Chapter 1: Predict the Class of a Flower from the Iris Dataset

#	Observations related to flower species:	Number of observations
1	**Iris-setosa**	50
2	**Iris-virginica**	50
3	**Iris-versicolor**	50
Total number of observations :---------- 150		

Feature Table	
Feature # 1	**Sepal Length**
Feature # 2	**Sepal Width**
Feature # 3	**Petal Length**
Feature # 4	**Petal Width**

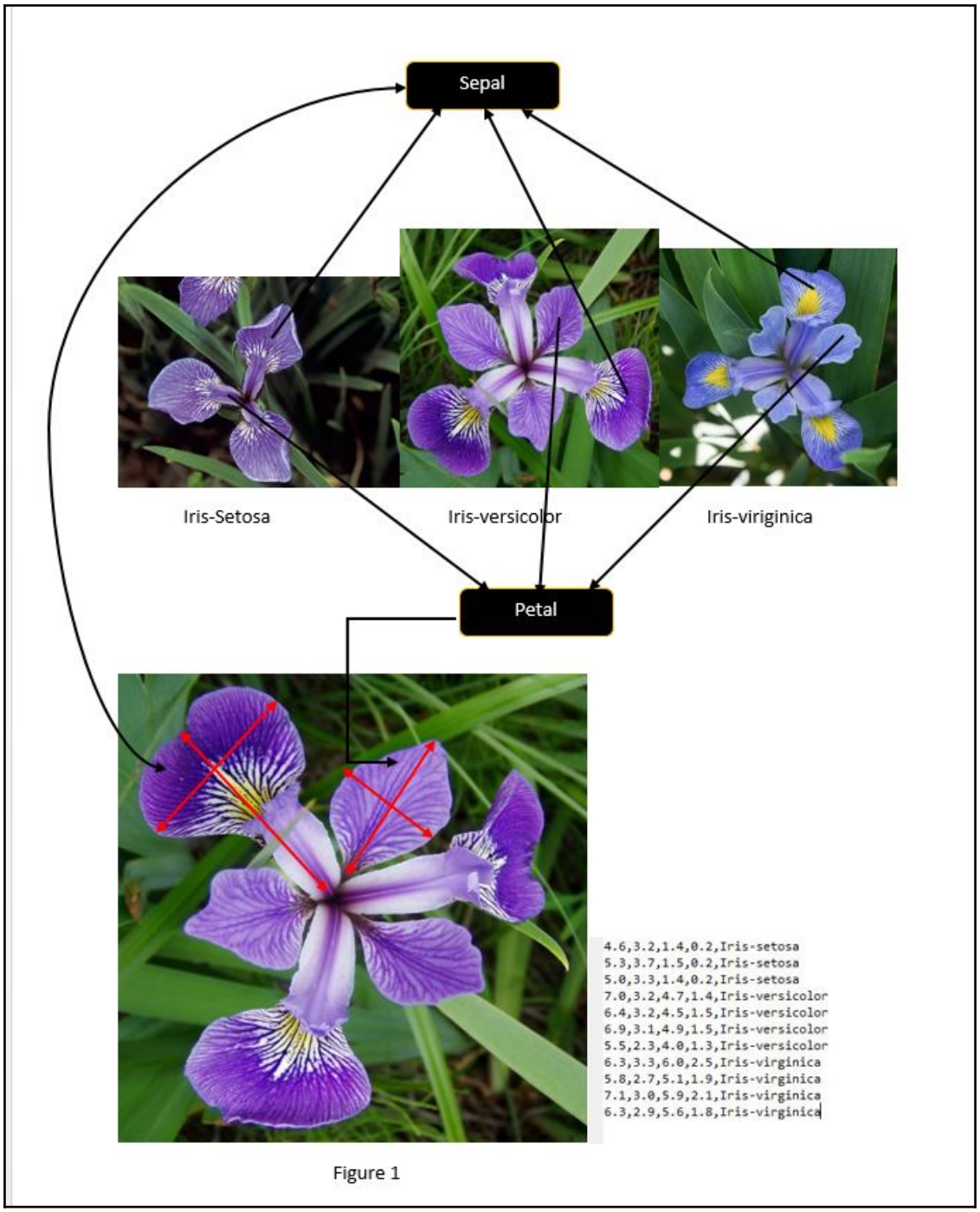
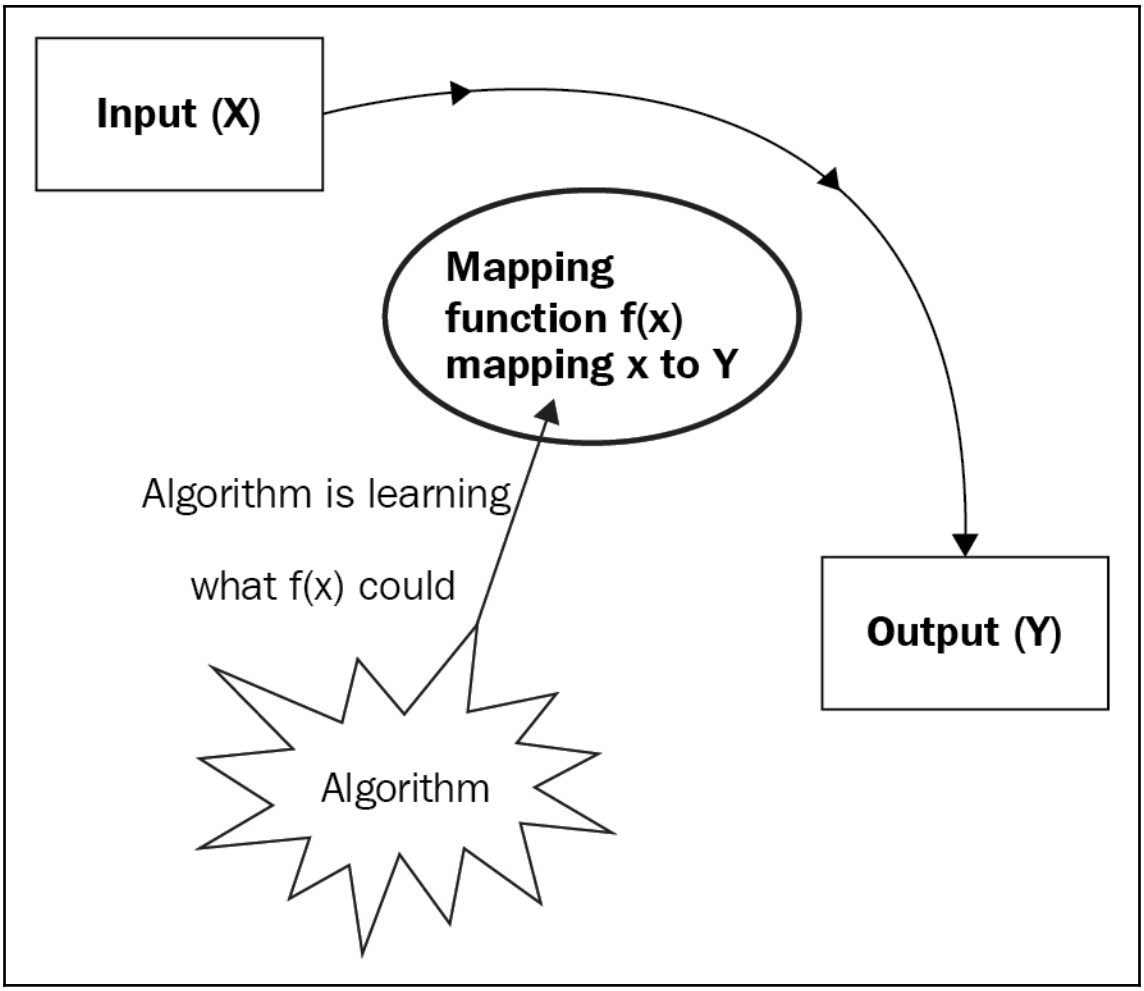

Sepal
Iris-Setosa
Iris-versicolor
Iris-viriginica
Petal
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica

Figure 1



Observed Features X:

Sepal Length	Petal Length	Sepal Width	Petal Width

Category Labels Y:

Iris Setosa	Iris Versicolor	Iris Virginica

High-Level Formulation of the Iris Classification Problem

Using spark's default log4j profile: org/apache/spark/10g4j-defaults.properties

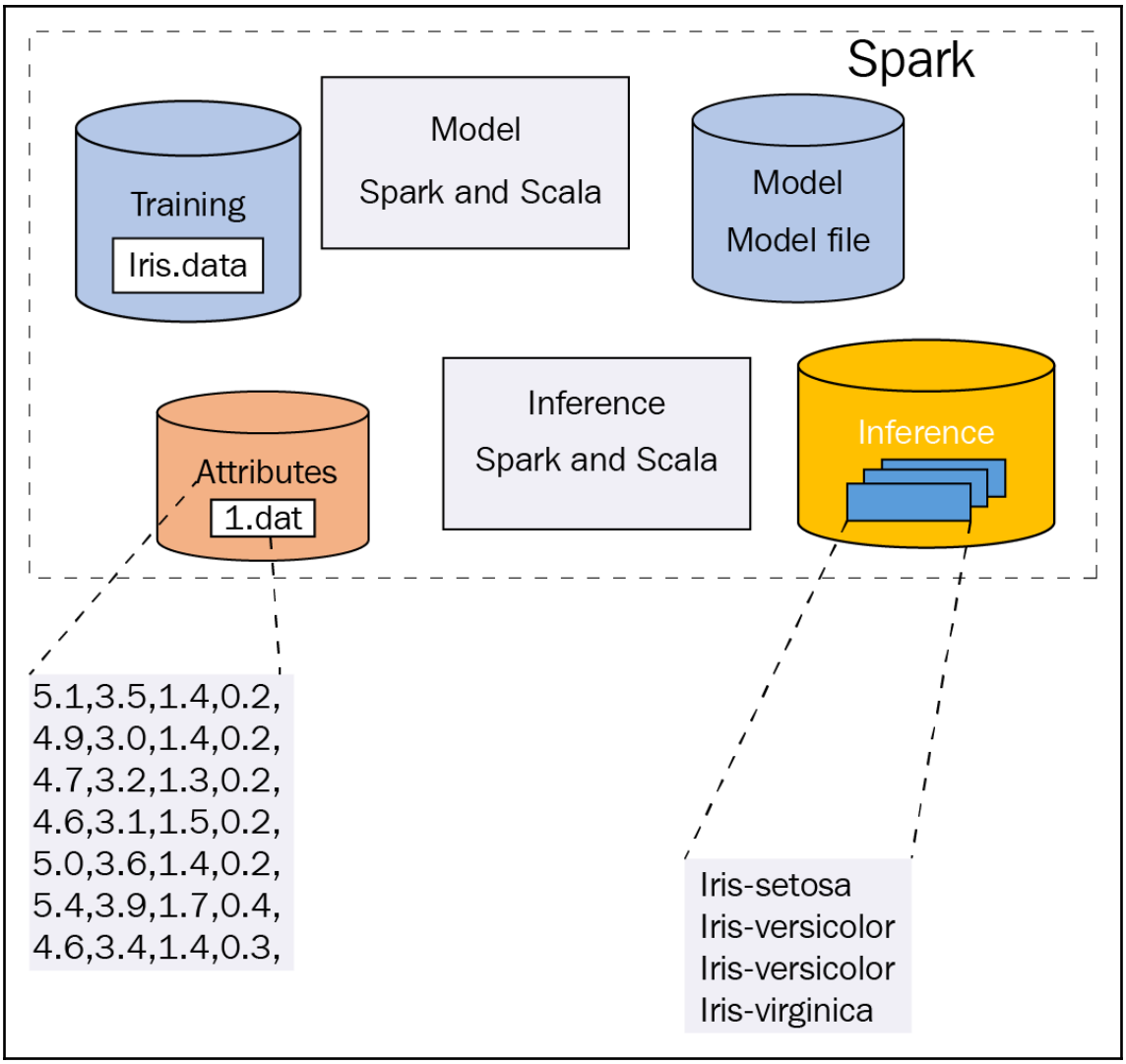
------- default log level to "WARN".

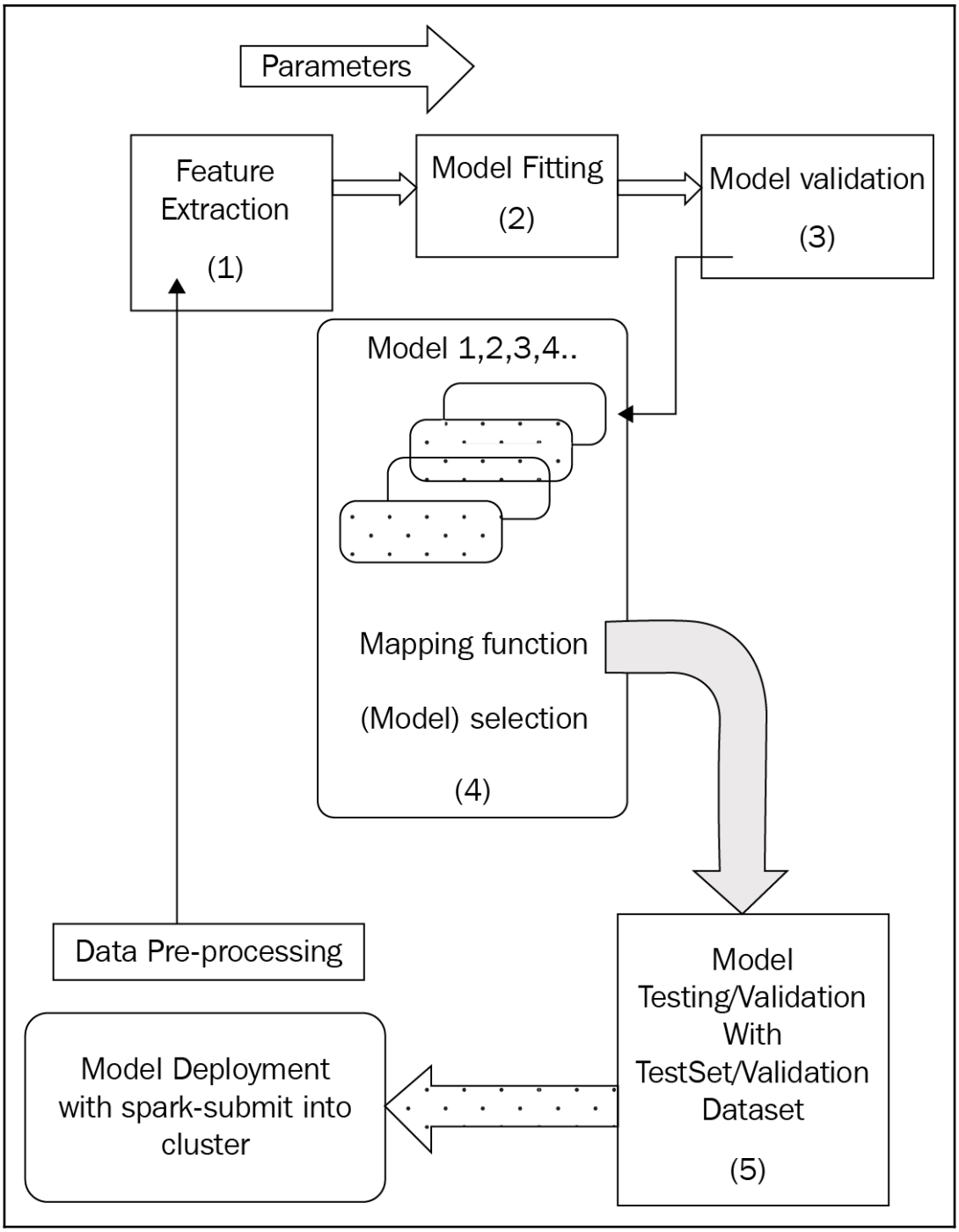
18/01/05 12:01:34 ERROR Shell: Failed to locate the Winutils binary in the hadoop binary path

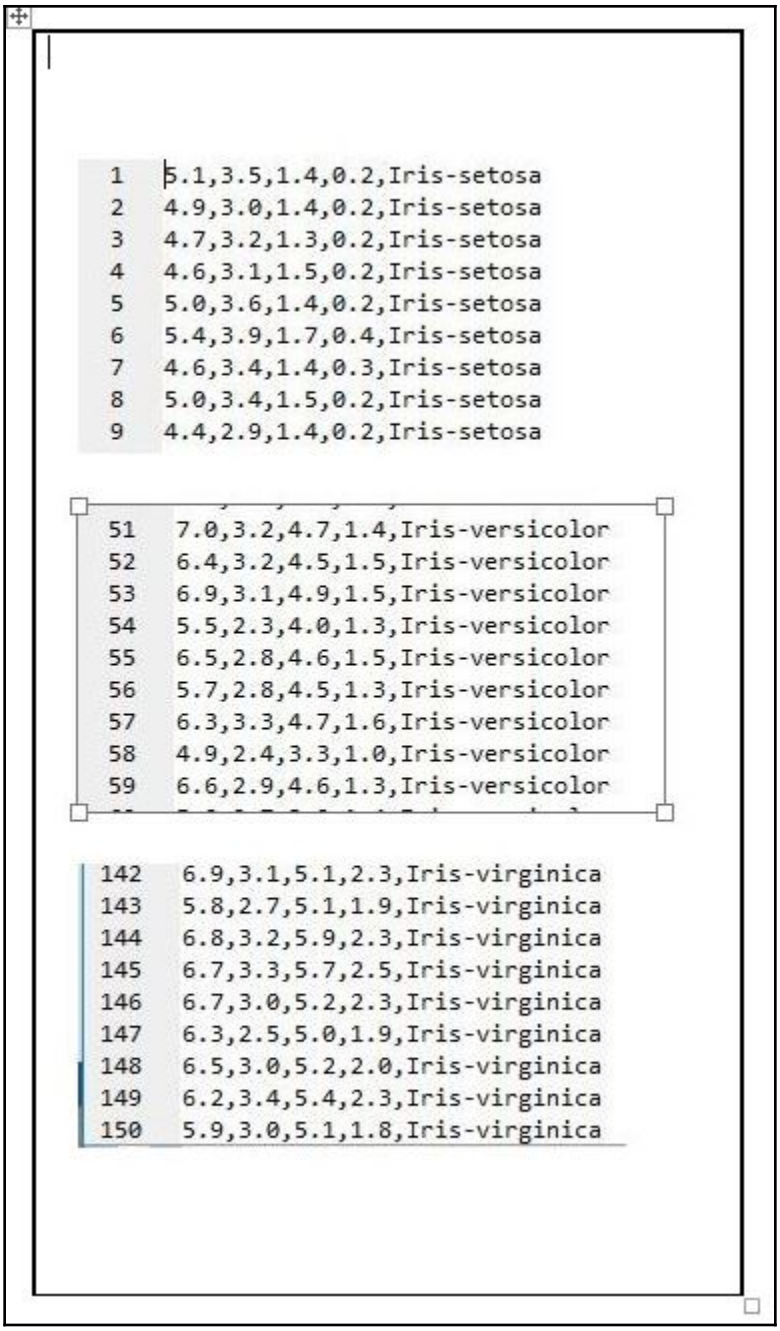
java.io.IOException: could not locate executable null\bin\winutils.exe in the Hadoop binaries.

```
scala> object DataReader {
     | def main(args: Array[String]): Unit = {
     |  val datasrc = Source.fromFile("iris.csv")
     | try datasrc.getLines.foreach(println) finally datasrc.close()
     | }
     | }
defined object DataReader

scala> DataReader.main("")
<console>:27: error: type mismatch;
 found   : String("")
 required: Array[String]
       DataReader.main("")
                       ^

scala> DataReader.main(Array(""))
java.io.FileNotFoundException: iris.csv (The system cannot find the file specified)
  at java.io.FileInputStream.open0(Native Method)
  at java.io.FileInputStream.open(FileInputStream.java:195)
  at java.io.FileInputStream.<init>(FileInputStream.java:138)
  at scala.io.Source$.fromFile(Source.scala:91)
  at scala.io.Source$.fromFile(Source.scala:76)
  at scala.io.Source$.fromFile(Source.scala:54)
  at DataReader$.main(<console>:26)
  ... 48 elided

scala> DataReader.main(Array(""))
sepal_length,sepal_width,petal_length,petal_width,species
5.1,3.5,1.4,0.2,setosa
4.9,3.0,1.4,0.2,setosa
4.7,3.2,1.3,0.2,setosa
4.6,3.1,1.5,0.2,setosa
5.0,3.6,1.4,0.2,setosa
5.4,3.9,1.7,0.4,setosa
```

```
+---+-------------+------------+-------------+------------+-----------+
| Id|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|    Species|
+---+-------------+------------+-------------+------------+-----------+
|  1|          5.1|         3.5|          1.4|         0.2|Iris-setosa|
|  2|          4.9|         3.0|          1.4|         0.2|Iris-setosa|
|  3|          4.7|         3.2|          1.3|         0.2|Iris-setosa|
|  4|          4.6|         3.1|          1.5|         0.2|Iris-setosa|
|  5|          5.0|         3.6|          1.4|         0.2|Iris-setosa|
|  6|          5.4|         3.9|          1.7|         0.4|Iris-setosa|
|  7|          4.6|         3.4|          1.4|         0.3|Iris-setosa|
|  8|          5.0|         3.4|          1.5|         0.2|Iris-setosa|
|  9|          4.4|         2.9|          1.4|         0.2|Iris-setosa|
| 10|          4.9|         3.1|          1.5|         0.1|Iris-setosa|
| 11|          5.4|         3.7|          1.5|         0.2|Iris-setosa|
| 12|          4.8|         3.4|          1.6|         0.2|Iris-setosa|
| 13|          4.8|         3.0|          1.4|         0.1|Iris-setosa|
| 14|          4.3|         3.0|          1.1|         0.1|Iris-setosa|
| 15|          5.8|         4.0|          1.2|         0.2|Iris-setosa|
| 16|          5.7|         4.4|          1.5|         0.4|Iris-setosa|
| 17|          5.4|         3.9|          1.3|         0.4|Iris-setosa|
| 18|          5.1|         3.5|          1.4|         0.3|Iris-setosa|
| 19|          5.7|         3.8|          1.7|         0.3|Iris-setosa|
| 20|          5.1|         3.8|          1.5|         0.3|Iris-setosa|
+---+-------------+------------+-------------+------------+-----------+
only showing top 20 rows

irisDataFrame: Unit = ()
```

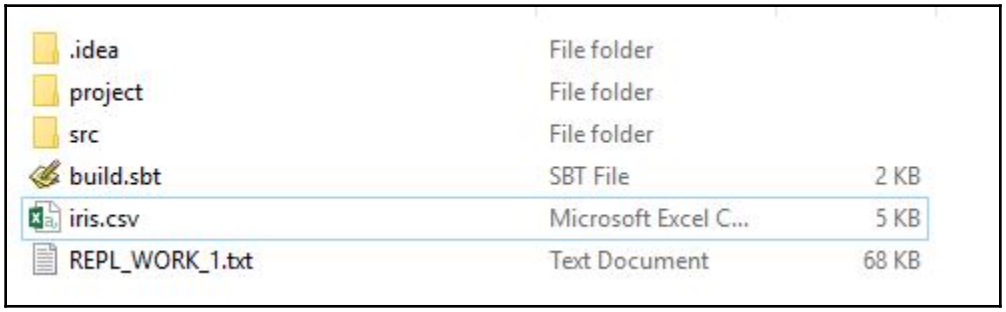

Spark
Training
Iris.data
Model
Spark and Scala
Model
Model file
Attributes
1.dat
Inference
Spark and Scala
Inference
5.1,3.5,1.4,0.2,
4.9,3.0,1.4,0.2,
4.7,3.2,1.3,0.2,
4.6,3.1,1.5,0.2,
5.0,3.6,1.4,0.2,
5.4,3.9,1.7,0.4,
4.6,3.4,1.4,0.3,
Iris-setosa
Iris-versicolor
Iris-versicolor
Iris-virginica


Parameters
Feature Extraction
(1)
Model Fitting
(2)
Model validation
(3)
Model 1,2,3,4..
Mapping function
(Model) selection
(4)
Data Pre-processing
Model Testing/Validation With TestSet/Validation Dataset
(5)
Model Deployment with spark-submit into cluster

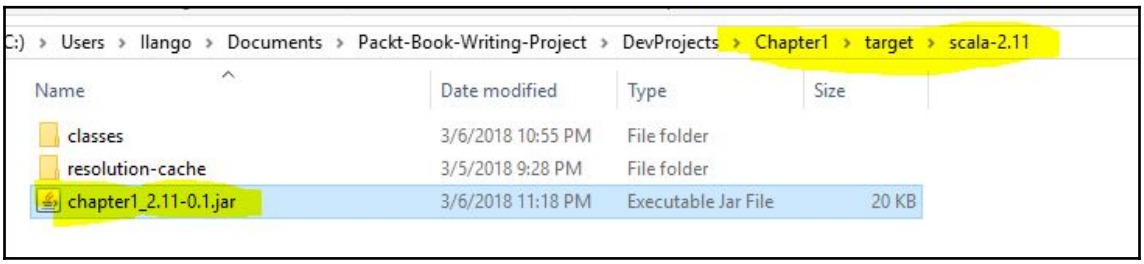

1 5.1,3.5,1.4,0.2,Iris-setosa
2 4.9,3.0,1.4,0.2,Iris-setosa
3 4.7,3.2,1.3,0.2,Iris-setosa
4 4.6,3.1,1.5,0.2,Iris-setosa
5 5.0,3.6,1.4,0.2,Iris-setosa
6 5.4,3.9,1.7,0.4,Iris-setosa
7 4.6,3.4,1.4,0.3,Iris-setosa
8 5.0,3.4,1.5,0.2,Iris-setosa
9 4.4,2.9,1.4,0.2,Iris-setosa
51 7.0,3.2,4.7,1.4,Iris-versicolor
52 6.4,3.2,4.5,1.5,Iris-versicolor
53 6.9,3.1,4.9,1.5,Iris-versicolor
54 5.5,2.3,4.0,1.3,Iris-versicolor
55 6.5,2.8,4.6,1.5,Iris-versicolor
56 5.7,2.8,4.5,1.3,Iris-versicolor
57 6.3,3.3,4.7,1.6,Iris-versicolor
58 4.9,2.4,3.3,1.0,Iris-versicolor
59 6.6,2.9,4.6,1.3,Iris-versicolor
142 6.9,3.1,5.1,2.3,Iris-virginica
143 5.8,2.7,5.1,1.9,Iris-virginica
144 6.8,3.2,5.9,2.3,Iris-virginica
145 6.7,3.3,5.7,2.5,Iris-virginica
146 6.7,3.0,5.2,2.3,Iris-virginica
147 6.3,2.5,5.0,1.9,Iris-virginica
148 6.5,3.0,5.2,2.0,Iris-virginica
149 6.2,3.4,5.4,2.3,Iris-virginica
150 5.9,3.0,5.1,1.8,Iris-virginica

```
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License.  You may obtain a copy of the License at
#
#    http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#

# Default system properties included when running spark-submit.
# This is useful for setting default environmental settings.

# Example:
# spark.master                     spark://master:7077
# spark.eventLog.enabled           true
# spark.eventLog.dir               hdfs://namenode:8021/directory
# spark.serializer                 org.apache.spark.serializer.KryoSerializer
# spark.driver.memory              5g
# spark.executor.extraJavaOptions  -XX:+PrintGCDetails -Dkey=value -Dnumbers="one two three"
spark.debug.maxToStringFields    150
```

summary	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
count	150	150	150	150	150	150
mean	75.5	5.843333333333335	3.0540000000000007	3.7586666666666693	1.1986666666666672	null
stddev	43.445367992456916	0.8280661279778637	0.4335943113621375	1.764420419952262	0.7631607417008414	null
min	1	4.3	2.0	1.0	0.1	Iris-setosa
max	150	7.9	4.4	6.9	2.5	Iris-virginica



```
libraryDependencies ++= Seq(
  "org.apache.spark" %% "spark-core" % "2.2.1",
  "org.apache.spark" %% "spark-mllib" % "2.2.1",
  "org.apache.spark" %% "spark-sql" % "2.2.1",
  // Last stable release
  "org.scalanlp" %% "breeze" % "0.13.2",

  // Native libraries are not included by default. add this if you want them (as of 0.7)
  // Native libraries greatly improve performance, but increase jar sizes.
  // It also packages various blas implementations, which have licenses that may or may not
  // be compatible with the Apache License. No GPL code, as best I know.

  "org.scalanlp" %% "breeze-natives" % "0.13.2",

  // The visualization library is distributed separately as well.
  // It depends on LGPL code
  "org.scalanlp" %% "breeze-viz" % "0.13.2",
  "joda-time" % "joda-time" % "2.9.9",
  "org.scalatest"   %% "scalatest"    % "3.0.1"   % "test",
  "org.slf4j" % "slf4j-api" % "1.7.22",
  "org.slf4j" % "slf4j-simple" % "1.7.22"

)
```

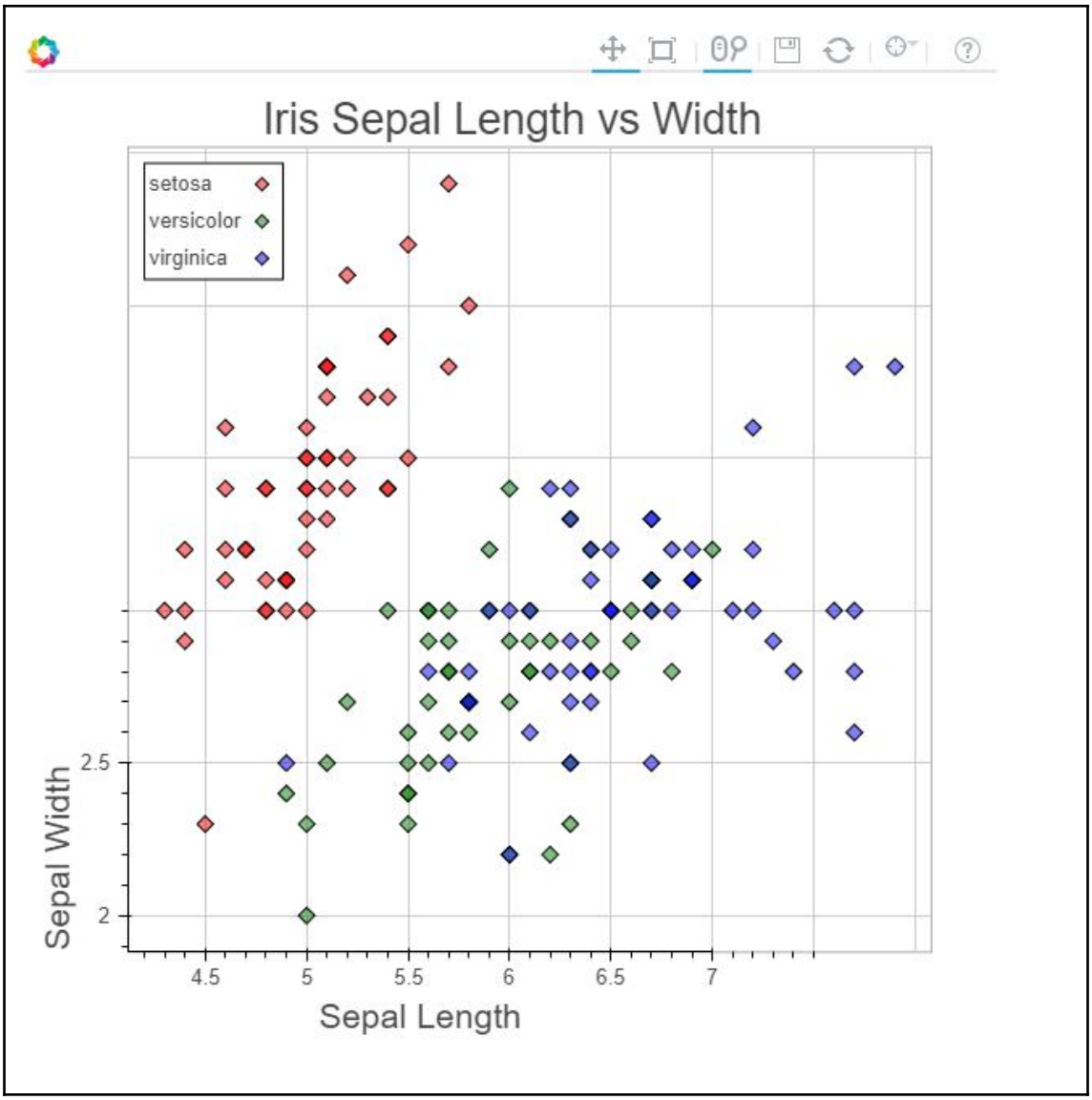


Chapter 2: Build a Breast Cancer Prognosis Pipeline with the Power of Spark and Scala

Breast Cancer Wisconsin Dataset at a glance	
Classification technique	Multivariate
Total no of instances	699
Number of Attributes	10
Attribute Data Type	Integer
Attribute Names (Columns)	Sample Code Number
Number of classes (labels)	2
Cell Nuclei measurements (Attributes)	Sample Code number
	Clump Thickness
	Uniformity of Cell Size
	Uniformity of Cell Shape
	Marginal Adhesion
	Single Epithelial Cell Size
	Bare Nuclei
	Bland Chromatin
	Normal Nucleoli
	Mitoses
	Class

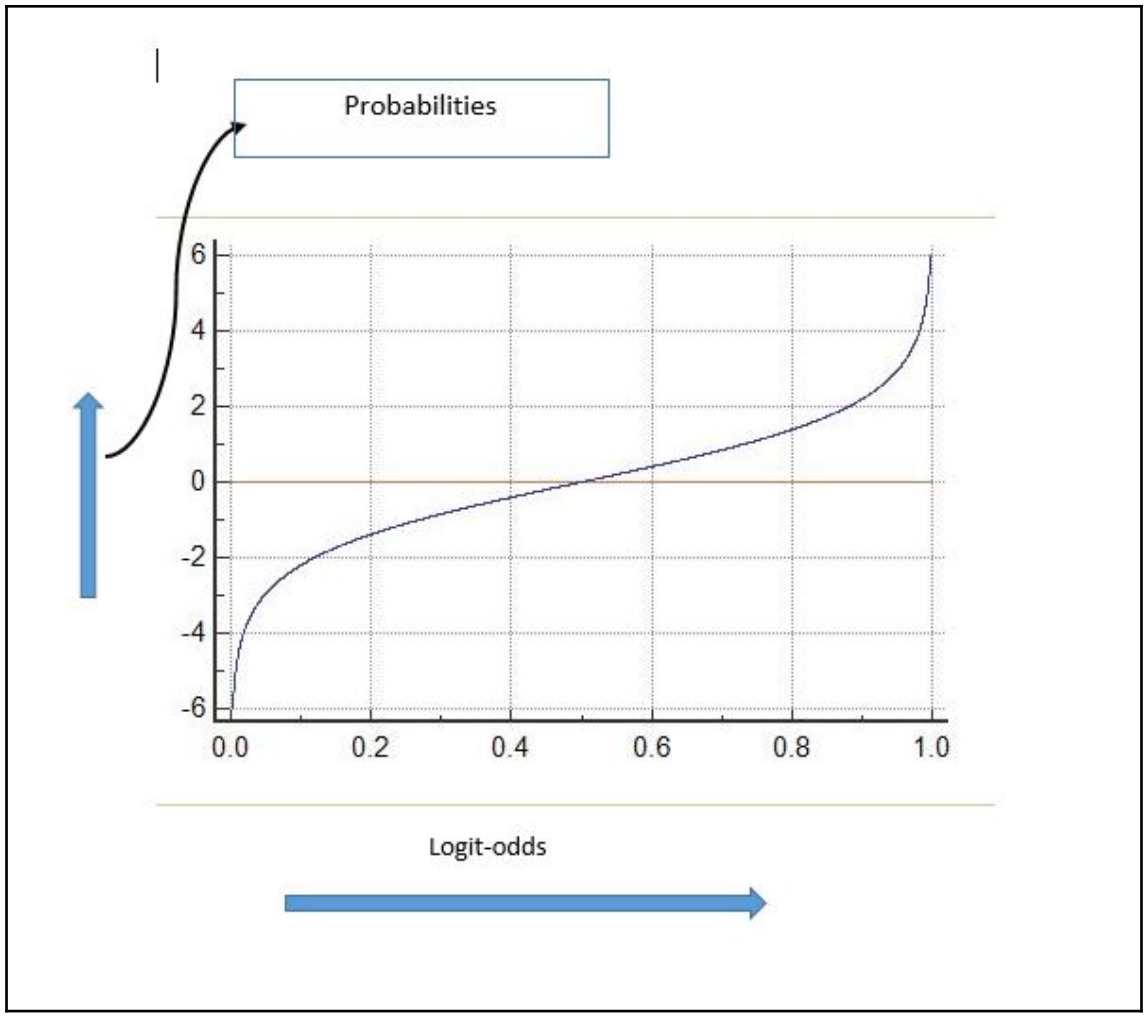
Logistic Regression

It is a Statistical Technique	The technique helps analyze a dataset
It operates well on a dataset that contains more than one independent variable.	Each of these independent or input attribute value determines a prediction outcome
Its prediction outcome variable (dependent) is dichotomous	A dichotomous (output) variable can be: 1) Alabama (Red State) or California (Blue State) 2) Giant or tiny
Its dichotomous variable is a dummy, and is only measured in one of two ways: 1) Nominal measurement 2) Ordinal measurement	Dichotomous variables are identified as having 2 states that are mutually exclusive. Each state is represented as either a 0 or a 1, lending it the nature of a dummy variable
Its dichotomous variable is also known as a binary variable. That explains the term "Binary Logistic Regression"	A binary variable represents two possible results or outcomes. The flip of a coin produces 1) Heads 2) Tails

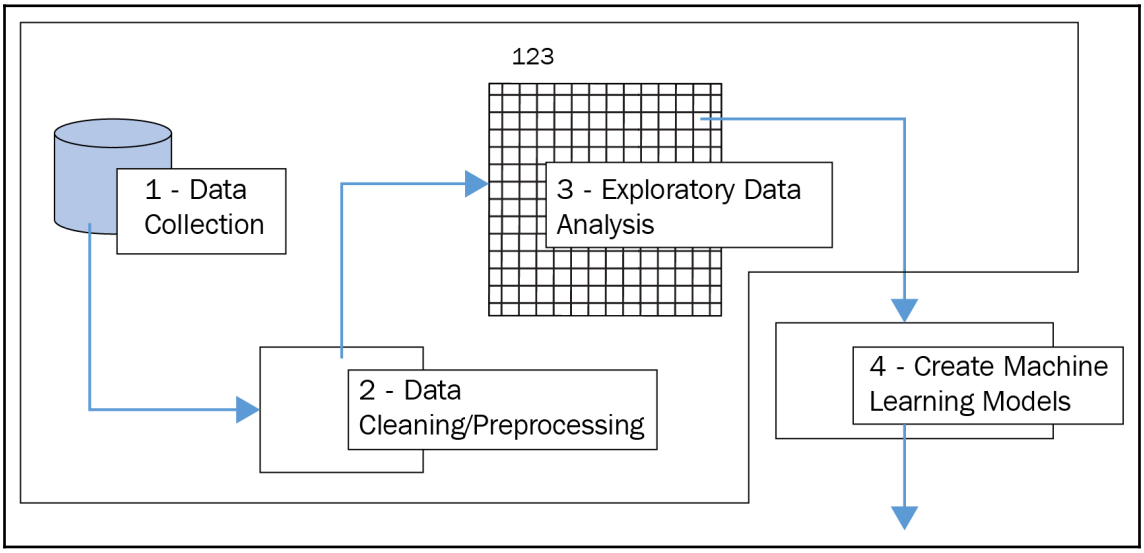

Iris Sepal Length vs Width
setosa
versicolor
virginica
Sepal Width
2.5
2
4.5
5
5.5
6
6.5
7
Sepal Length

Individual	2nd Date	Dating Workshop	Coolness Score
1	1	1	80
2	0	1	85
3	0	1	65
4	0	0	60
5	1	0	70
6	1	1	75
7	1	0	78
8	1	0	65
9	1	0	87
10	0	0	60
11	0	1	55
12	0	1	58
13	0	1	40
14	0	0	35
15	0	1	70
16	1	1	69
17	1	1	81
18	1	0	35
19	1	0	65
20	1	0	55

Correlation Coefficient	Date	Workshop	Cool
Date	1		
Workshop	-3.1	1	
Cool	0.47	-.17	1
Mean	0.49	.36	52.52
Standard Deviation	.50	.50	9.35


