cloudera®

Introduction to Apache Pig

Chapter 3

Course Chapters

- Introduction
- Apache Hadoop Fundamentals
- **Introduction to Apache Pig**
- Basic Data Analysis with Apache Pig
- Processing Complex Data with Apache Pig
- Multi-Dataset Operations with Apache Pig
- Apache Pig Troubleshooting and Optimization
- Introduction to Apache Hive and Impala
- Querying with Apache Hive and Impala
- Apache Hive and Impala Data Management
- Data Storage and Performance
- Relational Data Analysis with Apache Hive and Impala
- Complex Data with Apache Hive and Impala
- Analyzing Text with Apache Hive and Impala
- Apache Hive Optimization
- Apache Impala Optimization
- Extending Apache Hive and Impala
- Choosing the Best Tool for the Job
- Conclusion

Introduction to Apache Pig

In this chapter, you will learn

- **The key features Pig offers**
- **How organizations use Pig for data processing and analysis**
- **How to use Pig interactively and in batch mode**

Chapter Topics

Introduction to Apache Pig

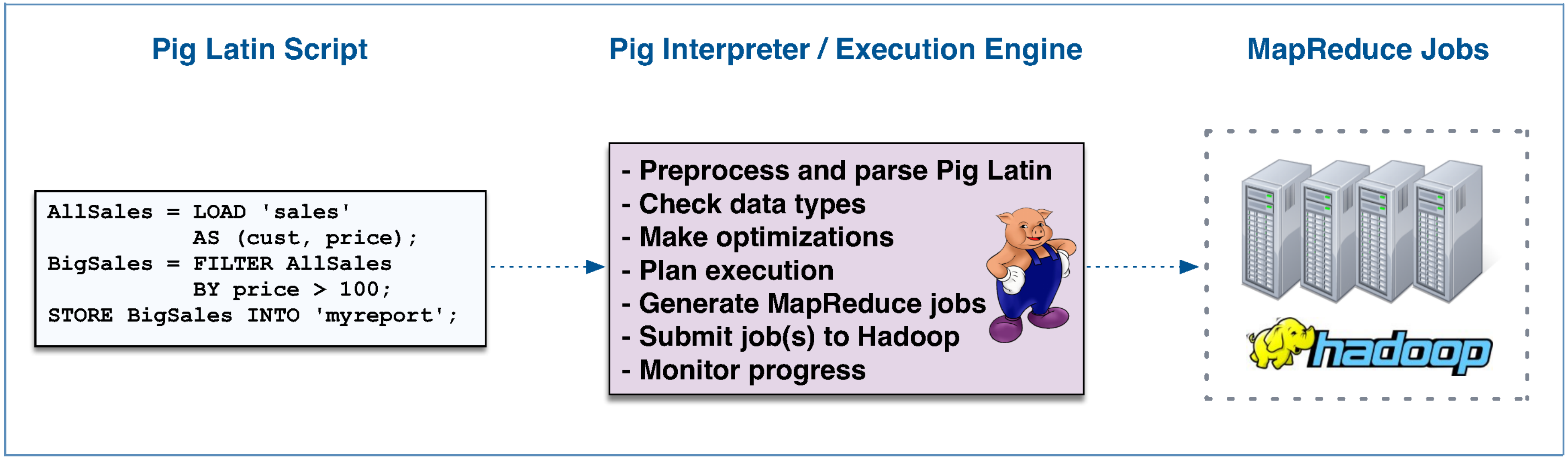
- **What Is Pig?**
- Pig Features
- Pig Use Cases
- Interacting with Pig
- Essential Points

Apache Pig Overview

- **Apache Pig is a platform for data analysis and processing on Hadoop**
  - It offers an alternative to writing MapReduce code directly
- **Originally developed as a research project at Yahoo**
  - Goals: flexibility, productivity, and maintainability
  - Now an open source Apache project

The Anatomy of Pig

- **Main components of Pig**
  - The data flow language (Pig Latin)
  - The interactive shell (Grunt) where you can type Pig Latin statements
  - The Pig interpreter and execution engine



Pig Latin Script

```
AllSales = LOAD 'sales'
           AS (cust, price);
BigSales = FILTER AllSales
           BY price > 100;
STORE BigSales INTO 'myreport';
```

Pig Interpreter / Execution Engine

- Preprocess and parse Pig Latin
- Check data types
- Make optimizations
- Plan execution
- Generate MapReduce jobs
- Submit job(s) to Hadoop
- Monitor progress

MapReduce Jobs

Chapter Topics

Introduction to Apache Pig

- What Is Pig?
- **Pig Features**
- Pig Use Cases
- Interacting with Pig
- Essential Points

