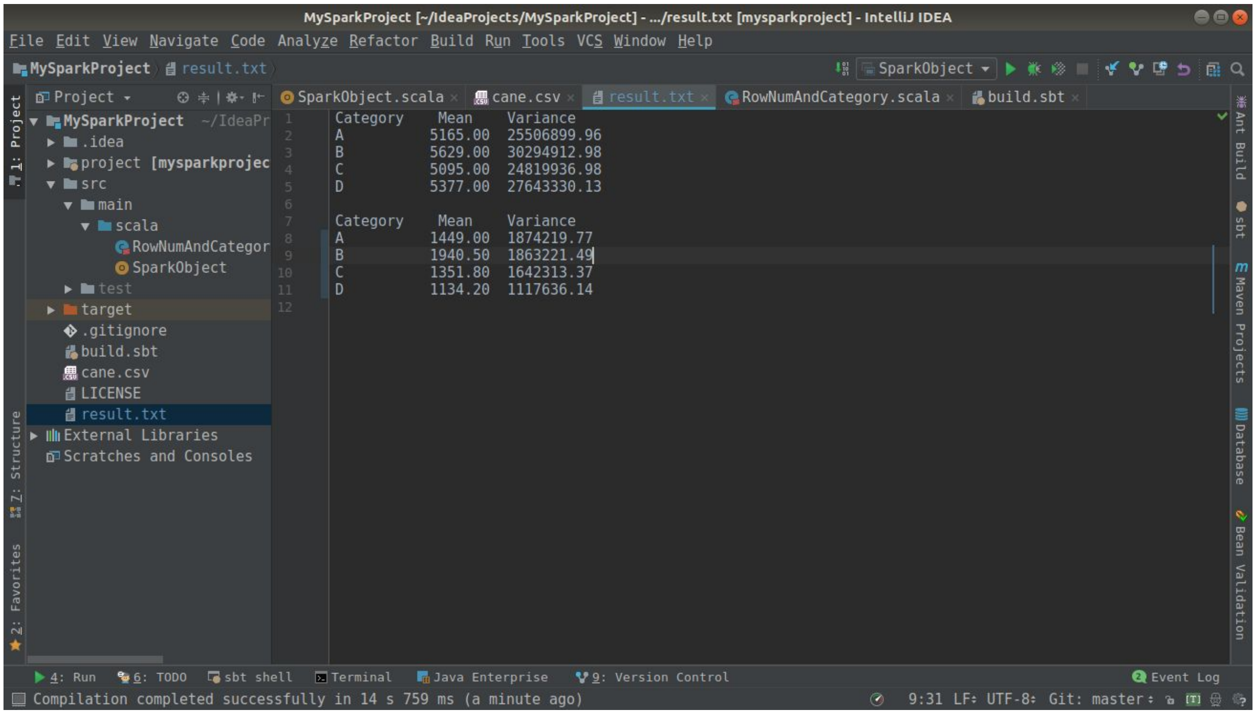


MySparkProject [~/IdeaProjects/MySparkProject] - .../src/main/scala/SparkObject.scala [mysparkproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
MySparkProject src main scala SparkObject.scala
SparkObject
Run: SparkObject
4: Run 6: TODO sbt shell Terminal Java Enterprise 9: Version Control
Event Log

```
18/06/16 01:45:25 INFO DAGScheduler: Parents of final stage: List()
18/06/16 01:45:25 INFO DAGScheduler: Missing parents: List()
18/06/16 01:45:25 INFO DAGScheduler: Submitting ResultStage 132 (MapPartitionsRDD[191] at map at SparkObject.scala:131), which has r
18/06/16 01:45:25 INFO MemoryStore: Block broadcast_133 stored as values in memory (estimated size 21.4 KB, free 590.8 MB)
18/06/16 01:45:25 INFO MemoryStore: Block broadcast_133_piece0 stored as bytes in memory (estimated size 8.6 KB, free 590.7 MB)
18/06/16 01:45:25 INFO BlockManagerInfo: Added broadcast_133_piece0 in memory on 172.19.142.23:38949 (size: 8.6 KB, free: 591.2 MB)
18/06/16 01:45:25 INFO SparkContext: Created broadcast 133 from broadcast at DAGScheduler.scala:996
18/06/16 01:45:25 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 132 (MapPartitionsRDD[191] at map at SparkObject.sc
18/06/16 01:45:25 INFO TaskSchedulerImpl: Adding task set 132.0 with 2 tasks
18/06/16 01:45:25 INFO TaskSetManager: Starting task 0.0 in stage 132.0 (TID 263, localhost, executor driver, partition 0, PROCESS_L
18/06/16 01:45:25 INFO TaskSetManager: Starting task 1.0 in stage 132.0 (TID 264, localhost, executor driver, partition 1, PROCESS_L
18/06/16 01:45:25 INFO Executor: Running task 0.0 in stage 132.0 (TID 263)
18/06/16 01:45:25 INFO Executor: Running task 1.0 in stage 132.0 (TID 264)
18/06/16 01:45:25 INFO HadoopRDD: Input split: file:/home/bakhodir/IdeaProjects/MySparkProject/cane.csv:2091+2092
18/06/16 01:45:25 INFO HadoopRDD: Input split: file:/home/bakhodir/IdeaProjects/MySparkProject/cane.csv:0+2091
18/06/16 01:45:25 INFO Executor: Finished task 1.0 in stage 132.0 (TID 264). 1123 bytes result sent to driver
18/06/16 01:45:25 INFO Executor: Finished task 0.0 in stage 132.0 (TID 263). 1123 bytes result sent to driver
18/06/16 01:45:25 INFO TaskSetManager: Finished task 1.0 in stage 132.0 (TID 264) in 18 ms on localhost (executor driver) (1/2)
