Understanding Pig Latin

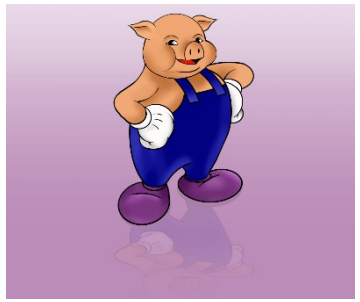
Thomas M. Henson

@henson_tm | www.thomashenson.com

Overview

- Install Sandbox
- Data Types & Expression
- Relational, Arithmetic, Boolean Operators
- Demo


Sandbox



HDFS

HIVE

APACHE HBASE

CentOS

Sandbox Options

cloudera®

- Founders Google, Facebook, and Yahoo developers
- Opensource contributors

- Founders were developers of Hadoop
- Pig contributors

- Hybrid Opensource
- Services & Products
- CDH 5.3

Sandbox Requirements

cloudera®

CDH 5.3	VirtualBox, VMware, or KV	RAM 4GB / File Size 3GB

Cloudera

Cloudera.com

VirtualBox

- Full Opensource
- Services
- HDP 2.2

Sandbox Requirements

HDP 2.2

VirtualBox, VMware, or HyperV

RAM/File Size 4GB

Hortonworks

Hortonworks.com

Expression

- End with semicolon
- Expressed as variable
- Step by step

```
a = LOAD ‘cereal.csv’ AS (name:chararray, calories:int);
b = FOREACH a GENERATE name;
DUMP b;
```

Example Expression

Data Types

- Numeric Types

```
Int 32-bit          → 5
Long 64-bit         → 5L
Float 32-bit float  → 5.5f
Double 64-bit       → 5.5
```

Numeric Types

4 different numeric types

Inherited from Java

Data Types

- Numeric Types
- Text

```
Chararray character string   → “some text”
```

Text Data Type

java.lang.string

Data Types

- Numeric Types
- Text
- Date

```
datetime  →  1981-07-26T00:00:00.000+00:00
```

DatetimeType

Data Types

- Numeric Types
- Text
- Date
- Binary

```
bytearray Byte array (blob)
```

Binary Data Type

Java class DataByteArray

Data Types

- Numeric Types
- Text
- Date
- Binary
- Complex

```
tuple  ordered list of fields → (7,26)
Bag    collection of tuples   → {(7,26), (9,5)}
Map    set of key value pairs → [somekey#somevalue]
```

Complex Data Type

Data Types

- Numeric Types
- Text
- Date
- Binary
- Complex

```
Addition        → +   → a + b
Subtraction     →  -  → a - b
Multiplication  → *   → a * b
Division        →  /  → a / b
```

Arithmetic Operators

```
Equal        → a == b
Not Equal    → a != b
Greater than → a > b , a >= b
Less than    → a < b , a <= b
```

Comparison Operators

```
AND   → a == 10 and b == 12
OR    → a == 10 or b == 12
```

Boolean Operators

Relational Operators

NASDAQ 100 Index

Date	Open	High	Low	Close	Volume	Adj Close
2015-03-06	44	45	42	45	190000	45
--	--	--	--	--	--	--
--	--	--	--	--	--	--
--	--	--	--	--	--	--
--	--	--	--	--	--	--

Relational Operators

- Limit

Limit

```
x = Limit stock 10;
```

Relational Operators

- Limit
- Group

Group

```
x = GROUP stock BY high;
```

Relational Operators

- Limit
- Group
- Filter

Filter

```
x = FILTER stock BY closing > 43;
```

Relational Operators

- Limit
- Group
- Filter
- Foreach

Foreach

```
x = FOREACH stock GENERATE (high, low, close);
```

Relational Operators

- Limit
- Group
- Filter
- Foreach

Example Data

- Stock Market Data
- NASDAQ 100 Index
- Yahoo
- .csv

Demo

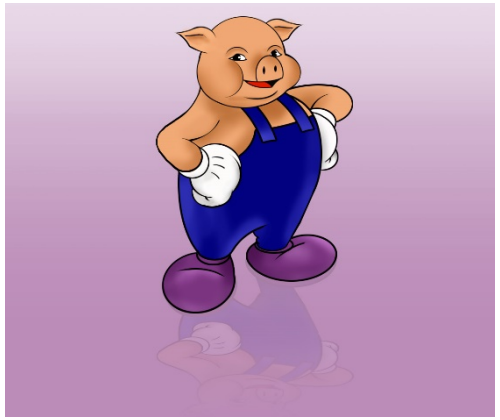
NASDAQ -100 Index

Daily 10/1/1985 – 03/08/2015

Pig Editor

Summary

That's all folks.



- Cloudera & Hortonworks
- Example Expressions & Data Types
- Pig Latin Syntax Basics
- Demo