Spark SQL + Pig-Latin
Combine Query Language and Data Flow Language for Data Science

Jeff Zhang (zjffdu@apache.org)
May 16, 2017
Who am I

- ASF Member, work in ASF for almost 8 years
- Committer of Apache Tez, Pig & Zeppelin
- Works in Hortonworks
Data Science

**Data Science**, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes and systems to **extract knowledge or insights** from data in various forms, either **structured or unstructured**.

- Describe what happens
- Explain what happens
- Predict what will happen
Data Munging

- Collect and Transform Server Log
  - User Agent Normalization
  - Robot Detection
  - Sessionize

- Move data from Database to HDFS

- Collect and Transform Social Media Data
Data Munging

Before Data Munging

After Data Munging
Data Analysis

- Combine different sources of data and apply statistics method, BI tools to get insight
  - Web Traffic Metrics
  - User Segmentation Analysis
  - A/B Test
# Data Munging vs Data Analysis

<table>
<thead>
<tr>
<th></th>
<th>Data Munging</th>
<th>Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Source</strong></td>
<td>Messy Structured / Unstructured Unorganized</td>
<td>Clean Structured Organized</td>
</tr>
<tr>
<td><strong>Stability</strong></td>
<td>Regular, Stable</td>
<td>Ad-hoc</td>
</tr>
<tr>
<td><strong>Tools</strong></td>
<td>Python, Spark, Hadoop and etc.</td>
<td>R, Python, SQL and etc.</td>
</tr>
</tbody>
</table>

---

Do you have to be full stack big data engineer to do data science?

What if you are a data analyst without much programming skills?
Data Science Infrastructure
What is Spark

**Apache Spark** is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads.
What is Apache Pig

- **Apache Pig** is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called **Pig Latin**. Pig can execute its jobs in MapReduce, Apache Tez, or Apache Spark

  - Ease of programming
  - Optimization opportunities
  - Extensibility
input = load '/path/to/file/' as (line: chararray);
words = foreach input generate flatten(TOKENIZE(line,' ')) as word;
grouped_words = group words by word;
wordcount = foreach grouped_words generate group as word, COUNT(words) as count;
ordered_wordcount = order wordcount by count desc;
store ordered_wordcount into '/path/to/store';
# Pig-Latin vs SQL

<table>
<thead>
<tr>
<th></th>
<th>SQL</th>
<th>Pig-Latin</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language Type</strong></td>
<td>Query Language</td>
<td>Data Flow Language</td>
</tr>
<tr>
<td></td>
<td>• de factor standard</td>
<td>more readable for long scripts</td>
</tr>
<tr>
<td></td>
<td>• unreadable for long script</td>
<td></td>
</tr>
<tr>
<td><strong>Data Source</strong></td>
<td>Structured Data</td>
<td>Structured / Unstructured</td>
</tr>
<tr>
<td><strong>Integration</strong></td>
<td>Integrated with most of BI Tools</td>
<td>Very few BI tools integrated with Pig-Latin</td>
</tr>
</tbody>
</table>

## Conclusion
- **Pig-Latin** for Data Munging
- **SQL** for Data Analysis
Pig-Latin + Spark SQL

Data Munging

Load → Store

Spark DataFrame Table

Spark SQL

Data Analysis
bankText = load 'bank.csv' using PigStorage('');
bank = foreach bankText generate $0 as age, $1 as job, $2 as marital, $3 as education, $5 as balance
bank = filter bank by age != ""age"
bank = foreach bank generate (int)age, REPLACE(job, '"', '') as job, REPLACE(marital, '"', '"')
       as marital, (int)(REPLACE(balance, '"', '"')) as balance;
store bank into 'bank' using SparkTableStorage();

Pig Latin

Spark Table (bank)  SQL
Integrate Spark into Pig

- Pig-Latin
- Logic Plan
- Physical Plan
- Execution Plan

Execution Engine
Where to run Pig-Latin & Spark SQL (Zeppelin)

Apache Zeppelin is a web-based notebook that enables interactive data analytics. You can make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more.
Zeppelin Architecture

Zeppelin Server

JVM

Spark Interpreter Group
- Scala
- Python
- R

JVM

Pig Interpreter Group
- Pig-Latin
- Spark SQL
Demo
Data Science Infrastructure (Recap)
Current Status & What’s Next

Status

– PIG-5080 (Support store alias as spark table)
– ZEPPELIN-2232 (Support Spark SQL for Pig Interpreter)

Next

– Integrate Spark MLlib in Pig
– Use DataFrame API instead of RDD API to integrate Spark with Pig
– Integrate Pig with other Spark APIs, like R, Python
Thank You