Pig Latin: Getting Started

Introducing Pig Latin

Thomas M. Henson

@henson_tm | www.thomashenson.com

Overview

What is Pig Latin

MapReduce and Pig

Pig Resources

SQL, HiveQL, and Pig Latin

What Do I Need to Know?

- Basic SQL Skills
  - Introduction to SQL – Jon Flanders
- Basic Hadoop Knowledge
- No Prior MapReduce Experience

Pig → Igpa

Stop → Topsa

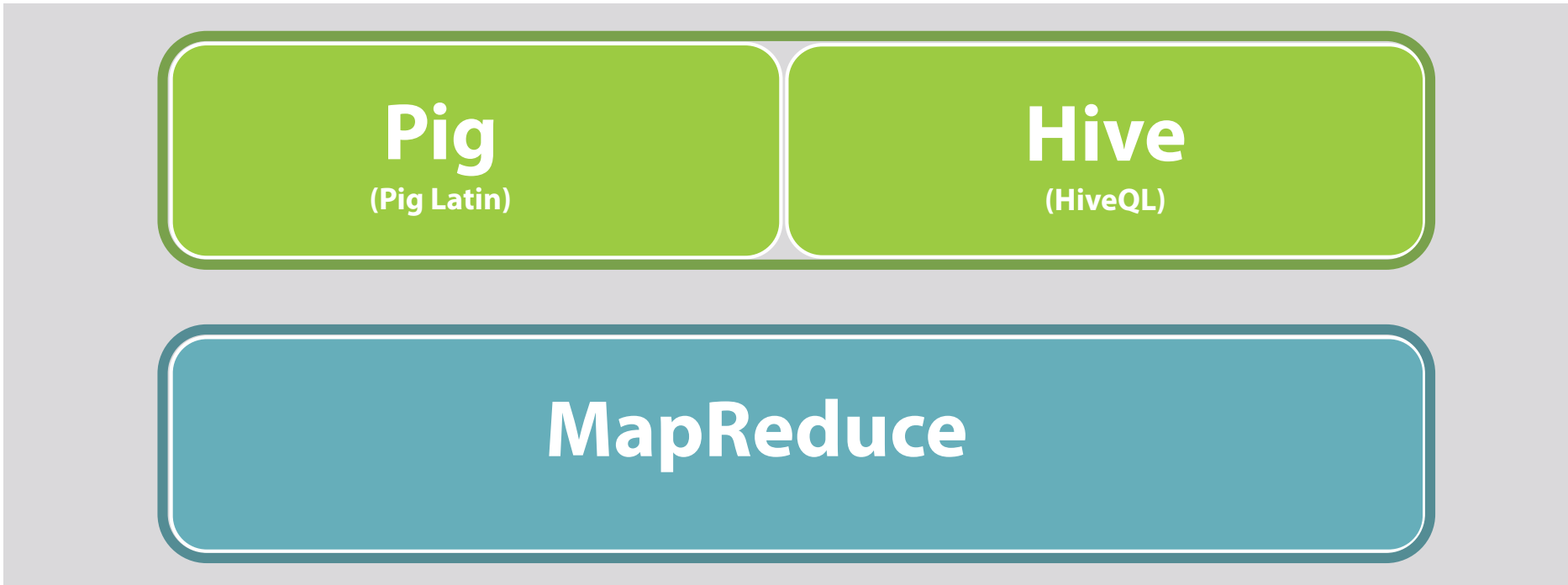
Pig Latin

Pig Latin is Pig's language that allows developers to express data flows

Pig

Pig is the application environment used to run Pig Latin and covert Pig Latin scripts into MapReduce jobs

What Is Pig?



Why Pig Over Mapreduce

Fewer lines of code

Quickly test queries

No Java experience

MapReduce Word Count

Example of word count script in Java

```
input_lines = LOAD '/tmp/word.txt' AS (line:chararray);
words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS word;
filtered_words = FILTER words BY word MATCHES '\\w+';
word_groups = GROUP filtered_words BY word;
word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS count, group AS word;
ordered_word_count = ORDER word_count BY count DESC;
STORE ordered_word_count INTO '/tmp/results.txt';
```

Pig Latin Word Count

7 Lines vs. 45 Lines

SQL like syntax

History of Pig

Yahoo

Large Datasets

SQL-Like

Apache

Pig Documentation

Where to find Pig Documentation

- **www.pig.apache.org**

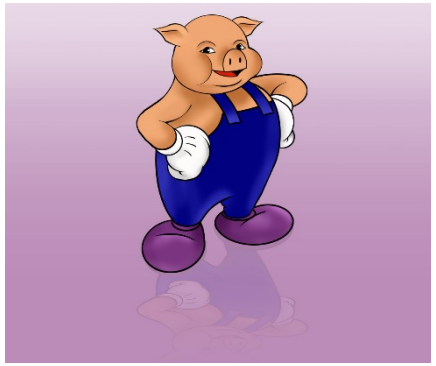
Comparing HiveQL and Pig Latin

HiveQL

- Declarative language based on SQL and schema bound

Pig Latin

- Procedural or data flow programming language with ability to declare schema at runtime



Example

Cereal Data

Return: name, calories, & protein

Protein > 4

id	name	calories	protein	fat	carbs
1	Bran	100	6	1	10
2	Flakes	120	2	3	15

SQL

```
SELECT name, calories, protein

FROM cereal WHERE protein > ‘4’;
```

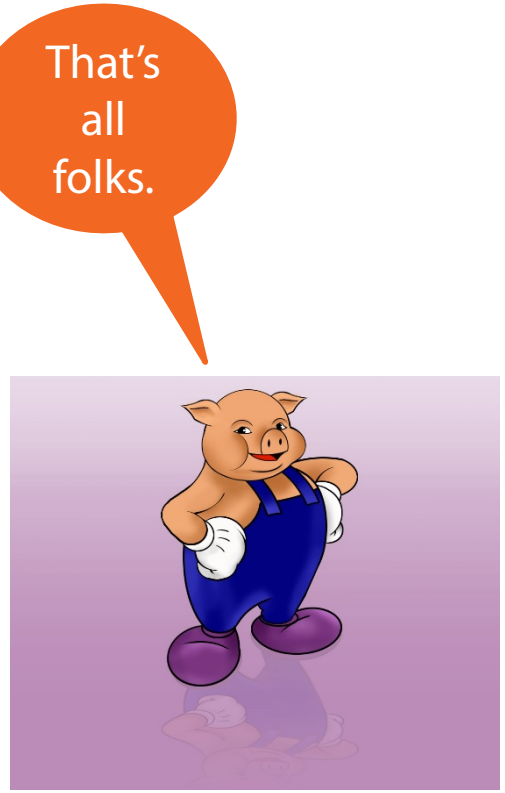
Pig Latin

```
cereal = FOREACH cereal GENERATE name, calories, protein;

cereal_filtered = FILTER cereal BY protein > 4;

DUMP cereal_filtered;
```

Summary



- Difference between Pig and Pig Latin
- Easier to write MapReduce
- Pig documentation
- Compared with SQL