18/06/16 01:45:25 INFO TaskSetManager: Finished task 0.0 in stage 132.0 (TID 263) in 18 ms on localhost (executor driver) (2/2)
18/06/16 01:45:25 INFO TaskSchedulerImpl: Removed TaskSet 132.0, whose tasks have all completed, from pool
18/06/16 01:45:25 INFO DAGScheduler: ResultStage 132 (count at SparkObject.scala:131) finished in 0.020 s
18/06/16 01:45:25 INFO DAGScheduler: Job 132 finished: count at SparkObject.scala:131, took 0.036832 s
18/06/16 01:45:25 INFO SparkContext: Invoking stop() from shutdown hook
18/06/16 01:45:25 INFO SparkUI: Stopped Spark web UI at http://172.19.142.23:4040
18/06/16 01:45:25 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/06/16 01:45:25 INFO MemoryStore: MemoryStore cleared
18/06/16 01:45:25 INFO BlockManager: BlockManager stopped
18/06/16 01:45:25 INFO BlockManagerMaster: BlockManagerMaster stopped
18/06/16 01:45:25 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/06/16 01:45:25 INFO SparkContext: Successfully stopped SparkContext
18/06/16 01:45:25 INFO ShutdownHookManager: Shutdown hook called
18/06/16 01:45:25 INFO ShutdownHookManager: Deleting directory /tmp/spark-3eda4f99-b959-4137-9ce4-c33cd66d12c2
```

```
Compilation completed successfully in 14 s 759 ms (a minute ago)        3491:1  LF:  UTF-8:  Git: master:
```

And here is the result file.



```
Category    Mean     Variance
A           5165.00  25506899.96
B           5629.00  30294912.98
C           5095.00  24819936.98
D           5377.00  27643330.13

Category    Mean     Variance
A           1449.00  1874219.77
B           1940.50  1863221.49
C           1351.80  1642313.37
D           1134.20  1117636.14
```

```
Compilation completed successfully in 14 s 759 ms (a minute ago)        9:31  LF:  UTF-8:  Git: master:
```

Conclusion

From doing the project I have learnt a lot of knowledge that are about Scala language, how it works briefly, what is Apache Spark open-source cluster-computing framework and integrating it with Scala. At first, I had a problem with importing Spark library inside Scala project and it took me much time and finally I realized that the problem was library version compatibility of Scala and Spark framework. It was really good experience doing the project and learnt new things.

Here, I have also provided github url to go my project. You can copy and work with it easily.
https://github.com/bakhodir10/Spark-Scala