Pig Features

- **Pig is an alternative to writing low-level MapReduce code in Java**
- **Many features enable sophisticated analysis and processing**
  - HDFS manipulation
  - UNIX shell commands
  - Relational operations
  - Positional references for fields
  - Common mathematical functions
  - Support for custom functions and data formats
  - Complex data structures

Chapter Topics

Introduction to Apache Pig

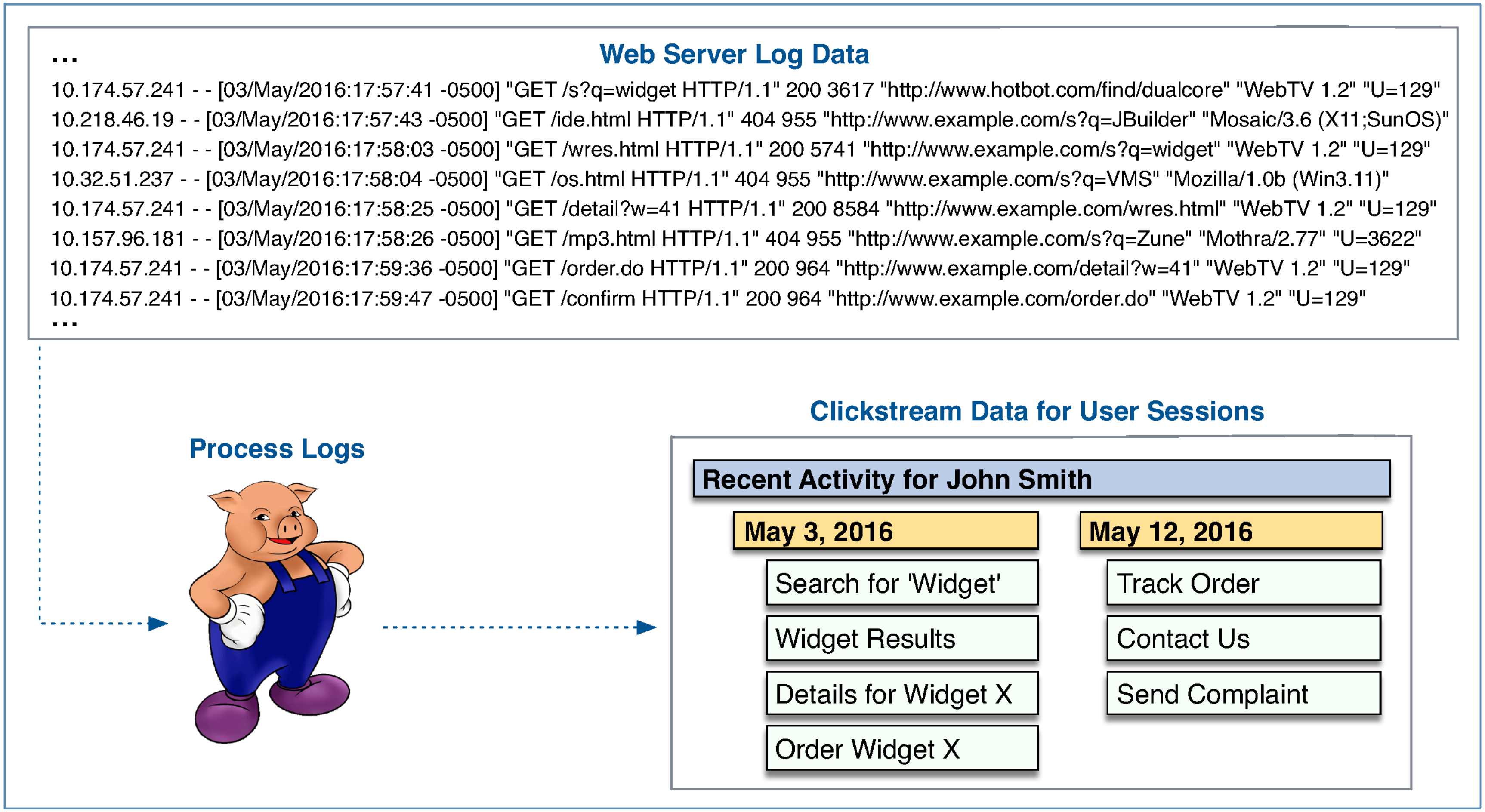
- What Is Pig?
- Pig Features
- **Pig Use Cases**
- Interacting with Pig
- Essential Points

How Are Organizations Using Pig?