Probabilities
6
4
2
0
-2
-4
-6
0.0
0.2
0.4
0.6
0.8
1.0
Logit-odds

$$p = \frac{1}{1 + e^{-L}}$$

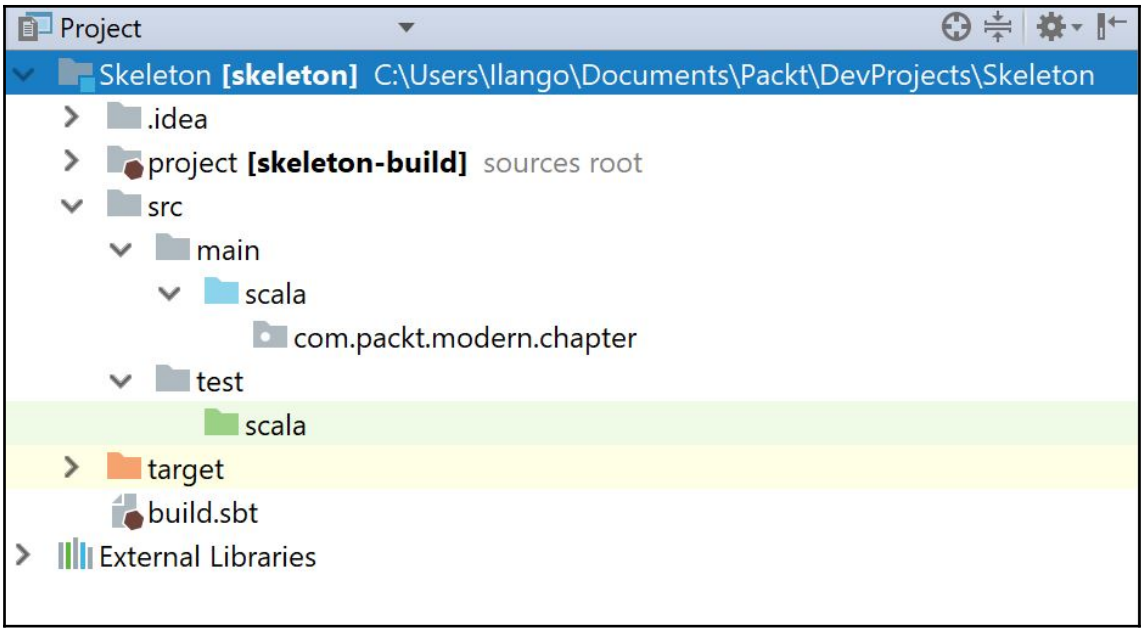

123
1 - Data Collection
2 - Data Cleaning/Preprocessing
3 - Exploratory Data Analysis
4 - Create Machine Learning Models

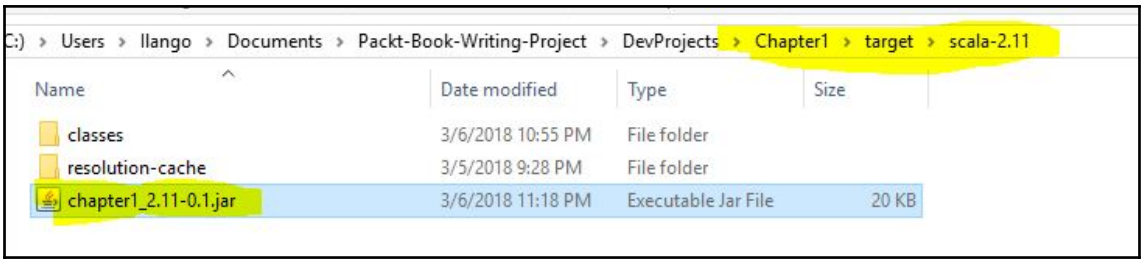
```
1000025,5,1,1,1,2,1,3,1,1,2
1002945,5,4,4,5,7,10,3,2,1,2
1015425,3,1,1,1,2,2,3,1,1,2
1016277,6,8,8,1,3,4,3,7,1,2
1017023,4,1,1,3,2,1,3,1,1,2
1017122,8,10,10,8,7,10,9,7,1,4
1018099,1,1,1,1,2,10,3,1,1,2
1018561,2,1,2,1,2,1,3,1,1,2
1033078,2,1,1,1,2,1,1,1,5,2
1033078,4,2,1,1,2,1,2,1,1,2
1035283,1,1,1,1,1,1,3,1,1,2
1036172,2,1,1,1,2,1,2,1,1,2
1041801,5,3,3,3,2,3,4,4,1,4
1043999,1,1,1,1,2,3,3,1,1,2
1044572,8,7,5,10,7,9,5,5,4,4
1047630,7,4,6,4,6,1,4,3,1,4
1048672,4,1,1,1,2,1,2,1,1,2
1049815,4,1,1,1,2,1,3,1,1,2
1050670,10,7,7,6,4,10,4,1,2,4
1050718,6,1,1,1,2,1,3,1,1,2
1054590,7,3,2,10,5,10,5,4,4,4
1054593,10,5,5,3,6,7,7,10,1,4
1056784,3,1,1,1,2,1,2,1,1,2
1057013,8,4,5,1,2,?,7,3,1,4
1059552,1,1,1,1,2,1,3,1,1,2
1065726,5,2,3,4,2,7,3,6,1,4
1066373,3,2,1,1,1,1,2,1,1,2
1066979,5,1,1,1,2,1,2,1,1,2
1067444,2,1,1,1,2,1,2,1,1,2
1070935,1,1,3,1,2,1,1,1,1,2
1070935,3,1,1,1,1,1,2,1,1,2
1071760,2,1,1,1,2,1,3,1,1,2
1072179,10,7,7,3,8,5,7,4,3,4
1074610,2,1,1,2,2,1,3,1,1,2
1075123,3,1,2,1,2,1,2,1,1,2
1079304,2,1,1,1,2,1,2,1,1,2
1080185,10,10,10,8,6,1,8,9,1,4
1081791,6,2,1,1,1,1,7,1,1,2
1084584,5,4,4,9,2,10,5,6,1,4
1091262,2,5,3,3,6,7,7,5,1,4
1096800,6,6,6,9,6,?,7,8,1,2
1099510,10,4,3,1,3,3,6,5,2,4
```

```
scala> dataFrame.show
+-------+---------------+---------------+----------------+-----------------+--------------+-------------+---------------+------
--------+-------+------+-----+
|     id|clump_thickness|size_uniformity|shape_uniformity|marginal_adhesion|epithelial_size|bare_nucleoli|bland_chromatin|normal
_nucleoli|mitoses|class|
+-------+---------------+---------------+----------------+-----------------+--------------+-------------+---------------+------
+---------+-------+-----+
|1000025|              5|              1|               1|                1|             2|            1|              3|
             1|      1|    2|
|1002945|              5|              4|               4|                5|             7|           10|              3|
             2|      1|    2|
|1015425|              3|              1|               1|                1|             2|            2|              3|
             1|      1|    2|
|1016277|              6|              8|               8|                1|             3|            4|              3|
             7|      1|    2|
|1017023|              4|              1|               1|                3|             2|            1|              3|
             1|      1|    2|
|1017122|              8|             10|              10|                8|             7|           10|              9|
             7|      1|    4|
|1018099|              1|              1|               1|                1|             2|           10|              3|
             1|      1|    2|
|1018561|              2|              1|               2|                1|             2|            1|              3|
             1|      1|    2|
|1033078|              2|              1|               1|                1|             2|            1|              1|
             1|      5|    2|
|1033078|              4|              2|               1|                1|             2|            1|              2|
             1|      1|    2|
|1035283|              1|              1|               1|                1|             1|            1|              3|
             1|      1|    2|
```

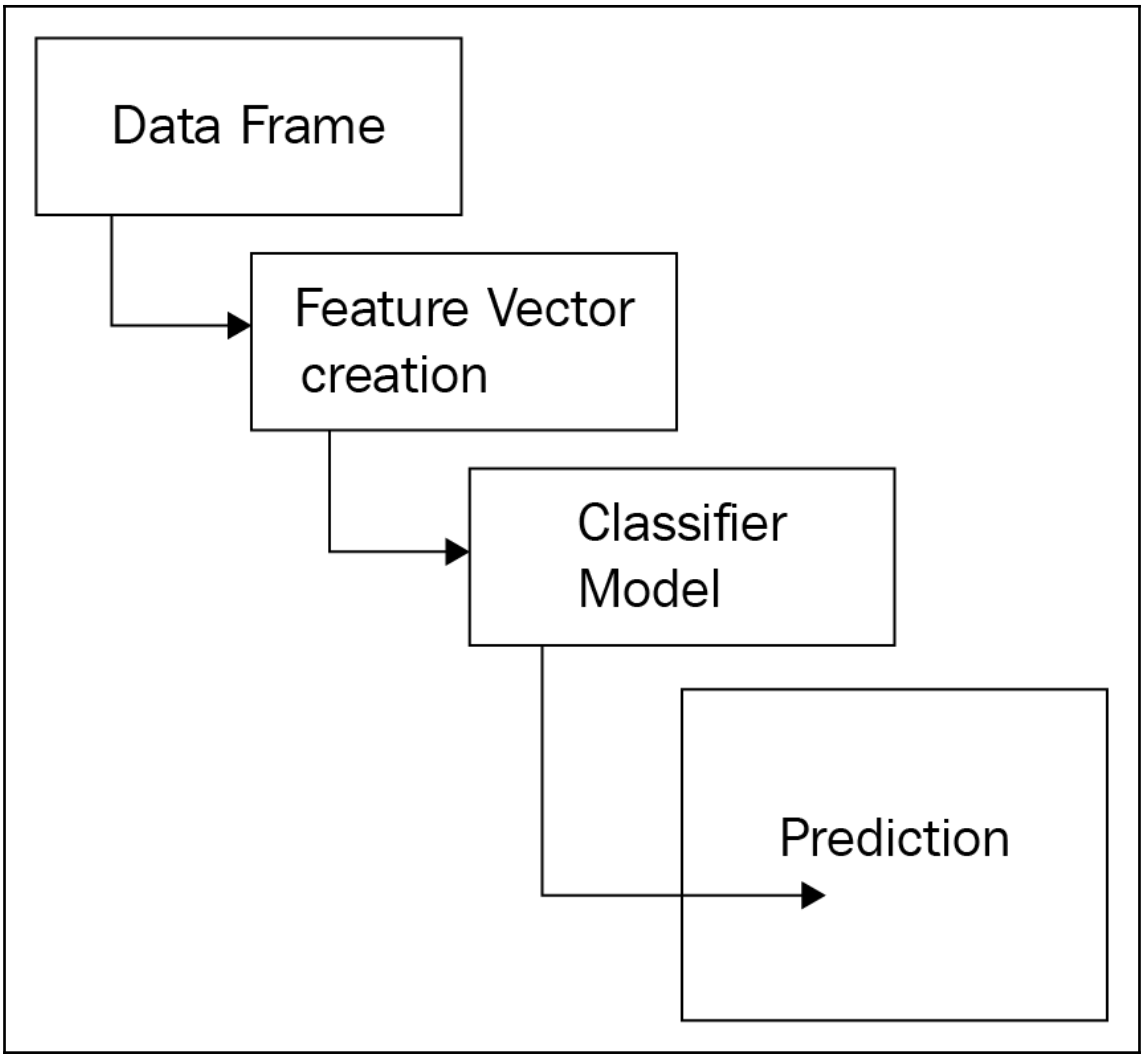
```
scala> stats.show
+-------+-----------------+----------------+-----------------+------------------+-----------------+-----------------+-----------------+---
-----------+-----------------+-----------------+-----------------+
|summary|               id|  clump_thickness|   size_uniformity|  shape_uniformity| marginal_adhesion|  epithelial_size|     bare_nucleoli|   |
nd_chromatin|   normal_nucleoli|          mitoses|             class|
+-------+-----------------+----------------+-----------------+------------------+-----------------+-----------------+-----------------+---
----------+-----------------+-----------------+-----------------+
|  count|              699|             699|              699|               699|              699|              699|              699|
              699|              699|              699|              699|
|   mean|1071704.0987124464|4.417739628040057|  3.13447782546495| 3.207439198855508|2.8068669527896994|3.216022889842632|3.5446559297218156|3.4
582403433476| 2.866952789699571|1.5894134477825466|2.6895565092989986|
| stddev| 617095.7298192448|2.8157406585949333|3.0514591099542003|2.9719127672157133| 2.855379239217023|2.214299886649047| 3.643857160492912|2.4
542523242512|3.0536338936127745| 1.7150779425067950.9512725322121839|
|    min|            61634|               1|                1|                 1|                1|                1|                1|
                1|                1|                1|                2|
|    max|         13454352|              10|               10|                10|               10|               10|                ?|
               10|               10|               10|                4|
+-------+-----------------+----------------+-----------------+------------------+-----------------+-----------------+-----------------+---
----------+-----------------+-----------------+-----------------+
```

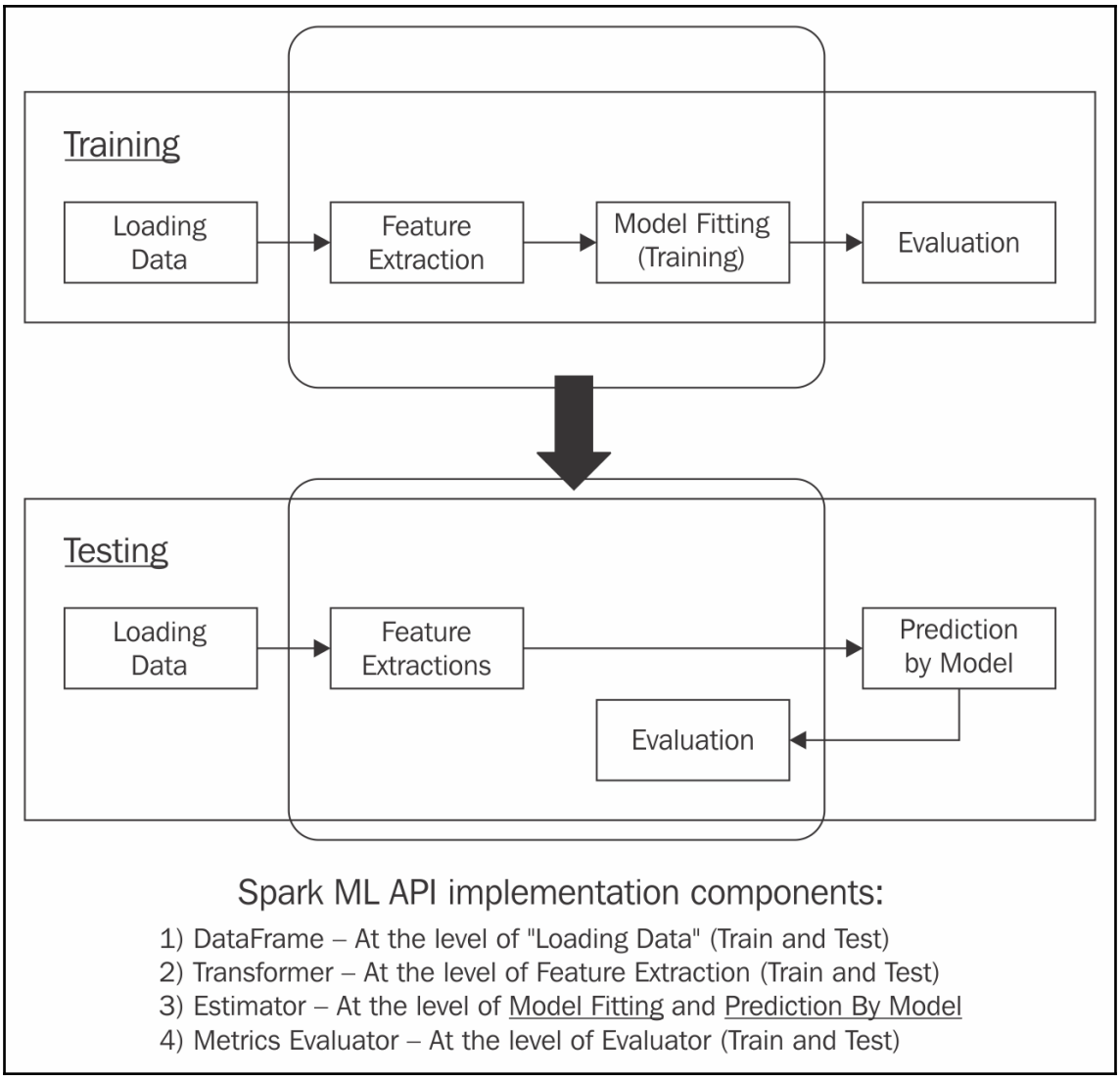
```
 2  1000025,5,1,1,1,2,1,3,1,1,2
 3  1002945,5,4,4,5,7,10,3,2,1,2
 4  1015425,3,1,1,1,2,2,3,1,1,2
 5  1016277,6,8,8,1,3,4,3,7,1,2
 6  1017023,4,1,1,3,2,1,3,1,1,2
 7  1017122,8,10,10,8,7,10,9,7,1,4
 8  1018099,1,1,1,1,2,10,3,1,1,2
 9  1018561,2,1,2,1,2,1,3,1,1,2
10  1033078,2,1,1,1,2,1,1,1,5,2
11  1033078,4,2,1,1,2,1,2,1,1,2
12  1035283,1,1,1,1,1,1,3,1,1,2
13  1036172,2,1,1,1,2,1,2,1,1,2
14  1041801,5,3,3,3,2,3,4,4,1,4
15  1043999,1,1,1,1,2,3,3,1,1,2
16  1044572,8,7,5,10,7,9,5,5,4,4
17  1047630,7,4,6,4,6,1,4,3,1,4
18  1048672,4,1,1,1,2,1,2,1,1,2
19  1049815,4,1,1,1,2,1,3,1,1,2
20  1050670,10,7,7,6,4,10,4,1,2,4
21  1050718,6,1,1,1,2,1,3,1,1,2
22  1054590,7,3,2,10,5,10,5,4,4,4
23  1054593,10,5,5,3,6,7,7,10,1,4
24  1056784,3,1,1,1,2,1,2,1,1,2
25  1057013,8,4,5,1,2,?,7,3,1,4
26  1059552,1,1,1,1,2,1,3,1,1,2
27  1065726,5,2,3,4,2,7,3,6,1,4
```

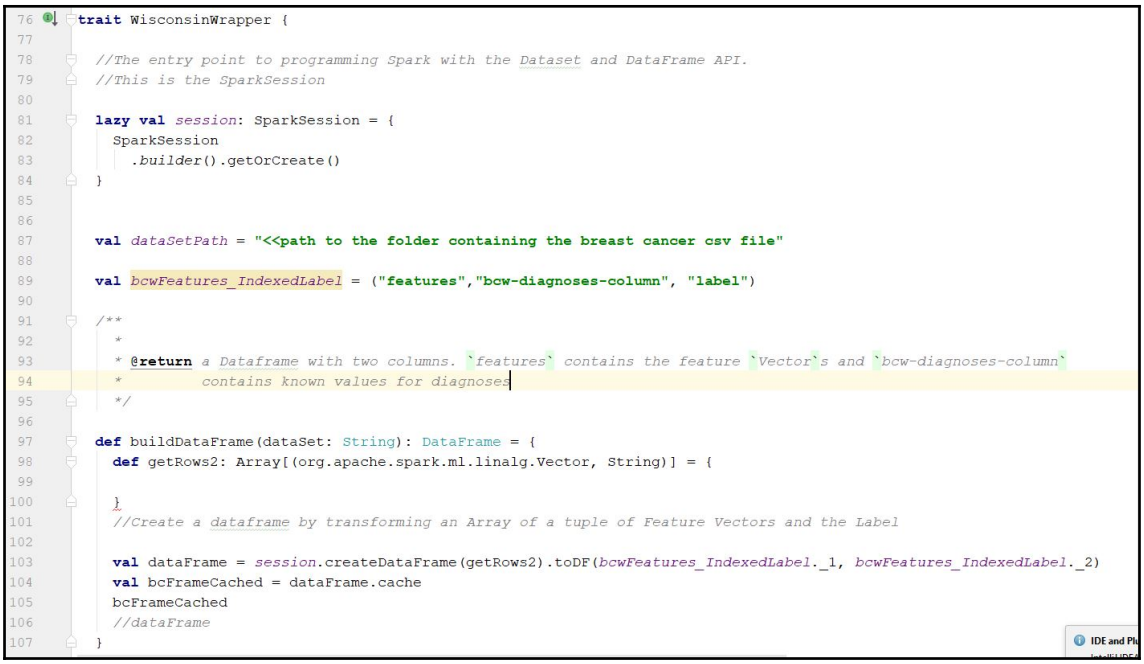

Project
Skeleton [skeleton] C:\Users\Ilango\Documents\Packt\DevProjects\Skeleton
.idea
project [skeleton-build] sources root
src
main
scala
com.packt.modern.chapter
test
scala
target
build.sbt
External Libraries


C:) › Users › Ilango › Documents › Packt-Book-Writing-Project › DevProjects › Chapter1 › target › scala-2.11
Name
Date modified
Type
Size
classes
3/6/2018 10:55 PM
File folder
resolution-cache
3/5/2018 9:28 PM
File folder
chapter1_2.11-0.1.jar
3/6/2018 11:18 PM
Executable Jar File
20 KB

Label *represented by*	class
Two possible values are:	Malignant → 1
	Benign → 0
Features	
clump_thickness	*Numerical Values that are placed in a Feature Vector later*
size_uniformity	
shape_uniformity	
marginal_adhesion	
epithelial_size	
bare_nucleoli	
bland_chromatin	
normal_nucleoli	
mitoses	


Data Frame
Feature Vector creation
Classifier Model
Prediction

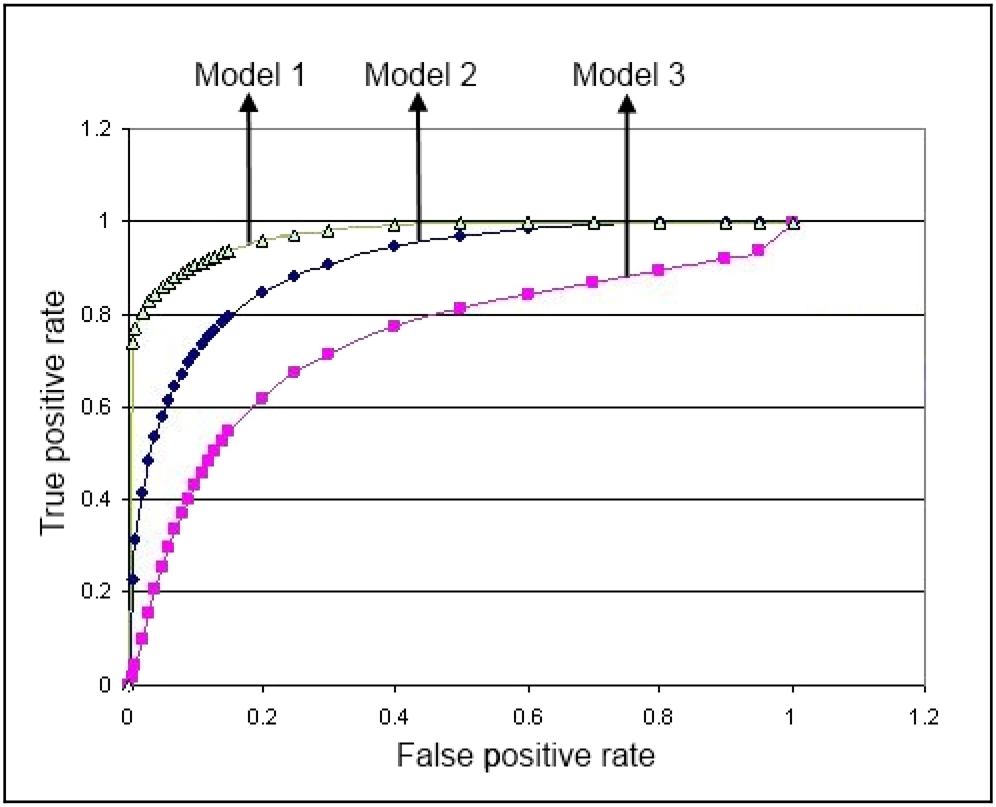

Training
Loading Data
Feature Extraction
Model Fitting (Training)
Evaluation
Testing
Loading Data
Feature Extractions
Prediction by Model
Evaluation
Spark ML API implementation components:
1) DataFrame – At the level of "Loading Data" (Train and Test)
2) Transformer – At the level of Feature Extraction (Train and Test)
3) Estimator – At the level of Model Fitting and Prediction By Model
4) Metrics Evaluator – At the level of Evaluator (Train and Test)



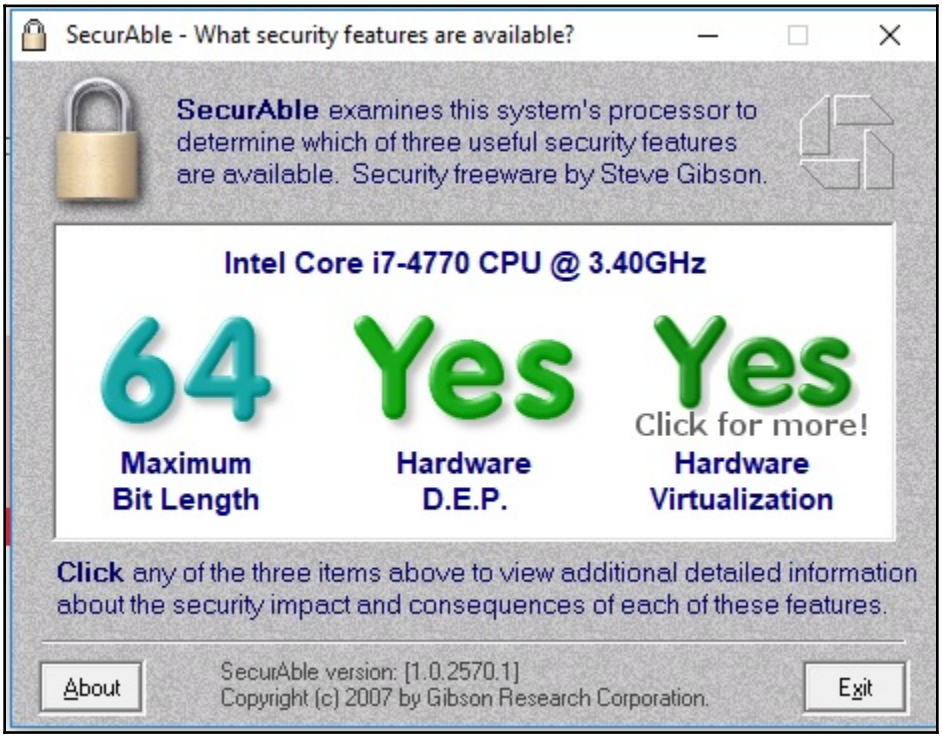
```
trait WisconsinWrapper {

  //The entry point to programming Spark with the Dataset and DataFrame API.
  //This is the SparkSession

  lazy val session: SparkSession = {
    SparkSession
      .builder().getOrCreate()
  }


  val dataSetPath = "<<path to the folder containing the breast cancer csv file"

  val bcwFeatures_IndexedLabel = ("features","bcw-diagnoses-column", "label")

  /**
    *
    * @return a Dataframe with two columns. `features` contains the feature `Vector`s and `bcw-diagnoses-column`
    *         contains known values for diagnoses
    */

  def buildDataFrame(dataSet: String): DataFrame = {
    def getRows2: Array[(org.apache.spark.ml.linalg.Vector, String)] = {

    }
    //Create a dataframe by transforming an Array of a tuple of Feature Vectors and the Label

    val dataFrame = session.createDataFrame(getRows2).toDF(bcwFeatures_IndexedLabel._1, bcwFeatures_IndexedLabel._2)
    val bcFrameCached = dataFrame.cache
    bcFrameCached
    //dataFrame
  }
```

```
+--------------------+--------------------+-----+
|            features|bcw-diagnoses-column|label|
+--------------------+--------------------+-----+
|[5.0,1.0,1.0,1.0,...|                   2|  0.0|
|[5.0,4.0,4.0,5.0,...|                   2|  0.0|
|[3.0,1.0,1.0,1.0,...|                   2|  0.0|
|[6.0,8.0,8.0,1.0,...|                   2|  0.0|
|[4.0,1.0,1.0,3.0,...|                   2|  0.0|
|[8.0,10.0,10.0,8....|                   4|  1.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|
|[2.0,1.0,2.0,1.0,...|                   2|  0.0|
|[2.0,1.0,1.0,1.0,...|                   2|  0.0|
|[4.0,2.0,1.0,1.0,...|                   2|  0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|
|[2.0,1.0,1.0,1.0,...|                   2|  0.0|
|[5.0,3.0,3.0,3.0,...|                   4|  1.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|
|[8.0,7.0,5.0,10.0...|                   4|  1.0|
|[7.0,4.0,6.0,4.0,...|                   4|  1.0|
|[4.0,1.0,1.0,1.0,...|                   2|  0.0|
|[4.0,1.0,1.0,1.0,...|                   2|  0.0|
|[10.0,7.0,7.0,6.0...|                   4|  1.0|
|[6.0,1.0,1.0,1.0,...|                   2|  0.0|
+--------------------+--------------------+-----+
only showing top 20 rows
```

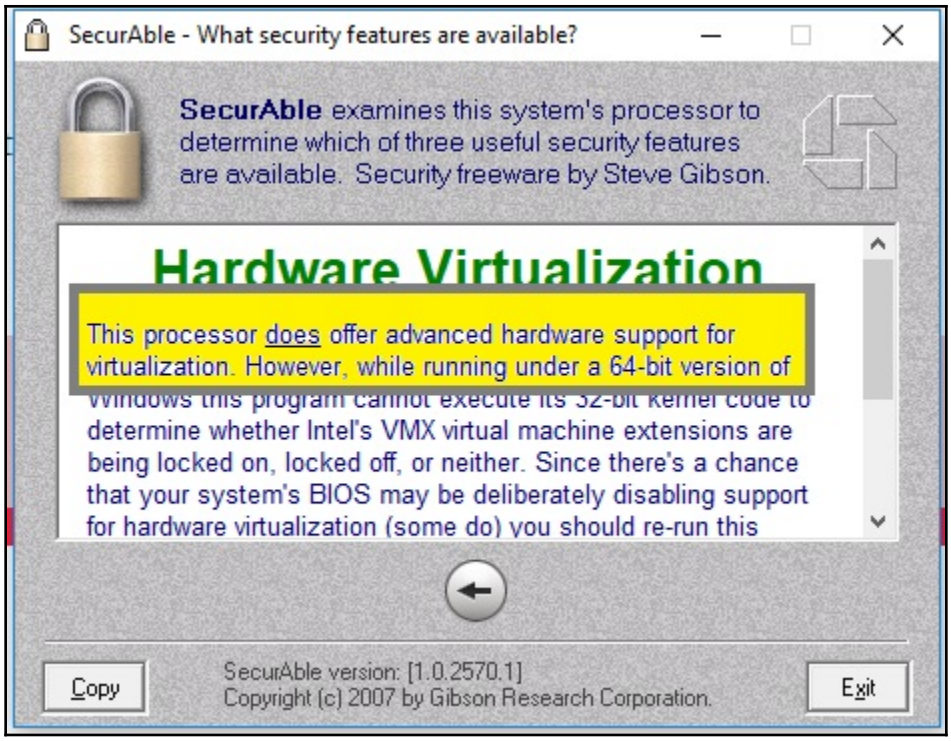
```
+--------------------+--------------------+-----+--------------------+--------------------+----------+
|            features|bcw-diagnoses-column|label|       rawPrediction|         probability|prediction|
+--------------------+--------------------+-----+--------------------+--------------------+----------+
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.43601098731156...|[0.99941071489731...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.08090265549135...|[0.99915969304957...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.02267364088016...|[0.99910935503926...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.41835610800472...|[0.9994002250538 6...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.41835610800472...|[0.99940022505386...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.00501876157331...|[0.99909350556530...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.00501876157331...|[0.99909350556530...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.00501876157331...|[0.99909350556530...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.00501876157331...|[0.99909350556530...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.00501876157331...|[0.99909350556530...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.00501876157331...|[0.99909350556530...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.00501876157331...|[0.99909350556530...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.59168141514191...|[0.99863014749668...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.59168141514191...|[0.99863014749668...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.59168141514191...|[0.99863014749668...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.59168141514191...|[0.99863014749668...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.59168141514191...|[0.99863014749668...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.59168141514191...|[0.99863014749668...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.97383168999539...|[0.99906481605134...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[5.70263257912325...|[0.99667393112928...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.08478285397673...|[0.99772791535642...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[5.6402584359674,...|[0.99646062136127...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[7.40070122869787...|[0.99938954859514...|       0.0|
|[1.0,1.0,1.0,1.0,...|                   2|  0.0|[6.49399751337237...|[0.99848979346810...|       0.0|
|[1.0,1.0,1.0,2.0,...|                   2|  0.0|[3.01813679889347...|[0.95338679396620...|       0.0|
+--------------------+--------------------+-----+--------------------+--------------------+----------+
only showing top 25 rows
```

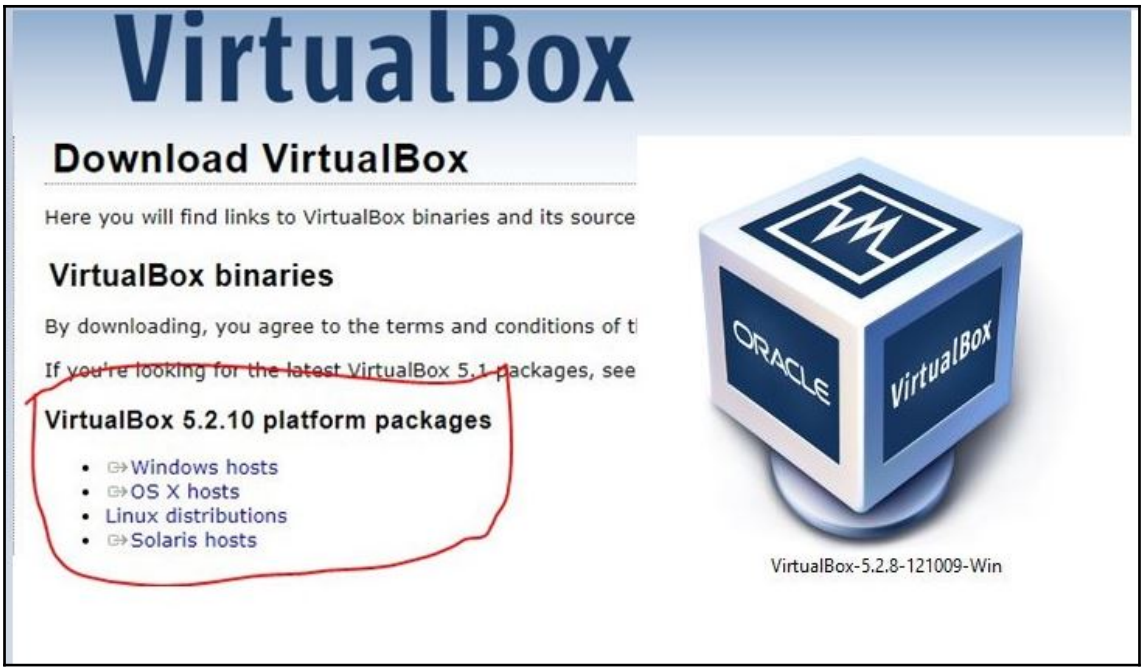


Chapter 3: Stock Price Predictions

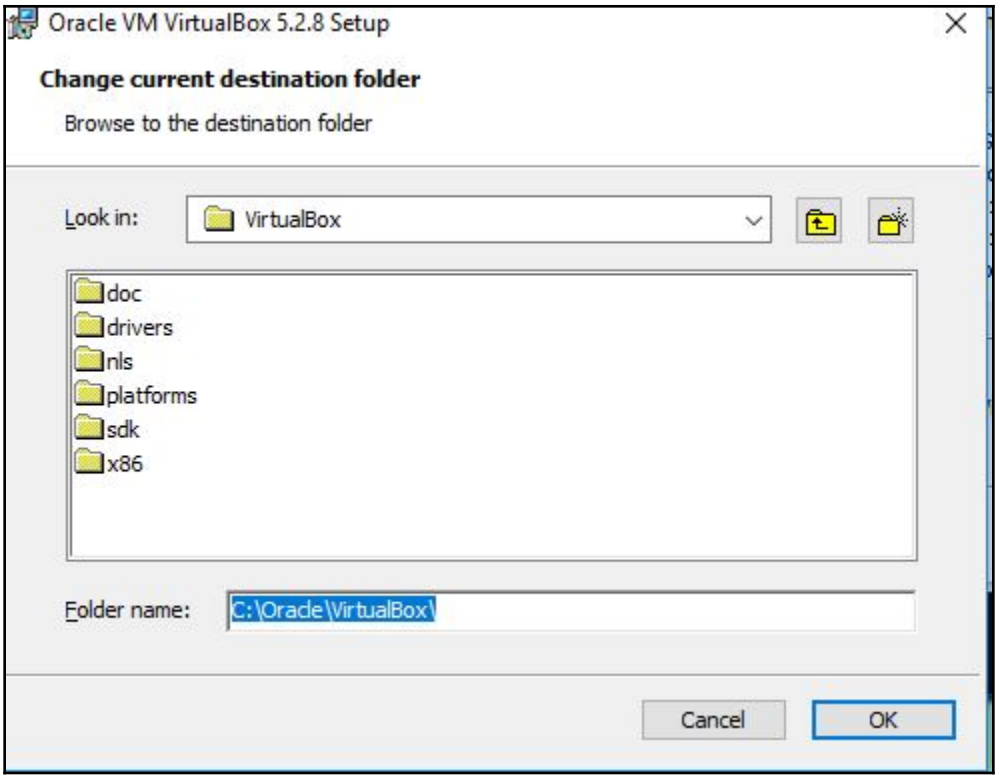


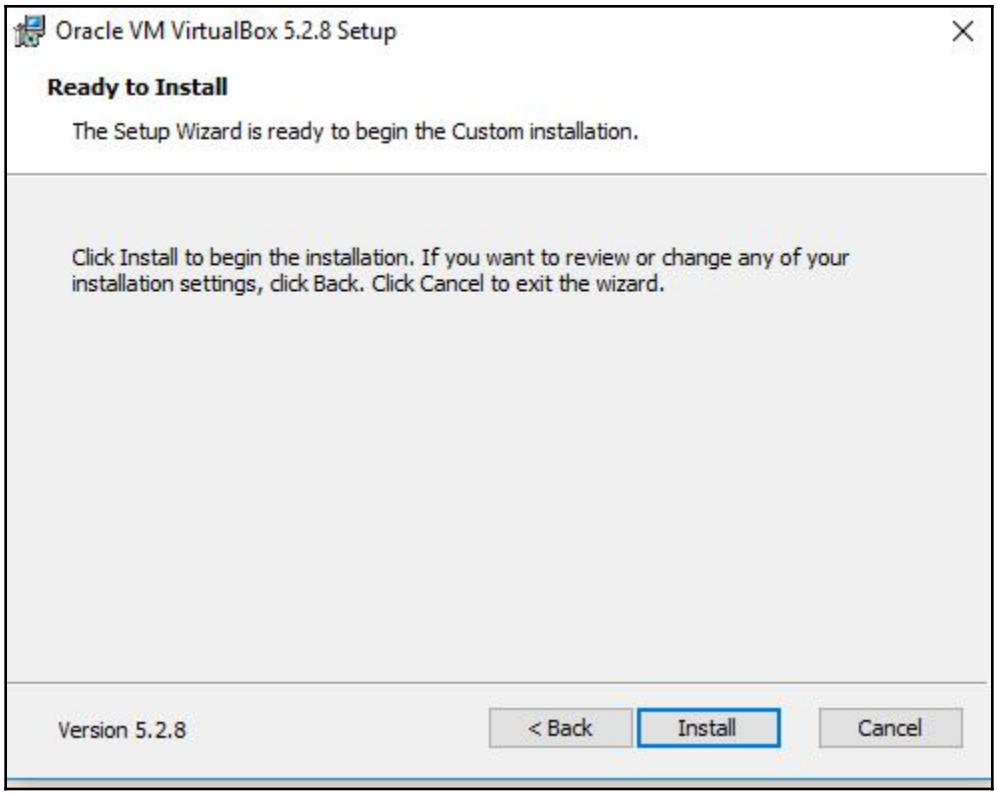

SecurAble - What security features are available?
SecurAble examines this system's processor to determine which of three useful security features are available. Security freeware by Steve Gibson.
64-Bit Processing Available
This processor does offer 64-bit modes of operation. This means that this system is able to run the significantly more secure 64-bit versions of Microsoft's Windows XP and Vista operating systems.
The biggest challenge for 64-bit Windows systems is the fact that existing 32-bit device drivers cannot be used by the 64-bit
Copy
SecurAble version: [1.0.2570.1]
Copyright (c) 2007 by Gibson Research Corporation.
Exit

