Data Science with Spark

Shelly Garion
IBM Research -- Haifa
Overview – Advanced Data Analysis Tools

- **Spark MLLib** – large scale machine learning
  - RDD based API
  - DataFrame based API

- **Spark GraphX** – graph-parallel processing

- **How to clean your data?**
- **How to combine it all?**
- **How to visualize it?**
Why Spark MLLib & GraphX?

Combined Analytics of Data

Analyze tabular data with SQL

Analyze graph data using GraphX graph analytics engine

Use same machine learning infrastructure

Use same solution for streaming data

Joseph Gonzalez, Reynold Xin, Ankur Dave, Daniel Crankshaw, Michael Franklin, and Ion Stoica,
“GRAPHX: UNIFIED GRAPH ANALYTICS ON SPARK”, spark summit July 2014
Machine Learning Algorithms

- **Classification**
  - Logistic regression
  - Linear support vector machine (SVM)
  - Naïve Bayes
  - Decision trees and forests

- **Regression**
  - Generalized linear regression (GLM)

- **Recommendation**
  - Alternating least squares (ALS)

- **Clustering**
  - K-means and Streaming K-means
  - Gaussian mixture
  - Latent Dirichlet allocation (LDA)

- **Dimensionality reduction**
  - Singular value decomposition (SVD)
  - Principal component analysis (PCA)

- **Feature extraction & selection**
  - Word2Vec

See: https://spark.apache.org/docs/latest/mllib-guide.html
Performance of MLLib

- It is built on Apache Spark, a fast and general engine for large-scale data processing.
- Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Logistic Regression

ALS Results

- [https://spark.apache.org/](https://spark.apache.org/)
Performance of MLLib

- Speed-up between MLLib versions

Figure 2: (a) Benchmarking results for ALS. (b) MLLib speedup between versions.

Example: K-Means Clustering (RDD based API)

Goal:
Segment tweets into clusters by geolocation using Spark MLLib K-means clustering

Example: K-Means Clustering (RDD based API)

To run the k-means algorithm in Spark, we need to first read the csv file

```scala
val sc = new SparkContext("local[4]", "kmeans")
// Load and parse the data, we only extract the latitude and longitude of each line
val data = sc.textFile(arg)
val parsedData = data.map {
  line =>
    Vectors.dense(line.split(',').slice(0, 2).map(_.toDouble))
}
```

Then we can run the spark kmeans algorithm:

```scala
val iterationCount = 100
val clusterCount = 10
val model = KMeans.train(parsedData, clusterCount, iterationCount)
```

Example: K-Means Clustering (RDD based API)

From the model we can get the cluster centers and group the tweets by cluster:

```scala
val clusterCenters = model.clusterCenters map (_.toArray)
val cost = model.computeCost(parsedData)
println("Cost: " + cost)
val tweetsByGoup = data
    .map {_.split(',').slice(0, 2).map(_.toDouble)}
    .groupBy{rdd => model.predict(Vectors.dense(rdd))}
    .collect()
sc.stop()
```

Example: K-Means Clustering (RDD based API)

Machine Learning Pipeline with Spark MLLib

Data pre-processing
Feature extraction
Model fitting
Model training
Validation
Model prediction
Spark MLLib Pipeline (DataFrame based API)

```scala
// create pipeline
tok = Tokenizer(in="text", out="words")
tf = HashingTF(in="words", out="features")
lr = LogisticRegression(maxIter=10, regParam=0.01)
pipeline = Pipeline(stages=[tok, tf, lr])

// train pipeline
df = sqlCtx.table("training")
model = pipeline.fit(df)

// make predictions
df = sqlCtx.read.json("/path/to/test")
model.transform(df)
  .select("id", "text", "prediction")
```

Spark MLLib Pipeline (DataFrame based API)

- **DataFrame:**
  - Use DataFrame from Spark SQL as ML dataset
  - Can have different columns storing text, feature vectors, true labels, and predictions

- **Transformer:**
  - A Transformer implements a method `transform()`
  - Algorithm that transforms one DataFrame to another DataFrame
    - Feature transformers (e.g., OneHotEncoder)
    - Trained ML models (e.g., LogisticRegressionModel)

