CAP4770/5771
Lab 4
Spark MLlib: Logistic Regression on AWS

University of Florida, CISE Department
TA: Xiaofeng Zhou

Outline

- What is Spark
- A simple example on AWS
    - Start a AWS EMR Spark cluster
    - Logistic regression: a simple example

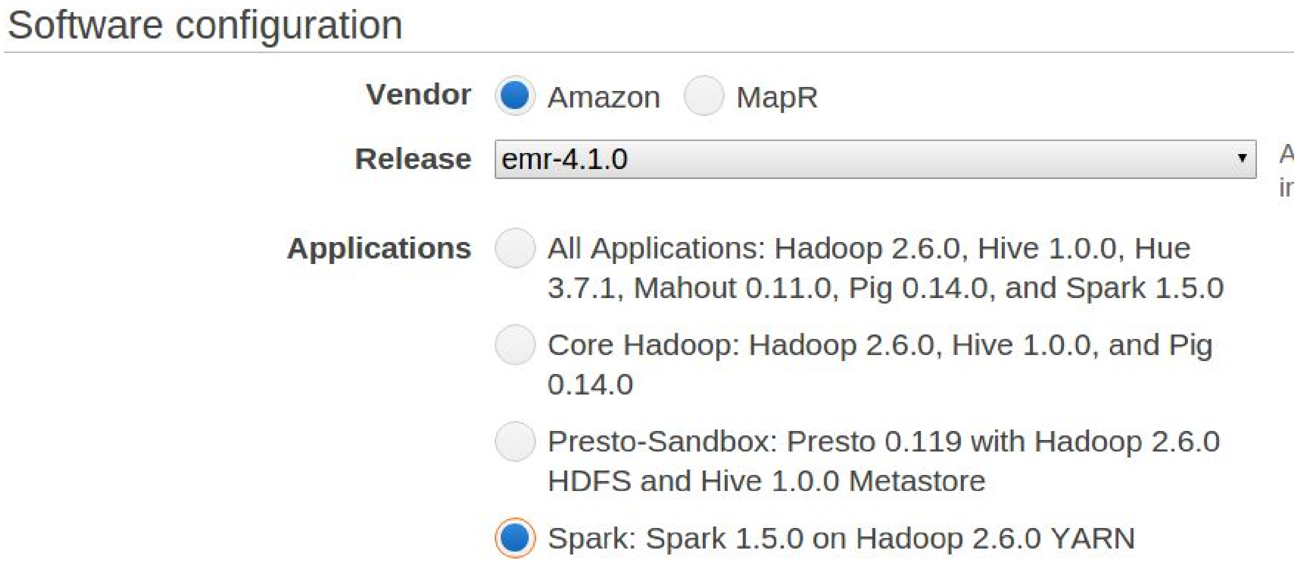
What is Spark

- Apache Spark is a fast and general engine for large-scale data processing.
  Spark over Hadoop
  - Spark stores data in-memory (as much as possible) whereas Hadoop stores data on disk.
  - Spark uses resilient distributed datasets (RDD), a clever way of guaranteeing fault tolerance that minimizes network I/O, whereas Hadoop uses replication to achieve fault tolerance.

  Faster, especially on iterative algorithms
- MLlib: Spark’s Machine Learning(ML) library
  - build on spark
  - supports Python, Scala, Java
  - broad cover of ML algorithms

A simple example on AWS

1. Firstly let’s create a Spark EMR cluster:



Software configuration

Vendor ● Amazon ○ MapR

Release emr-4.1.0 ▾

A
ir

Applications

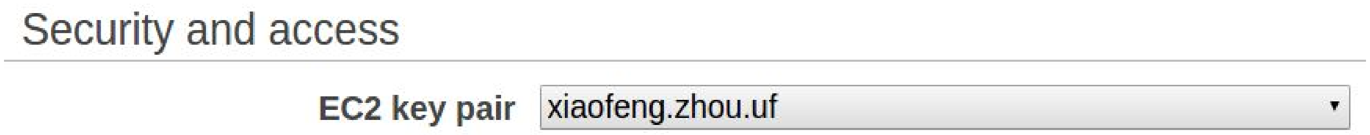
○ All Applications: Hadoop 2.6.0, Hive 1.0.0, Hue 3.7.1, Mahout 0.11.0, Pig 0.14.0, and Spark 1.5.0

○ Core Hadoop: Hadoop 2.6.0, Hive 1.0.0, and Pig 0.14.0

○ Presto-Sandbox: Presto 0.119 with Hadoop 2.6.0 HDFS and Hive 1.0.0 Metastore

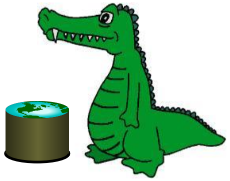
● Spark: Spark 1.5.0 on Hadoop 2.6.0 YARN

Choose your EC2 Key pair for ssh access



Security and access

EC2 key pair xiaofeng.zhou.uf ▾



A simple example on AWS - cont

Follow the example here(Python)

1. Download the input file in the example from here
2. put it in your bucket.
3. ssh into the Spark cluster master node
4. type “pyspark” to start coding!

Code Walk through

```
from pyspark.mllib.classification import
LogisticRegressionWithLBFGS, LogisticRegressionModel
from pyspark.mllib.regression import LabeledPoint
# Load and parse the data
def parsePoint(line):
    values = [float(x) for x in line.split(' ')]
    return LabeledPoint(values[0], values[1:])
data = sc.textFile("s3://uf-dsr-courses-ids/sample_svm_data.txt")
parsedData = data.map(parsePoint)
# Build the model
model = LogisticRegressionWithLBFGS.train(parsedData)
```

Code Walk through - cont

```
# Evaluating the model on training data
labelsAndPreds = parsedData.map(lambda p: (p.label, model.
predict(p.features)))
trainErr = labelsAndPreds.filter(lambda (v, p): v != p).
count() / float(parsedData.count())
print("Training Error = " + str(trainErr))
# Save and load model
model.save(sc, "s3://uf-dsr-courses-ids/savedModel")
sameModel = LogisticRegressionModel.load(sc, "savedModel")
```

Spark MLlib for NIST Pre-Pilot

1. For very small dataset, you can use Scikit, but for large dataset in the NIST project, you will need to use Spark MLlib.
2. Remember to shut EMR cluster down when you have finished using it.