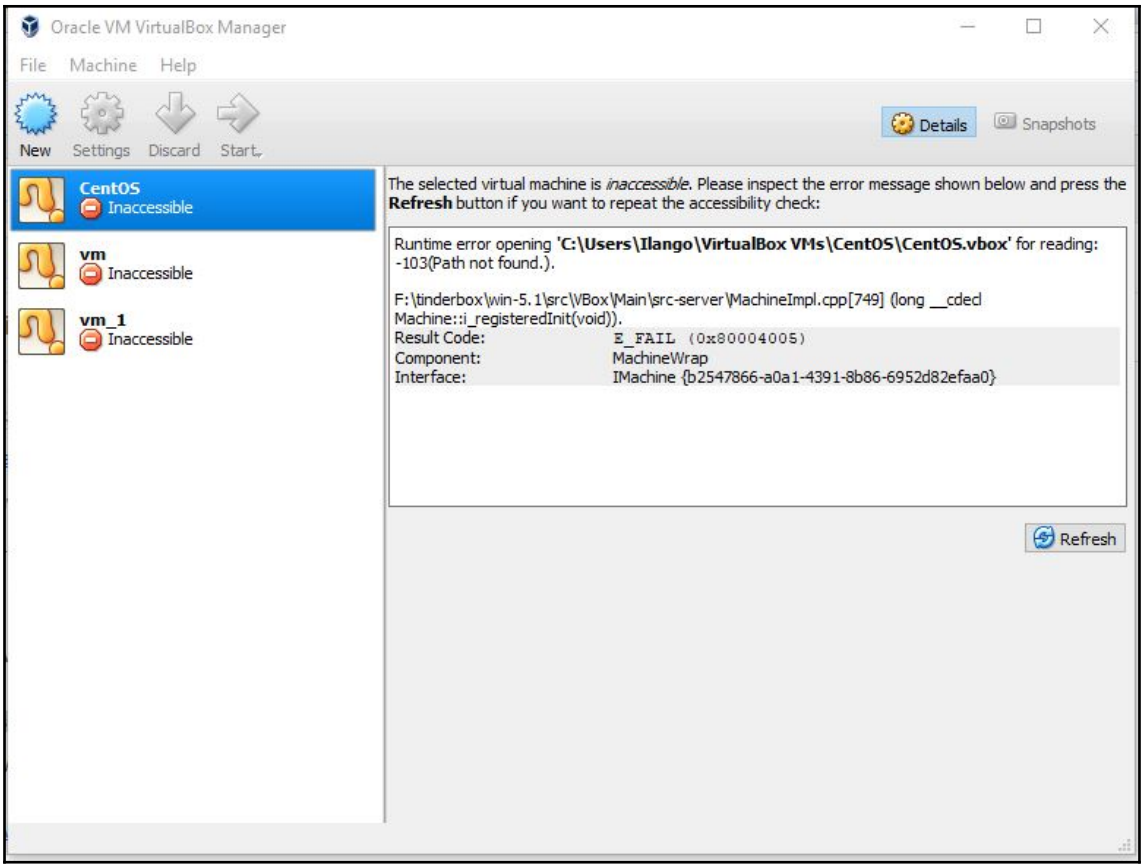

SecurAble - What security features are available?
SecurAble examines this system's processor to determine which of three useful security features are available. Security freeware by Steve Gibson.
Hardware Virtualization
This processor does offer advanced hardware support for virtualization. However, while running under a 64-bit version of
determine whether Intel's VMX virtual machine extensions are being locked on, locked off, or neither. Since there's a chance that your system's BIOS may be deliberately disabling support for hardware virtualization (some do) you should re-run this
Copy
SecurAble version: [1.0.2570.1]
Copyright (c) 2007 by Gibson Research Corporation.
Exit

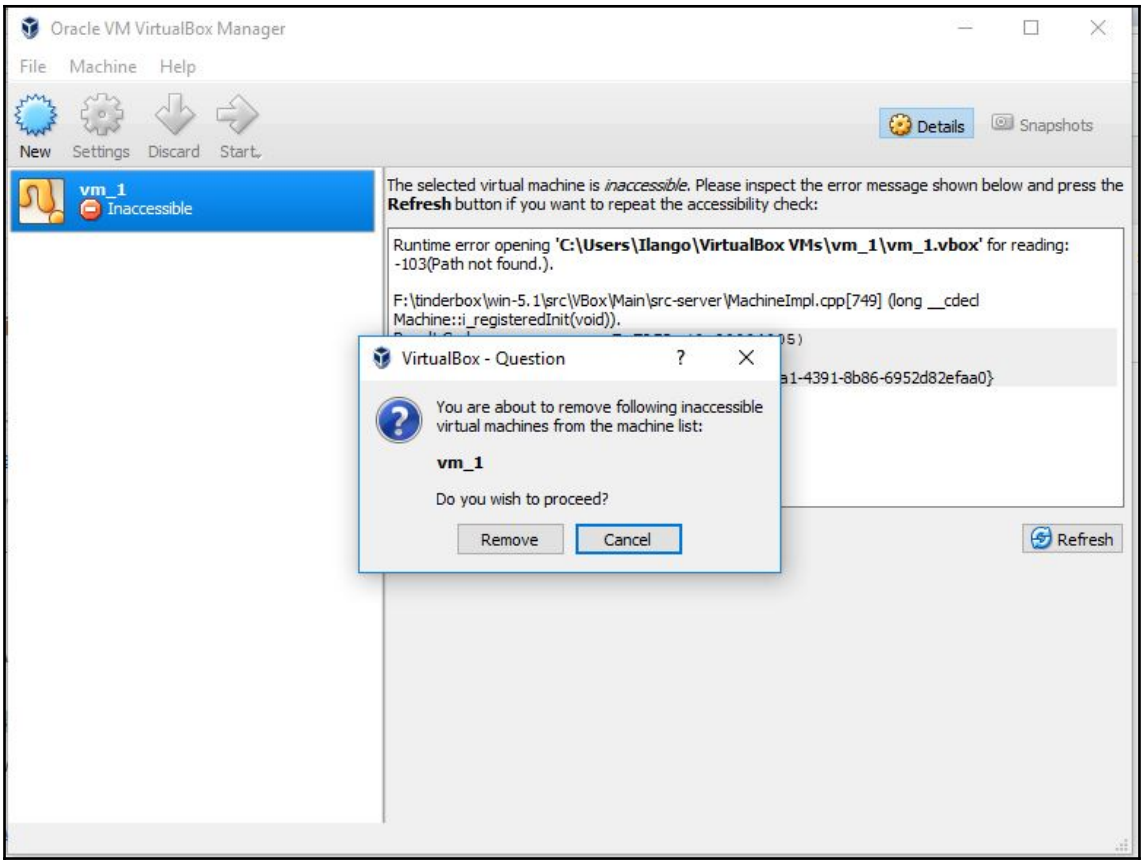

VirtualBox
Download VirtualBox
Here you will find links to VirtualBox binaries and its source
VirtualBox binaries
By downloading, you agree to the terms and conditions of t
If you're looking for the latest VirtualBox 5.1 packages, see
VirtualBox 5.2.10 platform packages
Windows hosts
OS X hosts
Linux distributions
Solaris hosts
ORACLE
VirtualBox
VirtualBox-5.2.8-121009-Win

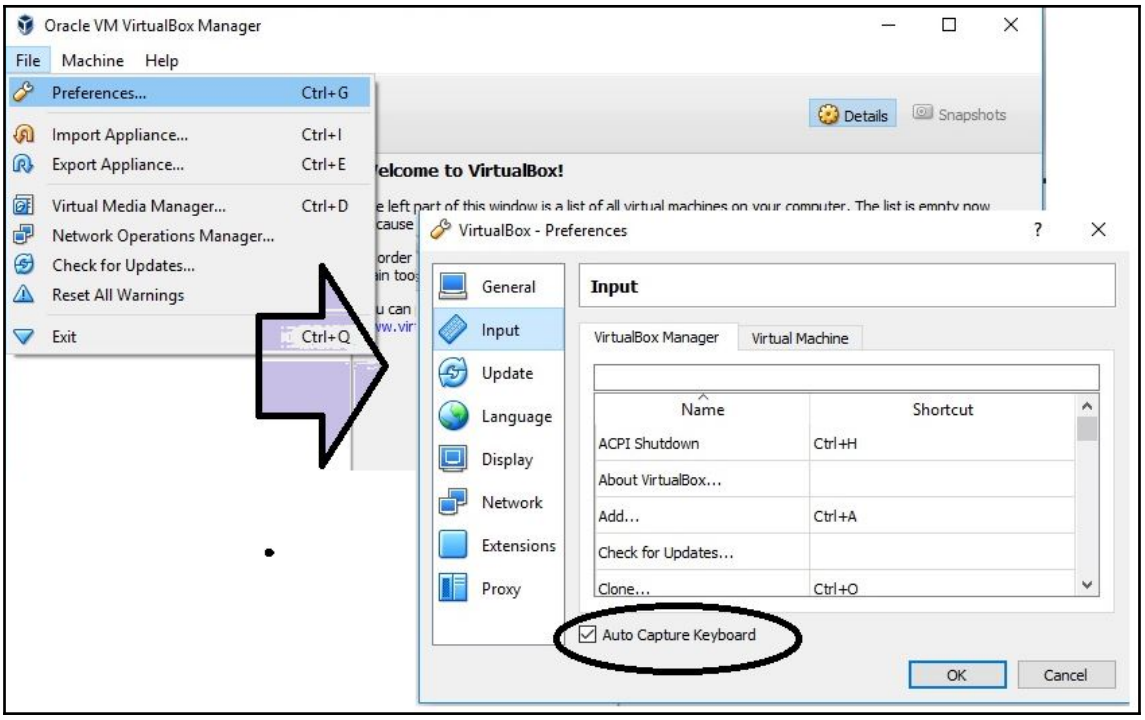

Oracle VM VirtualBox 5.2.8 Setup
Welcome to the Oracle VM VirtualBox 5.2.8 Setup Wizard
The Setup Wizard will install Oracle VM VirtualBox 5.2.8 on your computer. Click Next to continue or Cancel to exit the Setup Wizard.
Version 5.2.8
Next >
Cancel

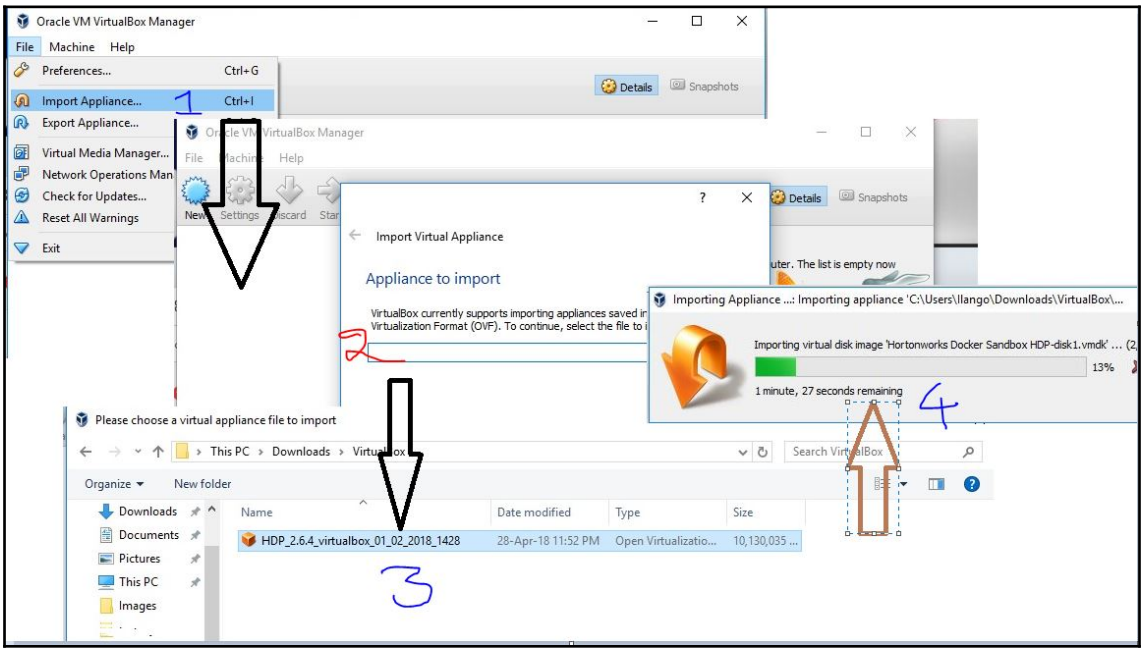

Oracle VM VirtualBox 5.2.8 Setup
Change current destination folder
Browse to the destination folder
Look in:
VirtualBox
doc
drivers
nls
platforms
sdk
x86
Folder name:
C:\Oracle\VirtualBox\
Cancel
OK

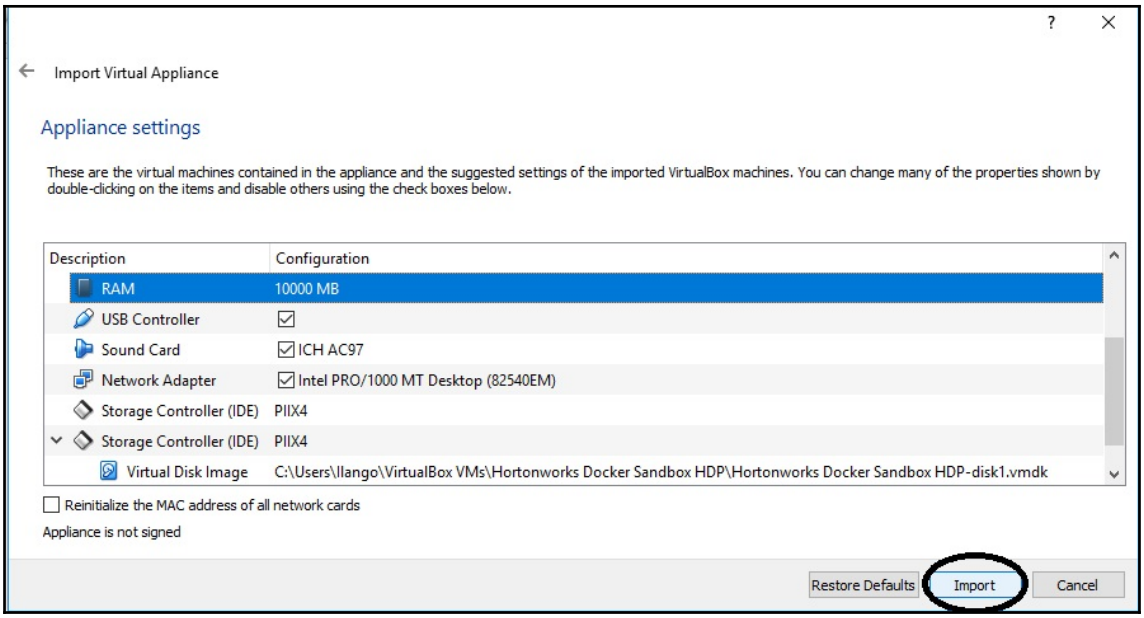

Oracle VM VirtualBox 5.2.8 Setup
Ready to Install
The Setup Wizard is ready to begin the Custom installation.
Click Install to begin the installation. If you want to review or change any of your installation settings, click Back. Click Cancel to exit the wizard.
Version 5.2.8
< Back
Install
Cancel

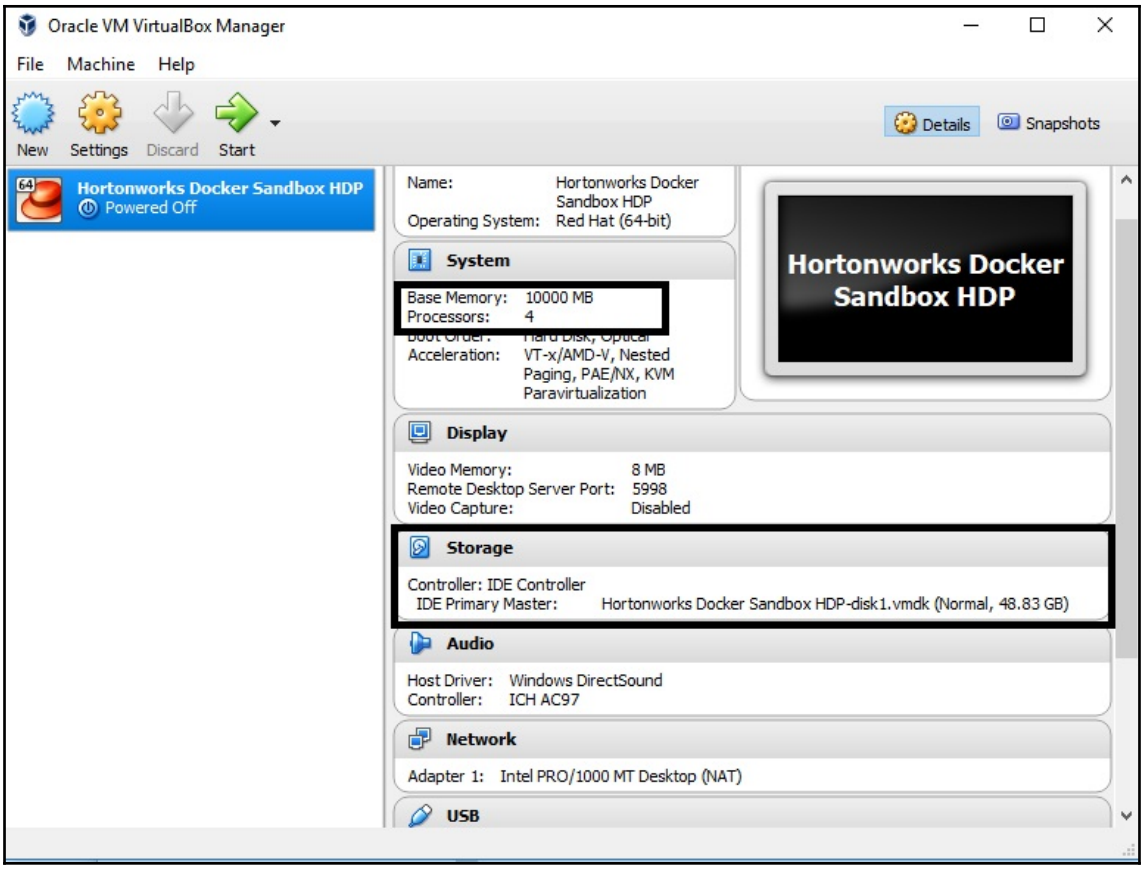

Oracle VM VirtualBox Manager
File
Machine
Help
New
Settings
Discard
Start
Details
Snapshots
CentOS
Inaccessible
vm
Inaccessible
vm_1
Inaccessible
The selected virtual machine is inaccessible. Please inspect the error message shown below and press the Refresh button if you want to repeat the accessibility check:
Runtime error opening 'C:\Users\Ilango\VirtualBox VMs\CentOS\CentOS.vbox' for reading: -103(Path not found.).
F:\tinderbox\win-5.1\src\VBox\Main\src-server\MachineImpl.cpp[749] (long __cdecl Machine::i_registeredInit(void)).
Result Code: E_FAIL (0x80004005)
Component: MachineWrap
Interface: IMachine {b2547866-a0a1-4391-8b86-6952d82efaa0}
Refresh

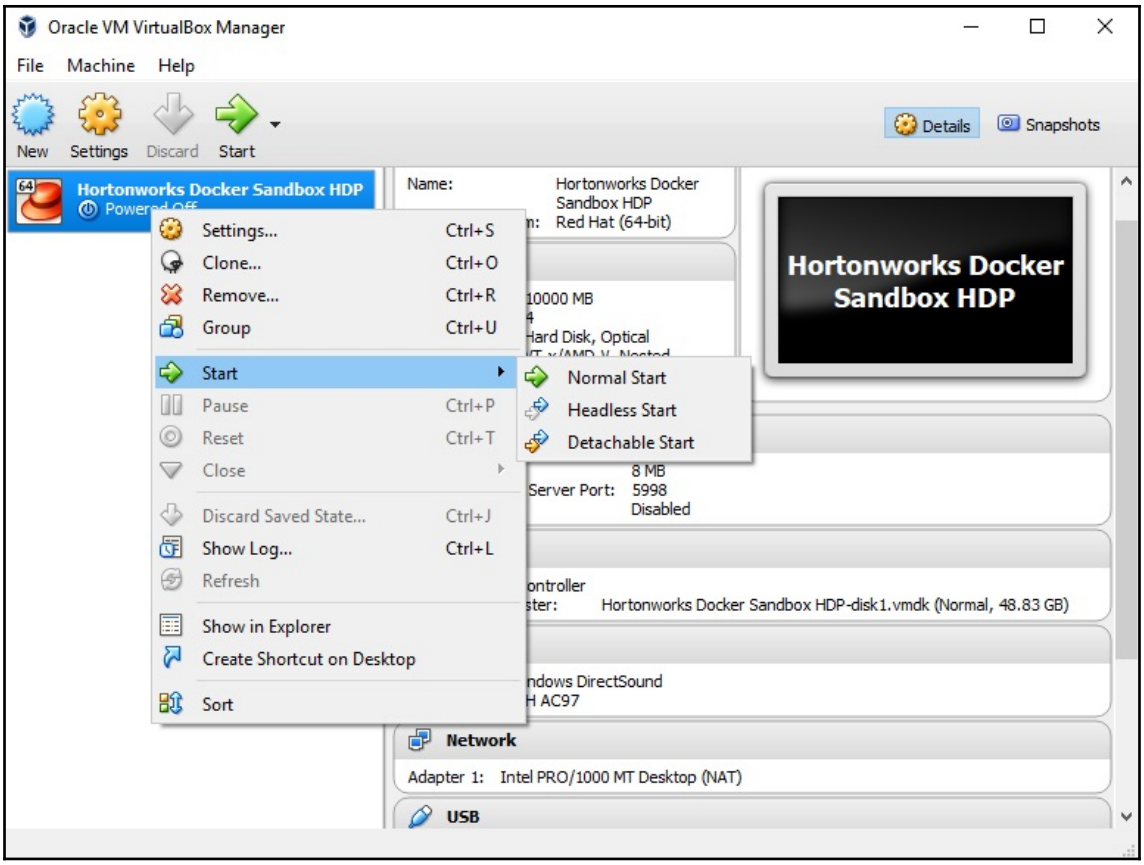

Oracle VM VirtualBox Manager
File
Machine
Help
New
Settings
Discard
Start
Details
Snapshots
vm_1
Inaccessible
The selected virtual machine is inaccessible. Please inspect the error message shown below and press the Refresh button if you want to repeat the accessibility check:
Runtime error opening 'C:\Users\Ilango\VirtualBox VMs\vm_1\vm_1.vbox' for reading:
-103(Path not found.).
F:\tinderbox\win-5.1\src\VBox\Main\src-server\MachineImpl.cpp[749] (long __cdecl
Machine::i_registeredInit(void)).
VirtualBox - Question
You are about to remove following inaccessible virtual machines from the machine list:
vm_1
Do you wish to proceed?
Remove
Cancel
Refresh

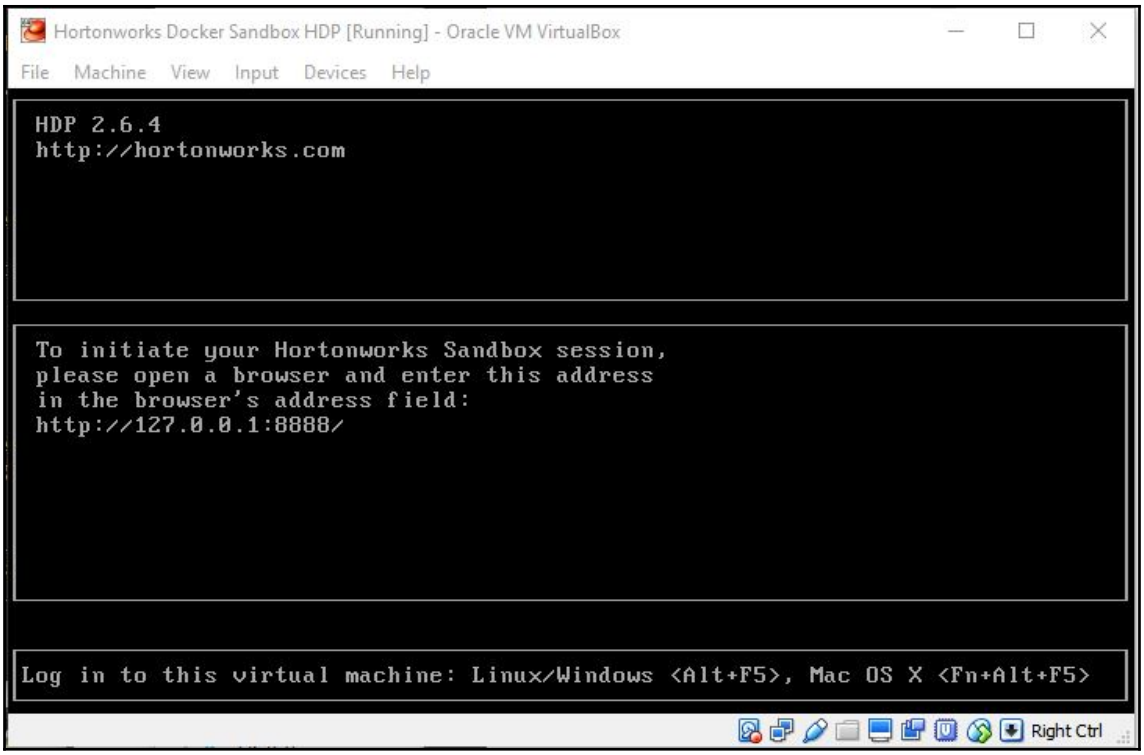

Oracle VM VirtualBox Manager
File
Machine
Help
Preferences... Ctrl+G
Import Appliance... Ctrl+I
Export Appliance... Ctrl+E
Virtual Media Manager... Ctrl+D
Network Operations Manager...
Check for Updates...
Reset All Warnings
Exit Ctrl+Q
Details
Snapshots
VirtualBox - Preferences
General
Input
Update
Language
Display
Network
Extensions
Proxy
Input
VirtualBox Manager
Virtual Machine
Name
Shortcut
ACPI Shutdown Ctrl+H
About VirtualBox...
Add... Ctrl+A
Check for Updates...
Clone... Ctrl+O
Auto Capture Keyboard
OK
Cancel

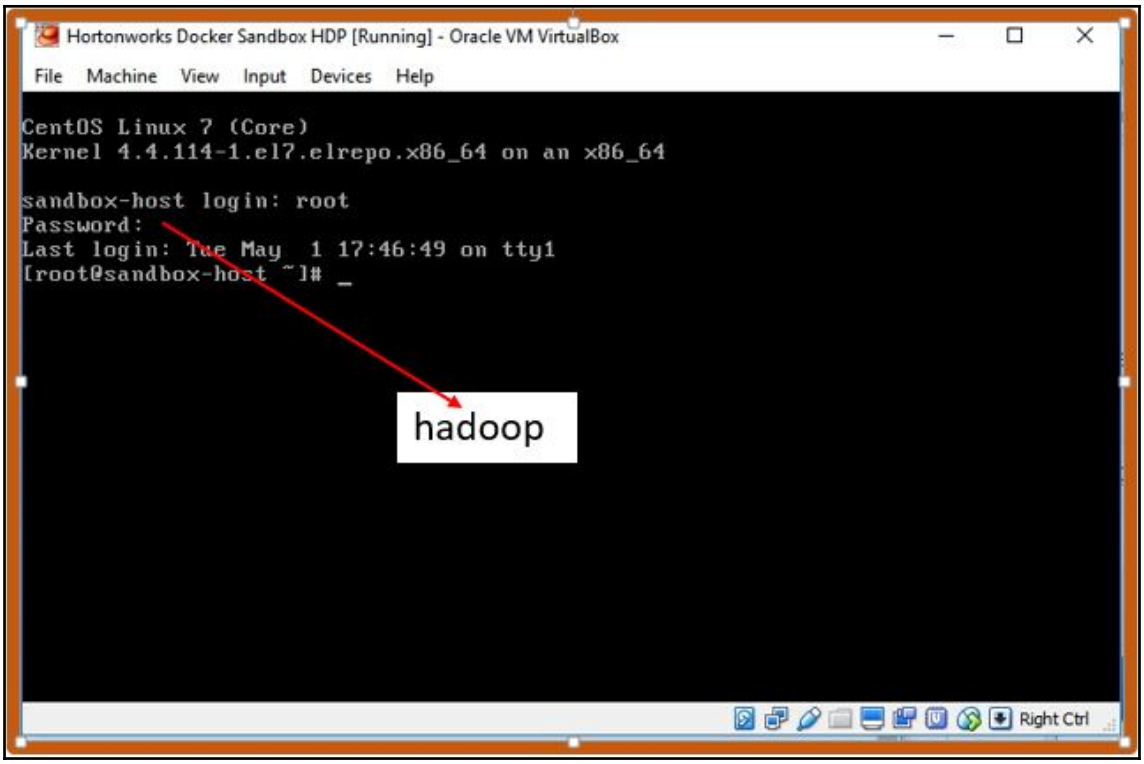

Oracle VM VirtualBox Manager
File
Machine
Help
Preferences... Ctrl+G
Import Appliance... Ctrl+I
Export Appliance...
Virtual Media Manager...
Network Operations Manager...
Check for Updates...
Reset All Warnings
Exit
1
Details
Snapshots
Import Virtual Appliance
Appliance to import
VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.
2
Importing Appliance ...: Importing appliance 'C:\Users\llango\Downloads\VirtualBox\...
Importing virtual disk image 'Hortonworks Docker Sandbox HDP-disk1.vmdk' ... (2
13%
1 minute, 27 seconds remaining
4
Please choose a virtual appliance file to import
This PC
Downloads
Search VirtualBox
Organize
New folder
Downloads
Documents
Pictures
This PC
Images
Name
Date modified
Type
Size
HDP_2.6.4_virtualbox_01_02_2018_1428
28-Apr-18 11:52 PM
Open Virtualizatio...
10,130,035 ...
3


Import Virtual Appliance
Appliance settings
These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.
Description
Configuration
RAM
10000 MB
USB Controller
Sound Card
ICH AC97
Network Adapter
Intel PRO/1000 MT Desktop (82540EM)
Storage Controller (IDE)
PIIX4
Storage Controller (IDE)
PIIX4
Virtual Disk Image
C:\Users\Ilango\VirtualBox VMs\Hortonworks Docker Sandbox HDP\Hortonworks Docker Sandbox HDP-disk1.vmdk
Reinitialize the MAC address of all network cards
Appliance is not signed
Restore Defaults
Import
Cancel

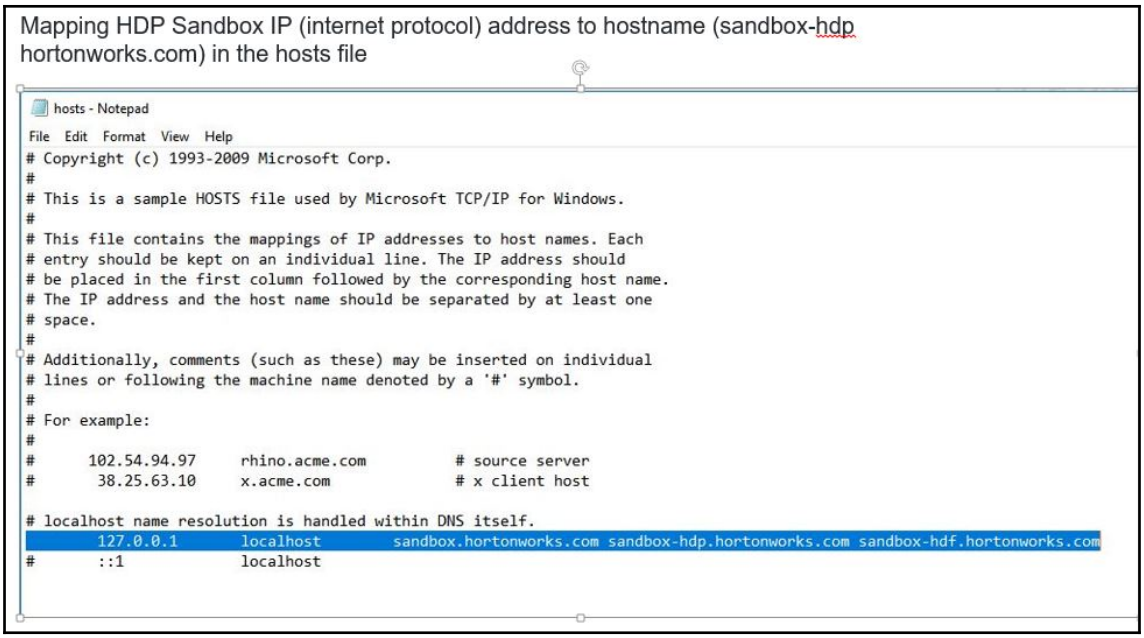

Oracle VM VirtualBox Manager
File
Machine
Help
New
Settings
Discard
Start
Details
Snapshots
Hortonworks Docker Sandbox HDP
Powered Off
Name: Hortonworks Docker Sandbox HDP
Operating System: Red Hat (64-bit)
System
Base Memory: 10000 MB
Processors: 4
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX, KVM Paravirtualization
Hortonworks Docker Sandbox HDP
Display
Video Memory: 8 MB
Remote Desktop Server Port: 5998
Video Capture: Disabled
Storage
Controller: IDE Controller
IDE Primary Master: Hortonworks Docker Sandbox HDP-disk1.vmdk (Normal, 48.83 GB)
Audio
Host Driver: Windows DirectSound
Controller: ICH AC97
Network
Adapter 1: Intel PRO/1000 MT Desktop (NAT)
USB

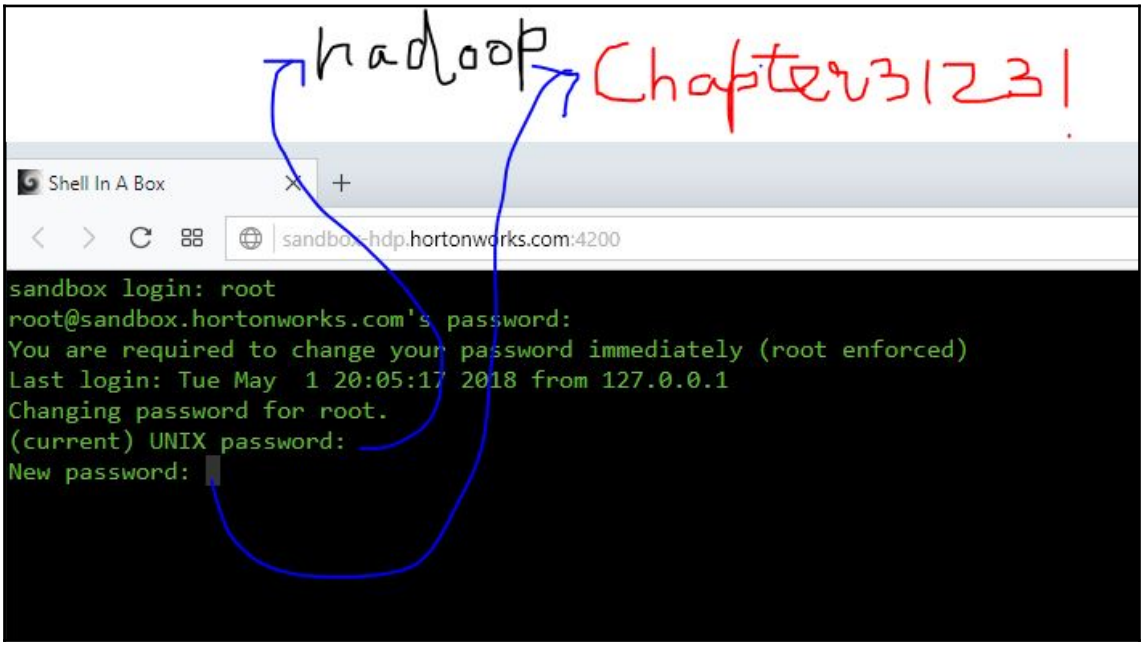

Oracle VM VirtualBox Manager
File
Machine
Help
New
Settings
Discard
Start
Details
Snapshots
Hortonworks Docker Sandbox HDP
Powered Off
Settings... Ctrl+S
Clone... Ctrl+O
Remove... Ctrl+R
Group Ctrl+U
Start
Pause Ctrl+P
Reset Ctrl+T
Close
Discard Saved State... Ctrl+J
Show Log... Ctrl+L
Refresh
Show in Explorer
Create Shortcut on Desktop
Sort
Normal Start
Headless Start
Detachable Start
Name: Hortonworks Docker Sandbox HDP
Red Hat (64-bit)
10000 MB
Hard Disk, Optical
8 MB
Server Port: 5998
Disabled
Hortonworks Docker Sandbox HDP-disk1.vmdk (Normal, 48.83 GB)
Network
Adapter 1: Intel PRO/1000 MT Desktop (NAT)
USB

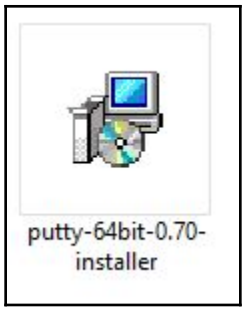

Hortonworks Docker Sandbox HDP [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
HDP 2.6.4
http://hortonworks.com
To initiate your Hortonworks Sandbox session,
please open a browser and enter this address
in the browser's address field:
http://127.0.0.1:8888/
Log in to this virtual machine: Linux/Windows <Alt+F5>, Mac OS X <Fn+Alt+F5>
Right Ctrl

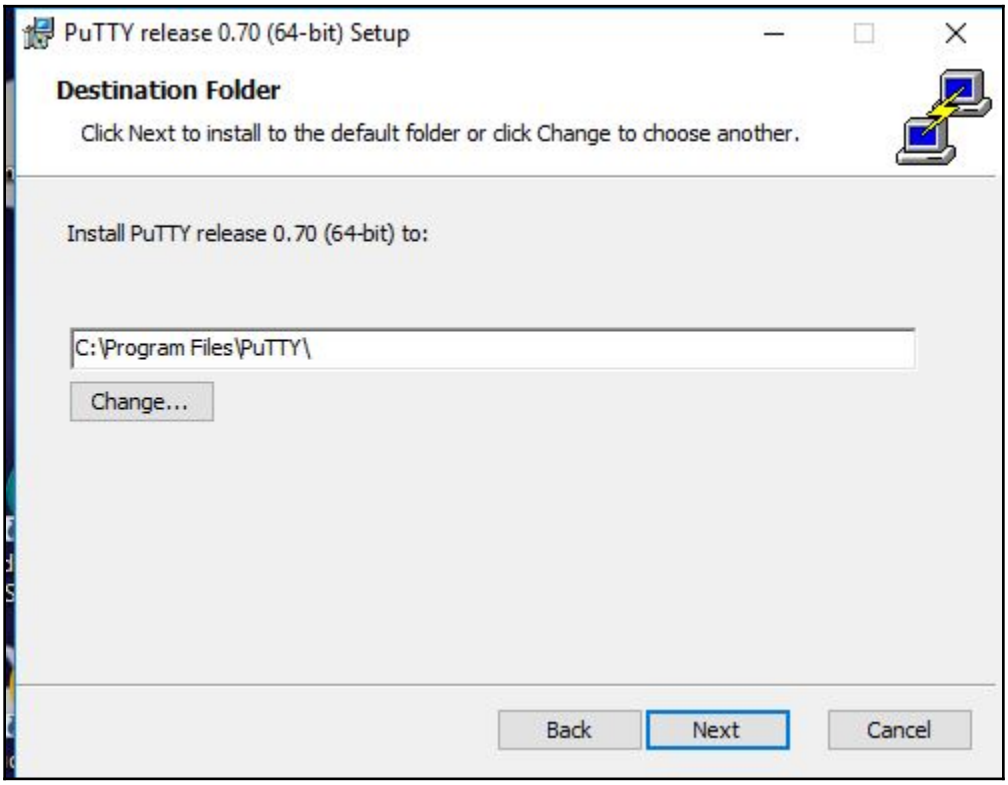

Hortonworks Docker Sandbox HDP [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
CentOS Linux 7 (Core)
Kernel 4.4.114-1.el7.elrepo.x86_64 on an x86_64
sandbox-host login: root
Password:
Last login: Tue May 1 17:46:49 on tty1
[root@sandbox-host ~]# _
hadoop
Right Ctrl

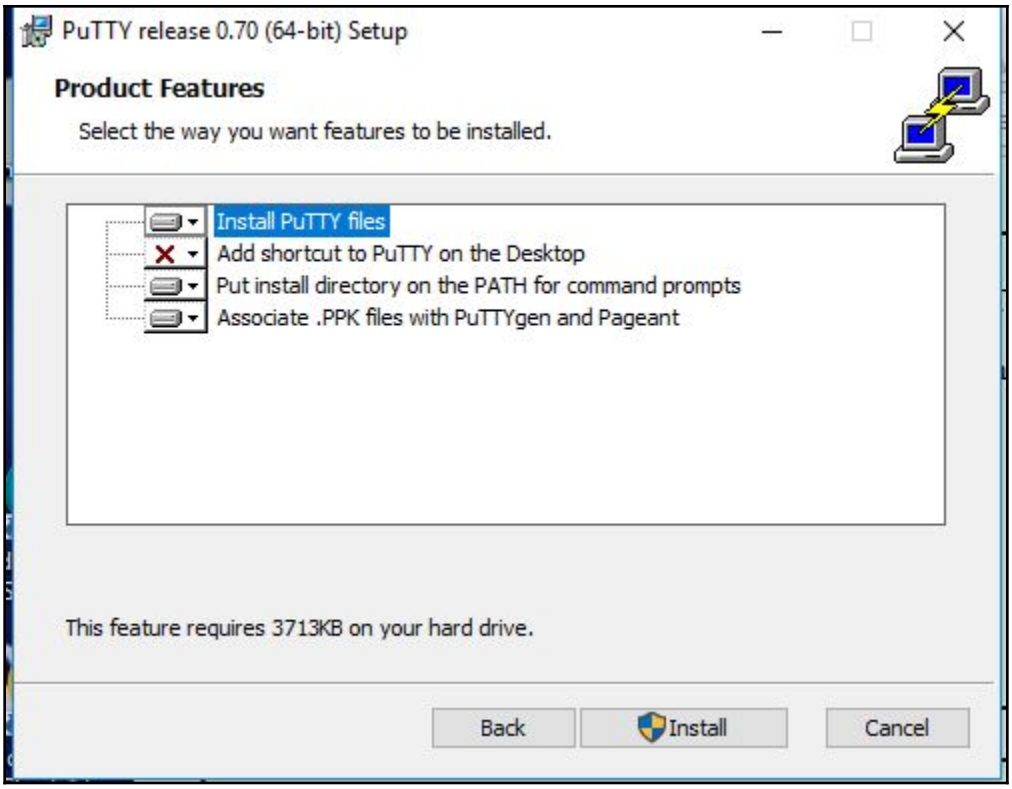
Mapping HDP Sandbox IP (internet protocol) address to hostname (sandbox-hdp hortonworks.com) in the hosts file

