



```
addSbtPlugin("com.eed319" % "sbt-assembly" % "0.14.10")
```

```
import org.apache.spark.mllib.recommendation.{ALS, MatrixFactorizationModel}
import org.apache.spark.rdd.RDD
import org.apache.spark.sql.{SparkSession}

object alsrecommendation {
  def main(args: Array[String]): Unit = {
    val spark: SparkSession = SparkSession
      .builder()
      .appName("ALS Movie Recommendation")
      .getOrCreate()

    spark.sparkContext.setLogLevel("WARN")
  }
}
```

```
import org.apache.spark.sql.{SparkSession}

object moviesimilarity {
  // Similarity Measures

  def main(args: Array[String]): Unit = {
    val spark: SparkSession = SparkSession
      .builder()
      .appName("Movie Similarity")
      .master("local[*]")
      .config("spark.sql.warehouse.dir", "C:\\sql")
      .getOrCreate()

    spark.sparkContext.setLogLevel("WARN")
  }
}
```

Datasets: ratings.csv & movies.csv

```
name := "spark_movie_recommend"
version := "0.1"
scalaVersion := "2.11.12"

// https://mvnrepository.com/artifact/org.apache.spark/spark-core
libraryDependencies += "org.apache.spark" %% "spark-core" % "2.4.6" % "provided"

// https://mvnrepository.com/artifact/org.apache.spark/spark-sql
libraryDependencies += "org.apache.spark" %% "spark-sql" % "2.4.6" % "provided"

// https://mvnrepository.com/artifact/org.apache.spark/spark-mllib
libraryDependencies += "org.apache.spark" %% "spark-mllib" % "2.4.6" % "compile"

mainClass in (assembly) := Some("alsrecommendation")
assemblyJarName in assembly := "spark_movie_recommend.jar"
```

```
assemblyMergeStrategy in assembly := {
  case PathList("META-INF", xs @ _*) =>
    xs map (_.toLowerCase) match {
      case "manifest.mf" :: Nil | "index.list" :: Nil | "dependencies" :: Nil =>
        MergeStrategy.discard
      case ps @ _ :: xs if ps.last.endsWith(".sf") || ps.last.endsWith(".dsa") =>
        MergeStrategy.discard
      case "plexus" :: xs =>
        MergeStrategy.discard
      case "services" :: xs =>
        MergeStrategy.filterDistinctLines
      case "spring.schemas" :: Nil | "spring.handlers" :: Nil =>
        MergeStrategy.filterDistinctLines
      case _ => MergeStrategy.first
    }
  case "application.conf" => MergeStrategy.concat
  case "reference.conf" => MergeStrategy.concat
  case _ => MergeStrategy.first
}
```



5

Generation

Name: My cluster

Logging: ☒ Logging

S3 folder: s3://aws-logs-704321693222-us-east-2/elasticmapreduce

Launch mode: ☒ Cluster ☐ Step execution

Software configuration

Release label: emr-5.31.0

Applications: ☒ Core Hadoop: Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

☐ HBase: HBase 1.4.13, Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Phoenix 4.14.3, and ZooKeeper 3.4.14

☐ Presto: Presto 0.230.3 with Hadoop 2.10.0 HDFS and Hive 2.3.7 Metastore

☐ Spark: Spark 2.4.6 on Hadoop 2.10.0 YARN and Zeppelin 0.8.2

☐ Use AWS Glue Data Catalog for table metadata

Hardware configuration

Instance type: m5.xlarge

Number of instances: 4 (1 master and 3 core nodes)

Cluster scaling: ☐ scale cluster nodes based on workload

Security and access

EC2 key pair: Choose an option

Permissions: ☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: EMR_DefaultRole

EC2 instance profile: EMR_EC2_DefaultRole

Cancel Create cluster

6

Create bucket

General configuration

Bucket name: mybucket

Bucket name must be unique and must not contain spaces or uppercase letters. See rules for bucket naming

Region: US East (Ohio) us-east-2

Copy settings from existing bucket - optional

Only the bucket settings in the following configuration are copied.

Choose bucket

hadoop

All Applications

Application History

ID	User	Name	Application Type	Queue	Application Priority	Start Time	Finish Time	State	Final Status	Progress	Tracking UI
application_1603672024603_0001	hadoop	ALS Movie Recommendation	SPARK	default	0	Tue Nov 17 22:41:57 2020	Tue Nov 17 00:00 2020	FINISHED	SUCCEEDED		Legacy

Showing 1 to 1 of 1 entries (filtered from 7 total entries)

Title	Genre	Year	Rating
Forrest Gump (1994)	Comedy	1994	8.8
Shogun: The Beginning (1980)	Drama	1980	8.8
Pat Paterson (1994)	Drama	1994	8.8
Wilson of the Woods (1991)	Drama	1991	8.8
Matrix, The (1999)	Action	1999	8.8
Star Wars: Episode IV - A New Hope (1977)	Adventure	1977	8.8
Schindler's List (1993)	Drama	1993	8.8
Braveheart (1995)	Adventure	1995	8.8
Trainspotting (1996)	Comedy	1996	8.8
Terminator 2: Judgment Day (1991)	Action	1991	8.8
Star Wars: Episode V - The Empire Strikes Back (1980)	Adventure	1980	8.8
Tomb Raider (1996)	Action	1996	8.8
Lord of the Rings: The Fellowship of the Ring, The (2001)	Adventure	2001	8.8
Unleashed, The (1995)	Drama	1995	8.8
Star Wars: Episode VI - Return of the Jedi (1983)	Adventure	1983	8.8
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)	Adventure	1981	8.8
American Beauty (1999)	Drama	1999	8.8
Godfather, The (1972)	Drama	1972	8.8
Lord of the Rings: The Two Towers, The (2002)	Adventure	2002	8.8

Rating: (UserID, MovieID, Rating)

Rating (610, 183947, 4.472530826402922)

Rating (610, 151989, 4.455056694885744)

Rating (610, 202231, 4.388460748498632)

Rating (610, 144202, 4.382982243183964)

Rating (610, 165559, 4.315681612768971)