- **Estimator:**
  - An Estimator implements a method `fit()`
  - Algorithm which can be fit on a DataFrame to produce a transformer
    - ML algorithms which trains on a DataFrame and produces a model (e.g., LogisticRegression)

- **Pipeline:**
  - Chains multiple Transformers and Estimators together to specify an ML workflow

https://spark.apache.org/docs/latest/ml-pipeline.html
Machine Learning Pipeline with Spark MLLib

Learning:

Pipeline (Estimator)

Pipeline.fit()

Raw text → Words → Feature vectors → Logistic Regression Model

Model:

PipelineModel (Transformer)

PipelineModel.transform()

Raw text → Words → Feature vectors → Predictions

https://spark.apache.org/docs/latest/ml-pipeline.html
Spark GraphX

Key idea

- Graphs are essential to analytics (e.g. social networks)

- Tables & Graphs are composable views of the same physical data

- Each view has its own operators that exploit the semantics of the view to achieve efficient execution

- Graph algorithms are based on Pregel API

---

Joseph Gonzalez, Reynold Xin, Daniel Crankshaw, Ankur Dave, Michael Franklin, and Ion Stoica, GraphX: Unifying Data-Parallel and Graph-Parallel Analytics, https://amplab.cs.berkeley.edu/wp-content/uploads/2014/02/graphx@strata2014_final.pdf
Spark GraphX
Main components

- **VertexRDD** maps IDs to vertex content
- **EdgeRDD** are of the form (ID1, ID2, ET)
- **Triplets** are a combination of Vertex & Edge RDDs

```python
def Graph(vertices: Table[ (Id, V) ],
           edges: Table[ (Id, Id, E) ])
// Table Views ---------------
def vertices: Table[ (Id, V) ]
def edges: Table[ (Id, Id, E) ]
def triplets: Table [ ((Id, V), (Id, V), E) ]```
Performance of GraphX

Live-Journal: 69 Million Edges

<table>
<thead>
<tr>
<th>System</th>
<th>Runtime (in seconds, PageRank for 10 iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mahout/Hadoop</td>
<td>1340</td>
</tr>
<tr>
<td>Naïve Spark</td>
<td>354</td>
</tr>
<tr>
<td>Giraph</td>
<td>207</td>
</tr>
<tr>
<td>GraphX</td>
<td>68</td>
</tr>
<tr>
<td>GraphLab</td>
<td>22</td>
</tr>
</tbody>
</table>

Joseph Gonzalez, Reynold Xin, Daniel Crankshaw, Ankur Dave, Michael Franklin, and Ion Stoica,
GraphX: Unifying Data-Parallel and Graph-Parallel Analytics,
Example - PageRank

Popular algorithm originally introduced by Google

Example - PageRank

PageRank Algorithm

• Start each page with a rank of 1

• On each iteration:

  A. \( \text{contrib} = \frac{\text{curRank}}{|\text{neighbors}|} \)

  B. \( \text{curRank} = 0.15 + 0.85 \sum_{\text{neighbors}} \text{contrib}_i \)

Example: PageRank
Spark GraphX

// get people with top-k pageranks
def findTopPageRank(allPeople: RDD[String], links: RDD[(String, String, Double)], k: Int) = {
  val versRDD = allPeople.map(p => (uid(p), p))
  val edgesRDD = links.map{ case (l, r, score) => Edge(uid(l), uid(r), score) }

  val g = Graph(versRDD, edgesRDD).cache
  val ranks = g.pageRank(0.001)

  ranks.vertices.top(k)(Ordering.by(_.2)).map(p => (fromUid(p._1), p._2))
}
Example: PageRank
How is it implemented in Pregel?

```java
def PageRank(v: Id, msgs: List[Double]) {
    // Compute the message sum
    var msgSum = 0
    for (m <- msgs) { msgSum += m }
    // Update the PageRank
    PR(v) = 0.15 + 0.85 * msgSum
    // Broadcast messages with new PR
    for (j <- OutNbrs(v)) {
        msg = PR(v) / NumLinks(v)
        send_msg(to=j, msg)
    }
    // Check for termination
    if (converged(PR(v))) voteToHalt(v)
}
```