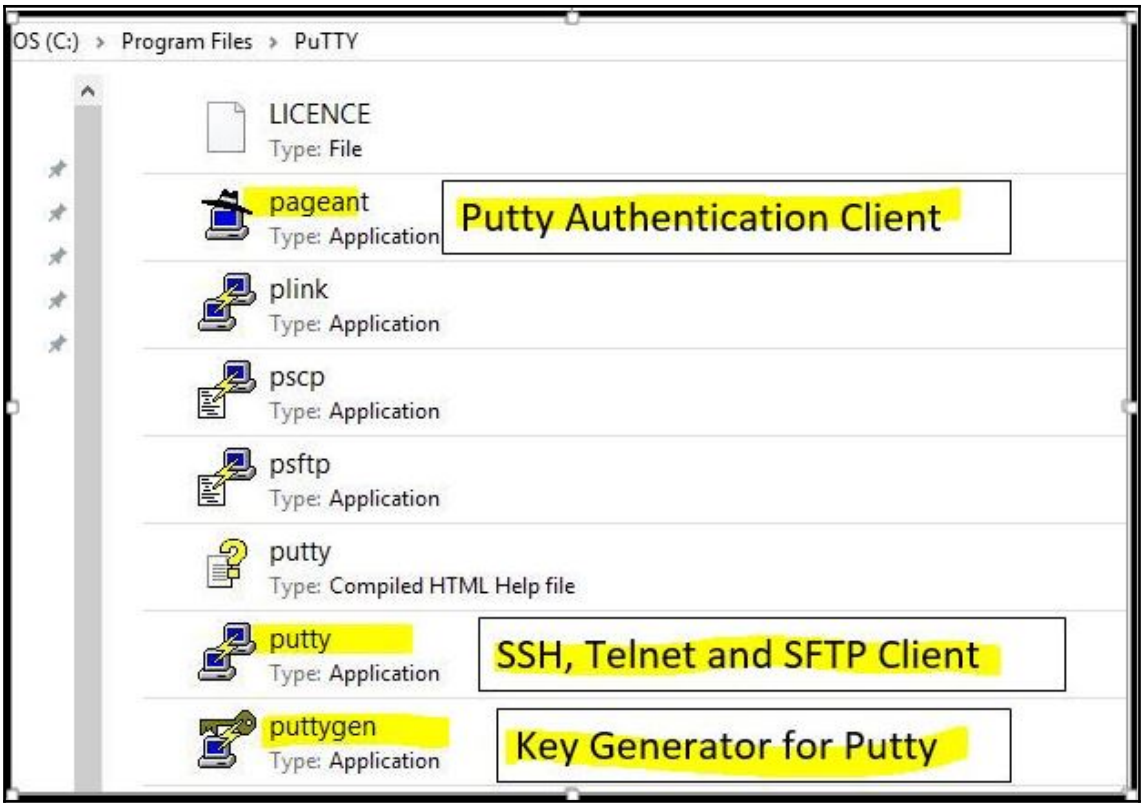


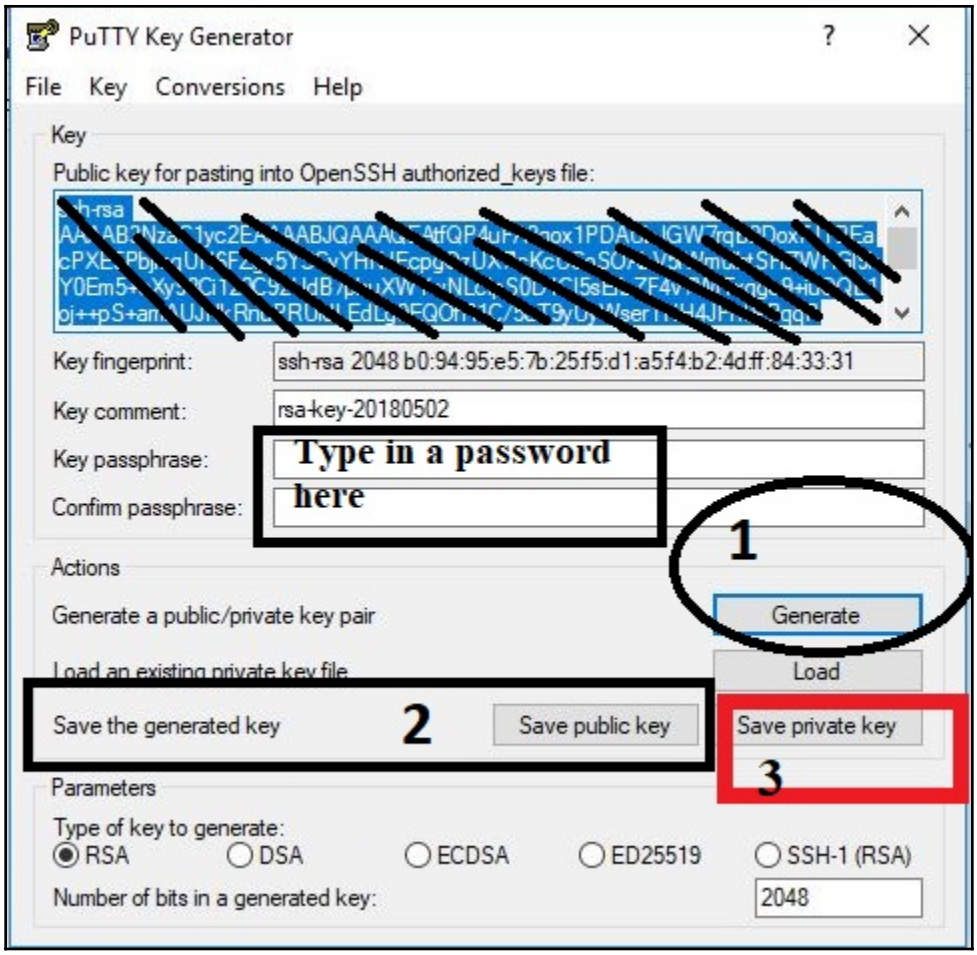


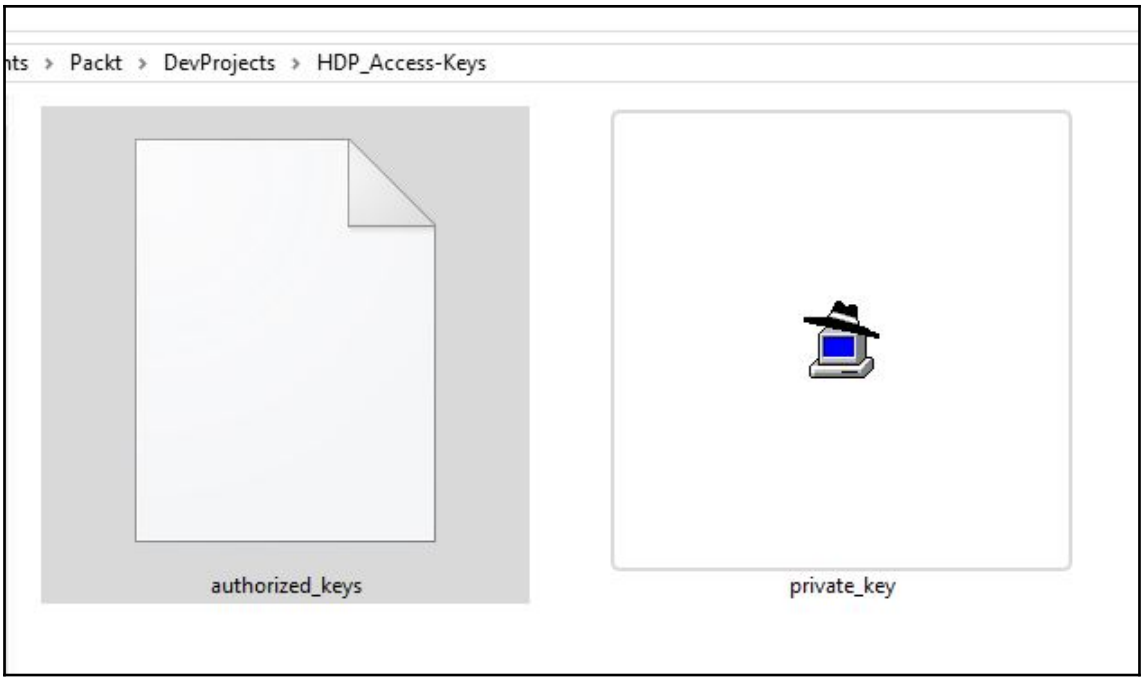

putty-64bit-0.70-
installer

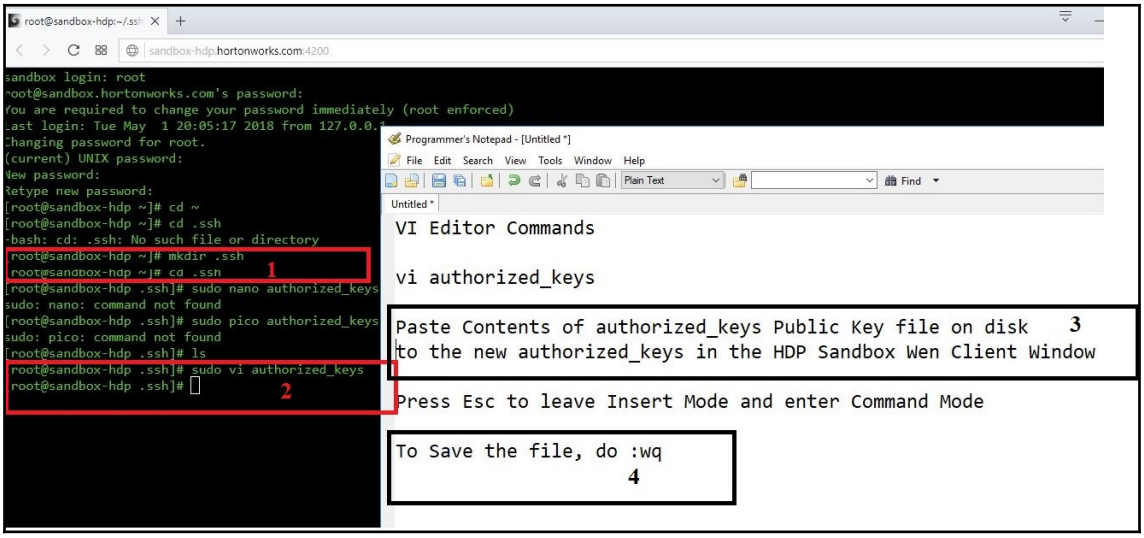

PuTTY release 0.70 (64-bit) Setup
Destination Folder
Click Next to install to the default folder or click Change to choose another.
Install PuTTY release 0.70 (64-bit) to:
C:\Program Files\PuTTY\
Change...
Back
Next
Cancel

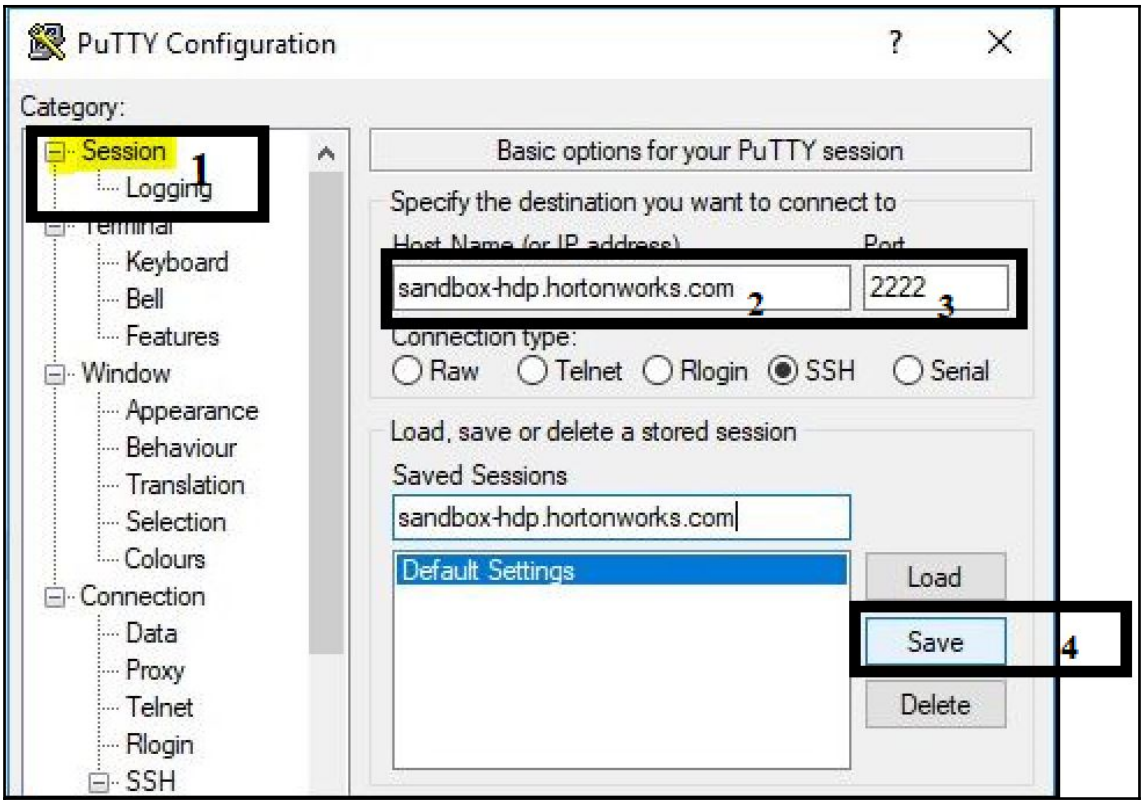

PuTTY release 0.70 (64-bit) Setup
Product Features
Select the way you want features to be installed.
Install PuTTY files
Add shortcut to PuTTY on the Desktop
Put install directory on the PATH for command prompts
Associate .PPK files with PuTTYgen and Pageant
This feature requires 3713KB on your hard drive.
Back
Install
Cancel

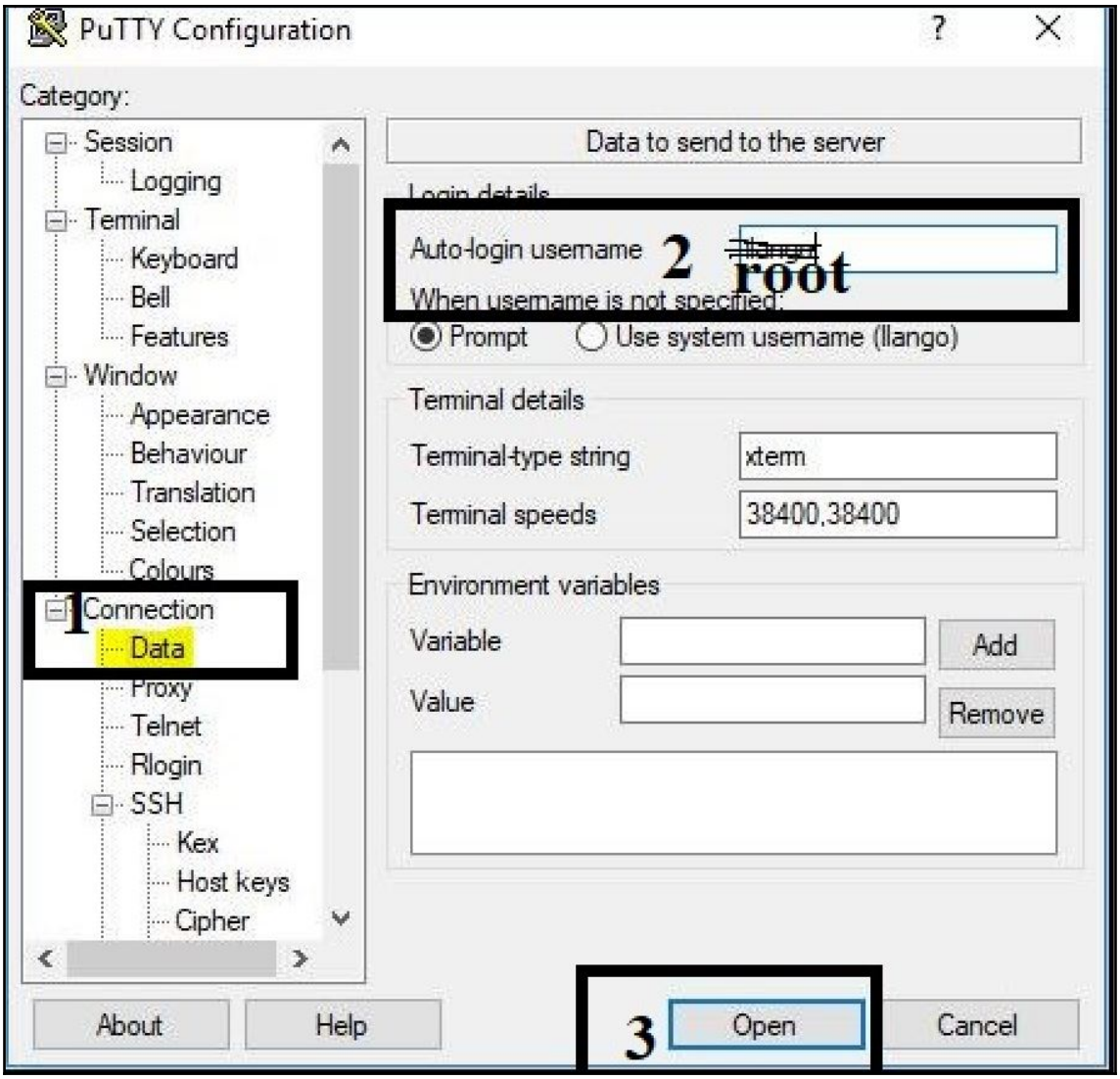

OS (C:) › Program Files › PuTTY
LICENCE
Type: File
pageant
Type: Application
Putty Authentication Client
plink
Type: Application
pscp
Type: Application
psftp
Type: Application
putty
Type: Compiled HTML Help file
putty
Type: Application
SSH, Telnet and SFTP Client
puttygen
Type: Application
Key Generator for Putty

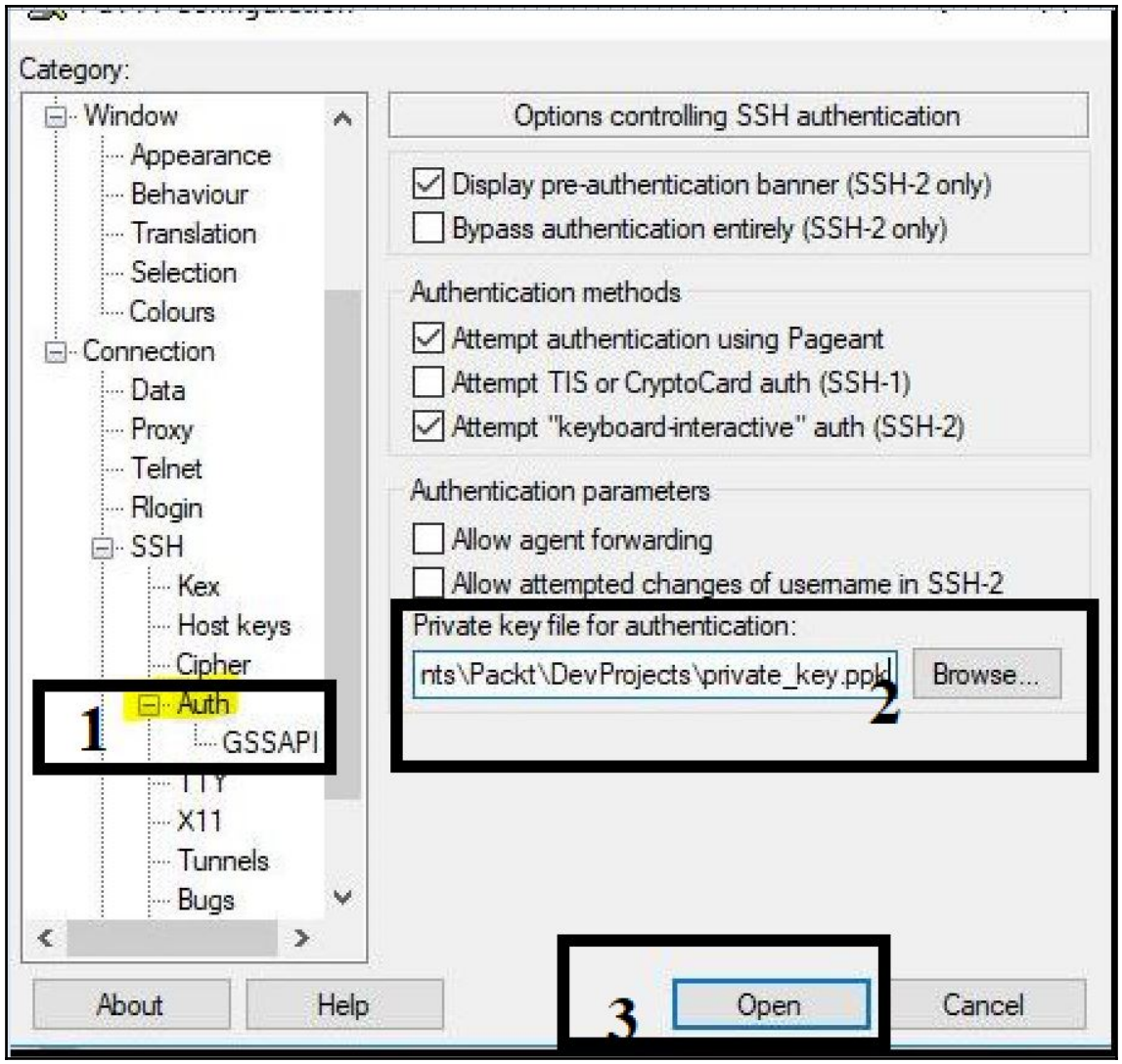

PuTTY Key Generator
File Key Conversions Help
Key
Public key for pasting into OpenSSH authorized_keys file:
ssh-rsa
Key fingerprint:
ssh-rsa 2048 b0:94:95:e5:7b:25f5:d1:a5f4:b2:4d:ff:84:33:31
Key comment:
rsa-key-20180502
Key passphrase:
Confirm passphrase:
Type in a password here
1
Actions
Generate a public/private key pair
Generate
Load an existing private key file
Load
Save the generated key
2
Save public key
Save private key
3
Parameters
Type of key to generate:
RSA
DSA
ECDSA
ED25519
SSH-1 (RSA)
Number of bits in a generated key:
2048

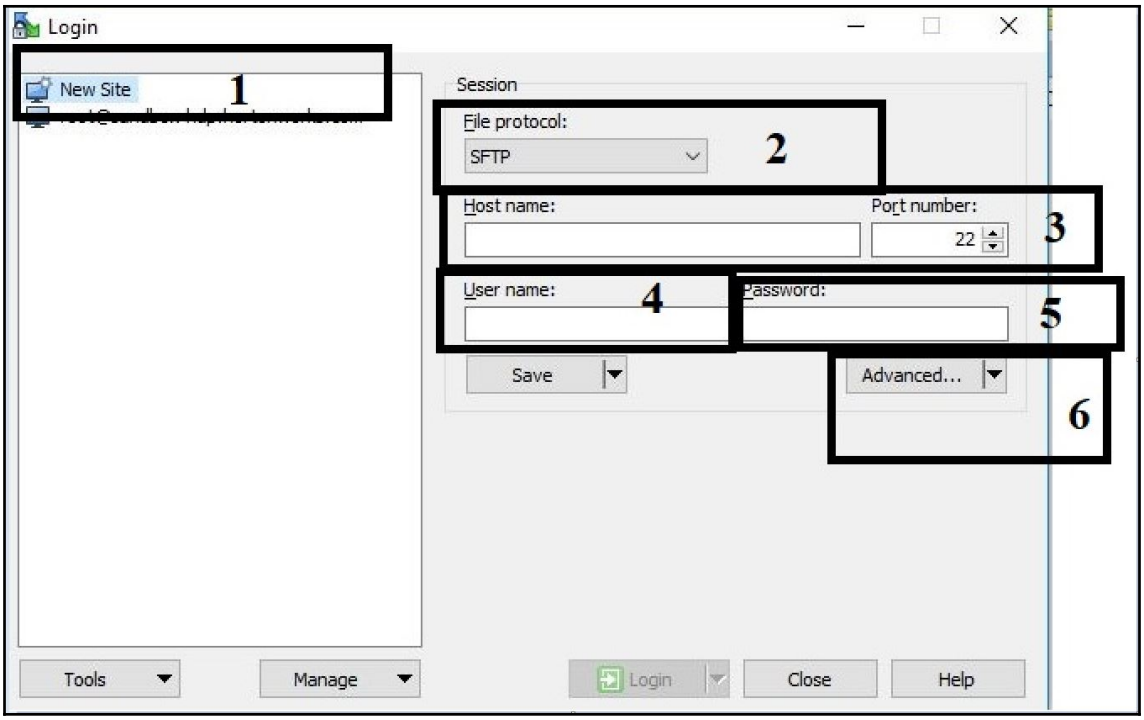

nts › Packt › DevProjects › HDP_Access-Keys
authorized_keys
private_key

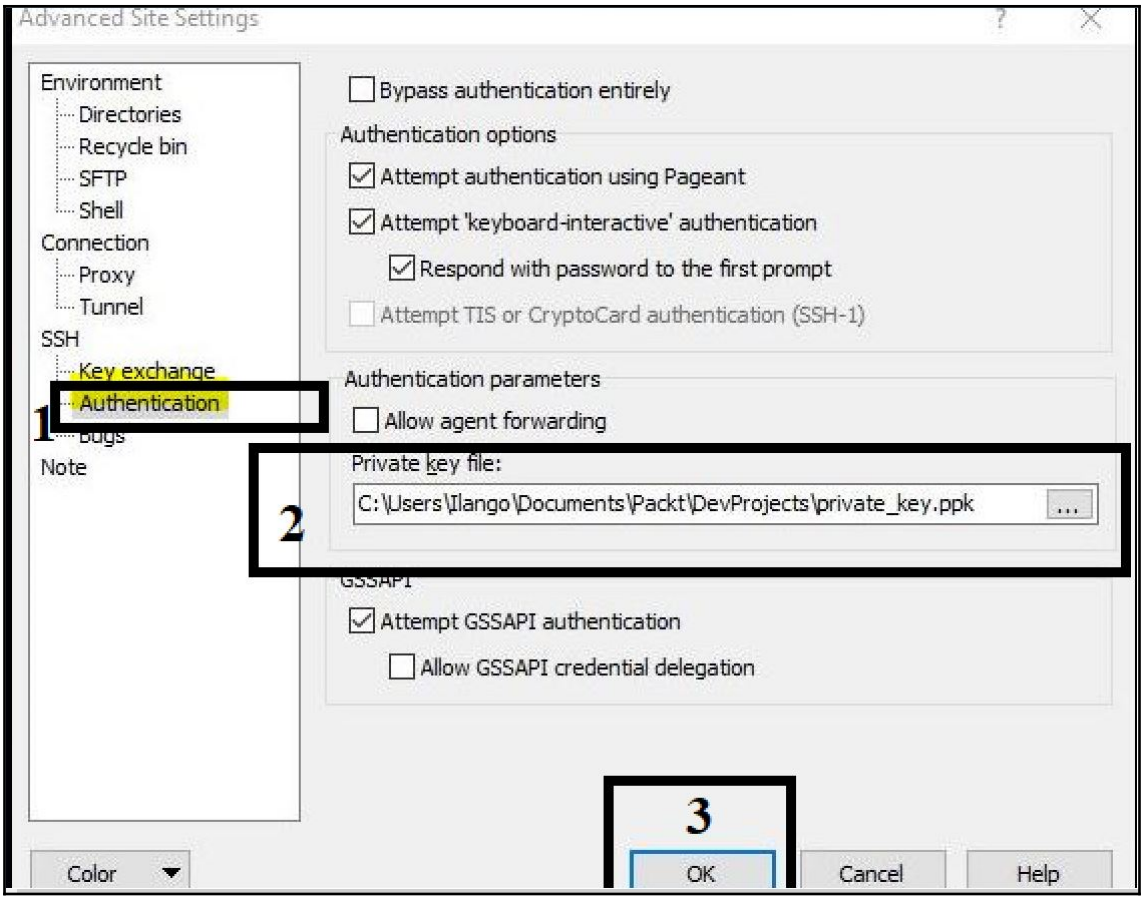

root@sandbox-hdp:~/ssh
sandbox-hdp.hortonworks.com:4200
sandbox login: root
root@sandbox.hortonworks.com's password:
You are required to change your password immediately (root enforced)
Last login: Tue May 1 20:05:17 2018 from 127.0.0.1
Changing password for root.
(current) UNIX password:
New password:
Retype new password:
[root@sandbox-hdp ~]# cd ~
[root@sandbox-hdp ~]# cd .ssh
-bash: cd: .ssh: No such file or directory
[root@sandbox-hdp ~]# mkdir .ssh
[root@sandbox-hdp ~]# cd .ssh
1
[root@sandbox-hdp .ssh]# sudo nano authorized_keys
sudo: nano: command not found
[root@sandbox-hdp .ssh]# sudo pico authorized_keys
sudo: pico: command not found
[root@sandbox-hdp .ssh]# ls
[root@sandbox-hdp .ssh]# sudo vi authorized_keys
[root@sandbox-hdp .ssh]#
2
Programmer's Notepad - [Untitled *]
File Edit Search View Tools Window Help
Plain Text
Find
Untitled *
VI Editor Commands
vi authorized_keys
Paste Contents of authorized_keys Public Key file on disk 3
to the new authorized_keys in the HDP Sandbox Wen Client Window
Press Esc to leave Insert Mode and enter Command Mode
To Save the file, do :wq
4

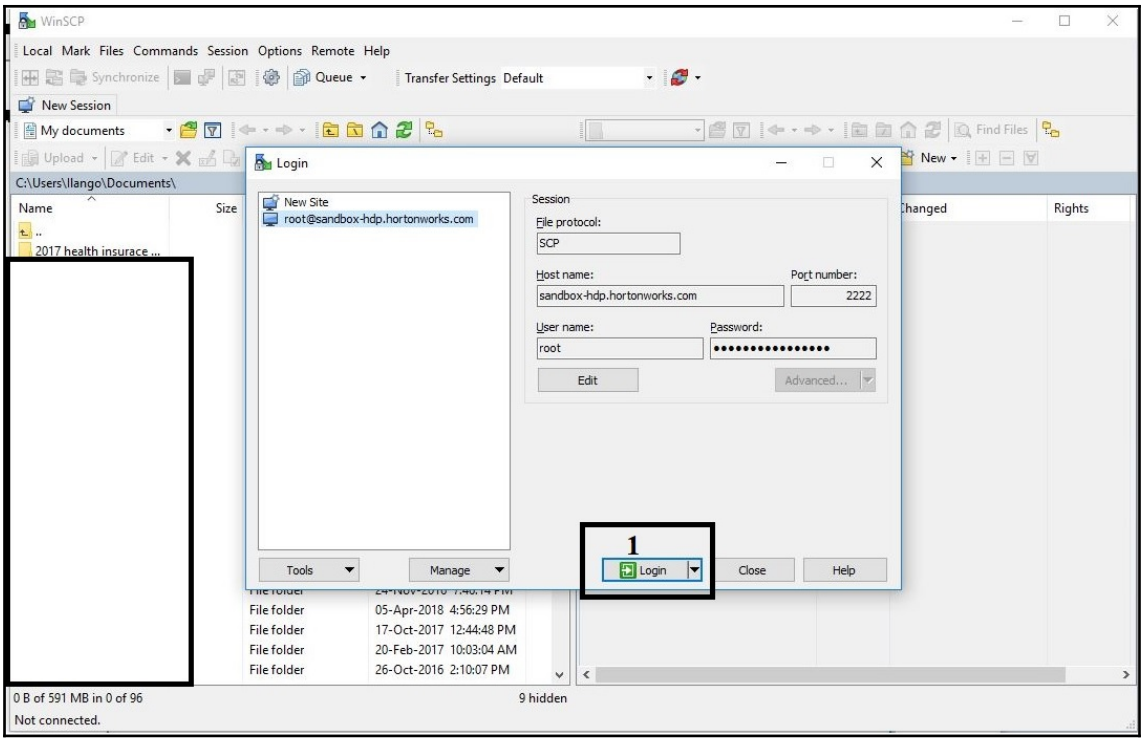

PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
1
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
sandbox-hdp.hortonworks.com
2
2222
3
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
sandbox-hdp.hortonworks.com
Default Settings
Load
Save
4
Delete

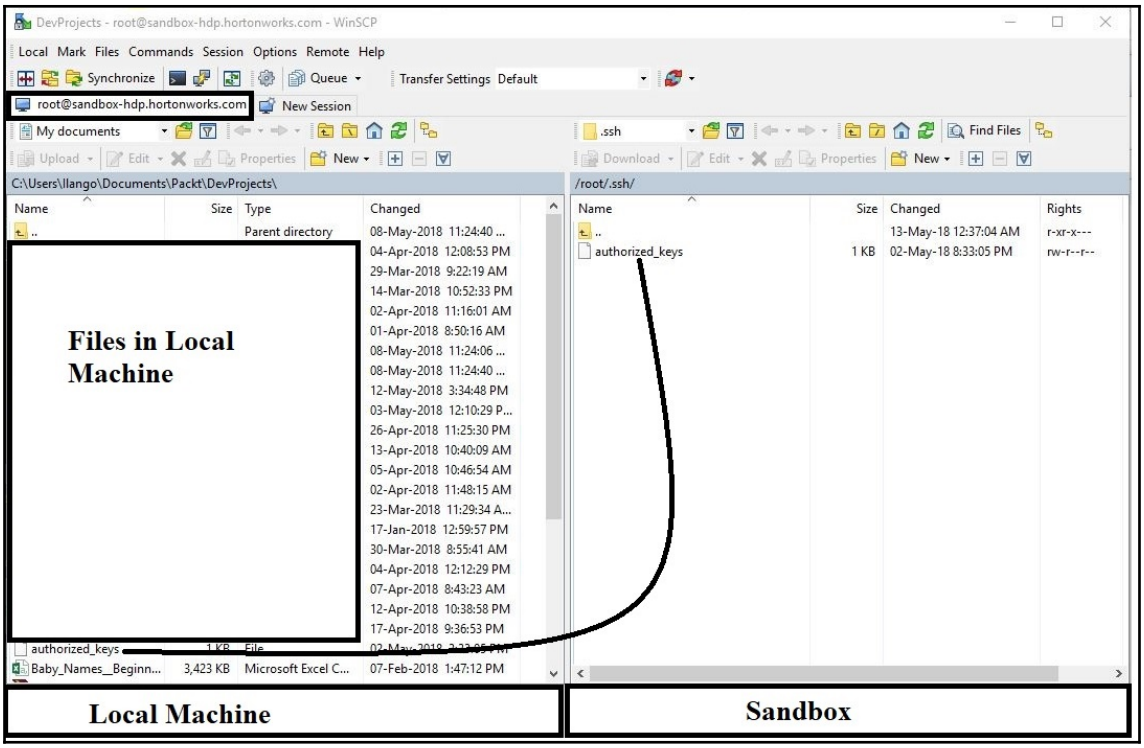

PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Kex
Host keys
Cipher
Data to send to the server
Login details
Auto-login username
root
When username is not specified:
Prompt
Use system username (llango)
Terminal details
Terminal-type string
xterm
Terminal speeds
38400,38400
Environment variables
Variable
Add
Value
Remove
About
Help
Open
Cancel
1
2
3


Category:
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Kex
Host keys
Cipher
Auth
GSSAPI
X11
Tunnels
Bugs
Options controlling SSH authentication
Display pre-authentication banner (SSH-2 only)
Bypass authentication entirely (SSH-2 only)
Authentication methods
Attempt authentication using Pageant
Attempt TIS or CryptoCard auth (SSH-1)
Attempt "keyboard-interactive" auth (SSH-2)
Authentication parameters
Allow agent forwarding
Allow attempted changes of username in SSH-2
Private key file for authentication:
nts\Packt\DevProjects\private_key.ppk
Browse...
1
2
3
About
Help
Open
Cancel


Login
New Site
1
Session
File protocol:
SFTP
2
Host name:
Port number:
22
3
User name:
4
Password:
5
Save
Advanced...
6
Tools
Manage
Login
Close
Help


Advanced Site Settings
Environment
Directories
Recycle bin
SFTP
Shell
Connection
Proxy
Tunnel
SSH
Key exchange
Authentication
Bugs
Note
1
Bypass authentication entirely
Authentication options
Attempt authentication using Pageant
Attempt 'keyboard-interactive' authentication
Respond with password to the first prompt
Attempt TIS or CryptoCard authentication (SSH-1)
Authentication parameters
Allow agent forwarding
Private key file:
C:\Users\Ilango\Documents\Packt\DevProjects\private_key.ppk
...
2
GSSAPI
Attempt GSSAPI authentication
Allow GSSAPI credential delegation
3
Color
OK
Cancel
Help


WinSCP
Local Mark Files Commands Session Options Remote Help
Synchronize
Queue
Transfer Settings Default
New Session
My documents
Upload
Edit
Find Files
New
C:\Users\llango\Documents\
Name
Size
Changed
Rights
..
2017 health insurace ...
Login
New Site
root@sandbox-hdp.hortonworks.com
Session
File protocol:
SCP
Host name:
Port number:
sandbox-hdp.hortonworks.com
2222
User name:
Password:
root
Edit
Advanced...
Tools
Manage
1
Login
Close
Help
File folder 05-Apr-2018 4:56:29 PM
File folder 17-Oct-2017 12:44:48 PM
File folder 20-Feb-2017 10:03:04 AM
File folder 26-Oct-2016 2:10:07 PM
0 B of 591 MB in 0 of 96
9 hidden
Not connected.