- **Many organizations use Pig for *data analysis***
  - Finding relevant records in a massive dataset
  - Querying multiple datasets
  - Calculating values from input data
- **Pig is also frequently used for *data processing***
  - Reorganizing an existing dataset
  - Joining data from multiple sources to produce a new dataset

Use Case: Web Log Sessionization

- **Pig can help you extract valuable information from web server log files**



Web Server Log Data

```
...
10.174.57.241 - - [03/May/2016:17:57:41 -0500] "GET /s?q=widget HTTP/1.1" 200 3617 "http://www.hotbot.com/find/dualcore" "WebTV 1.2" "U=129"
10.218.46.19 - - [03/May/2016:17:57:43 -0500] "GET /ide.html HTTP/1.1" 404 955 "http://www.example.com/s?q=JBuilder" "Mosaic/3.6 (X11;SunOS)"
10.174.57.241 - - [03/May/2016:17:58:03 -0500] "GET /wres.html HTTP/1.1" 200 5741 "http://www.example.com/s?q=widget" "WebTV 1.2" "U=129"
10.32.51.237 - - [03/May/2016:17:58:04 -0500] "GET /os.html HTTP/1.1" 404 955 "http://www.example.com/s?q=VMS" "Mozilla/1.0b (Win3.11)"
10.174.57.241 - - [03/May/2016:17:58:25 -0500] "GET /detail?w=41 HTTP/1.1" 200 8584 "http://www.example.com/wres.html" "WebTV 1.2" "U=129"
10.157.96.181 - - [03/May/2016:17:58:26 -0500] "GET /mp3.html HTTP/1.1" 404 955 "http://www.example.com/s?q=Zune" "Mothra/2.77" "U=3622"
10.174.57.241 - - [03/May/2016:17:59:36 -0500] "GET /order.do HTTP/1.1" 200 964 "http://www.example.com/detail?w=41" "WebTV 1.2" "U=129"
10.174.57.241 - - [03/May/2016:17:59:47 -0500] "GET /confirm HTTP/1.1" 200 964 "http://www.example.com/order.do" "WebTV 1.2" "U=129"
...
```

