Apache Pig for Data Science

Casey Stella

Hortonworks

April 9, 2014

Table of Contents

Introduction

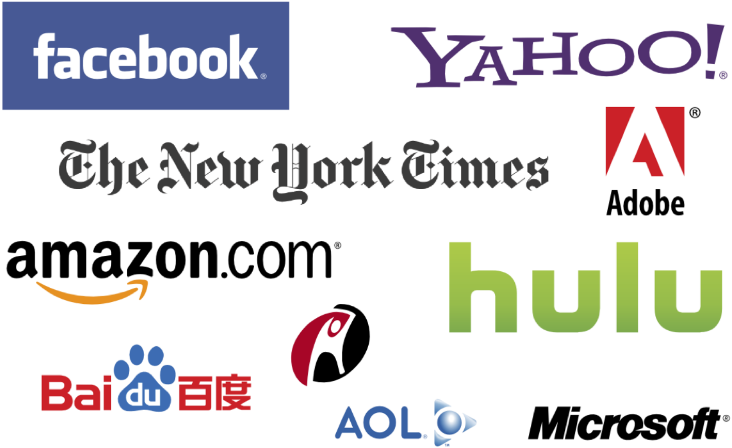
- I'm a Principal Architect at Hortonworks
- I work primarily on Data Science in the Hadoop Ecosystem
- Prior to this, I've spent my time and had a lot of fun
  - Doing data mining on medical data at Explorys using the Hadoop ecosystem
  - Doing signal processing on seismic data at Ion Geophysical using 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
- Reliable – Failover and redundant storage
- Vast – Many ecosystem projects surrounding data ingestion, processing and export
- Economical – Use commodity hardware

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

- Compiles scripting language into MapReduce operations
- Optimizes such that the minimal number of MapReduce jobs need be run
- Familiar relational primitives built in (i.e. [left, right, outer, cross] join)
- Extensible via User Defined Functions for special processing

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

D.J. Patil in *Data Jujitsu*

> *"80% of the work in any data project is in cleaning the data."*

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

A LinkedIn project called **datafu**[1] provides some of these tooling in the form of Pig UDFs:

- Computing quantiles of data
- Sampling
  - Uniform sampling by percentage (built into pig)
  - Uniform sampling by reservoir sampling
  - Weighted sampling without replacement
  - Random Sample with replacement

[1] http://github.com/linkedin/datafu

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.

Case Study: Bootstrapping

Datafu provides two tools which can be used together to provide that random sample with replacement:

- SimpleRandomSampleWithReplacementVote – Ranks for every position in a k-sample multiple candidates and a score
- SimpleRandomSampleWithReplacementElect – Chooses, for every position in a k-sample, the candidate with the lowest score

See the javadocs for SimpleRandomSampleWithReplacement for an example of generating a boostrap of the mean estimator. TODO: cite X Meng for this.

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

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
- 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
- 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
- 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

- Stanford CoreNLP integrated the work of Richard Socher, Alex Perelygin, et al using recursive deep neural networks to predict sentiment of movie reviews.
- Let's look at how to encapsulate this into a Pig UDF and run on some movie review data.

Questions & Bibliography

Thanks for your attention! Questions?

- Code and scripts for this talk at http://github.com/cestella/NLPWithMahout
- Find me at http://caseystella.com
- Twitter handle: @casey_stella
- Email address: cstella@hortonworks.com