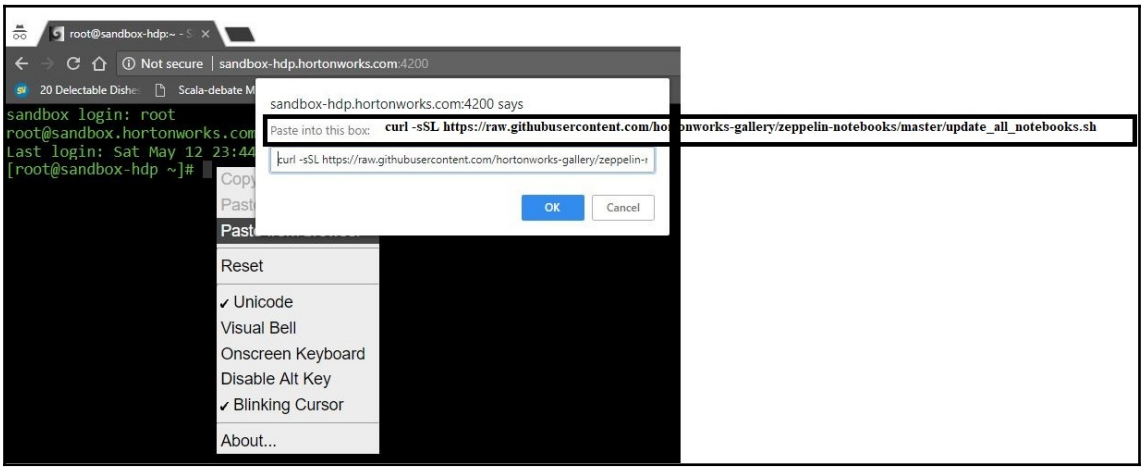



```
[root@sandbox-hdp ~]# python
Python 2.6.6 (r266:84292, Aug 18 2016, 15:13:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-17)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>>
[root@sandbox-hdp ~]# curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
```

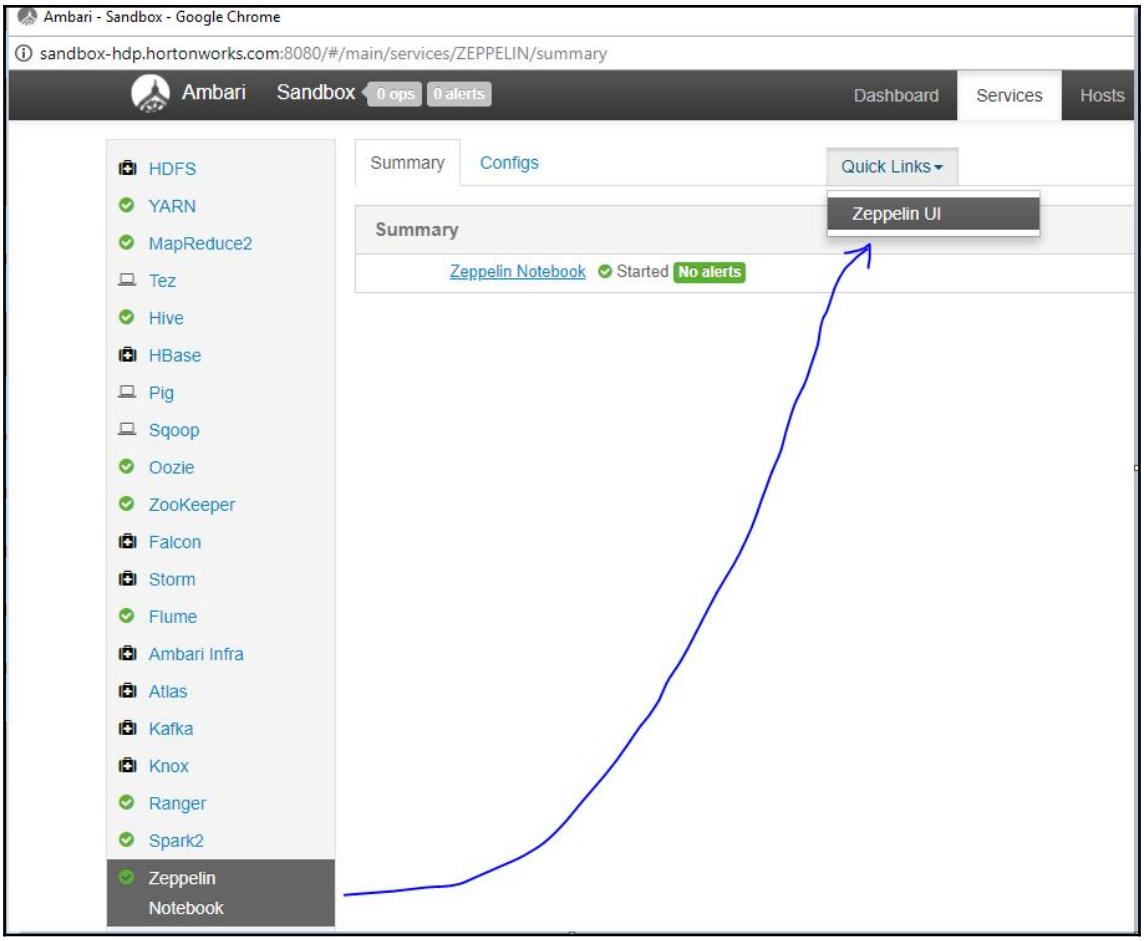
```
[root@sandbox-hdp bin]# bash Miniconda2-latest-Linux-x86_64.sh

Welcome to Miniconda2 4.5.1

In order to continue the installation process, please review the license
agreement.
Please, press ENTER to continue
>>>
```

```
sandbox login: root
root@sandbox.hortonworks.com's password:
root@sandbox.hortonworks.com's password:
Last login: Tue May  1 20:58:12 2018 from 127.0.0.1
[root@sandbox-hdp ~]# python
Python 2.7.14 |Anaconda, Inc.| (default, Mar 27 2018, 17:29:31)
[GCC 7.2.0] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

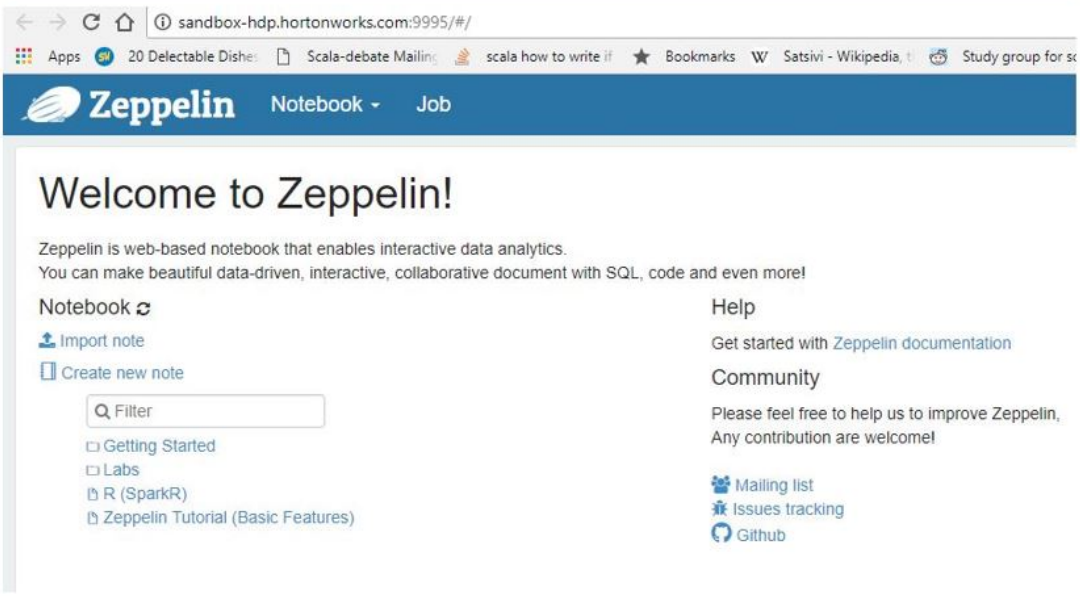
```
[root@sandbox-hdp ~]# curl -V/ --version
sh2/1.4.2.7 (x86_64-redhat-linux-gnu) libcurl/7.19.7 NSS/3.27.1 zlib/1.2.3 libidn/1.18 libs
Protocols: tftp ftp telnet dict ldap ldaps http file https ftps scp sftp

Features: GSS-Negotiate IDN IPv6 Largefile NTLM SSL libz
[root@sandbox-hdp ~]# curl -V/ --version
sh2/1.4.2.7 (x86_64-redhat-linux-gnu) libcurl/7.19.7 NSS/3.27.1 zlib/1.2.3 libidn/1.18 libs
Protocols: tftp ftp telnet dict ldap ldaps http file https ftps scp sftp

Features: GSS-Negotiate IDN IPv6 Largefile NTLM SSL libz
[root@sandbox-hdp ~]#
```

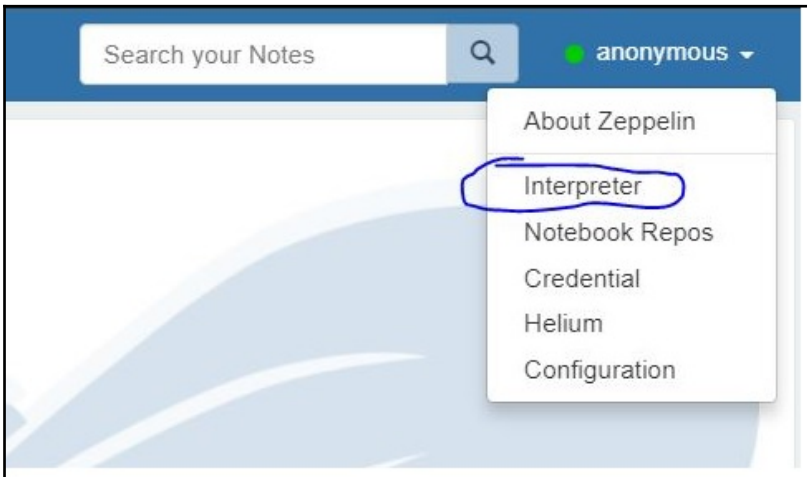


```
[root@sandbox-hdp ~]# curl -sSL https://raw.githubusercontent.com/hortonworks-gallery/zeppelin-not
ebooks/master/update_all_notebooks.sh
if [ -d "/usr/hdp/current/zeppelin-server/notebook" ]; then
  NOTES_DIR="/usr/hdp/current/zeppelin-server/notebook"
else
  NOTES_DIR="/usr/hdp/current/zeppelin-server/lib/notebook"
fi

cd $NOTES_DIR
OLD_DIR=`date +%Y%m%d-%H%M%S`
mkdir old_${OLD_DIR}
mv 2* old_${OLD_DIR}/
rm -rf zeppelin-notebooks/
git clone -q --progress https://github.com/hortonworks-gallery/zeppelin-notebooks.git
/bin/mv -f zeppelin-notebooks/* ./
chown -R zeppelin:hadoop *
echo Restarting Apache Zeppelin...
/usr/hdp/current/zeppelin-server/lib/bin/zeppelin-daemon.sh restart &> /dev/null
echo Done!
[root@sandbox-hdp ~]# sudo -u zeppelin -E sh
sh-4.1$
```

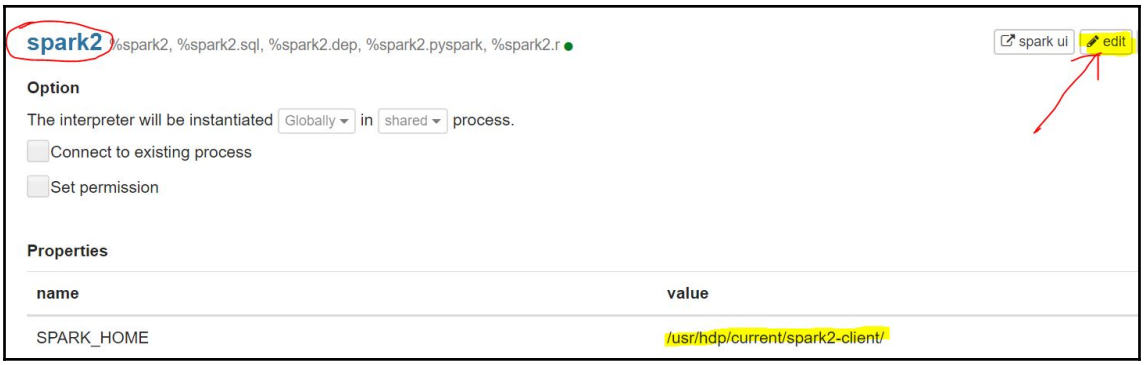

Ambari - Sandbox - Google Chrome
sandbox-hdp.hortonworks.com:8080/#/main/services/ZEPPELIN/summary
Ambari
Sandbox
0 ops
0 alerts
Dashboard
Services
Hosts
HDFS
YARN
MapReduce2
Tez
Hive
HBase
Pig
Sqoop
Oozie
ZooKeeper
Falcon
Storm
Flume
Ambari Infra
Atlas
Kafka
Knox
Ranger
Spark2
Zeppelin Notebook
Summary
Configs
Quick Links
Zeppelin UI
Summary
Zeppelin Notebook
Started
No alerts

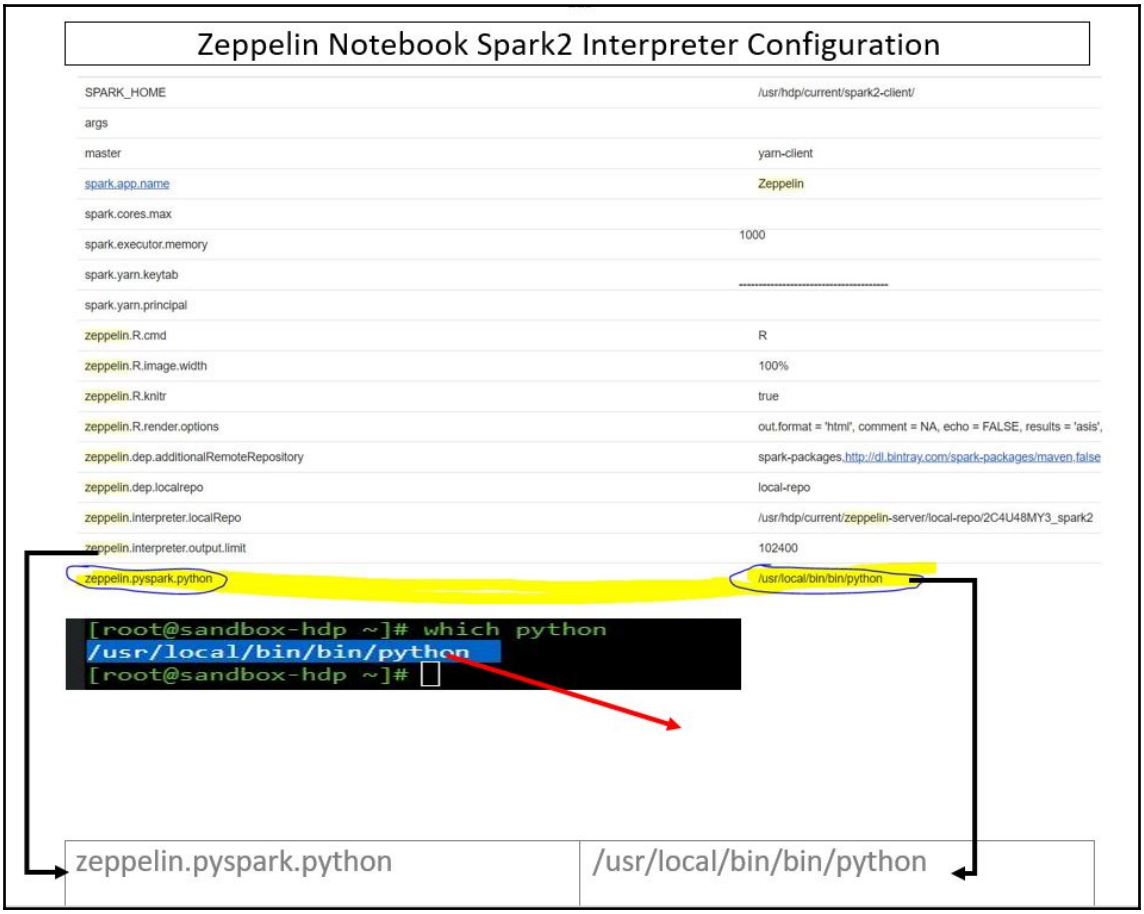
Zeppelin UI

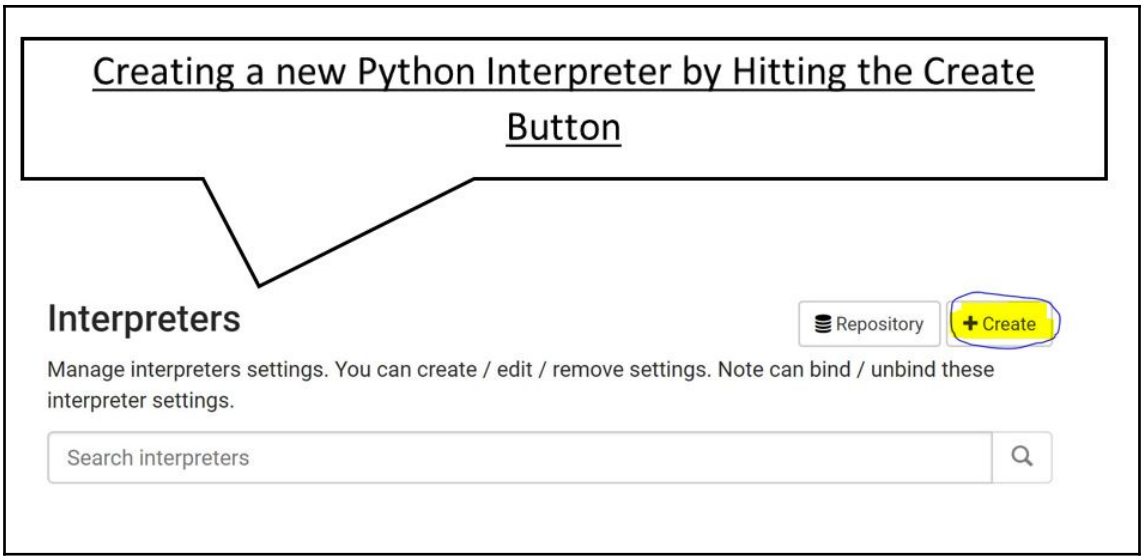


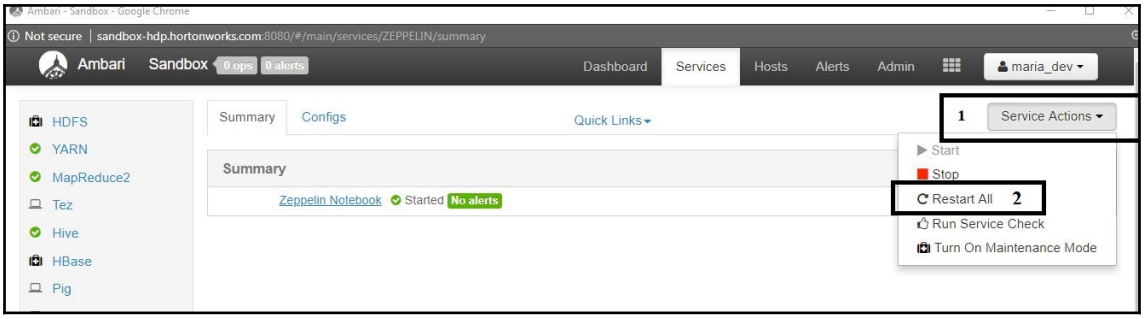


http://sandbox-hdp.hortonworks.com:9995/#/

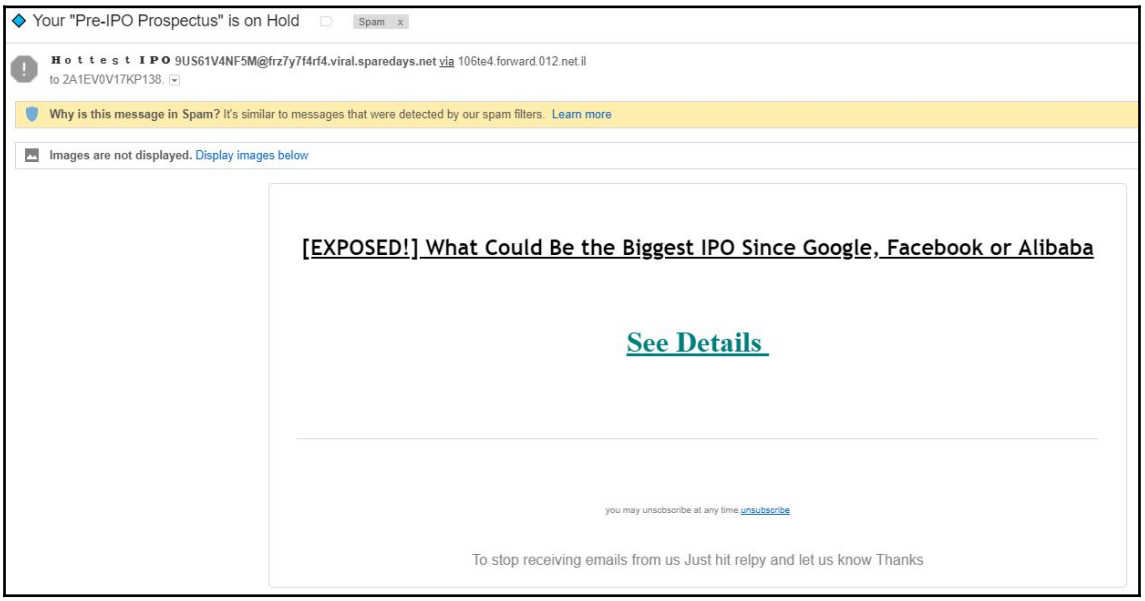

spark2
%spark2, %spark2.sql, %spark2.dep, %spark2.pyspark, %spark2.r
spark ui
edit
Option
The interpreter will be instantiated Globally in shared process.
Connect to existing process
Set permission
Properties
name
value
SPARK_HOME
/usr/hdp/current/spark2-client/


Zeppelin Notebook Spark2 Interpreter Configuration
SPARK_HOME
/usr/hdp/current/spark2-client/
args
master
yarn-client
spark.app.name
Zeppelin
spark.cores.max
spark.executor.memory
1000
spark.yarn.keytab
spark.yarn.principal
zeppelin.R.cmd
R
zeppelin.R.image.width
100%
zeppelin.R.knitr
true
zeppelin.R.render.options
out.format = 'html', comment = NA, echo = FALSE, results = 'asis',
zeppelin.dep.additionalRemoteRepository
spark-packages,http://dl.bintray.com/spark-packages/maven,false
zeppelin.dep.localrepo
local-repo
zeppelin.interpreter.localRepo
/usr/hdp/current/zeppelin-server/local-repo/2C4U48MY3_spark2
zeppelin.interpreter.output.limit
102400
zeppelin.pyspark.python
/usr/local/bin/bin/python
[root@sandbox-hdp ~]# which python
/usr/local/bin/bin/python
[root@sandbox-hdp ~]#
zeppelin.pyspark.python
/usr/local/bin/bin/python


Creating a new Python Interpreter by Hitting the Create Button
Interpreters
Repository
+ Create
Manage interpreters settings. You can create / edit / remove settings. Note can bind / unbind these interpreter settings.
Search interpreters


Ambari - Sandbox - Google Chrome
Not secure | sandbox-hdp.hortonworks.com:8080/#/main/services/ZEPPELIN/summary
Ambari
Sandbox
0 ops
0 alerts
Dashboard
Services
Hosts
Alerts
Admin
maria_dev
HDFS
YARN
MapReduce2
Tez
Hive
HBase
Pig
Summary
Configs
Quick Links
1
Service Actions
Start
Stop
Restart All
2
Run Service Check
Turn On Maintenance Mode
Summary
Zeppelin Notebook
Started
No alerts

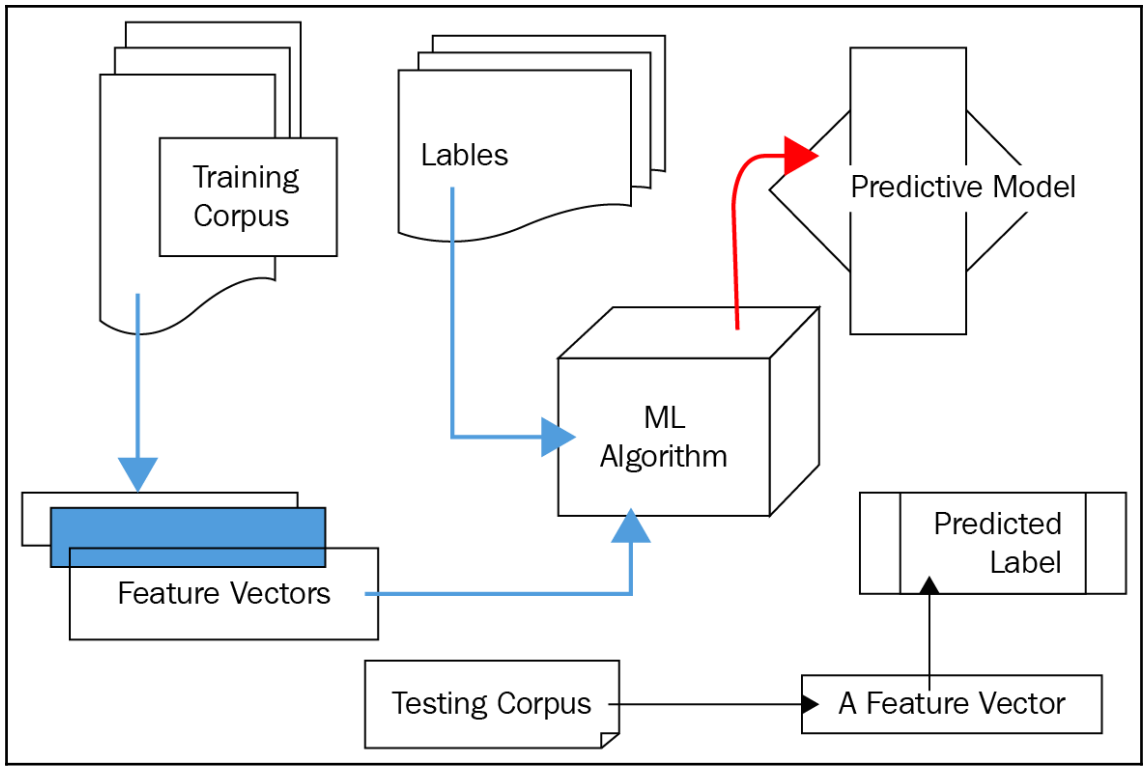
Chapter 4: Building a Spam Classification Pipeline



Choose the most appropriate answer:

What is Classification:	1) Supervised Learning Technique 2) Unsupervised Learning Technique 3) Semi-Supervised

# 1	Read in both Ham and Spam datasets
# 2	Create a Ham dataset that contains enough samples of spam and regular e-mails
# 3	A Feature Extraction step converts the email text corpus into useful features for training e.g. remove stop words, words frequency. Then we are able to evaluate these features with attribute selection technique.
# 4	Our combined dataset is split in two in a 80-20 ration as: • Training set, and a Testing set
# 5	We then set up a Pipeline consisting of multiple stages: • Tokenize the sentence (message features) of the dataset • Convert each dataset's words into feature vectors • Train a Regression Model
# 6	Make predictions on the test dataset
# 7	We finally use the trained model to classify Spam and Ham in new e-mails from the "wild"


Training
Corpus
Lables
Predictive Model
ML
Algorithm
Feature Vectors
Predicted
Label
Testing Corpus
A Feature Vector

Punctuation Marks

Terminal Marks	Pausing Marks	Dashes and Hyphens	Others
The period . Question Mark ? Exclamation Point !	Comma , Semicolon ; Colon :	Hypen (-)	Angle Brackets <> Slash / Parantheses ( ) Braces { } Apostrophe ‘ Square Brackets [ ] Quotation Marks “
The list is only representative, it covers commonly used punctuation marks			

#	Stop Word	
1	A	All these words are so-called Function words. They have limited meaning, and occur throughout a corpus or text document frequently **Other stop words are:** **1) At** **2) For** **3) To** **4) On** **5) With** 6) **From**
2	About	
3	Also	
4	And	
5	Another	
6	The	
7	Is	
8	It	

Note:

The above list is by means complete. It is however representative of what our Stop Word Remover will get rid of.

A regular expression is a ASCII character-set based combination of alphanumeric characters and symbols. The purpose of a regex is to serve as a "matching template" in a corpus text	
Regular Expression	**What it does, when used in the classifier: Finds, matches and used in code to remove:**
[0-9]	**Single Digit Numbers between 0 and 9, 0 and 9 included**
1[0-9][1-9][0-9]	**Matches a number in the range**: 1010-1999
/[^A-Za-z0-9]+/g	Notes: / - **Indicates the start of an expression** Caret ^ and $ are anchors marking start and end of an expression (a sequence of characters) A-Z **matches a character in the range** A-Z a-z **matches a character in the range** a-z 0-9 **matches a character (number) in the range of numbers** 0-9 + **Matches one or more of characters matched by the previous regular expression tokens** / **indicates the end of the regular expression**
This is only a small representative set of a vast vocabulary of regular expressions. It is intended to inform the user about the potential of regular expressions. Sophisticated Spam Classifiers may use complex regular expressions as required to stop Spam.	

Implementation Infrastructure – Recommended Prerequisites (Software Development Environment)

Choice # 1

The **Hortonworks Development Platform Sandbox 2.6.4** on your host machine where we may develop and deploy with the comprehensive environment provided by the sandbox.

- The Zeppelin Notebook is an important piece of this infrastructure

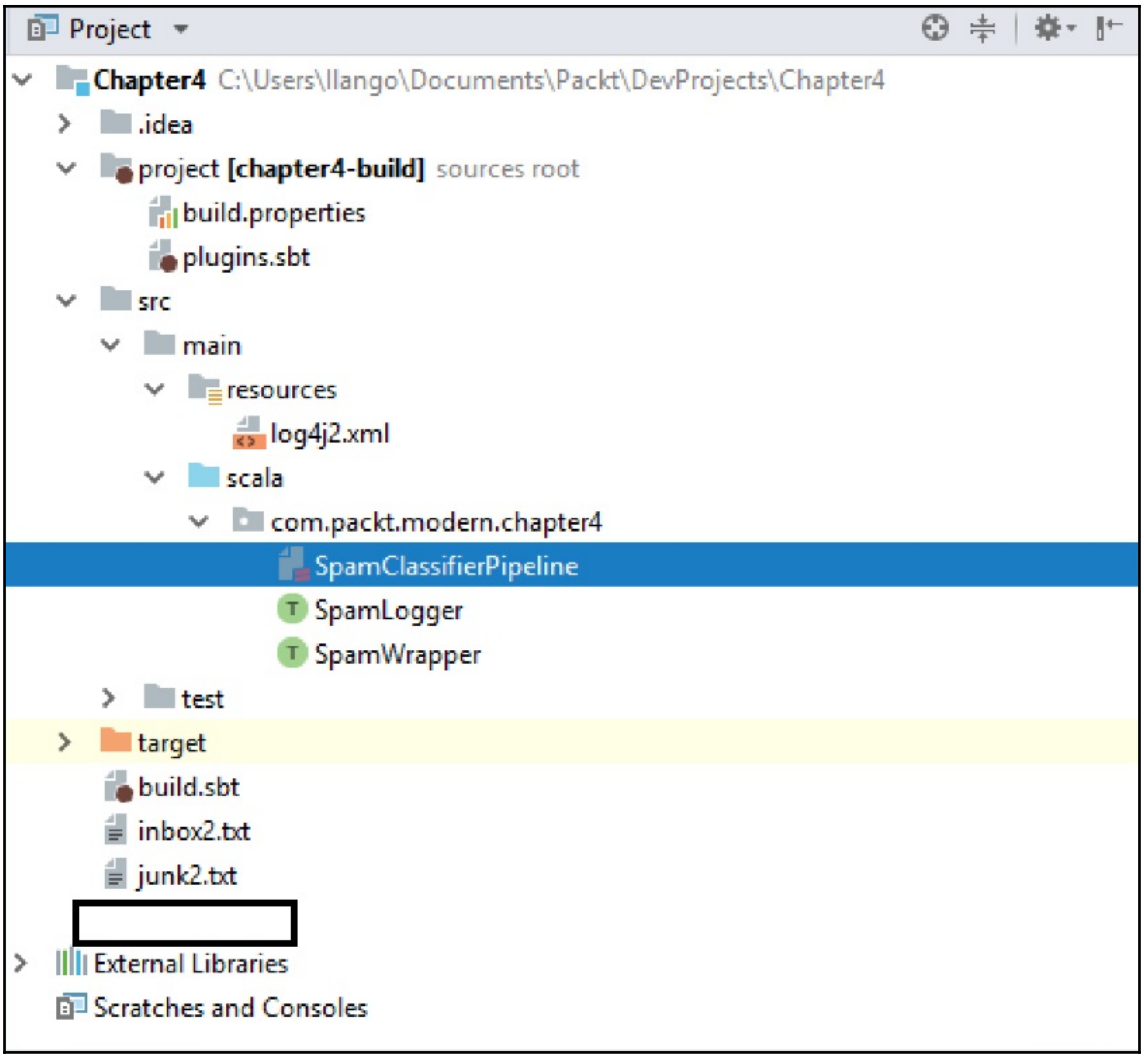
OR

Choice # 2

IntelliJ Integrated Development Environment + Command Line (or terminal) + SBT commands on the command line (or terminal + the Spark submit command on the command line (or terminal).

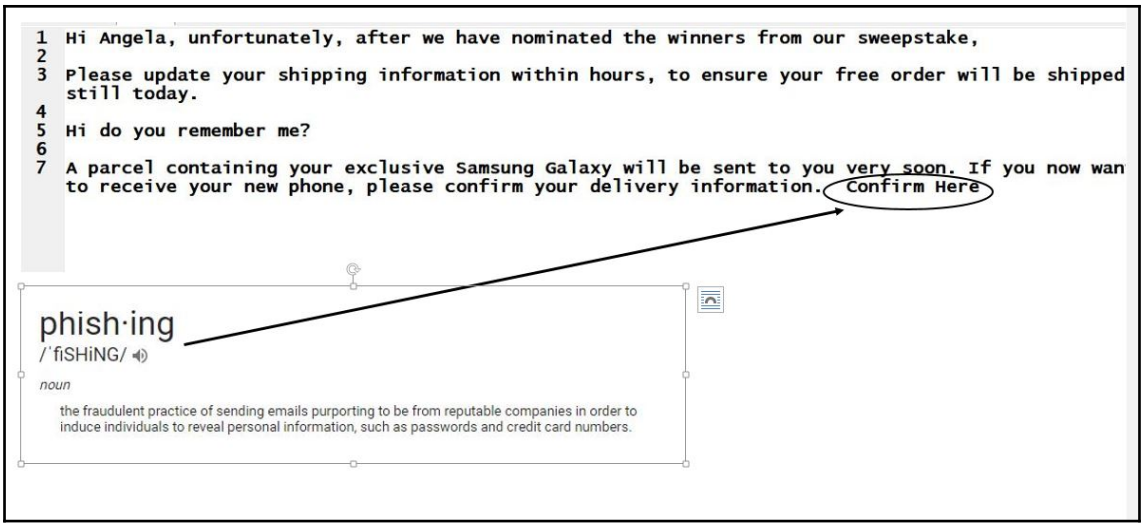
- The Spam Classification Pipeline is an IntelliJ SBT project. Just make sure this project is set up to work with the following versions of software.

JDK 1.8_172; (IntelliJ Community Edition 2018.1 and system-wide)	Scala 2.11.12; (IntelliJ and system-wide
`IntelliJ set up with Scala Plugin`	
A build.sbt file set up: **Spark 2.3.1, Log4j Scala 2.11 API** • **What is new**: Log4j Scala 2.11 is a Scala wrapper of the popular Logging implementation - Apache Log4j	


Project
Chapter4 C:\Users\llango\Documents\Packt\DevProjects\Chapter4
.idea
project [chapter4-build] sources root
build.properties
plugins.sbt
src
main
resources
log4j2.xml
scala
com.packt.modern.chapter4
SpamClassifierPipeline
SpamLogger
SpamWrapper
test
target
build.sbt
inbox2.txt
junk2.txt
External Libraries
Scratches and Consoles

```
libraryDependencies ++= Seq(
  "org.apache.spark" %% "spark-core" % "2.3.1",
  "org.apache.spark" %% "spark-mllib" % "2.3.1",
  "org.apache.spark" %% "spark-sql" % "2.3.1",
  "org.apache.logging.log4j" %% "log4j-api-scala" % "11.0",
  "org.apache.logging.log4j" % "log4j-scala" % "11.0" pomOnly(),
  "org.apache.logging.log4j" % "log4j-api" % "2.11.0",
  "org.apache.logging.log4j" % "log4j-core" % "2.11.0" % Runtime
)
```

```
package com.packt.modern.chapter4

import org.apache.spark.sql.SparkSession
import org.apache.logging.log4j.scala.Logging
import org.apache.logging.log4j.Level


trait SpamWrapper extends Logging{

  //The entry point to programming Spark with the Dataset and DataFrame API.
  //This is the SparkSession

  lazy val session: SparkSession = {
    SparkSession
      .builder()
      .master( master = "local")
      .appName( name = "spam-classifier-pipeline")
      .getOrCreate()
  }

    //logger.getLogger("org").setLevel(Level.OFF)
  //logger.getLogger("akka").setLevel(Level.OFF)
    logger.apply(Level.OFF, "org")
    logger.apply(Level.OFF, "akka")

}
```


1 Hi Angela, unfortunately, after we have nominated the winners from our sweepstake,
2
3 Please update your shipping information within hours, to ensure your free order will be shipped still today.
4
5 Hi do you remember me?
6
7 A parcel containing your exclusive Samsung Galaxy will be sent to you very soon. If you now wan to receive your new phone, please confirm your delivery information. Confirm Here
phish·ing
/ˈfiSHiNG/
noun
the fraudulent practice of sending emails purporting to be from reputable companies in order to induce individuals to reveal personal information, such as passwords and credit card numbers.

Hi there,

Thank you for being part of the Lightbend community. We wanted to let you know that we have made some updates to Lightbend's Privacy Policy and Cookie Policy in preparation for the EU's Data Protection Regulation that will be effective on May 25, 2018.

Changes include:

- Greater transparency into what type of data is being collected, processed and disclosed
- Detailed instructions on how to access, correct or delete your personal information
- The new Cookie Policy explains what type of cookies are being used as part of our services and allows users to more easily restrict sharing of personal data

At Lightbend, we are strongly committed to protecting the privacy of personal data that we maintain about Lightbend clients, employees and other individuals interested in the Reactive and Fast Data journey.

Thank you for your continued support.

If you have any questions, please contact us at privacy@lightbend.com.

Regards,

Lightbend Team

Hi Fellow Little Hills Toastmasters:
This coming Tuesday we are going to take some time in the meeting to get everyone up and running in Pathways. If you are interested in being a part of this, please bring your laptop or tablet. I look forward to seeing you there.
Thanks, Kat
User Name: xyzstl
Dear Ilango,
Here's a summary of your Barclaycard Ring Mastercard® account activity in the past week. To view your purchases at any time, please log in to www.BarclaycardUS.com
and select View activity and statements. Questions? Please log in to our servicing site
and send a secure message. Please remember to review your statement to see
transactions, payments, and other important account information and disclosures.
Your Weekly Snapshot since 06/17/2018
Recent Payments & Credits: $0.00
Recent Charges: $0.00
Statement Balance: $0.00
Top 10 Resume Writing Tips for 2018 - DailyWritingTips

bitte achten Sie auf den
Zahlungseingang.
Mit besten Grüßen
Kavin Kumar Krishnamurthy
we are having issue with our
subscribers Ilangowrites
If you would like to unsubscribe,
please use this link
Re your message. click here
Unsubsrcibe From Mailing List
There's a small issue, Angela
Hi Angela, unfortunately, after we
have nominated the winners from our
sweepstake, we realized that there is
an issue related for your order.
Please update your shipping
information within hours, to ensure
your free order will be shipped still
today.
Hi do you remember me?
Delivery Status:
Confirmation Pending
Shipping Contractor:
UPS
Order destination:
Angela Baggett
A parcel containing your exclusive
Samsung Galaxy will be sent to you
very soon. If you now want to receive
your new phone, please confirm your
delivery information.
Confirm Here

```
package com.packt.modern.chapter4

import org.apache.spark.ml.classification.NaiveBayes
import org.apache.spark.ml.{Pipeline, PipelineStage}
import org.apache.spark.sql.{DataFrame, DataFrameNaFunctions, Row, SparkSession}
import org.apache.spark.ml.feature.{HashingTF, IDF, Normalizer, Tokenizer}
import org.apache.spark.rdd.RDD
import org.apache.spark.sql.types._
import org.apache.spark.sql.functions.explode
```

```
+--------------------+-----+--------------------+----------------------+--------------------+
| lowerCasedSentences|label|    mailFeatureWords|noStopWordsMailFeatures|   mailFeatureHashes|
+--------------------+-----+--------------------+----------------------+--------------------+
|this coming tuesd...|  0.0|[this, coming, tu...|   [coming, tuesday,...|(10000,[380,855,1...|
|pin free dialing ...|  0.0|[pin, free, diali...|   [pin, free, diali...|(10000,[1073,1097...|
|regards support team|  0.0|[regards, support...|   [regards, support...|(10000,[468,695,9...|
|           thankskat|  0.0|         [thankskat]|           [thankskat]|(10000,[5652],[1.0])|
|speed dialing let...|  0.0|[speed, dialing, ...|   [speed, dialing, ...|(10000,[1097,3245...|
|keep your user in...|  0.0|[keep, your, user...|   [keep, user, info...|(10000,[2904,5813...|
| user name ilangostl|  0.0|[user, name, ilan...|   [user, name, ilan...|(10000,[15,742,58...|
|now your family m...|  0.0|[now, your, famil...|   [family, member, ...|(10000,[1094,1181...|
|click on link bel...|  0.0|[click, on, link,...|   [click, link, ent...|(10000,[847,1719,...|
|hi fellow little ...|  0.0|[hi, fellow, litt...|   [hi, fellow, litt...|(10000,[1960,3391...|
|for every person ...|  0.0|[for, every, pers...|   [every, person, r...|(10000,[855,1073,...|
|we look forward t...|  0.0|[we, look, forwar...|   [look, forward, s...|(10000,[7923,9504...|
|thank you for cho...|  0.0|[thank, you, for,...|   [thank, choosing,...|(10000,[763,768,1...|
|anbei die steuerb...|  0.0|[anbei, die, steu...|   [anbei, die, steu...|(10000,[1409,1576...|
|we are having iss...|  0.0|[we, are, having,...|              [issue]|(10000,[6748],[1.0])|
|re your message c...|  0.0|[re, your, messag...|   [re, message, click]|(10000,[1719,2425...|
|hi do you remembe...|  0.0|[hi, do, you, rem...|        [hi, remember]|(10000,[1960,5685...|
|                ups|  0.0|               [ups]|                 [ups]|(10000,[2525],[1.0])|
|        confirm here|  0.0|     [confirm, here]|             [confirm]|(10000,[5943],[1.0])|
|      angela baggett|  0.0|   [angela, baggett]|     [angela, baggett]|(10000,[6290,9622...|
+--------------------+-----+--------------------+----------------------+--------------------+
only showing top 20 rows
```

```
+--------------------+-----+
| lowerCasedSentences|label|
+--------------------+-----+
|pin free dialing ...|  0.0|
|regards support team|  0.0|
|this coming tuesd...|  0.0|
|speed dialing let...|  0.0|
| user name ilangostl|  0.0|
|for every person ...|  0.0|
|hi fellow little ...|  0.0|
|thank you for cho...|  0.0|
|we look forward t...|  0.0|
|anbei die steuerb...|  0.0|
|      angela baggett|  0.0|
|        confirm here|  0.0|
|hi do you remembe...|  0.0|
|re your message c...|  0.0|
|                 ups|  0.0|
|we are having iss...|  0.0|
|weve received you...|  0.0|
|content of your o...|  0.0|
|delivery address ...|  0.0|
|   order destination|  0.0|
+--------------------+-----+
only showing top 20 rows
```