Process Logs

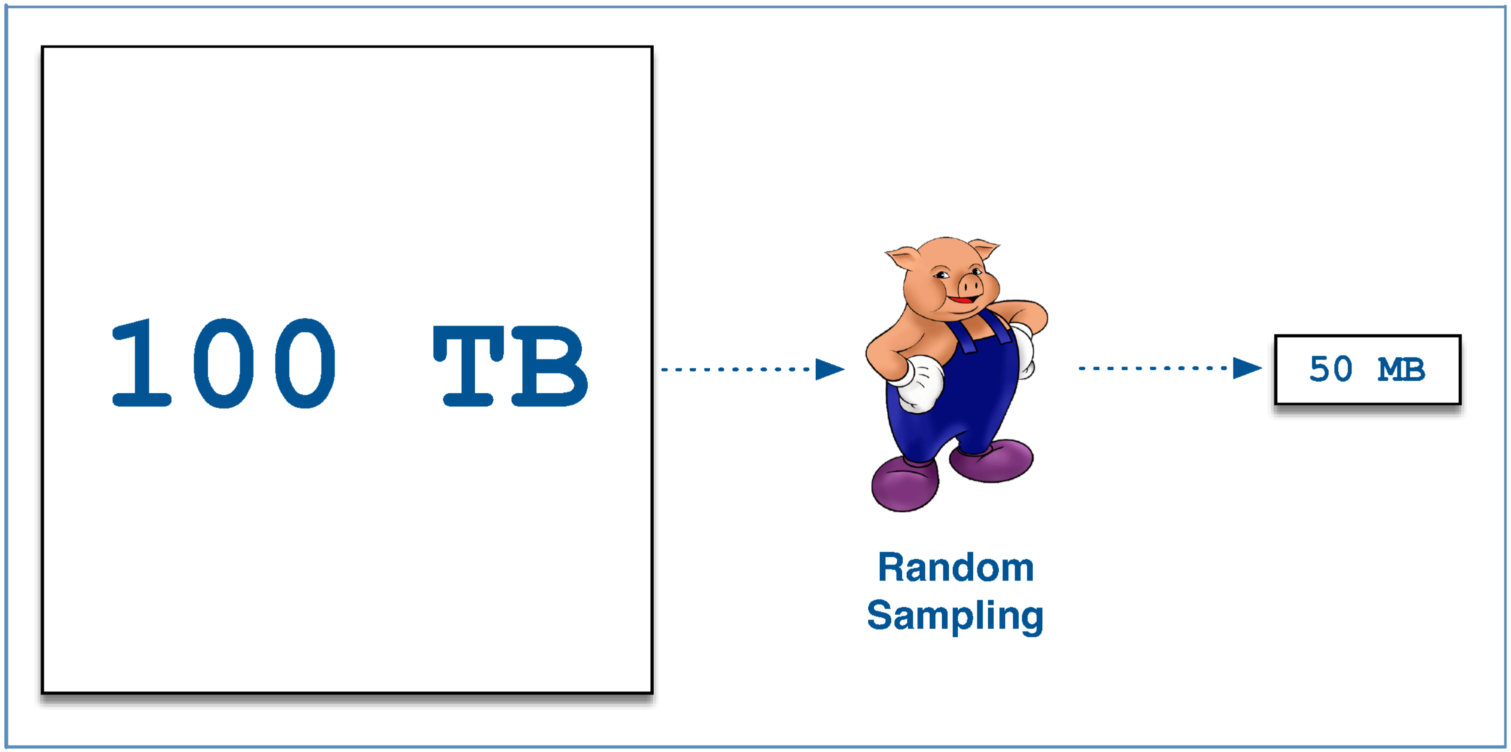
Clickstream Data for User Sessions

Recent Activity for John Smith

May 3, 2016	**May 12, 2016**
Search for 'Widget'	Track Order
Widget Results	Contact Us
Details for Widget X	Send Complaint
Order Widget X	

