Apache Pig for Data Science

Casey Stella

Hortonworks

April 9, 2014
Table of Contents

Preliminaries
   Apache Hadoop
   Apache Pig

Pig in the Data Science Toolbag
   Understanding Your Data
   Machine Learning with Pig
   Applying Models with Pig

Unstructured Data Analysis with Pig

Questions & Bibliography
Introduction

• I’m a Principal Architect at Hortonworks
• I work primarily doing Data Science in the Hadoop Ecosystem
• Prior to this, I’ve spent my time and had a lot of fun
  ◦ Doing data mining on medical data at Explorys using the Hadoop ecosystem
  ◦ Doing signal processing on seismic data at Ion Geophysical using MapReduce
  ◦ Being a graduate student in the Math department at Texas A&M in algorithmic complexity theory
• I’m going to talk about Apache Pig’s role for doing scalable data science.
Apache Hadoop: What is it?

Hadoop is a distributed storage and processing system

- Scalable – Efficiently store and process data
Apache Hadoop: What is it?

Hadoop is a distributed storage and processing system

- Scalable – Efficiently store and process data
- Reliable – Failover and redundant storage
Apache Hadoop: What is it?

Hadoop is a distributed storage and processing system

- Scalable – Efficiently store and process data
- Reliable – Failover and redundant storage
- Vast – Many ecosystem projects surrounding data ingestion, processing and export
Apache Hadoop: What is it?

Hadoop is a distributed storage and processing system

- Scalable – Efficiently store and process data
- Reliable – Failover and redundant storage
- Vast – Many ecosystem projects surrounding data ingestion, processing and export
- Economical – Use commodity hardware and open source software
Apache Hadoop: What is it?

Hadoop is a distributed storage and processing system

- Scalable – Efficiently store and process data
- Reliable – Failover and redundant storage
- Vast – Many ecosystem projects surrounding data ingestion, processing and export
- Economical – Use commodity hardware and open source software
- Not a one-trick-pony – Not just MapReduce anymore
Apache Hadoop: Who is using it?
Apache Pig: What is it?

Pig is a high level scripting language for operating on large datasets inside Hadoop
Apache Pig: What is it?

Pig is a high level scripting language for operating on large datasets inside Hadoop

- Compiles scripting language into MapReduce operations
Apache Pig: What is it?

Pig is a high level scripting language for operating on large datasets inside Hadoop

- Compiles scripting language into MapReduce operations
- Optimizes such that the minimal number of MapReduce jobs need be run
Apache Pig: What is it?

Pig is a high level scripting language for operating on large datasets inside Hadoop

- Compiles scripting language into MapReduce operations
- Optimizes such that the minimal number of MapReduce jobs need be run
- Familiar relational primitives available
Apache Pig: What is it?

Pig is a high level scripting language for operating on large datasets inside Hadoop

- Compiles scripting language into MapReduce operations
- Optimizes such that the minimal number of MapReduce jobs need be run
- Familiar relational primitives available
- Extensible via User Defined Functions and Loaders for customized data processing and formats
Apache Pig: An Familiar Example

SENTENCES = load '...' as (sentence:chararray);
WORDS = foreach SENTENCES
    generate flatten(TOKENIZE(sentence)) as word;
WORD_GROUPS = group WORDS by word;
WORD_COUNTS = foreach WORD_GROUPS
    generate group as word, COUNT(WORDS);
store WORD_COUNTS into '...';
Understanding Data

“80% of the work in any data project is in cleaning the data.”

— D.J. Patel in Data Jujitsu
Understanding Data

A core pre-requisite to analyzing data is understanding data’s shape and distribution. This requires (among other things):

• Computing distribution statistics on data
• Sampling data
Understanding Data: Datafu

An Apache Incubating project called datafu\(^1\) provides some of these tooling in the form of Pig UDFs:

- Computing quantiles of data
- Sampling
  - Bernoulli sampling by probability (built into pig)
  - Simple Random Sample
  - Reservoir sampling
  - Weighted sampling without replacement
  - Random Sample with replacement

\(^1\)http://datafu.incubator.apache.org/
Understanding Data: Datafu

An Apache Incubating project called datafu\(^1\) provides some of these tooling in the form of Pig UDFs:

- Computing quantiles of data
- Sampling
  - Bernoulli sampling by probability (built into pig)

\(^1\)http://datafu.incubator.apache.org/
An Apache Incubating project called datafu\textsuperscript{1} provides some of these tooling in the form of Pig UDFs:

- Computing quantiles of data
- Sampling
  - Bernoulli sampling by probability (built into pig)
  - Simple Random Sample

\textsuperscript{1}http://datafu.incubator.apache.org/
An Apache Incubating project called **datafu**\(^1\) provides some of these tooling in the form of Pig UDFs:

- Computing quantiles of data
- **Sampling**
  - Bernoulli sampling by probability (built into pig)
  - Simple Random Sample
  - Reservoir sampling

\(^{1}\)http://datafu.incubator.apache.org/
An Apache Incubating project called datafu\(^1\) provides some of these tooling in the form of Pig UDFs:

- Computing quantiles of data
- Sampling
  - Bernoulli sampling by probability (built into pig)
  - Simple Random Sample
  - Reservoir sampling
  - Weighted sampling without replacement