```
+--------------------+-----+--------------------+----------------------+--------------------+
| lowerCasedSentences|label|    mailFeatureWords|noStopWordsMailFeatures|   mailFeatureHashes|
+--------------------+-----+--------------------+----------------------+--------------------+
|keep your user in...|  0.0|[keep, your, user...|   [keep, user, info...|(10000,[2904,5813...|
|          thankskat|  0.0|         [thankskat]|           [thankskat]|(10000,[5652],[1.0])|
|click on link bel...|  0.0|[click, on, link,...|   [click, link, ent...|(10000,[847,1719,...|
|now your family m...|  0.0|[now, your, famil...|   [family, member, ...|(10000,[1094,1181...|
|das guthaben wird...|  0.0|[das, guthaben, w...|   [das, guthaben, w...|(10000,[568,2306,...|
|     delivery status|  0.0|  [delivery, status]|     [delivery, status]|(10000,[7128,7497...|
|subscribers ilang...|  0.0|[subscribers, ila...|   [subscribers, ila...|(10000,[1999,6012...|
|hi angela unfortu...|  0.0|[hi, angela, unfo...|   [hi, angela, unfo...|(10000,[84,721,10...|
+--------------------+-----+--------------------+----------------------+--------------------+
```

```
+--------------------+--------------------+--------------------+----------+----------+
|             mailIDF|            features|       rawPrediction|probability|prediction|
+--------------------+--------------------+--------------------+----------+----------+
|(10000,[2904,5813...|(10000,[2904,5813...|[-25.87894074578371]|     [1.0]|       0.0|
|(10000,[5652],[3....|(10000,[5652],[1.0])| [-9.21781850738124]|     [1.0]|       0.0|
|(10000,[847,1719,...|(10000,[847,1719,...|[-21.815534062393...|     [1.0]|       0.0|
|(10000,[1094,1181...|(10000,[1094,1181...|[-28.681677648570...|     [1.0]|       0.0|
|(10000,[568,2306,...|(10000,[568,2306,...|[-30.354807519098...|     [1.0]|       0.0|
|(10000,[7128,7497...|(10000,[7128,7497...|[-12.508199802474...|     [1.0]|       0.0|
|(10000,[1999,6012...|(10000,[1999,6012...|[-13.03596394863227]|     [1.0]|       0.0|
|(10000,[84,721,10...|(10000,[84,721,10...|[-40.21026716158472]|     [1.0]|       0.0|
+--------------------+--------------------+--------------------+----------+----------+
```

```
Displaying Predictions as below:
+--------------------+----------+
| lowerCasedSentences|prediction|
+--------------------+----------+
|keep your user in...|       0.0|
|           thankskat|       0.0|
|click on link bel...|       0.0|
|now your family m...|       0.0|
|das guthaben wird...|       0.0|
|     delivery status|       0.0|
|subscribers ilang...|       0.0|
|hi angela unfortu...|       0.0|
+--------------------+----------+
```

Chapter 5: Build a Fraud Detection System

Testing Dataset

Transaction | Distance_In_Miles_From_Cardholder_Address | Label

First 10 rows

```
1    429.2146,28.5524,0
2    473.838,31.7216,0
3    92.2235,4.6191,1
4    477.5542,31.9855,0
5    339.856,22.2061,0
6    77.7948,3.5944,0
7    166.4641,9.8917,0
8    297.9719,19.2315,0
9    499.1783,33.5212,1
10   502.7954,33.7781,1
```

Last 10 rows

```
990    429.9087,28.6017,1
991    185.5397,11.2465,0
992    429.1245,28.546,0
993    416.646,27.6598,0
994    447.6093,29.8588,0
995    277.7619,17.7962,0
996    341.4741,22.321,0
997    495.9383,33.2911,1
998    247.5424,15.65,0
999    59.4092,2.2887,0
1000   454.7074,30.3629,0
```

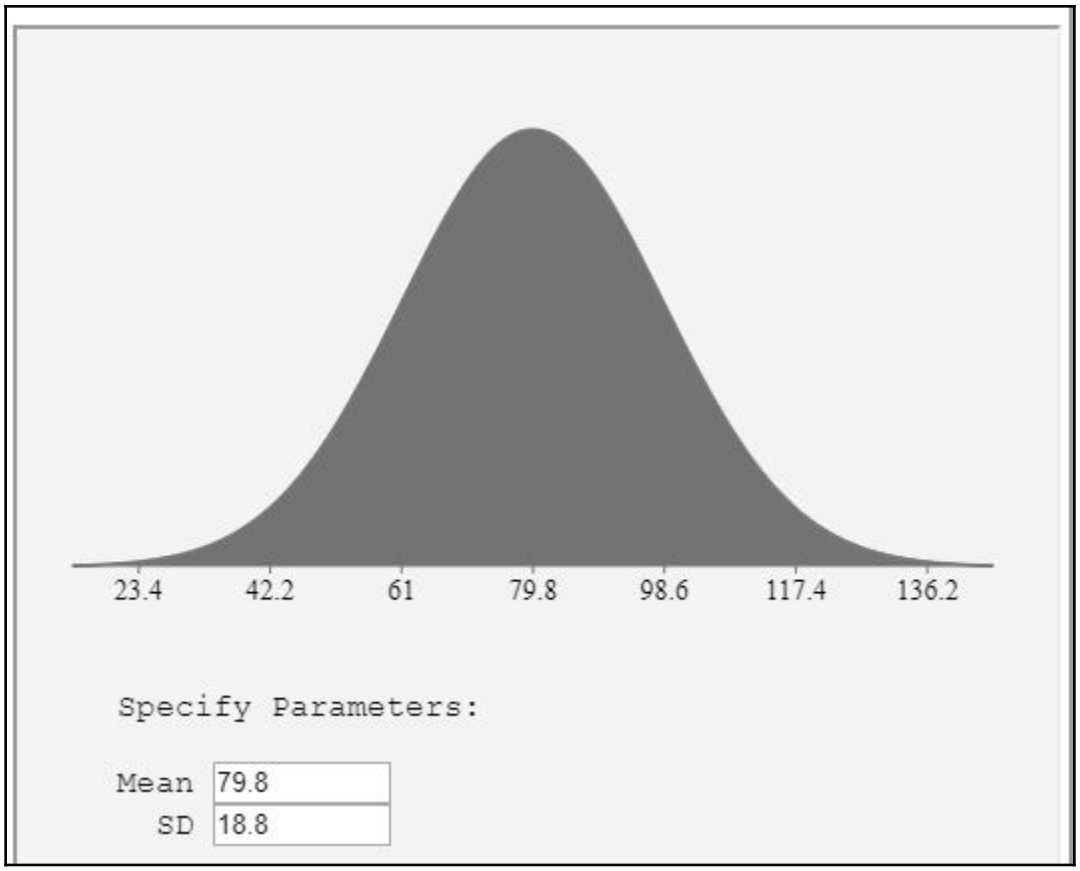
Testing Dataset view in Excel (Comma Separated File)

	Transaction ($)	Distance_In_Miles_From_Cardholder_Address	Label
1	429.2146	28.5524	0
2	473.838	31.7216	0
3	92.2235	4.6191	1
4	477.5542	31.9855	0
5	339.856	22.2061	0
6	77.7948	3.5944	0
7	166.4641	9.8917	0
8	297.9719	19.2315	0
9	499.1783	33.5212	1
10	502.7954	33.7781	1

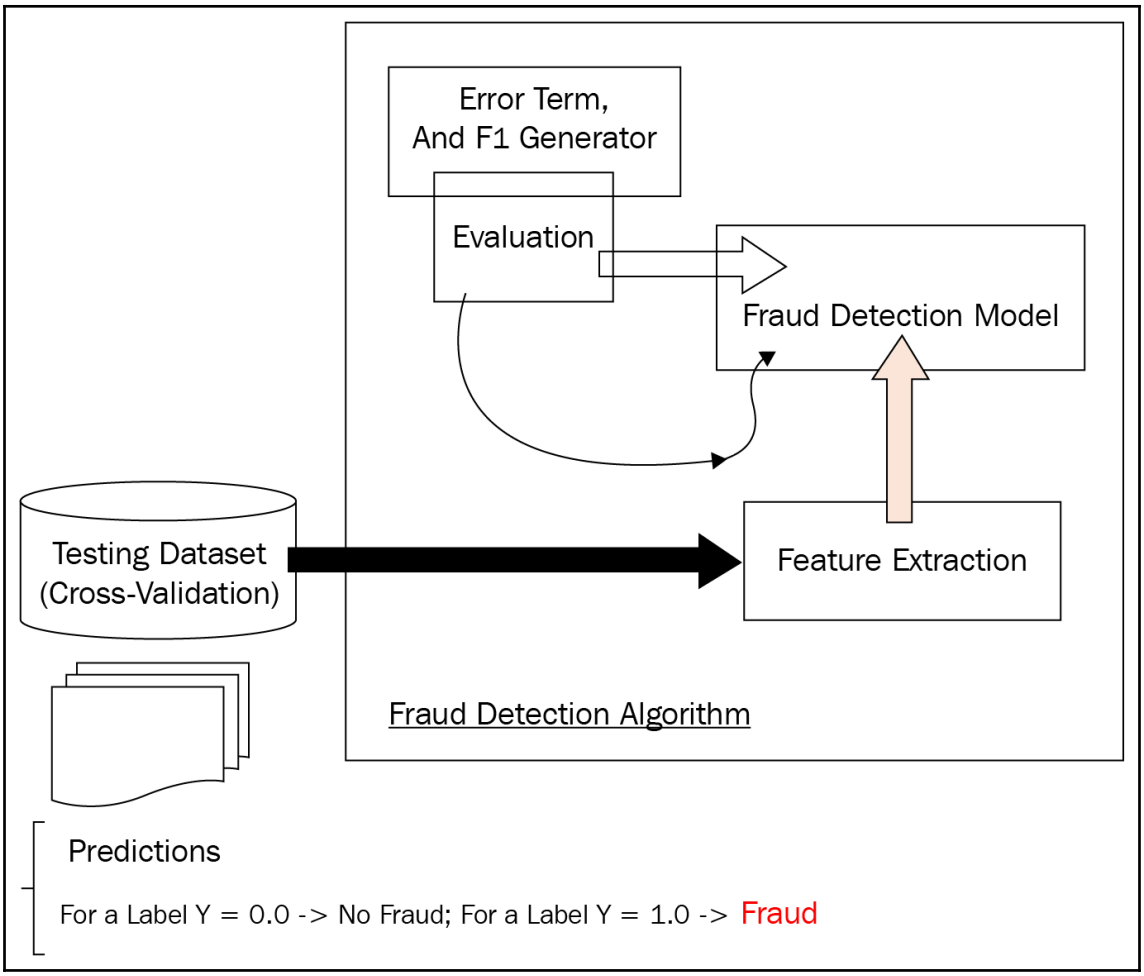
Training Dataset

```
1    399.2146,34.3974
```

```
990    399.9087,27.5344
```


23.4
42.2
61
79.8
98.6
117.4
136.2
Specify Parameters:
Mean 79.8
SD 18.8

Property #	Classification	Anomaly Identification
1		It is not immediately apparent upfront or by casual inspection that data is either "regular", "expected", "within a certain range" or something that is unexpected represents a significant deviation The proportion of data with an unexpected value in comparison with those with regular or expected value is really small
2	Both Classification and Anomaly identification tasks perform classification in a sense, but the differences set them apart	
	Categories in a Classification task are clear cut. For example, a Breast Cancer sample is either Benign or Malignant	As opposed to a Classification task, there are no distinct categories available. In the case of finance transaction, it is harder to have an algorithm make straightforward predictions, because finance transactions can be arbitrary occurrences between one person and the next. Spending habits of different people make an Anomaly Identification task not so trivial.
3	The number of samples that are "positive" make up much higher proportion of the data. For example, we may have many breast cancer samples that are "malignant".	The number of anomalies indicating "Potential Fraud" are small.


Error Term,
And F1 Generator
Evaluation
Fraud Detection Model
Testing Dataset
(Cross-Validation)
Feature Extraction
Fraud Detection Algorithm
Predictions
For a Label Y = 0.0 -> No Fraud; For a Label Y = 1.0 -> Fraud

Understand what mathematical equation our Fraud Detection System will be based upon.

Choose the right features that might most meaningfully represent samples that indicate outliers

It is suggested you break up you dataset into: 1) Training (55%) 2) Testing (25%)

Calculate the following stats: Mean (mu or μ) and Standard Deviation (σ)

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

$$\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

Implement our Model, which is simply the probability function Y= p(X)

Note: We (naturally need the values of Mean and Standard Deviation from the previous step)

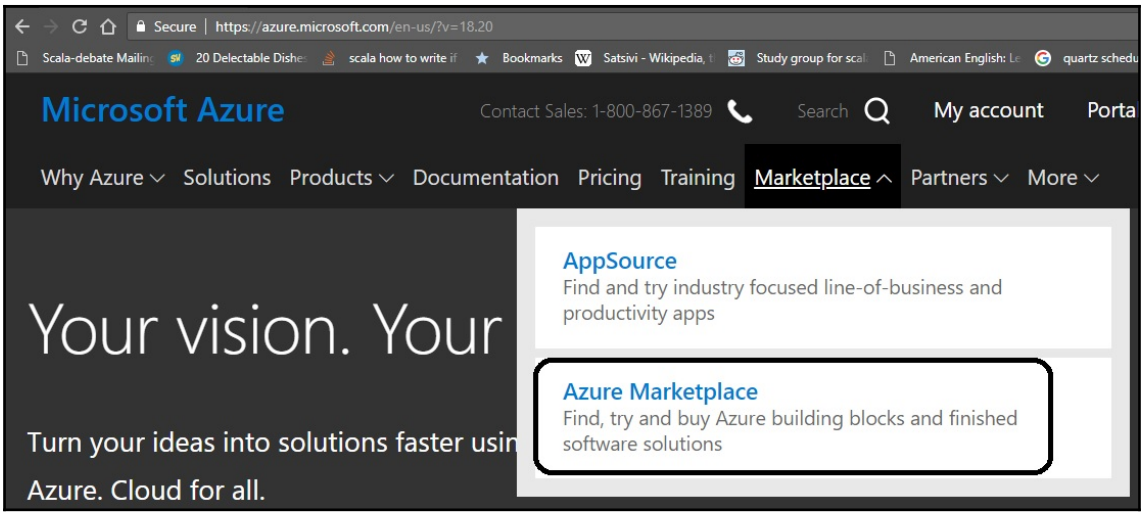
$$y = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma}$$

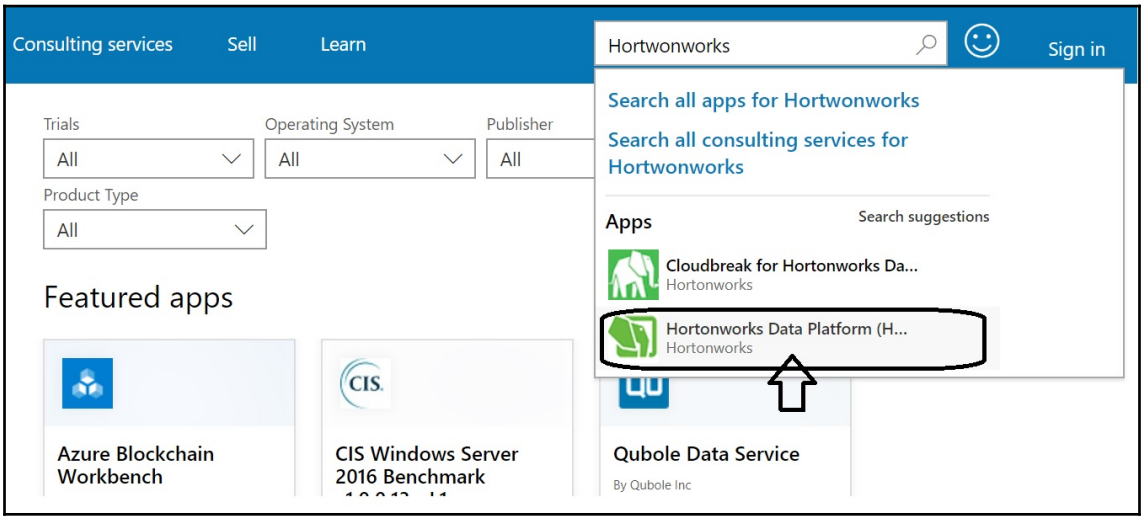
μ = Mean
σ = Standard Deviation
$\pi \approx 3.14159$
$e \approx 2.71828$

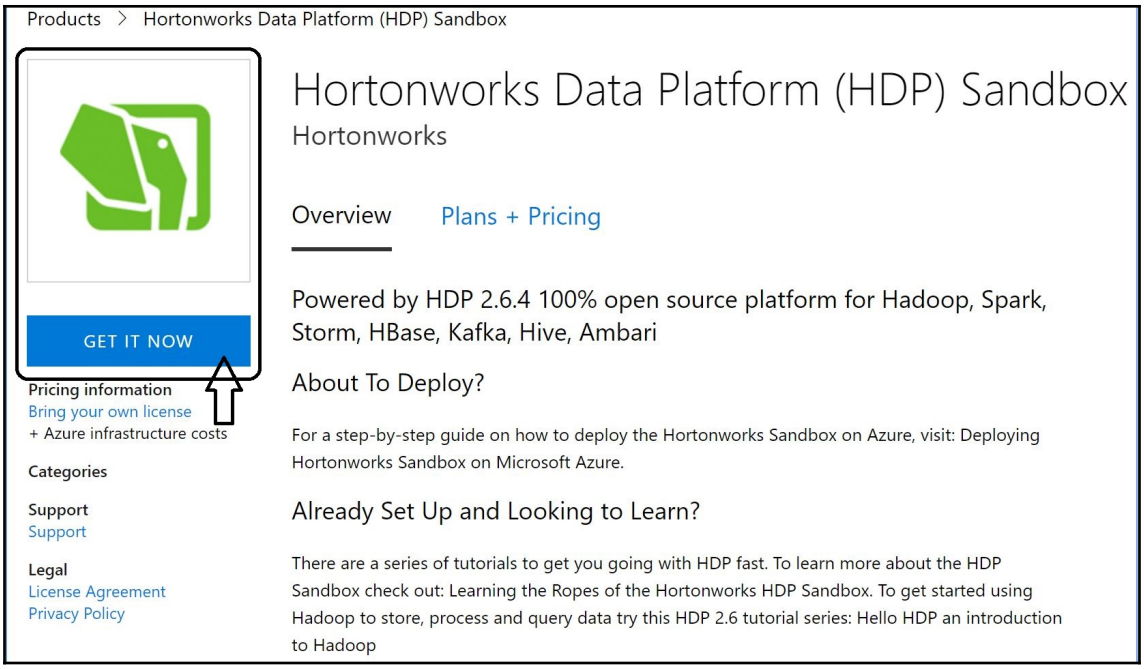
Generate your predictions Y, after calculating probability (density) values for each testing set datapoint (Transaction or Distance feature)

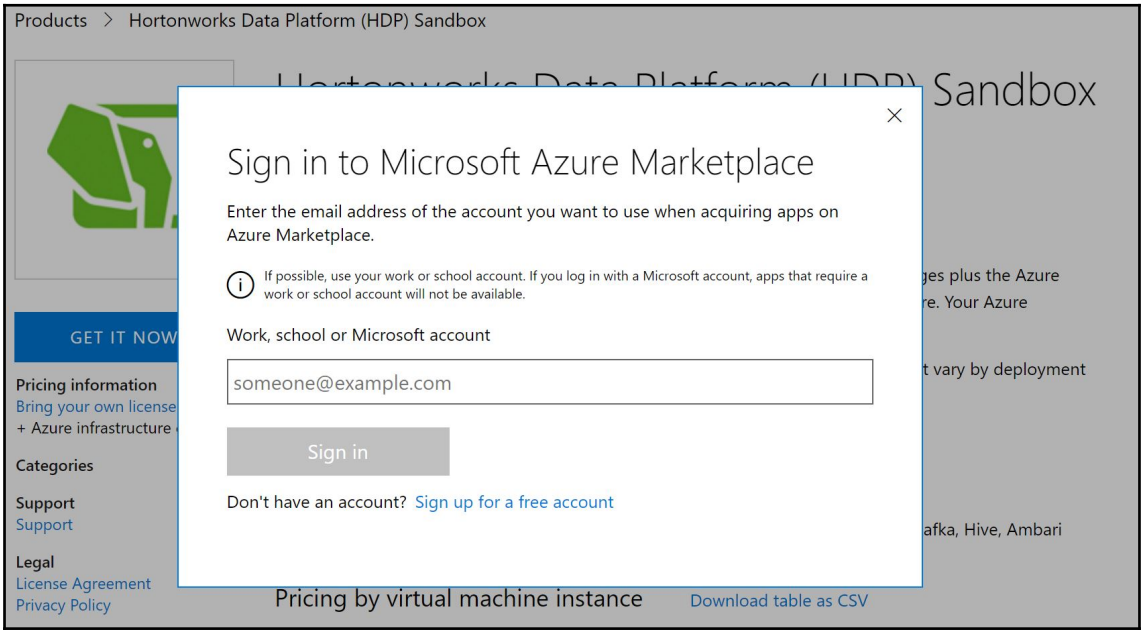
Compute FPs, FNs, TPs, and TNs, to come up with model performance Evaluation (double) metrics: 1) Precision, 2) Recall, and the 3) F1 measure

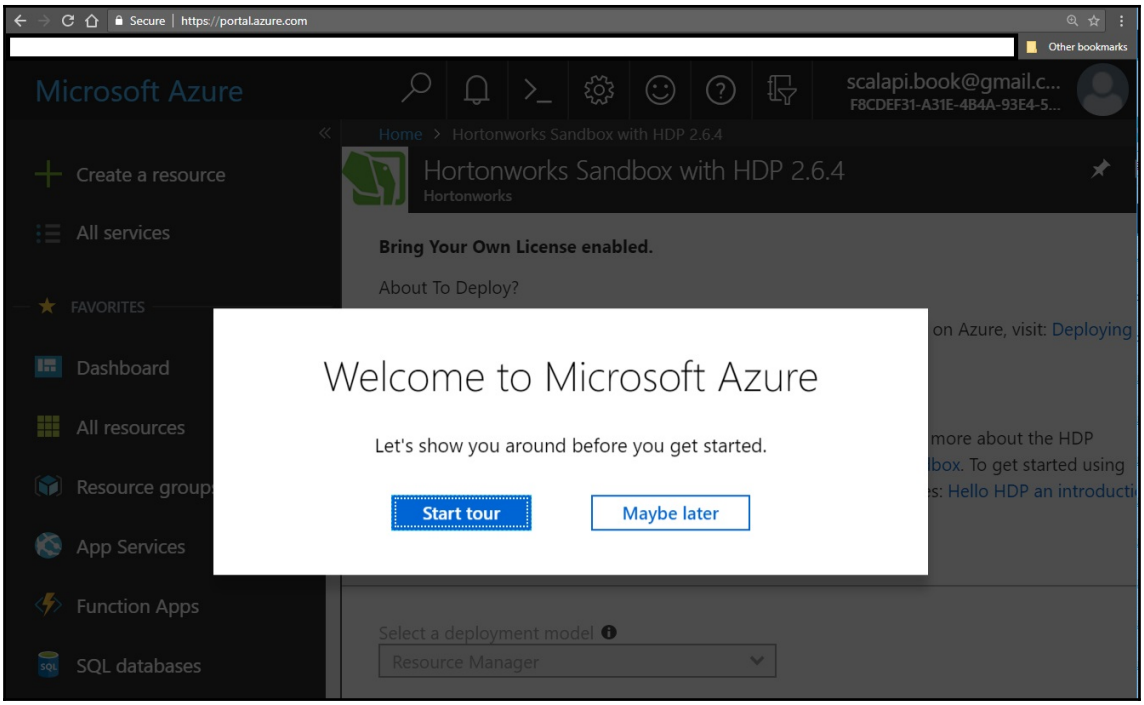
Run the algorithm over a range of Error Terms to come up with Best Error Term (Epsilon). Optimize our model with the Best Error Term

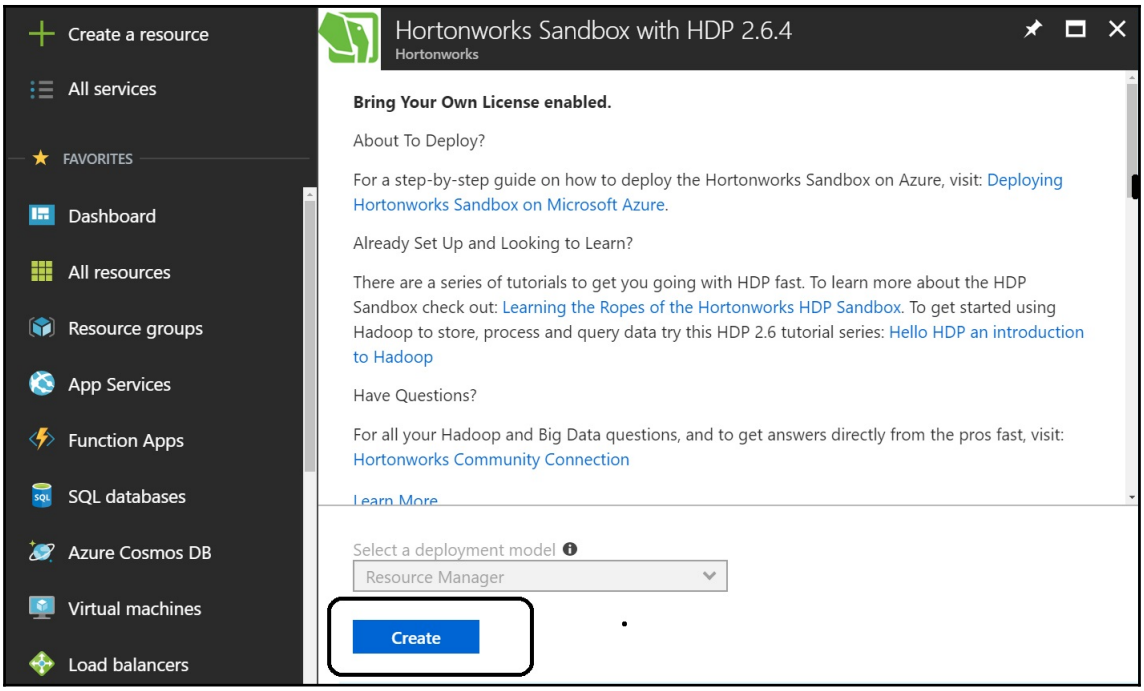

Secure | https://azure.microsoft.com/en-us/?v=18.20
Microsoft Azure
Contact Sales: 1-800-867-1389
Search
My account
Why Azure
Solutions
Products
Documentation
Pricing
Training
Marketplace
Partners
More
Your vision. Your
Turn your ideas into solutions faster usin
Azure. Cloud for all.
AppSource
Find and try industry focused line-of-business and productivity apps
Azure Marketplace
Find, try and buy Azure building blocks and finished software solutions

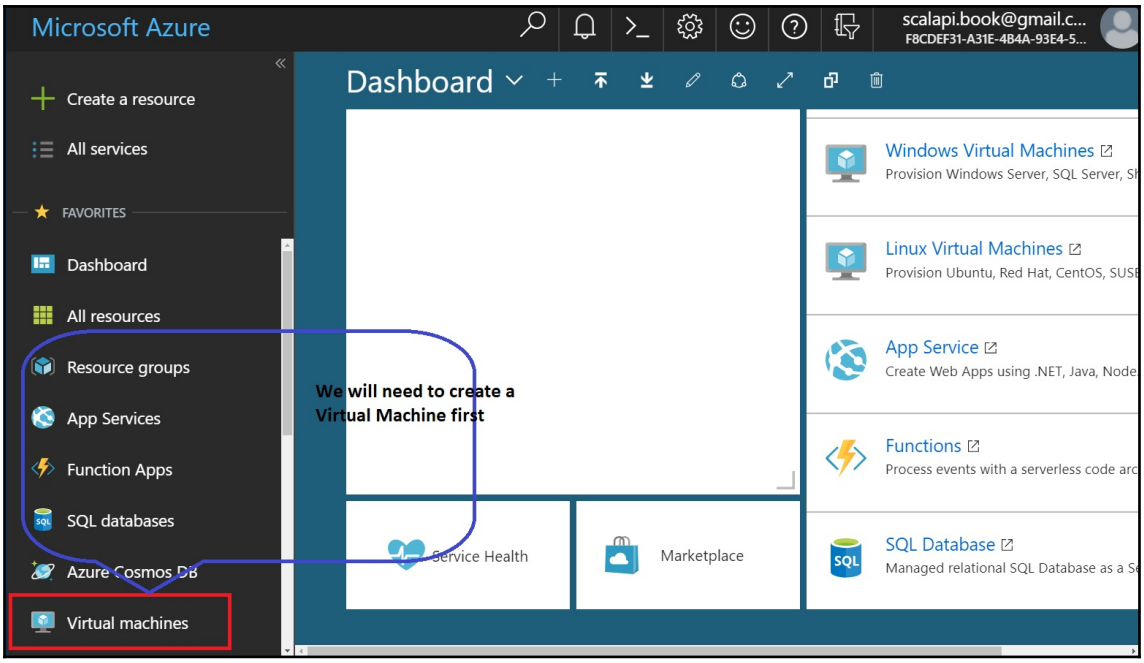

Consulting services
Sell
Learn
Hortwonworks
Sign in
Search all apps for Hortwonworks
Search all consulting services for Hortwonworks
Apps
Search suggestions
Cloudbreak for Hortonworks Da...
Hortonworks
Hortonworks Data Platform (H...
Hortonworks
Trials
All
Operating System
All
Publisher
All
Product Type
All
Featured apps
Azure Blockchain Workbench
CIS Windows Server 2016 Benchmark
Qubole Data Service
By Qubole Inc


Products > Hortonworks Data Platform (HDP) Sandbox
GET IT NOW
Pricing information
Bring your own license
+ Azure infrastructure costs
Categories
Support
Support
Legal
License Agreement
Privacy Policy
Hortonworks Data Platform (HDP) Sandbox
Hortonworks
Overview
Plans + Pricing
Powered by HDP 2.6.4 100% open source platform for Hadoop, Spark, Storm, HBase, Kafka, Hive, Ambari
About To Deploy?
For a step-by-step guide on how to deploy the Hortonworks Sandbox on Azure, visit: Deploying Hortonworks Sandbox on Microsoft Azure.
Already Set Up and Looking to Learn?
There are a series of tutorials to get you going with HDP fast. To learn more about the HDP Sandbox check out: Learning the Ropes of the Hortonworks HDP Sandbox. To get started using Hadoop to store, process and query data try this HDP 2.6 tutorial series: Hello HDP an introduction to Hadoop


Products > Hortonworks Data Platform (HDP) Sandbox
Sign in to Microsoft Azure Marketplace
Enter the email address of the account you want to use when acquiring apps on Azure Marketplace.
If possible, use your work or school account. If you log in with a Microsoft account, apps that require a work or school account will not be available.
Work, school or Microsoft account
someone@example.com
Sign in
Don't have an account? Sign up for a free account
Pricing by virtual machine instance
Download table as CSV


Secure | https://portal.azure.com
Other bookmarks
Microsoft Azure
scalapi.book@gmail.c...
F8CDEF31-A31E-4B4A-93E4-5...
Home > Hortonworks Sandbox with HDP 2.6.4
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
Function Apps
SQL databases
Hortonworks Sandbox with HDP 2.6.4
Hortonworks
Bring Your Own License enabled.
About To Deploy?
on Azure, visit: Deploying
more about the HDP
box. To get started using
s: Hello HDP an introducti
Welcome to Microsoft Azure
Let's show you around before you get started.
Start tour
Maybe later
Select a deployment model
Resource Manager


Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
Function Apps
SQL databases
Azure Cosmos DB
Virtual machines
Load balancers
Hortonworks Sandbox with HDP 2.6.4
Hortonworks
Bring Your Own License enabled.
About To Deploy?
For a step-by-step guide on how to deploy the Hortonworks Sandbox on Azure, visit: Deploying Hortonworks Sandbox on Microsoft Azure.
Already Set Up and Looking to Learn?
There are a series of tutorials to get you going with HDP fast. To learn more about the HDP Sandbox check out: Learning the Ropes of the Hortonworks HDP Sandbox. To get started using Hadoop to store, process and query data try this HDP 2.6 tutorial series: Hello HDP an introduction to Hadoop
Have Questions?
For all your Hadoop and Big Data questions, and to get answers directly from the pros fast, visit: Hortonworks Community Connection
Learn More
Select a deployment model
Resource Manager
Create


Microsoft Azure
scalapi.book@gmail.c...
F8CDEF31-A31E-4B4A-93E4-5...
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
Function Apps
SQL databases
Azure Cosmos DB
Virtual machines
Dashboard
We will need to create a Virtual Machine first
Service Health
Marketplace
Windows Virtual Machines
Provision Windows Server, SQL Server,
Linux Virtual Machines
Provision Ubuntu, Red Hat, CentOS,
App Service
Create Web Apps using .NET, Java, Node
Functions
Process events with a serverless code arc
SQL Database
Managed relational SQL Database as a

```
409.7061,27.1669,0
430.4786,28.6422,0
455.6604,30.4306,0
71.3736,3.1384,0
225.8935,14.1124,0
157.3365,9.2435,0
422.0336,28.0424,1
241.3928,15.2132,0
476.2173,31.8905,0
119.105,6.5283,1
159.2634,9.3803,0
101.3141,5.2648,0
96.6736,4.9352,0
```

Chapter 6: Build Flights Performance Prediction Model

Type of Data	Dataset Name	Fields
Airports	Airports.csv	IATA – International Airport Code AIRPORT – Name of Airport CITY – City of Airport STATE – State of Airport COUNTRY – Country of Airport LATITUDE – Longitude Reading LONGITUDE – Latitude Reading

View of Data

IATA	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
00M	Thigpen	Bay Springs	MS	USA	31.95376472	-89.2345
00R	Livingston Municipal	Livingston	TX	USA	30.68586111	-95.01793
00V	Meadow Lake	Colorado Springs	CO	USA	38.94574889	-104.5699
01G	Perry-Warsaw	Perry	NY	USA	42.74134667	-78.05208

Carrier Codes	AirlineCarriers.csv	CODE – Airline Carrier Code DESCRIPTION – Airline Carrier Name

View of Data:

CODE	DESCRIPTION
02Q	Titan Airways
04Q	Tradewind Aviation
05Q	Comlux Aviation, AG
06Q	Master Top Linhas Aereas Ltd.
07Q	Flair Airlines Ltd.
09Q	Swift Air, LLC

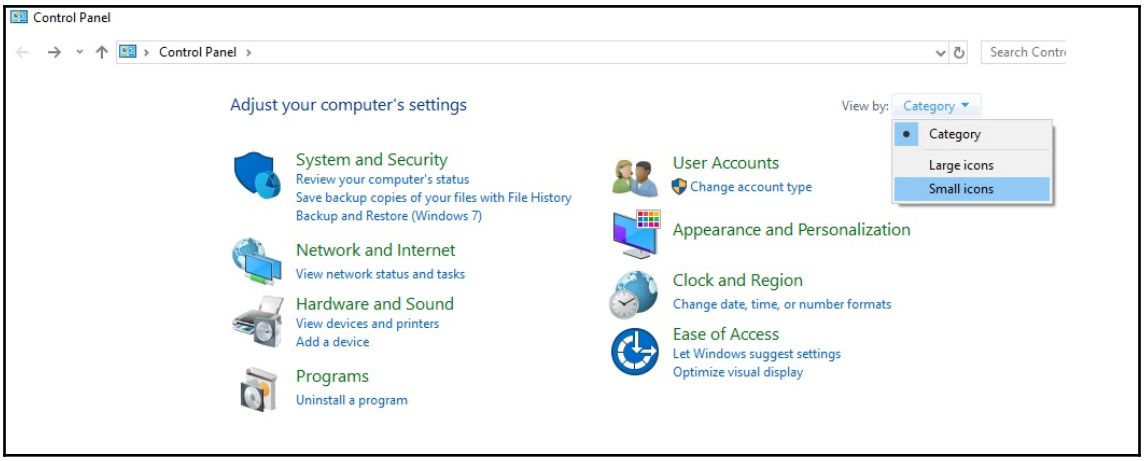
FlightYear */* 1 */*	**FlightMonth** */* 2 */*
FlightDayOfmonth */* 3 */*	**FlightDayOfweek** */* 4 */*
FlightDepTime */* 5 */*	**FlightCrsDeptime** */* 6 */*
FlightArrtime */* 7 */*	**FlightCrsArrTime** */* 8 */*
FlightUniqueCarrier */* 9 */*	**FlightNumber** */* 10 */*
FlightTailNumber */* 11 */*	**FlightActualElapsedTime** */* 12 */*
FlightCrsElapsedTime */* 13*/*	**FlightAirTime** */* 14 */*
FlightArrDelay */* 15 */*	**FlightDepDelay** */* 16 */*
FlightOrigin */* 17 */*	**FlightDest** */* 18 */*
FlightDistance */* 19 */*	**FlightTaxiin** */* 20 */*
FlightTaxiout */* 21 */*	**FlightCancelled** */* 22 */*
FlightCancellationCode */* 23 */*	**FlightDiverted** */* 24 */*
FlightCarrierDelay */* 25 */*	**FlightWeatherDelay** */* 26 */*
FlightNasDelay */* 27 */*	**FlightSecuritDelay** */* 28 */*

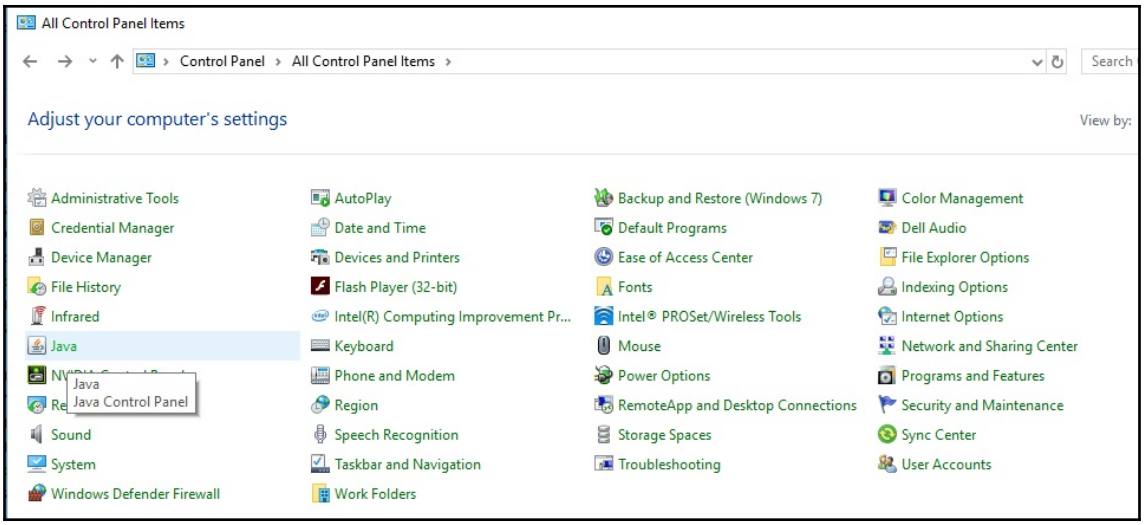
FlightLateAircraftDelay /* 29 */

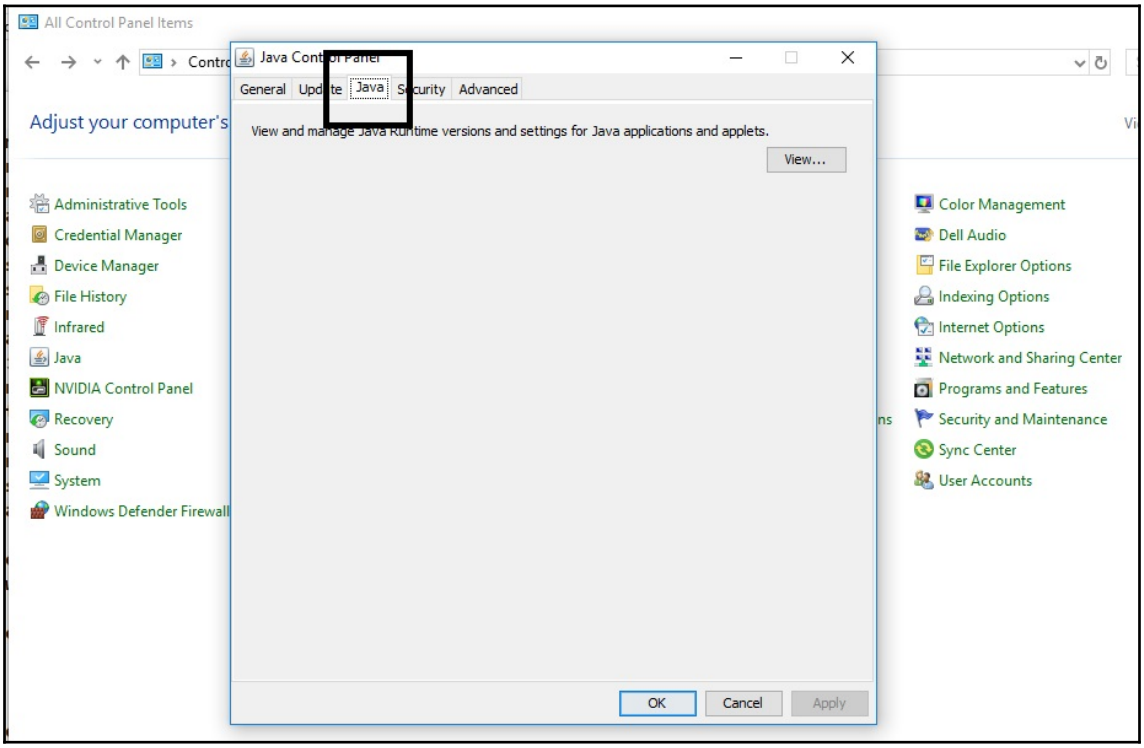
```
18/06/29 00:14:49 ERROR SparkUncaughtExceptionHandler: Uncaught exception in thread Thread[Executor task launch worker f
or task 1,5,run-main-group-0]
java.lang.OutOfMemoryError: GC overhead limit exceeded
        at java.nio.ByteBuffer.wrap(ByteBuffer.java:373)
        at org.apache.hadoop.io.Text.decode(Text.java:389)
        at org.apache.hadoop.io.Text.toString(Text.java:280)
        at org.apache.spark.SparkContext$$anonfun$textFile$1$$anonfun$apply$8.apply(SparkContext.scala:825)
        at org.apache.spark.SparkContext$$anonfun$textFile$1$$anonfun$apply$8.apply(SparkContext.scala:825)
        at scala.collection.Iterator$$anon$11.next(Iterator.scala:410)
        at scala.collection.Iterator$class.foreach(Iterator.scala:891)
        at scala.collection.AbstractIterator.foreach(Iterator.scala:1334)
        at scala.collection.generic.Growable$class.$plus$plus$eq(Growable.scala:59)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuffer.scala:104)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuffer.scala:48)
        at scala.collection.TraversableOnce$class.to(TraversableOnce.scala:310)
        at scala.collection.AbstractIterator.to(Iterator.scala:1334)
        at scala.collection.TraversableOnce$class.toBuffer(TraversableOnce.scala:302)
        at scala.collection.AbstractIterator.toBuffer(Iterator.scala:1334)
        at scala.collection.TraversableOnce$class.toArray(TraversableOnce.scala:289)
        at scala.collection.AbstractIterator.toArray(Iterator.scala:1334)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$12.apply(RDD.scala:939)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$12.apply(RDD.scala:939)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:2074)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:2074)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:87)
        at org.apache.spark.scheduler.Task.run(Task.scala:109)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:345)
```

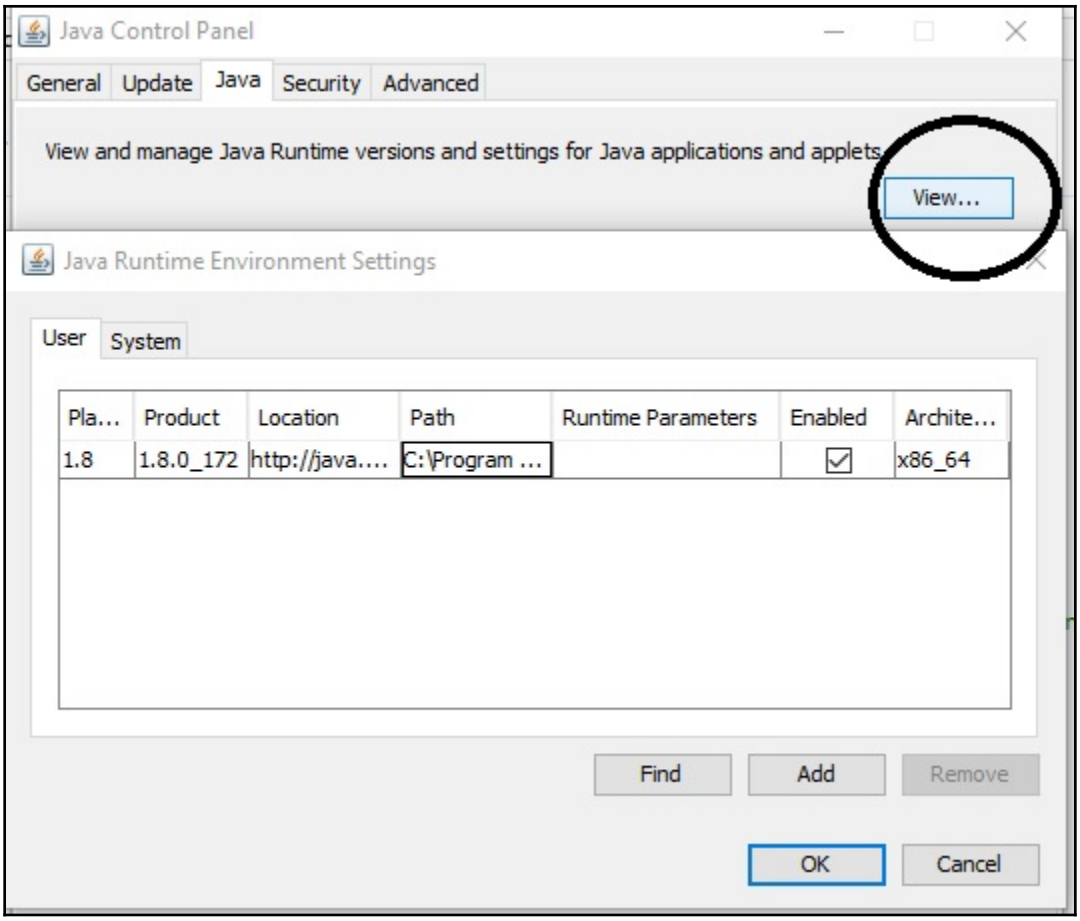
```
C:\Users\Ilango\Documents\Packt\DevProjects\Chapter63>set SBT_OPTS="-XX:MaxPermSize=1G -Xmx2G"