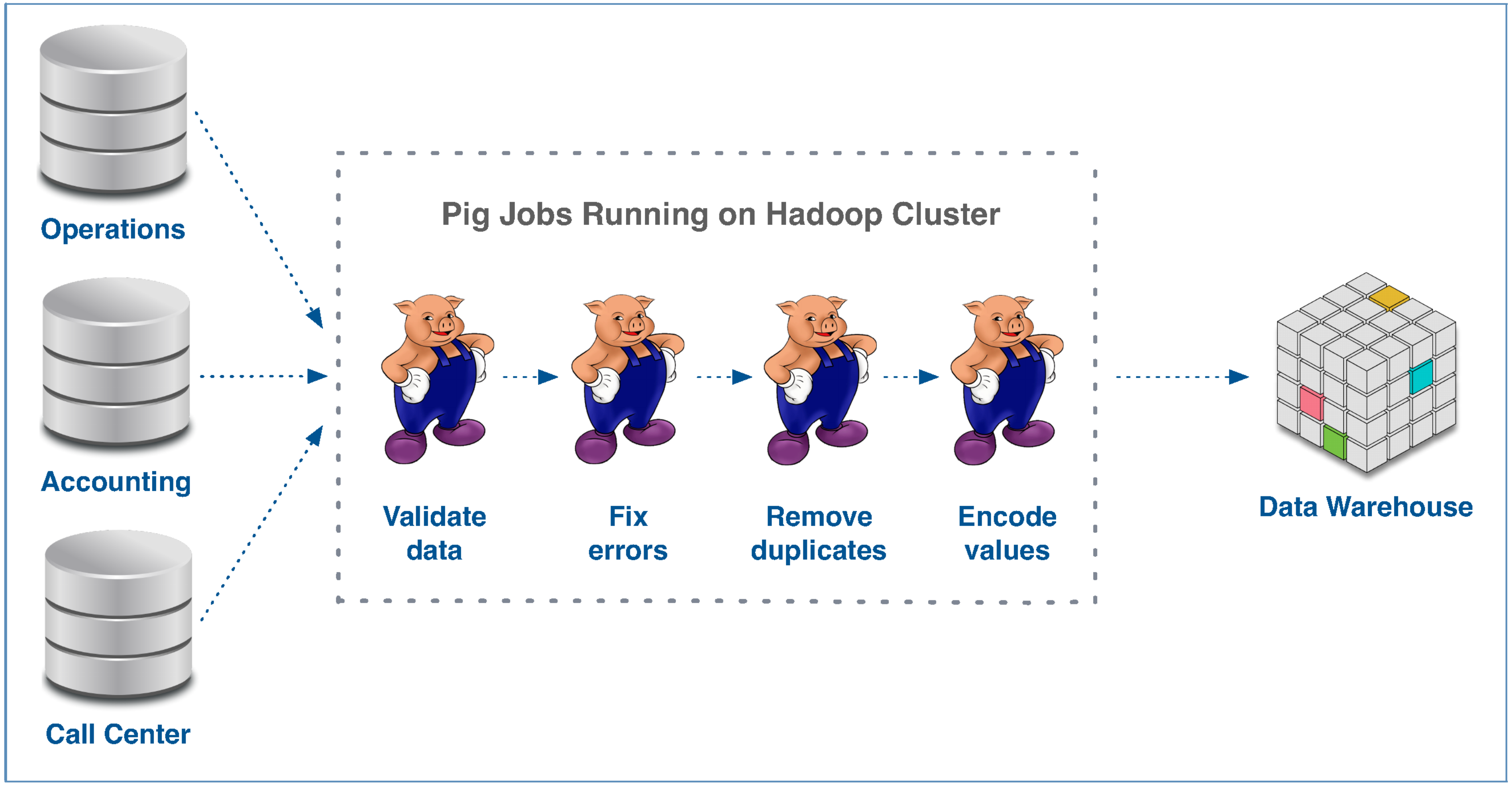
Use Case: Data Sampling

- **Sampling can help you explore a representative portion of a large dataset**
  - Allows you to examine this portion with tools that do not scale as well
  - Supports faster iterations during development of analysis jobs



Use Case: ETL Processing

- **Pig is also widely used for Extract, Transform, and Load (ETL) processing**



Chapter Topics

Introduction to Apache Pig

- What Is Pig?
- Pig Features
- Pig Use Cases
- **Interacting with Pig**
- Essential Points

Using Pig Interactively

- **You can use Pig interactively in the Grunt shell**
  - Pig interprets each Pig Latin statement as you type it
  - Execution is delayed until output is required
  - Very useful for ad hoc data inspection

- **Example of how to start, use, and exit Grunt**

```
$ pig
grunt> allsales = LOAD 'sales' AS (name, price);
grunt> bigsales = FILTER allsales BY price > 999;
grunt> STORE bigsales INTO 'myreport';
grunt> quit;
```

- **Use `pig -e` to execute a Pig Latin statement from the UNIX shell**

Interacting with HDFS

- **You can manipulate HDFS with Pig, using the `fs` command**

```
grunt> fs -mkdir sales/;
grunt> fs -put europe.txt sales/;
grunt> allsales = LOAD 'sales' AS (name, price);
grunt> bigsales = FILTER allsales BY price > 999;
grunt> STORE bigsales INTO 'myreport';
grunt> fs -getmerge myreport/ bigsales.txt;
```

Interacting with UNIX

- **The sh command lets you run UNIX programs from Pig**

```
grunt> sh date;
Wed Nov 12 06:39:13 PST 2016
grunt> sh ls;                        -- lists local files
grunt> fs -ls;                       -- lists HDFS files
```

Running Pig Scripts

- **A Pig script is simply Pig Latin code stored in a text file**
  - By convention, these files have the `.pig` extension
- **You can run a Pig script from within the Grunt shell using `run`**
  - This is useful for automation and batch execution

```
grunt> run salesreport.pig;
```

- **It is common to run a Pig script directly from the UNIX shell**

```
$ pig salesreport.pig
```

MapReduce and Local Modes

- **As described earlier, Pig turns Pig Latin into MapReduce jobs**
  - Pig submits those jobs for execution on the Hadoop cluster
- **It is also possible to run Pig in “local mode” using the −x flag**
  - This runs jobs on the *local machine* instead of the cluster
  - Local mode uses the local filesystem instead of HDFS
  - Can be helpful for testing before deploying a job to production

```
$ pig -x local                       -- interactive

$ pig -x local salesreport.pig   -- batch
```

Client-Side Log Files

- **If a job fails, Pig may produce a log file to explain why**
  - These log files are typically produced in your current working directory on the local (client) machine

Chapter Topics

Introduction to Apache Pig

- What Is Pig?
- Pig Features
- Pig Use Cases
- Interacting with Pig
- **Essential Points**

Essential Points

- **Pig offers an alternative to writing MapReduce code directly**
  - Pig interprets Pig Latin code in order to create MapReduce jobs
  - It then submits these jobs to the Hadoop cluster
- **You can execute Pig Latin code interactively through the Grunt shell**
  - Pig delays job execution until output is required
- **It is also common to store Pig Latin code in a script for batch execution**
  - Allows for automation and code reuse

Bibliography

The following offer more information on topics discussed in this chapter

- **Apache Pig website**
  - `http://pig.apache.org/`
- **Process a Million Songs with Apache Pig**
  - `http://tiny.cloudera.com/dac03a`
- **Powered by Pig**
  - `http://tiny.cloudera.com/poweredbypig`
- **LinkedIn: User Engagement Powered By Apache Pig and Hadoop**
  - `http://tiny.cloudera.com/dac03c`
- ***Programming Pig* (O'Reilly book)**
  - `http://tiny.cloudera.com/programmingpig`