\(^1\)http://datafu.incubator.apache.org/
Understanding Data: Datafu

An Apache Incubating project called datafu\(^1\) provides some of these tooling in the form of Pig UDFs:

- Computing quantiles of data
- Sampling
  - Bernoulli sampling by probability (built into pig)
  - Simple Random Sample
  - Reservoir sampling
  - Weighted sampling without replacement
  - Random Sample with replacement

\(^1\)http://datafu.incubator.apache.org/
Case Study: Bootstrapping

Bootstrapping is a resampling technique which is intended to measure accuracy of sample estimates. It does this by measuring an estimator (such as mean) across a set of random samples with replacement from an original (possibly large) dataset.
Datafu provides two tools which can be used together to provide that random sample with replacement:

- **SimpleRandomSampleWithReplacementVote** – Ranks multiple candidates for each position in a sample
- **SimpleRandomSampleWithReplacementElect** – Chooses, for each position in the sample, the candidate with the lowest score

The datafu docs provide an example\(^2\) of generating a bootstrap of the mean estimator.

\(^2\)http://datafu.incubator.apache.org/docs/datafu/guide/sampling.html
What is Machine Learning?

Machine learning is the study of systems that can learn from data. The general tasks fall into one of two categories:
What is Machine Learning?

Machine learning is the study of systems that can learn from data. The general tasks fall into one of two categories:

- **Unsupervised Learning**
  - Clustering
  - Outlier detection
  - Market Basket Analysis

- **Supervised Learning**
  - Classification
  - Regression
  - Recommendation
What is Machine Learning?

Machine learning is the study of systems that can learn from data. The general tasks fall into one of two categories:

- **Unsupervised Learning**
  - Clustering
  - Outlier detection
  - Market Basket Analysis

- **Supervised Learning**
  - Classification
  - Regression
  - Recommendation
Building Machine Learning Models with Pig

Machine Learning at scale in Hadoop generally falls into two categories:

• Build one large model on all (or almost all) of the data
• Sample the large dataset and build the model based on that sample
Building Machine Learning Models with Pig

Machine Learning at scale in Hadoop generally falls into two categories:

- Build one large model on all (or almost all) of the data
- Sample the large dataset and build the model based on that sample

Pig can assist in intelligently sampling down the large data into a training set. You can then use your favorite ML algorithm (which can be run on the JVM) to generate a machine learning model.
Applying Models with Pig

Pig shines at batch application of an existing ML model. This generally is of the form:

- Train a model out-of-band
Applying Models with Pig

Pig shines at batch application of an existing ML model. This generally is of the form:

- Train a model out-of-band

- Write a UDF in Java or another JVM language which can apply the model to a data point
Applying Models with Pig

Pig shines at batch application of an existing ML model. This generally is of the form:

- Train a model out-of-band
- Write a UDF in Java or another JVM language which can apply the model to a data point
- Call the UDF from a pig script to distribute the application of the model across your data in parallel
What is Natural Language Processing?

- Natural language processing is the field of Computer Science, Linguistics & Math that covers computer understanding and manipulation of human language.
  - Historically, linguists hand-coded rules to accomplish much analysis
  - Most modern approaches involves using Machine Learning
What is Natural Language Processing?

- Natural language processing is the field of Computer Science, Linguistics & Math that covers computer understanding and manipulation of human language.
  - Historically, linguists hand-coded rules to accomplish much analysis
  - Most modern approaches involves using Machine Learning

- Mature field with many useful libraries on the JVM
  - Apache OpenNLP
  - Stanford CoreNLP
  - MALLET
Natural Language Processing with Large Data

• Generally low-volume, complex analysis
  ○ Big companies often don’t have a ton of natural language data
  ○ Dropped previously because they were unable to analyze
Natural Language Processing with Large Data

• Generally low-volume, complex analysis
  ◦ Big companies often don’t have a ton of natural language data
  ◦ Dropped previously because they were unable to analyze

• Sometimes high-volume, complex analysis
  ◦ Search Engines
  ◦ Social media content analysis
Natural Language Processing with Large Data

• Generally low-volume, complex analysis
  ○ Big companies often don’t have a ton of natural language data
  ○ Dropped previously because they were unable to analyze

• Sometimes high-volume, complex analysis
  ○ Search Engines
  ○ Social media content analysis

• Typically many small-data problems in parallel
  ○ Often requires only the context of a single document
  ○ Ideal for encapsulating as Pig UDFs
Natural Language Processing: Demo


- There is a large set of IMDB movie reviews used to analyze sentiment analysis [1].

- Let’s look at how to encapsulate this into a Pig UDF and run on some movie review data.
Results

- Executing on a sample of size 1022 Positive and Negative documents.
- Overall Accuracy of 77.2%

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>367</td>
<td>114</td>
</tr>
<tr>
<td>Negative</td>
<td>119</td>
<td>422</td>
</tr>
<tr>
<td>Total</td>
<td>486</td>
<td>536</td>
</tr>
</tbody>
</table>
Questions

Thanks for your attention! Questions?

- Code & scripts for this talk available on my github presentation page.\(^3\)
- Find me at http://caseystella.com
- Twitter handle: @casey_stella
- Email address: cstella@hortonworks.com

\(^3\)http://github.com/cestella/presentations/
Bibliography