C:\Users\Ilango\Documents\Packt\DevProjects\Chapter63>sbt
"C:\Users\Ilango\.sbt\preloaded\org.scala-sbt\sbt\"1.0.4"\jars\sbt.jar"
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=1G -Xmx2G; support was removed in 8.0
[info] Loading settings from idea.sbt ...
[info] Loading global plugins from C:\Users\Ilango\.sbt\1.0\plugins
[info] Loading project definition from C:\Users\Ilango\Documents\Packt\DevProjects\Chapter63\project
[info] Loading settings from build.sbt ...
[info] Set current project to Chapter63 (in build file:/C:/Users/Ilango/Documents/Packt/DevProjects/Chapter63/)
[info] sbt server started at 127.0.0.1:5191
sbt:Chapter63>
```

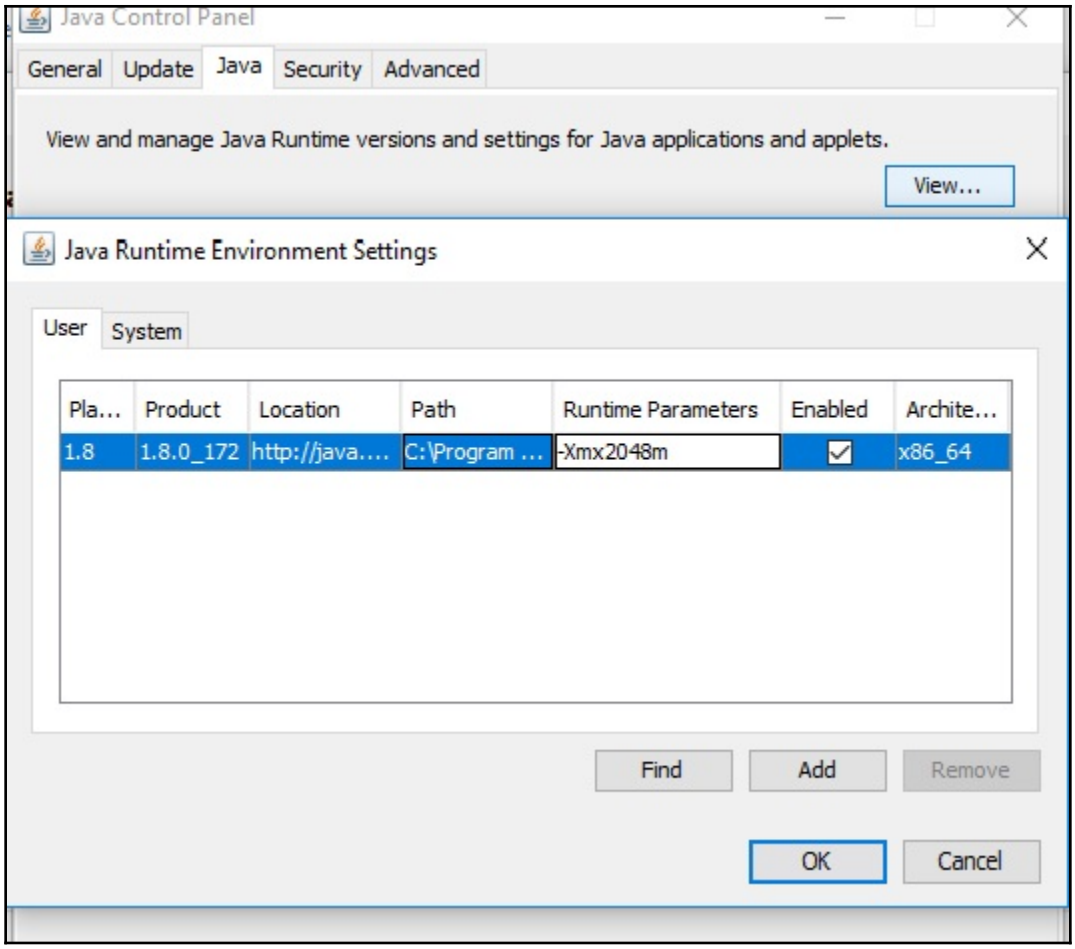
```
C:\Users\Ilango\Documents\Packt\DevProjects\Chapter63>scala
Welcome to Scala 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_172).
Type in expressions for evaluation. Or try :help.

scala> import java.lang.Runtime
import java.lang.Runtime

scala> val jvmMemoryStats = Runtime.getRuntime.totalMemory / (1024 * 1024)
jvmMemoryStats: Long = 175

scala> val jvmMemoryStats = Runtime.getRuntime
jvmMemoryStats: Runtime = java.lang.Runtime@251c4280

scala> val totalMemory = jvmMemoryStats.totalMemory/(1024 * 1024)
totalMemory: Long = 177

scala> val maxMemory = jvmMemoryStats.maxMemory/(1024 * 1024)
maxMemory: Long = 1820

scala> val freeMemory = jvmMemoryStats.freeMemory/(1024 * 1024)
freeMemory: Long = 76

scala>
```



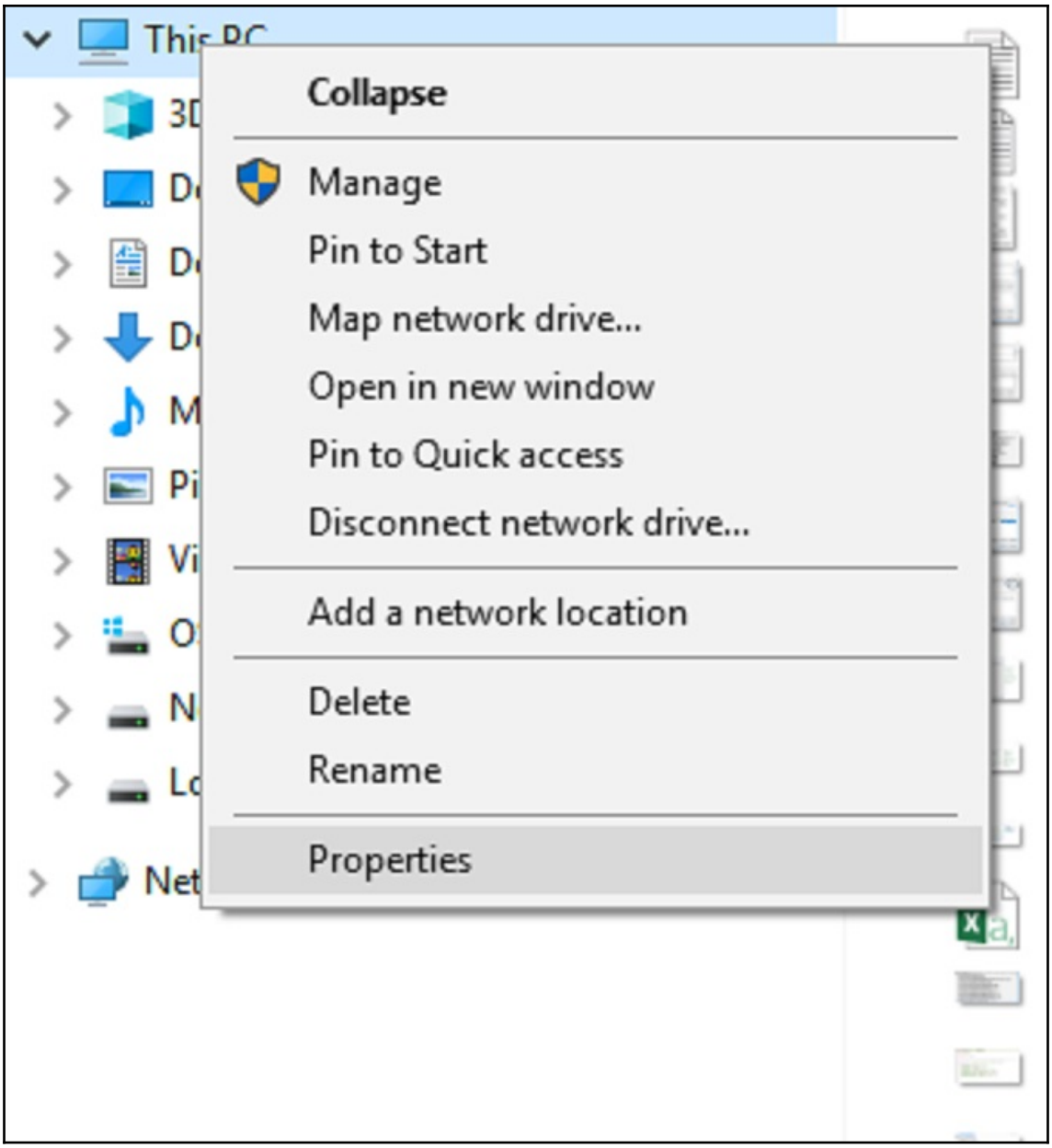

All Control Panel Items
Control Panel › All Control Panel Items ›
Search
Adjust your computer's settings
View by:
Administrative Tools
AutoPlay
Backup and Restore (Windows 7)
Color Management
Credential Manager
Date and Time
Default Programs
Dell Audio
Device Manager
Devices and Printers
Ease of Access Center
File Explorer Options
File History
Flash Player (32-bit)
Fonts
Indexing Options
Infrared
Intel(R) Computing Improvement Pr...
Intel® PROSet/Wireless Tools
Internet Options
Java
Keyboard
Mouse
Network and Sharing Center
Java
Java Control Panel
Phone and Modem
Power Options
Programs and Features
Region
RemoteApp and Desktop Connections
Security and Maintenance
Sound
Speech Recognition
Storage Spaces
Sync Center
System
Taskbar and Navigation
Troubleshooting
User Accounts
Windows Defender Firewall
Work Folders

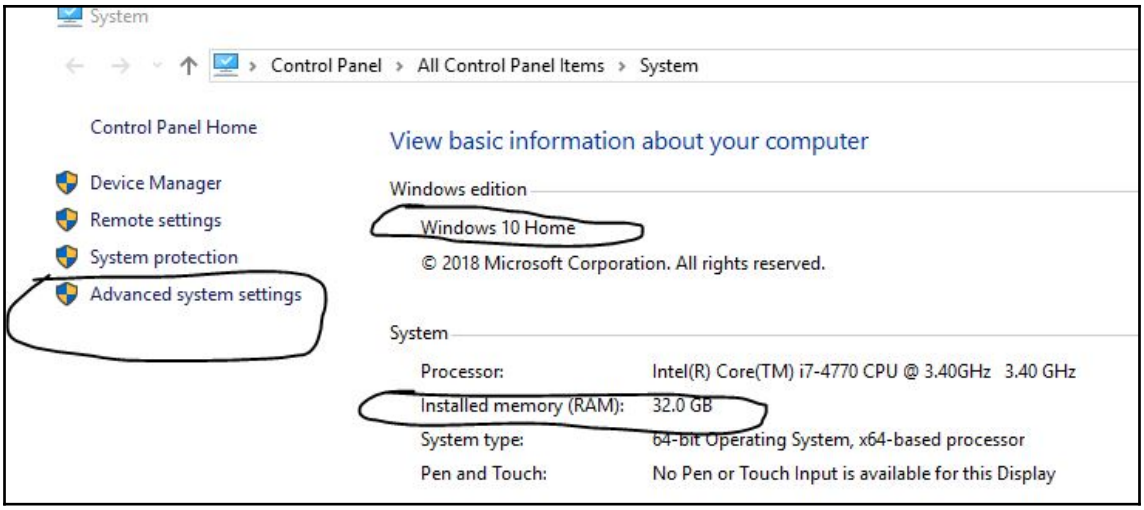

All Control Panel Items
Java Control Panel
General
Java
Security
Advanced
View and manage Java Runtime versions and settings for Java applications and applets.
View...
OK
Cancel
Apply
Adjust your computer's
Administrative Tools
Credential Manager
Device Manager
File History
Infrared
Java
NVIDIA Control Panel
Recovery
Sound
System
Windows Defender Firewall
Color Management
Dell Audio
File Explorer Options
Indexing Options
Internet Options
Network and Sharing Center
Programs and Features
Security and Maintenance
Sync Center
User Accounts

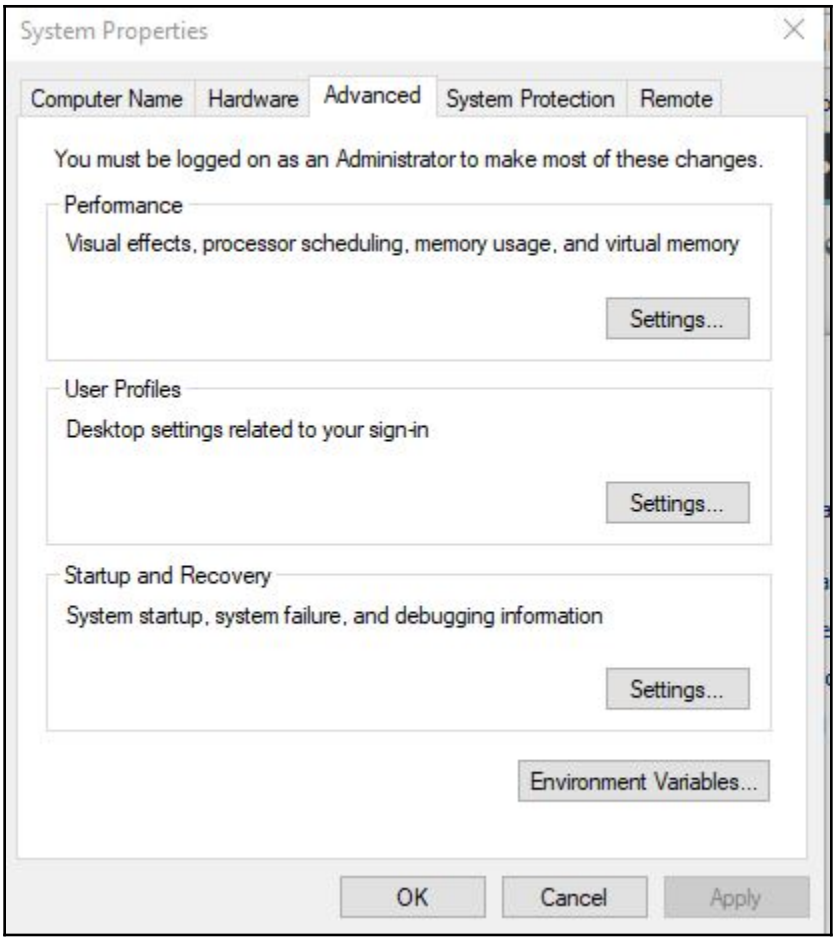

Java Control Panel
General
Update
Java
Security
Advanced
View and manage Java Runtime versions and settings for Java applications and applets.
View...
Java Runtime Environment Settings
User
System
Pla...
Product
Location
Path
Runtime Parameters
Enabled
Archite...
1.8
1.8.0_172
http://java....
C:\Program ...
x86_64
Find
Add
Remove
OK
Cancel

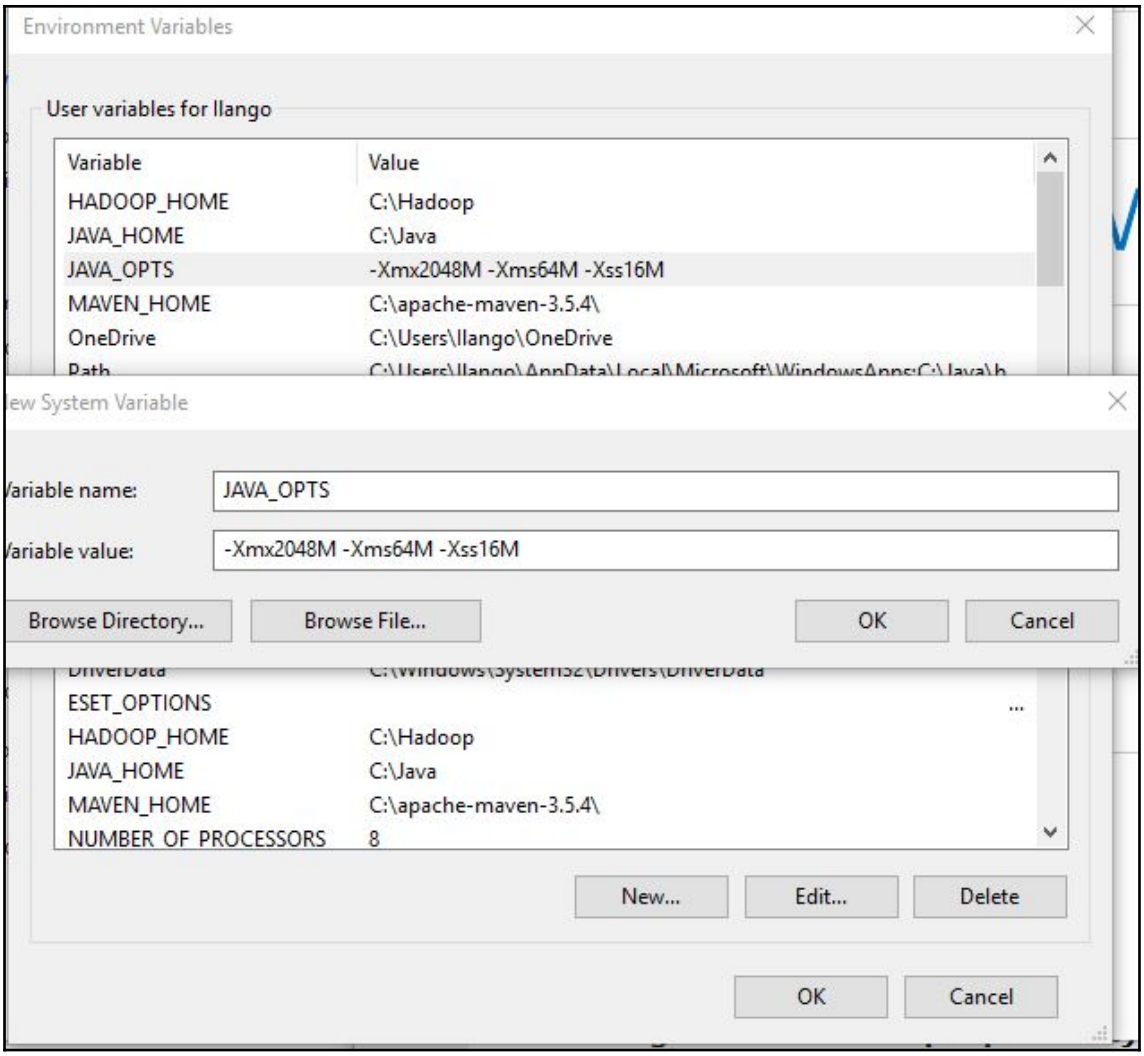

Java Control Panel
General
Update
Java
Security
Advanced
View and manage Java Runtime versions and settings for Java applications and applets.
View...
Java Runtime Environment Settings
User
System
Pla...
Product
Location
Path
Runtime Parameters
Enabled
Archite...
1.8
1.8.0_172
http://java....
C:\Program ...
-Xmx2048m
x86_64
Find
Add
Remove
OK
Cancel

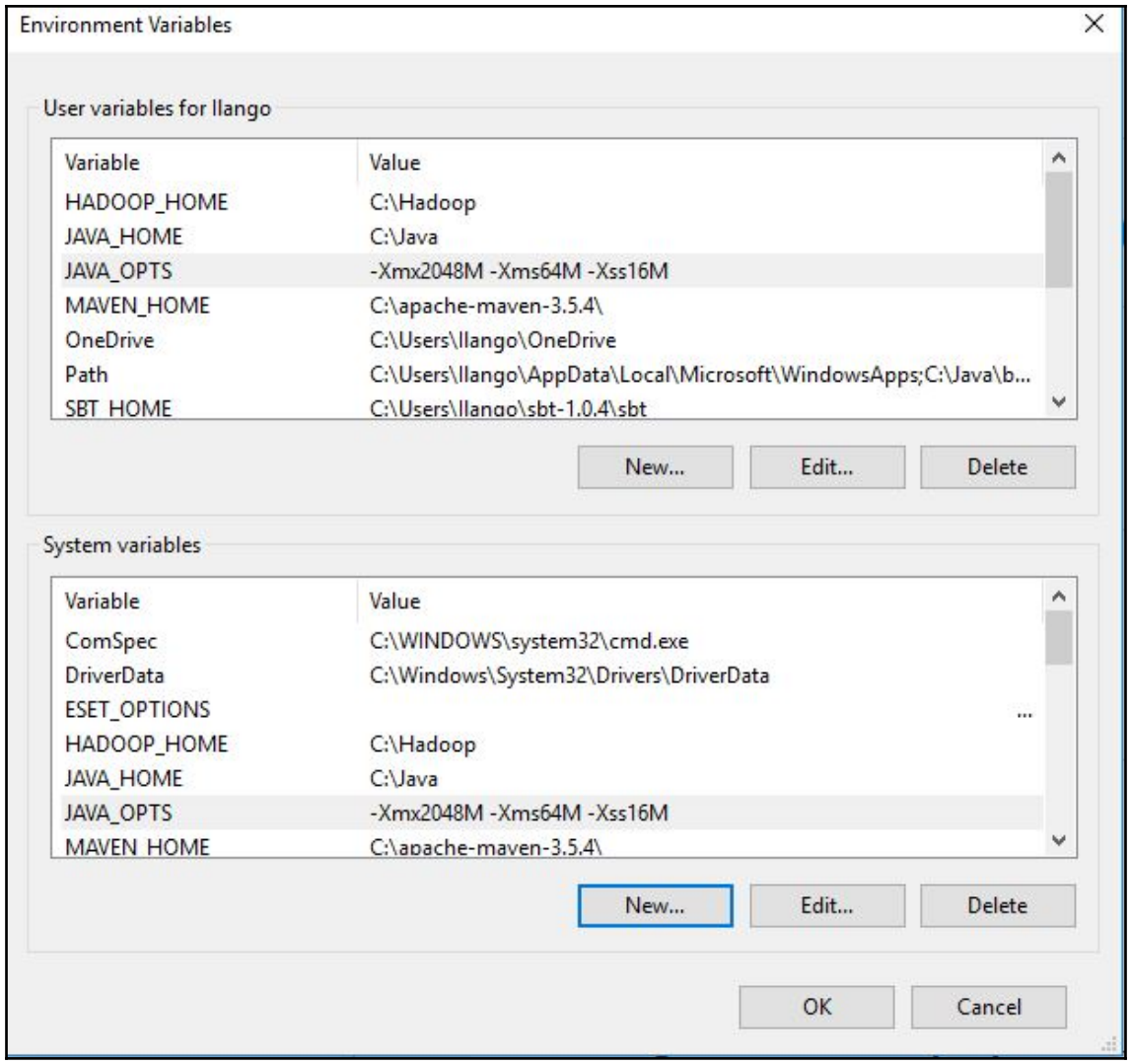
```
C:\Users\Ilango\Documents\Packt\DevProjects\Chapter63>java -X
    -Xmixed           mixed mode execution (default)
    -Xint             interpreted mode execution only
    -Xbootclasspath:<directories and zip/jar files separated by ;>
                      set search path for bootstrap classes and resources
    -Xbootclasspath/a:<directories and zip/jar files separated by ;>
                      append to end of bootstrap class path
    -Xbootclasspath/p:<directories and zip/jar files separated by ;>
                      prepend in front of bootstrap class path
    -Xdiag            show additional diagnostic messages
    -Xnoclassgc       disable class garbage collection
    -Xincgc           enable incremental garbage collection
    -Xloggc:<file>    log GC status to a file with time stamps
    -Xbatch           disable background compilation
    -Xms<size>        set initial Java heap size
    -Xmx<size>        set maximum Java heap size
    -Xss<size>        set java thread stack size
    -Xprof            output cpu profiling data
```


Collapse
Manage
Pin to Start
Map network drive...
Open in new window
Pin to Quick access
Disconnect network drive...
Add a network location
Delete
Rename
Properties

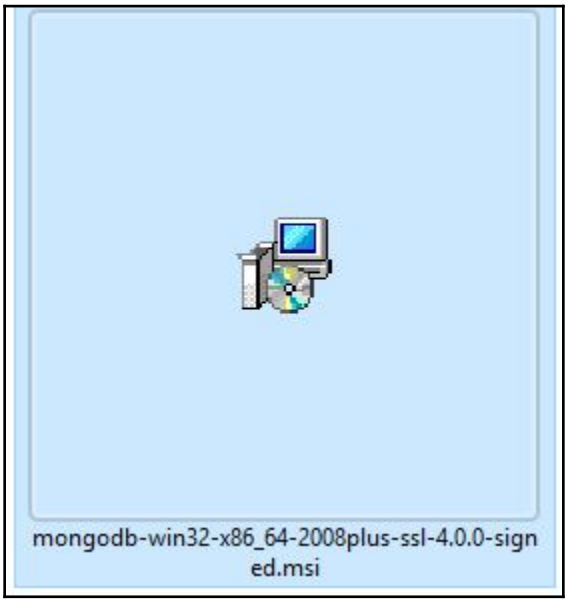

System
Control Panel > All Control Panel Items > System
Control Panel Home
Device Manager
Remote settings
System protection
Advanced system settings
View basic information about your computer
Windows edition
Windows 10 Home
© 2018 Microsoft Corporation. All rights reserved.
System
Processor: Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz 3.40 GHz
Installed memory (RAM): 32.0 GB
System type: 64-bit Operating System, x64-based processor
Pen and Touch: No Pen or Touch Input is available for this Display

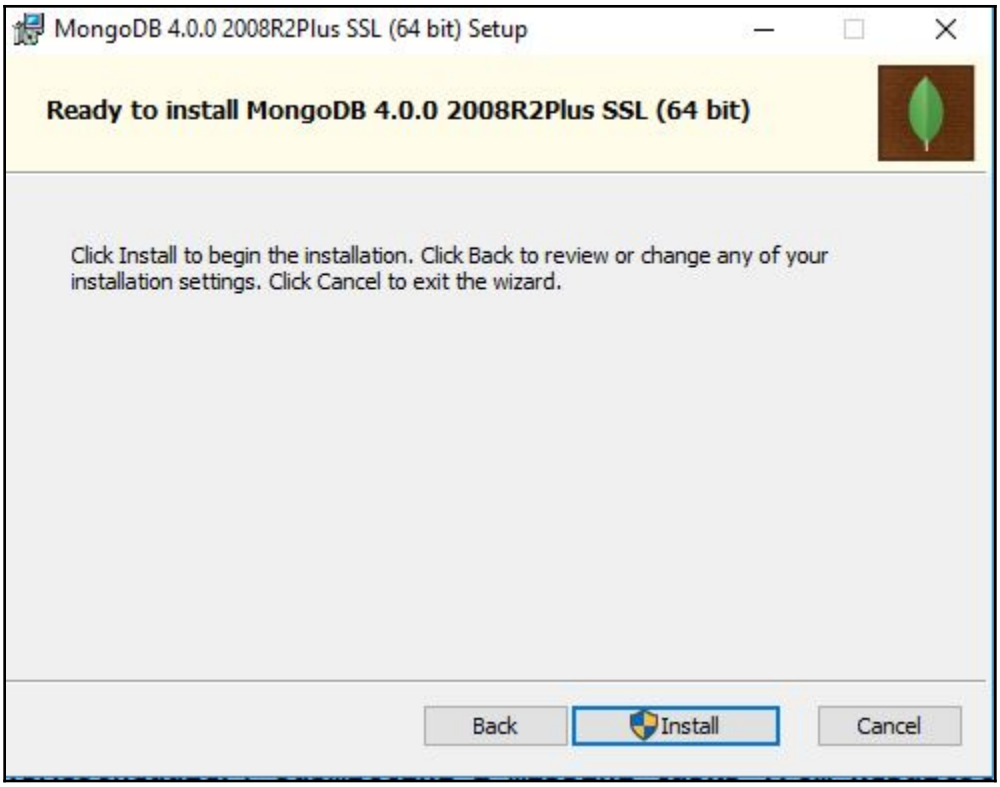

System Properties
Computer Name
Hardware
Advanced
System Protection
Remote
You must be logged on as an Administrator to make most of these changes.
Performance
Visual effects, processor scheduling, memory usage, and virtual memory
Settings...
User Profiles
Desktop settings related to your sign-in
Settings...
Startup and Recovery
System startup, system failure, and debugging information
Settings...
Environment Variables...
OK
Cancel
Apply

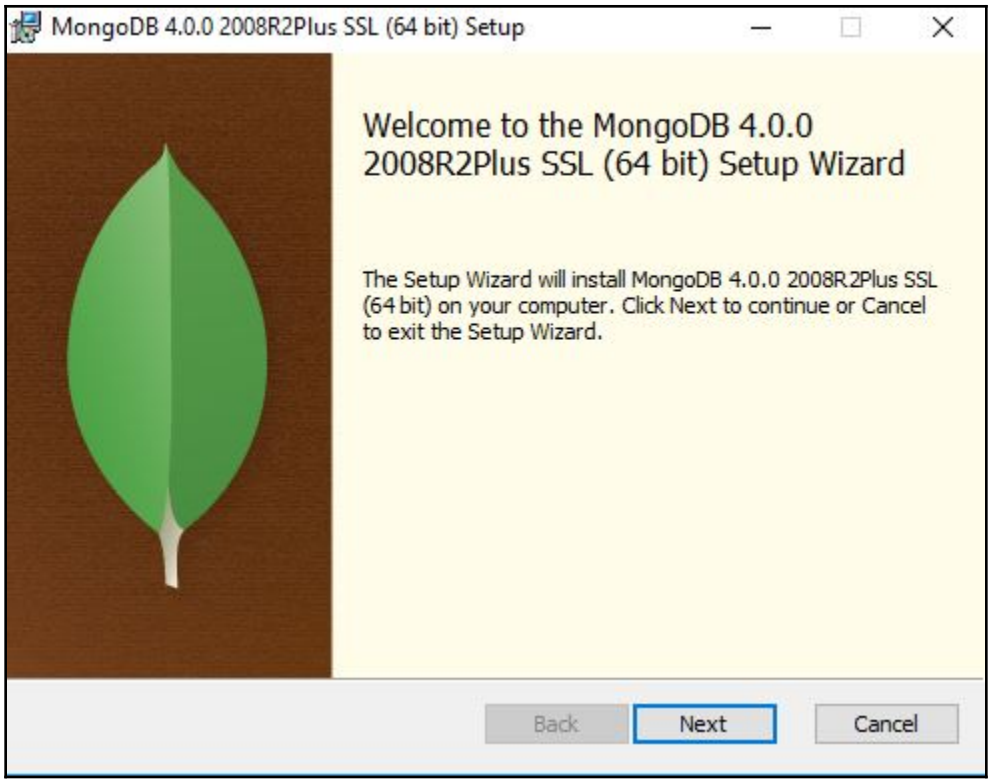

Environment Variables
User variables for llango
Variable
Value
HADOOP_HOME
C:\Hadoop
JAVA_HOME
C:\Java
JAVA_OPTS
-Xmx2048M -Xms64M -Xss16M
MAVEN_HOME
C:\apache-maven-3.5.4\
OneDrive
C:\Users\llango\OneDrive
New System Variable
Variable name:
JAVA_OPTS
Variable value:
-Xmx2048M -Xms64M -Xss16M
Browse Directory...
Browse File...
OK
Cancel
ESET_OPTIONS
HADOOP_HOME
C:\Hadoop
JAVA_HOME
C:\Java
MAVEN_HOME
C:\apache-maven-3.5.4\
NUMBER_OF_PROCESSORS
8
New...
Edit...
Delete
OK
Cancel

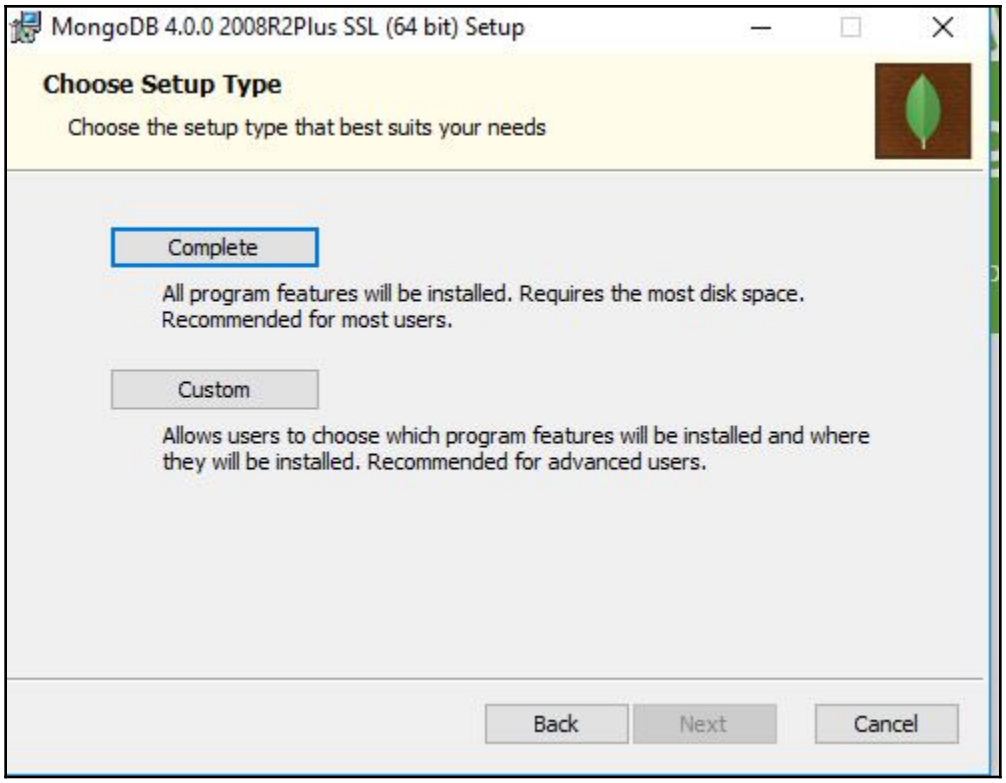

Environment Variables
User variables for llango
Variable
Value
HADOOP_HOME
C:\Hadoop
JAVA_HOME
C:\Java
JAVA_OPTS
-Xmx2048M -Xms64M -Xss16M
MAVEN_HOME
C:\apache-maven-3.5.4\
OneDrive
C:\Users\llango\OneDrive
Path
C:\Users\llango\AppData\Local\Microsoft\WindowsApps;C:\Java\b...
SBT_HOME
C:\Users\llango\sbt-1.0.4\sbt
New...
Edit...
Delete
System variables
Variable
Value
ComSpec
C:\WINDOWS\system32\cmd.exe
DriverData
C:\Windows\System32\Drivers\DriverData
ESET_OPTIONS
HADOOP_HOME
C:\Hadoop
JAVA_HOME
C:\Java
JAVA_OPTS
-Xmx2048M -Xms64M -Xss16M
MAVEN_HOME
C:\apache-maven-3.5.4\
New...
Edit...
Delete
OK
Cancel

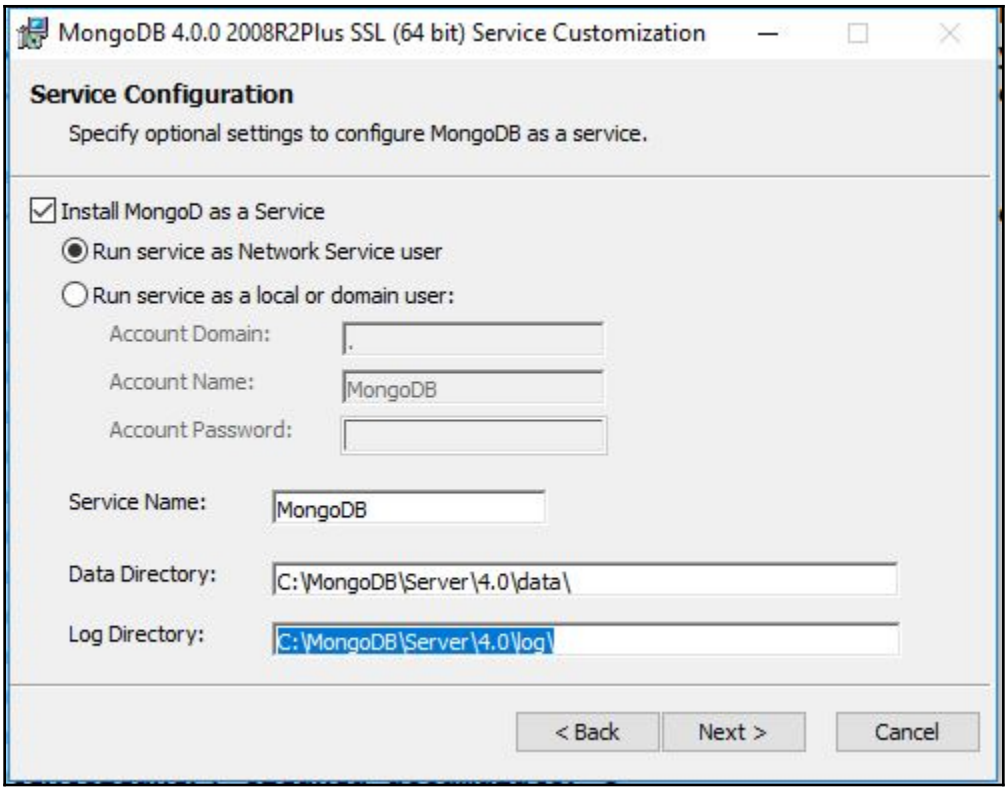
```
HADOOP_HOME                C:\Hadoop
HOMEDRIVE                  C:
HOMEPATH                   \Users\Ilango
JAVA_HOME                  C:\Java
JAVA_OPTS                  -Xmx2048M -Xms64M -Xss16M
LOCALAPPDATA               C:\Users\Ilango\AppData\Local
LOGONSERVER                \\LIVINGROOM
MAVEN_HOME                 C:\apache-maven-3.5.4\
NUMBER_OF_PROCESSORS       8
OneDrive                   C:\Users\Ilango\OneDrive
OS                         Windows_NT
Path                       C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\WINDOWS\system32;C:\WINDO.
PATHEXT                    .COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.CPL
PROCESSOR_ARCHITECTURE     AMD64
PROCESSOR_IDENTIFIER       Intel64 Family 6 Model 60 Stepping 3, GenuineIntel
PROCESSOR_LEVEL            6
PROCESSOR_REVISION         3c03
ProgramData                C:\ProgramData
ProgramFiles               C:\Program Files
ProgramFiles(x86)          C:\Program Files (x86)
ProgramW6432               C:\Program Files
PSModulePath               C:\Users\Ilango\Documents\WindowsPowerShell\Modules;C:\Program Files\WindowsPowerShel..
PUBLIC                     C:\Users\Public
SBT_HOME                   C:\Users\Ilango\sbt-1.0.4\sbt
SCALA_HOME                 C:\Program Files (x86)\scala
SPARK_HOME                 C:\spark-2.3.1-bin-hadoop2.7
```

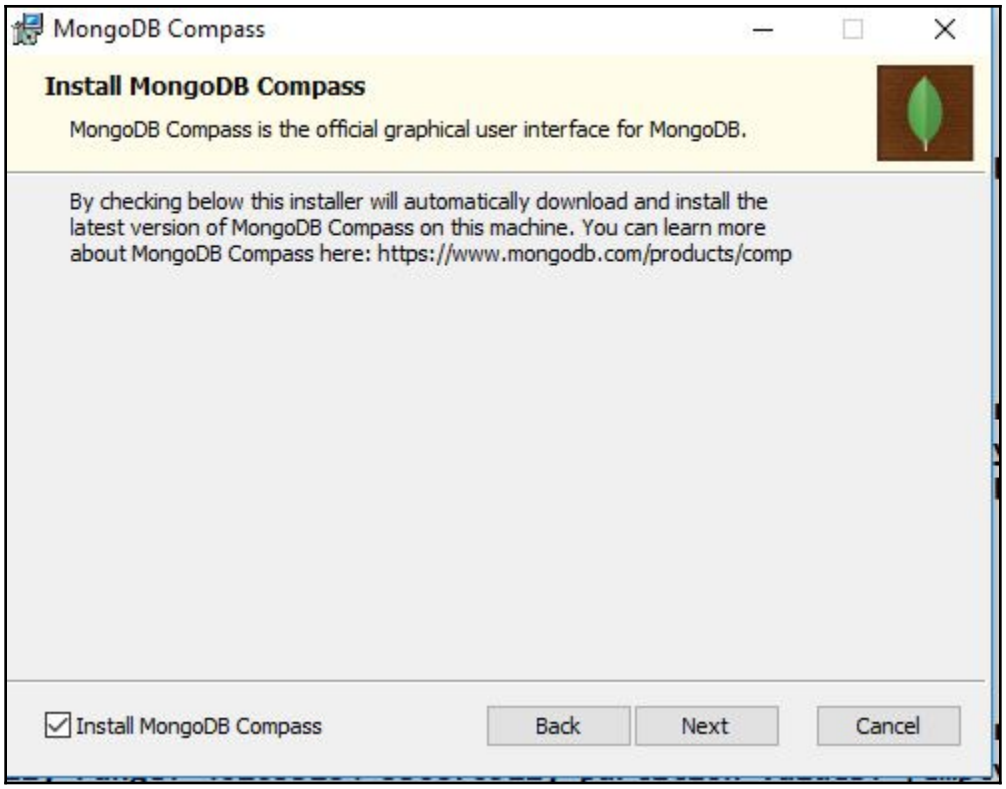


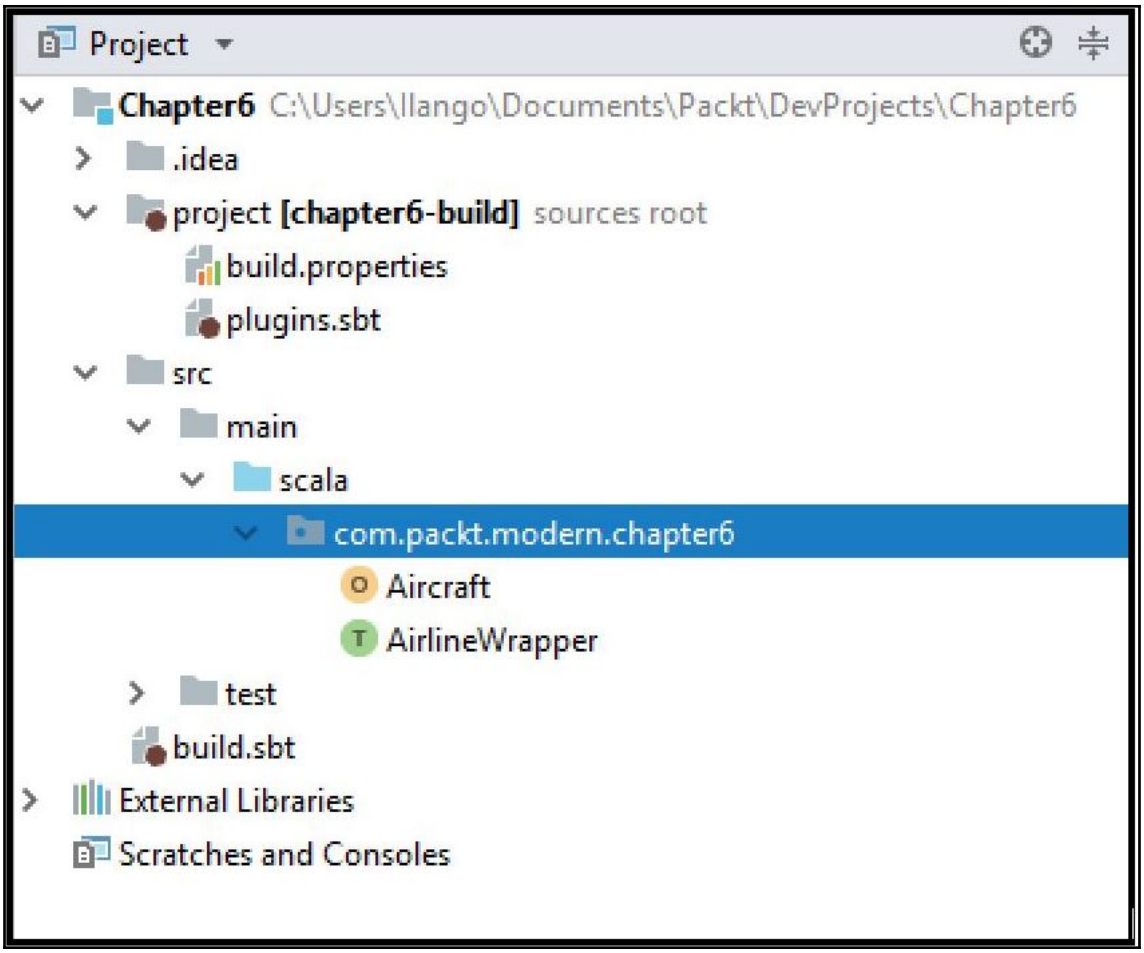

MongoDB 4.0.0 2008R2Plus SSL (64 bit) Setup
Ready to install MongoDB 4.0.0 2008R2Plus SSL (64 bit)
Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.
Back
Install
Cancel


MongoDB 4.0.0 2008R2Plus SSL (64 bit) Setup
Welcome to the MongoDB 4.0.0 2008R2Plus SSL (64 bit) Setup Wizard
The Setup Wizard will install MongoDB 4.0.0 2008R2Plus SSL (64 bit) on your computer. Click Next to continue or Cancel to exit the Setup Wizard.
Back
Next
Cancel


MongoDB 4.0.0 2008R2Plus SSL (64 bit) Setup
Choose Setup Type
Choose the setup type that best suits your needs
Complete
All program features will be installed. Requires the most disk space.
Recommended for most users.
Custom
Allows users to choose which program features will be installed and where
they will be installed. Recommended for advanced users.
Back
Next
Cancel

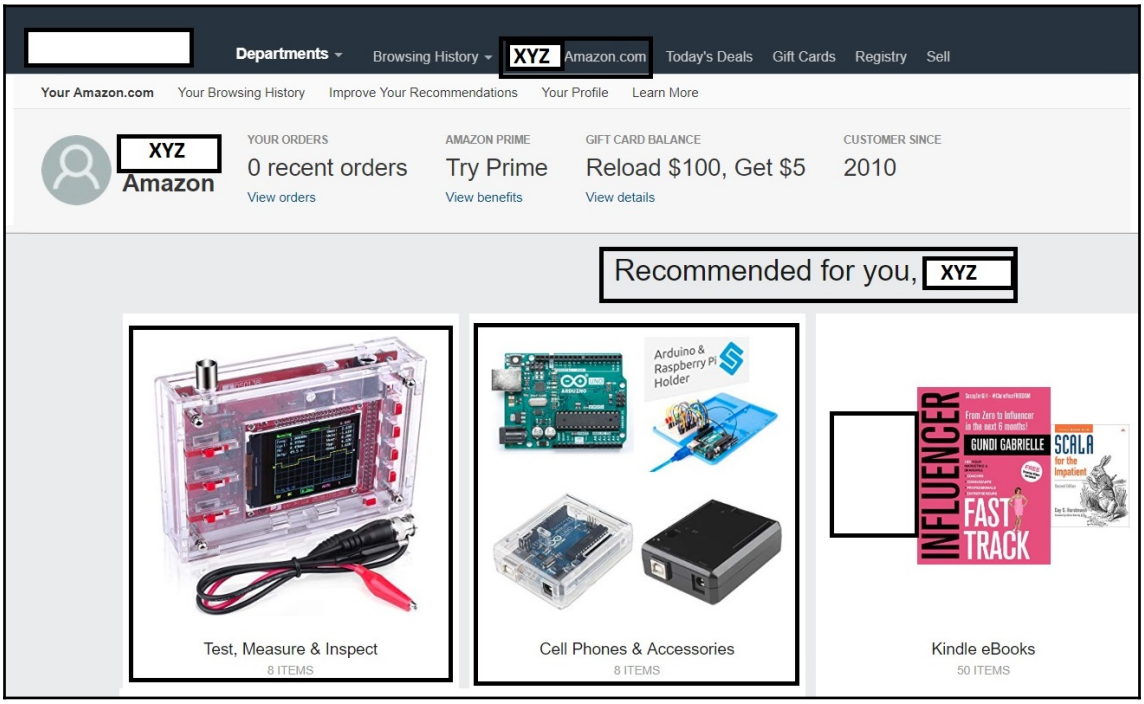

MongoDB 4.0.0 2008R2Plus SSL (64 bit) Service Customization
Service Configuration
Specify optional settings to configure MongoDB as a service.
Install MongoD as a Service
Run service as Network Service user
Run service as a local or domain user:
Account Domain:
.
Account Name:
MongoDB
Account Password:
Service Name:
MongoDB
Data Directory:
C:\MongoDB\Server\4.0\data\
Log Directory:
C:\MongoDB\Server\4.0\log\
< Back
Next >
Cancel


MongoDB Compass
Install MongoDB Compass
MongoDB Compass is the official graphical user interface for MongoDB.
By checking below this installer will automatically download and install the latest version of MongoDB Compass on this machine. You can learn more about MongoDB Compass here: https://www.mongodb.com/products/comp
Install MongoDB Compass
Back
Next
Cancel


Project
Chapter6 C:\Users\llango\Documents\Packt\DevProjects\Chapter6
.idea
project [chapter6-build] sources root
build.properties
plugins.sbt
src
main
scala
com.packt.modern.chapter6
Aircraft
AirlineWrapper
test
build.sbt
External Libraries
Scratches and Consoles

```
        Year: integer (nullable = true)
|-- Quarter: integer (nullable = true)
|-- Month: integer (nullable = true)
|-- DayofMonth: integer (nullable = true)
|-- DayOfWeek: integer (nullable = true)
|-- FlightDate: string (nullable = true)
|-- UniqueCarrier: string (nullable = true)
|-- AirlineID: integer (nullable = true)
|-- Carrier: string (nullable = true)
|-- TailNum: string (nullable = true)
|-- FlightNum: integer (nullable = true)
|-- OriginAirportID: integer (nullable = true)
|-- OriginAirportSeqID: integer (nullable = true)
|-- OriginCityMarketID: integer (nullable = true)
|-- Origin: string (nullable = true)
|-- OriginCityName: string (nullable = true)
|-- OriginState: string (nullable = true)
|-- OriginStateFips: integer (nullable = true)
|-- OriginStateName: string (nullable = true)
|-- OriginWac: integer (nullable = true)
|-- DestAirportID: integer (nullable = true)
|-- DestAirportSeqID: integer (nullable = true)
|-- DestCityMarketID: integer (nullable = true)
|-- Dest: string (nullable = true)
|-- DestCityName: string (nullable = true)
|-- DestState: string (nullable = true)
|-- DestStateFips: integer (nullable = true)
|-- DestStateName: string (nullable = true)
|-- DestWac: integer (nullable = true)
|-- CRSDepTime: integer (nullable = true)
|-- DepTime: integer (nullable = true)
|-- DepDelay: integer (nullable = true)
|-- DepDelayMinutes: integer (nullable = true)
|-- DepDel15: integer (nullable = true)
|-- DepartureDelayGroups: integer (nullable = true)
|-- DepTimeBlk: string (nullable = true)
|-- TaxiOut: integer (nullable = true)
|-- WheelsOff: integer (nullable = true)
|-- WheelsOn: integer (nullable = true)
|-- TaxiIn: integer (nullable = true)
|-- CRSArrTime: integer (nullable = true)
|-- ArrTime: integer (nullable = true)
```

Chapter 7: Building a Recommendation Engine



Your recently viewed items and featured recommendations

Inspired by your browsing history

Inspired by your purchases

Frequently bought together

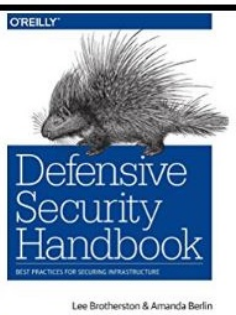

Total price: $54.77

Add both to Cart

Add both to List

- **This item:** Definitive Guide to LEGO MINDSTORMS, Second Edition by Dave Baum Paperback $29.99
- Extreme Mindstorms: an Advanced Guide to Lego Mindstorms by Michael Gasperi Paperback $24.78

Customers who bought this item also bought



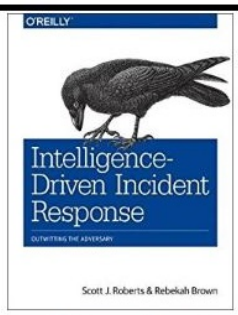
Defensive Security Handbook: Best Practices for Securing Infrastructure

› Lee Brotherston

39

Paperback

$28.95 prime



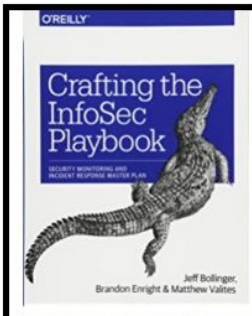
Intelligence-Driven Incident Response: Outwitting the Adversary

› Scott J. Roberts

7

Paperback

$48.47 prime



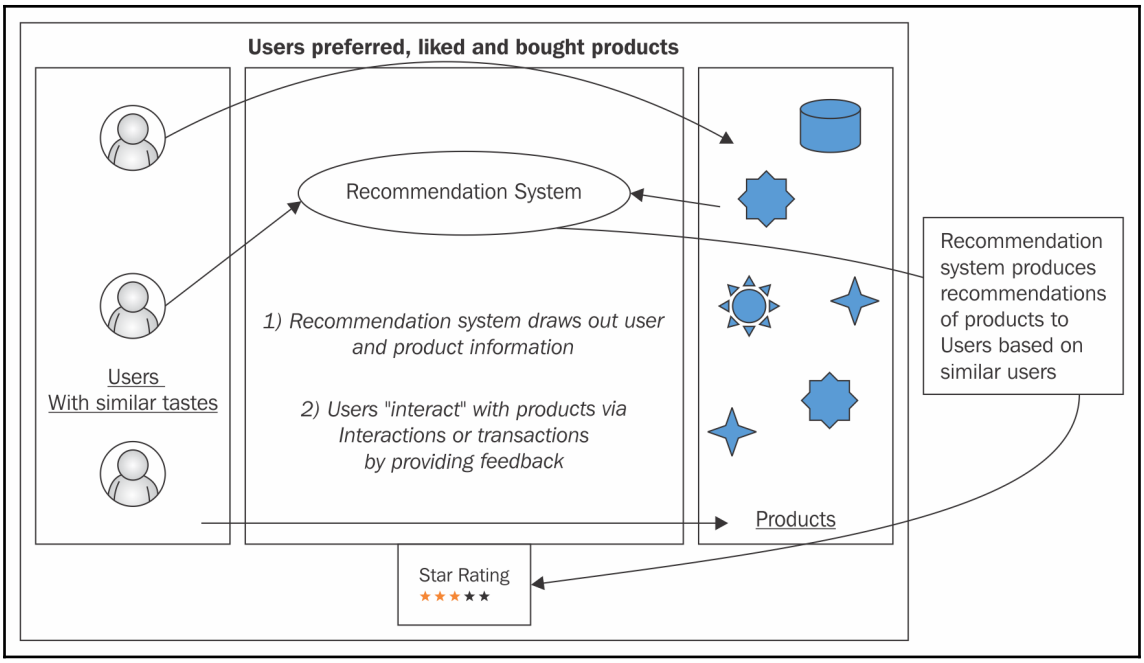
Crafting the InfoSec Playbook: Security Monitoring and Incident…

› Jeff Bollinger

9

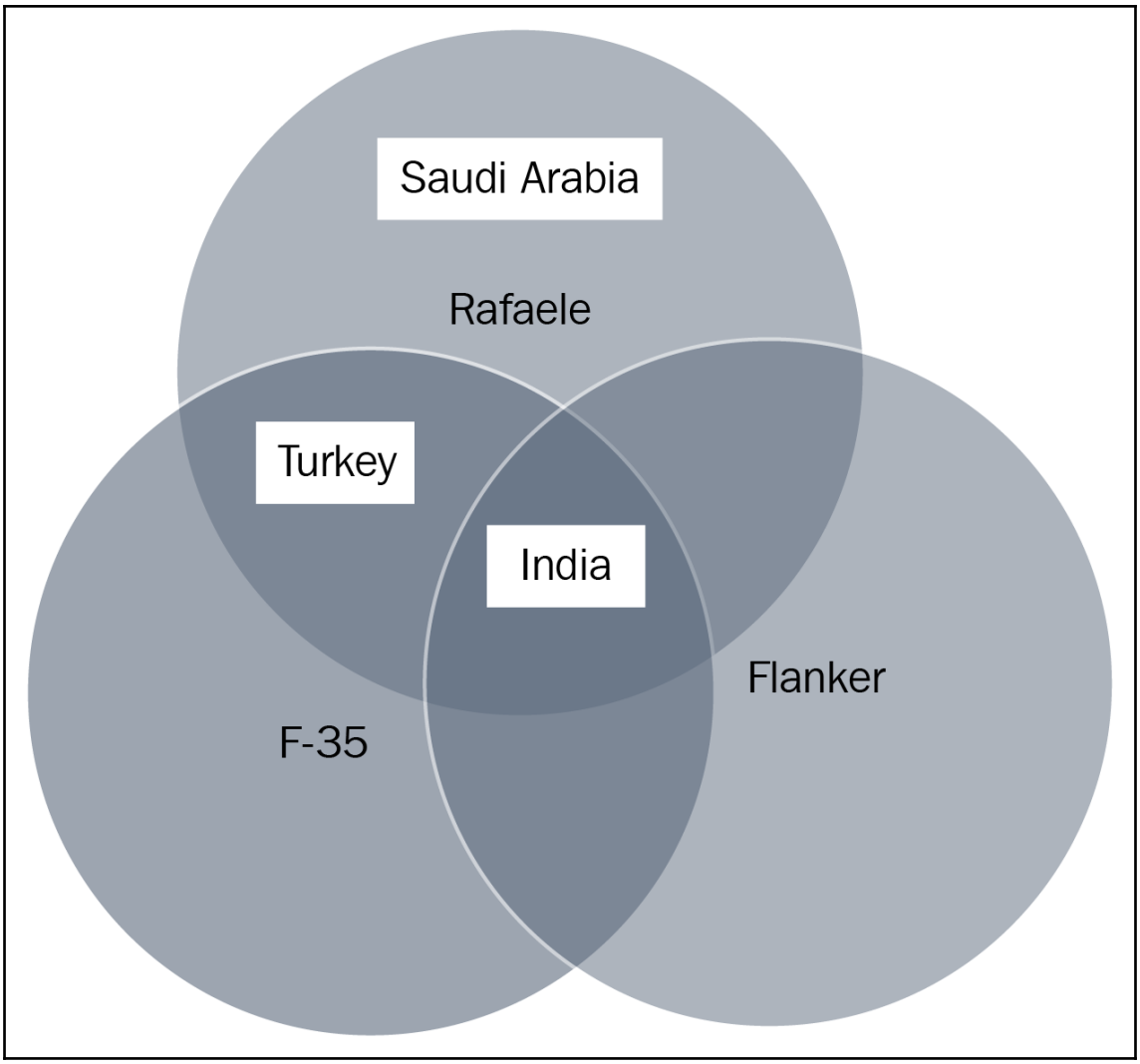
Paperback

$41.27 prime



Ratings are on a scale of 1 – 10 with 1 being deplorable to 10 being awesome

Fighter Aircraft		**Rafaele Air Superority Fighter**	**SU-35 Flanker**	**The F-35 Joint Strike Fighter**
Country/Customer		**Product 1**	**Product 2**	**Product 3**
U1	**India**	**Rating = 7**	**Rating = 8**	**Rating = 7**
U2	**Turkey**	**Rating = 6**		**Rating = 7**
U3	**Saudi Arabia**	**Rating = 7**		


Saudi Arabia
Rafaele
Turkey
India
Flanker
F-35

CustomerId	CustomerName	ItemId	ItemName	ItemUnitPrice	OrderSize	AmountPaid
1	Australia	217	WeaponsSystem217	2	25	50
1	Australia	183	WeaponsSystem183	6	9	64
2	Seychelles	355	WeaponsSystem355	3	20	60
3	Fiji	99	WeaponsSystem99	5	3	15
3	Fiji	217	WeaponsSystem217	2	10	20
4	Turkey	160	WeaponsSystem160	5	20	100
4	Turkey	45	WeaponsSystem45	3	10	40
5	Jordan	100	WeaponsSystem100	2	4	8
5	Jordan	57	WeaponsSystem57	3	5	15
6	SKorea	81	WeaponsSystem81	5	8	40
6	SKorea	217	WeaponsSystem217	2	26	52
7	Djibouti	107	WeaponsSystem107	3	15	45
7	Djibouti	30	WeaponsSystem30	4	4	16
7	Djibouti	355	WeaponsSystem355	3	5	15
8	India	217	WeaponsSystem217	2	36	72
8	India	99	WeaponsSystem99	5	120	600
8	India	45	WeaponsSystem45	3	20	60

```
root
 |-- sCustomerId: integer (nullable = true)
 |-- sCustomerName: string (nullable = true)
 |-- sItemId: integer (nullable = true)
 |-- sItemName: string (nullable = true)
 |-- sItemUnitPrice: double (nullable = true)
 |-- sOrderSize: double (nullable = true)
 |-- sAmountPaid: double (nullable = true)
```

CustomerId	CustomerName	ItemId	ItemName
1	Australia	99	WeaponsSystem99
1	Australia	101	WeaponsSystem101
2	Seychelles	89	WeaponsSystem89
2	Seychelles	217	WeaponsSystem217
3	Fiji	160	WeaponsSystem160
3	Fiji	217	WeaponsSystem217
4	Turkey	183	WeaponsSystem183
4	Turkey	100	WeaponsSystem100
4	Turkey	160	WeaponsSystem160
8	India	217	WeaponSystem217
5	Jordan	99	WeaponsSystem99
5	Jordan	355	WeaponsSystem355
6	SKorea	107	WeaponsSystem107
6	SKorea	217	WeaponsSystem217
7	Djibouti	57	WeaponsSystem57
7	Djibouti	355	WeaponsSystem355
7	Djibouti	183	WeaponsSystem183
8	India	45	WeaponsSystem45
8	India	81	WeaponsSystem81
8	India	160	WeaponsSystem160
8	India	355	WeaponsSystem355

sCustomerId	sCustomerName	sItemId	sItemName	sItemUnitPrice	sOrderSize	sAmountPaid
1	Australia	217	WeaponsSystem217	2.0	25.0	50.0
1	Australia	183	WeaponsSystem183	6.0	9.0	64.0
2	Seychelles	355	WeaponsSystem355	3.0	20.0	60.0
3	Fiji	99	WeaponsSystem99	5.0	3.0	15.0
3	Fiji	217	WeaponsSystem217	2.0	10.0	20.0
4	Turkey	160	WeaponsSystem160	5.0	20.0	100.0
4	Turkey	45	WeaponsSystem45	3.0	10.0	40.0
5	Jordan	100	WeaponsSystem100	2.0	4.0	8.0

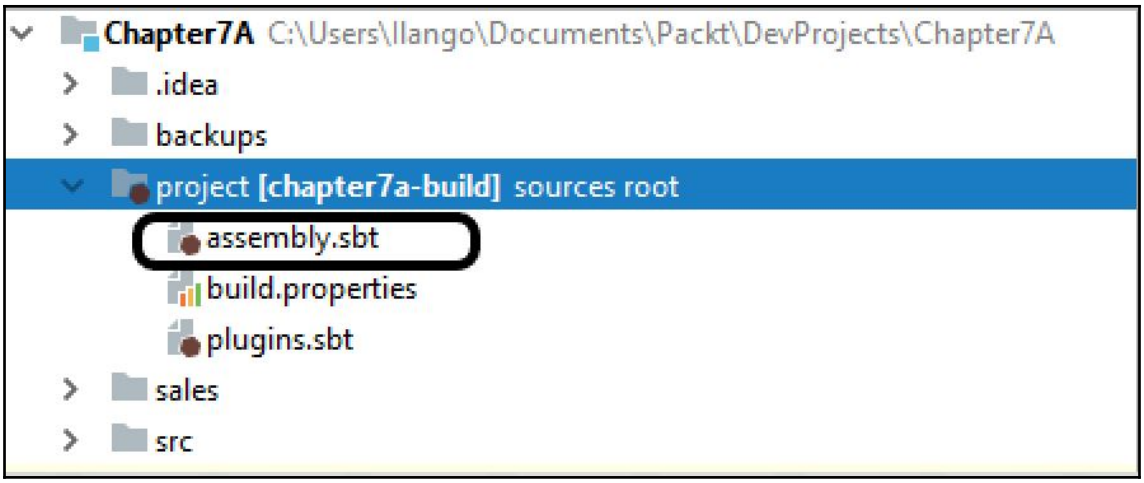
sCustomerId	sCustomerName	sItemId	sItemName
1	Australia	99	WeaponsSystem99
1	Australia	101	WeaponsSystem101
2	Seychelles	89	WeaponsSystem89
2	Seychelles	217	WeaponsSystem217
3	Fiji	160	WeaponsSystem160
3	Fiji	217	WeaponsSystem217
4	Turkey	183	WeaponsSystem183
4	Turkey	100	WeaponsSystem100
4	Turkey	160	WeaponsSystem160
8	India	217	WeaponSystem217
5	Jordan	99	WeaponsSystem99
5	Jordan	355	WeaponsSystem355
6	SKorea	107	WeaponsSystem107
6	SKorea	217	WeaponsSystem217
7	Djibouti	57	WeaponsSystem57
7	Djibouti	355	WeaponsSystem355
7	Djibouti	183	WeaponsSystem183
8	India	45	WeaponsSystem45
8	India	81	WeaponsSystem81
8	India	160	WeaponsSystem160

only showing top 20 rows

```
The Customer-Weapons System dataframe as tuple pairs looks like:
+----+----+
|user|item|
+----+----+
|   1|  99|
|   1| 101|
|   2|  89|
|   2| 217|
|   3| 160|
|   3| 217|
|   4| 183|
|   4| 100|
|   4| 160|
|   8| 217|
|   5|  99|
|   5| 355|
|   6| 107|
|   6| 217|
|   7|  57|
|   7| 355|
|   7| 183|
|   8|  45|
|   8|  81|
|   8| 160|
+----+----+
only showing top 20 rows
```

```
Customer Nation: 1 Weapons System:  99 Rating: 110.64294367754968
Customer Nation: 2 Weapons System:  217 Rating: -14.59353061514684
Customer Nation: 8 Weapons System:  45 Rating: 58.8411030583238
Customer Nation: 6 Weapons System:  217 Rating: 27.983562575277176
Customer Nation: 7 Weapons System:  183 Rating: -4.752616315700221
Customer Nation: 5 Weapons System:  355 Rating: -0.022699539162764973
Customer Nation: 6 Weapons System:  107 Rating: -6.146977037630051
Customer Nation: 8 Weapons System:  217 Rating: 70.75430361741672
Customer Nation: 3 Weapons System:  160 Rating: -2.7402710706968243
Customer Nation: 3 Weapons System:  217 Rating: 7.4330505088971295
Customer Nation: 4 Weapons System:  100 Rating: 0.018719075478031554
Customer Nation: 8 Weapons System:  160 Rating: 149.8431054419417
Customer Nation: 4 Weapons System:  160 Rating: 80.74557079776903
Customer Nation: 7 Weapons System:  355 Rating: 14.726831549178502
Customer Nation: 8 Weapons System:  81 Rating: 68.79163228044008
Customer Nation: 8 Weapons System:  355 Rating: -33.267337863617
Customer Nation: 5 Weapons System:  99 Rating: 0.13603012624246374
Customer Nation: 7 Weapons System:  57 Rating: -0.025637651994121942
Customer Nation: 4 Weapons System:  183 Rating: 5.413882517433166
```

```
C:\Users\Ilango\Documents\Packt\DevProjects\Chapter7A>sbt compile
"C:\Users\Ilango\.sbt\preloaded\org.scala-sbt\sbt\"1.0.4"\jars\sbt.jar"
[info] Loading settings from idea.sbt ...
[info] Loading global plugins from C:\Users\Ilango\.sbt\1.0\plugins
[info] Loading settings from assembly.sbt,plugins.sbt ...
[info] Loading project definition from C:\Users\Ilango\Documents\Packt\DevProjects\Chapter7A\project
[info] Loading settings from build.sbt ...
[info] Set current project to Chapter7A (in build file:/C:/Users/Ilango/Documents/Packt/DevProjects/Chapter7A/)
[info] Executing in batch mode. For better performance use sbt's shell
[success] Total time: 1 s, completed Jul 16, 2018 6:54:25 PM
```

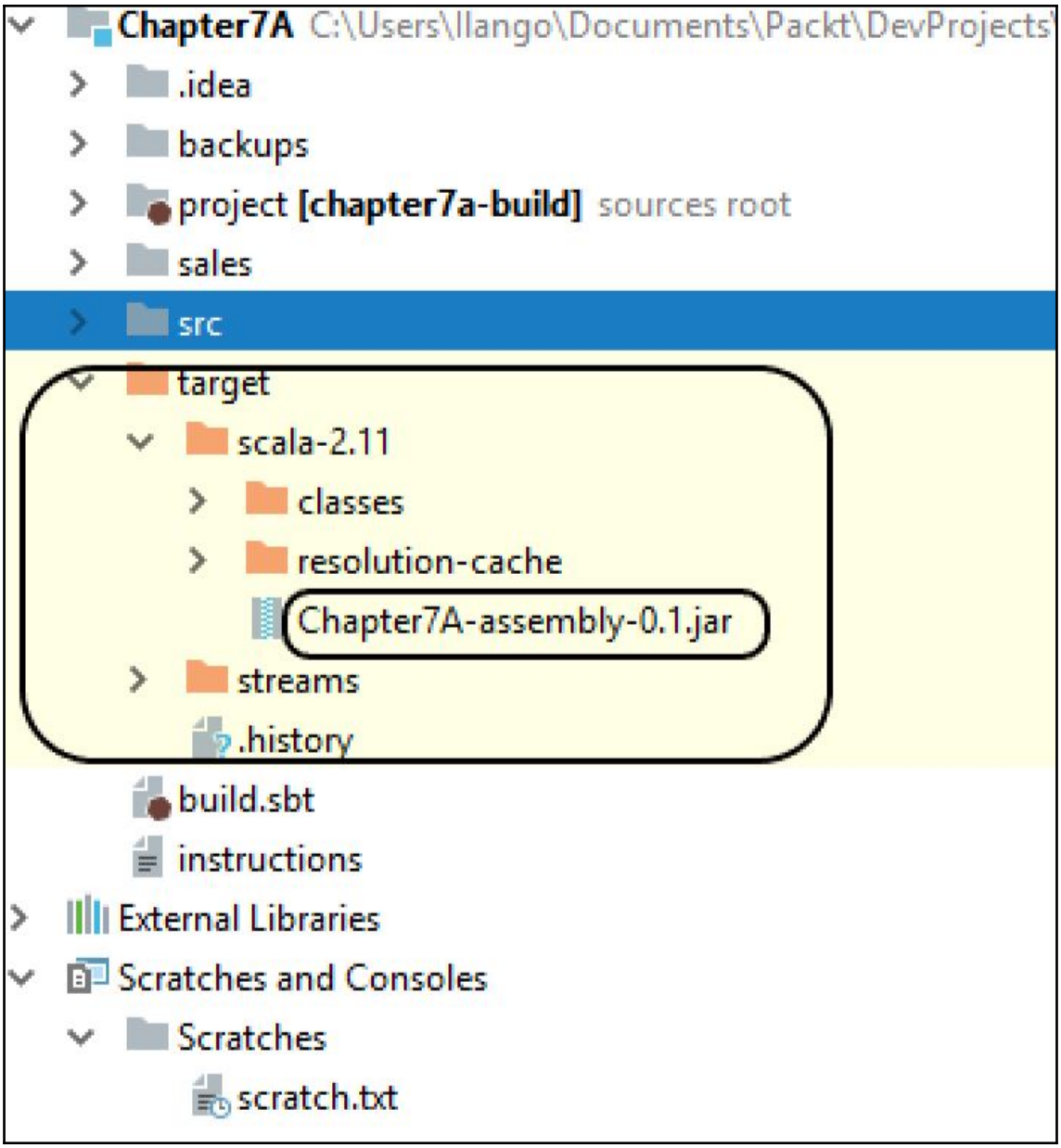



```
addSbtPlugin( dependency = "com.eed3si9n" % "sbt-assembly" % "0.14.7")
```

```
C:\Users\Ilango\Documents\Packt\DevProjects\Chapter7A>sbt assembly
"C:\Users\Ilango\.sbt\preloaded\org.scala-sbt\sbt\"1.0.4"\jars\sbt.jar"
[info] Loading settings from idea.sbt ...
[info] Loading global plugins from C:\Users\Ilango\.sbt\1.0\plugins
[info] Loading settings from assembly.sbt,plugins.sbt ...
[info] Loading project definition from C:\Users\Ilango\Documents\Packt\DevProjects\Chapter7A\project
[info] Loading settings from build.sbt ...
[info] Set current project to Chapter7A (in build file:/C:/Users/Ilango/Documents/Packt/DevProjects/Chapter7A/)
[error] 112 errors were encountered during merge
[error] java.lang.RuntimeException: deduplicate: different file contents found in the following:
[error] C:\Users\Ilango\.ivy2\cache\org.apache.arrow\arrow-vector\jars\arrow-vector-0.8.0.jar:git.properties
[error] C:\Users\Ilango\.ivy2\cache\org.apache.arrow\arrow-format\jars\arrow-format-0.8.0.jar:git.properties
[error] C:\Users\Ilango\.ivy2\cache\org.apache.arrow\arrow-memory\jars\arrow-memory-0.8.0.jar:git.properties
[error] deduplicate: different file contents found in the following:
```

```
mainClass in (Compile, run) := Some("com.packt.modern.chapter7.RecSystem")

libraryDependencies ++= Seq(
  "org.apache.spark" %% "spark-core" % "2.3.1",
  "org.apache.spark" %% "spark-mllib" % "2.3.1",
  "org.apache.spark" %% "spark-sql" % "2.3.1"
)


resolvers += "Sonatype Releases" at "https://oss.sonatype.org/content/repositories/releases/"
resolvers += "Sonatype Snapshots" at "http://oss.sonatype.org/content/repositories/snapshots"
fork in run := true
fork in test := true


assemblyMergeStrategy in assembly := {
  case PathList("META-INF", xs @ _*) => MergeStrategy.discard
  case x => MergeStrategy.first
}
```

```
C:\Users\Ilango\Documents\Packt\DevProjects\Chapter7A>sbt assembly
"C:\Users\Ilango\.sbt\preloaded\org.scala-sbt\sbt\"1.0.4"\jars\sbt.jar"
[info] Loading settings from idea.sbt ...
[info] Loading global plugins from C:\Users\Ilango\.sbt\1.0\plugins
[info] Loading settings from assembly.sbt, plugins.sbt ...
[info] Loading project definition from C:\Users\Ilango\Documents\Packt\DevProjects\Chapter7A\project
[info] Loading settings from build.sbt ...
[info] Set current project to Chapter7A (in build file:/C:/Users/Ilango/Documents/Packt/DevProjects/Chapter7A/)
[info] Strategy 'discard' was applied to 520 files (Run the task at debug level to see details)
[info] Strategy 'first' was applied to 438 files (Run the task at debug level to see details)
[info] Packaging C:\Users\Ilango\Documents\Packt\DevProjects\Chapter7A\target\scala-2.11\Chapter7A-assembly-0.1.jar
[info] Done packaging.
[success] Total time: 121 s, completed Jul 16, 2018 6:56:40 PM
```




```
C:\Users\Ilango\Documents\Packt\DevProjects\Chapter7A>spark-submit --class "com.packt.modern.chapter7.RecSys
tem" --master local[2] --deploy-mode client --driver-memory 16g --num-executors 2 --executor-memory 2g --exe
cutor-cores 2 "target\\scala-2.11\\Chapter7A-assembly-0.1.jar"
```

Command-Line Parameter	Value	Explanation
--class	com.packt.modern.chapter7.RecSystem	The entry point to the application
--master	The default value: Local[2]	The Master URL, which defaults to local[*]
--deploy-mode	The default value client	Either deploy on worker-nodes in a cluster or locally
--driver-memory	4g or 8g or 16g	This value at any rate cannot exceed the total RAM on your machine
--num-executors	2	The number of executors to be created
--executor-memory	2g	Memory allocated to each executor
--executor-cores	2	The number of concurrent threads available for every executor