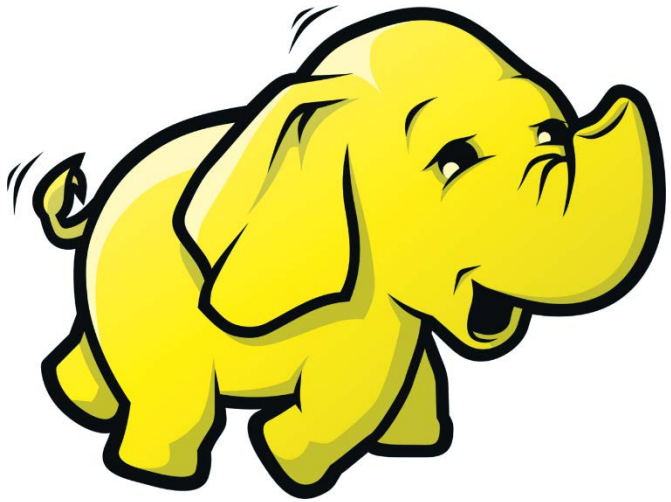
Querying Data with Pig and Hive

Thomas M. Henson

@henson_tm | www.thomashenson.com

Overview



Learning about Hive

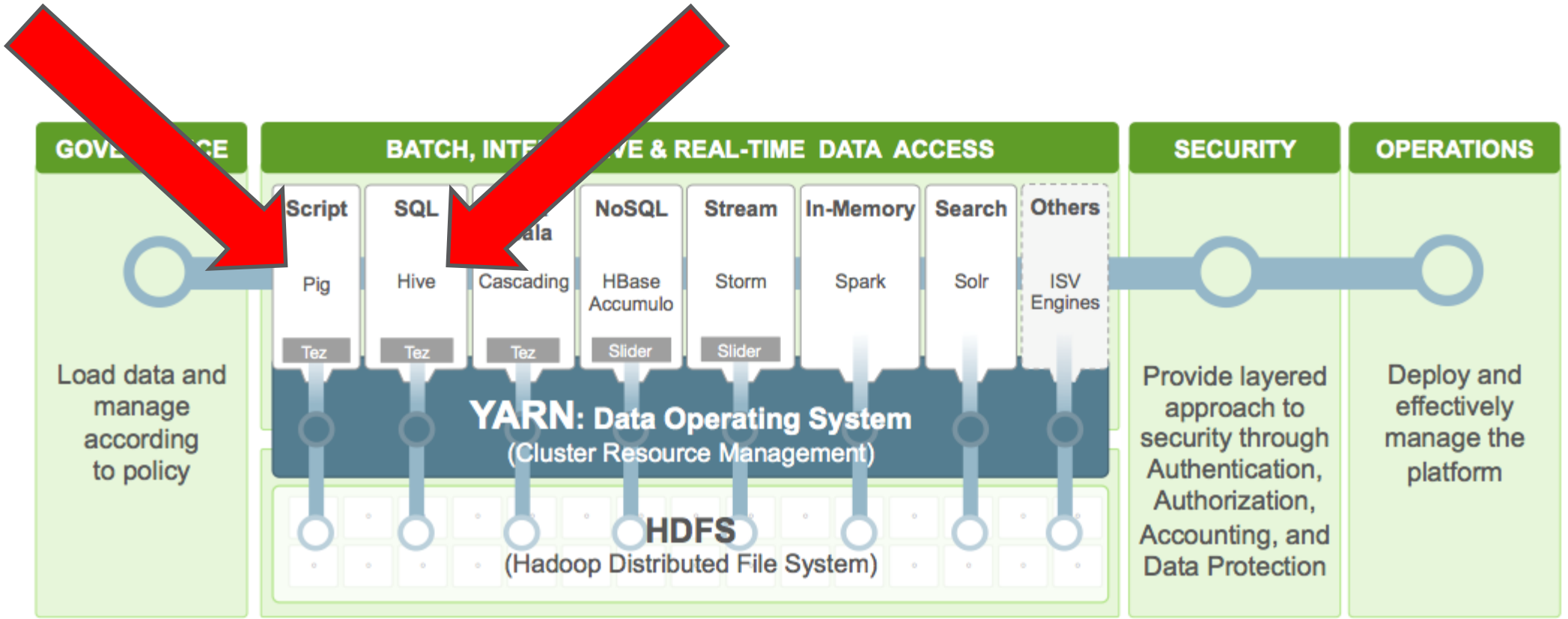
Querying data in HDFS with Hive

Defining Apache Pig

Using Apache Pig on data in HDFS

Discover resources to learn more

Hive and Pig in Hadoop Stack



From Hortonworks: http://hortonworks.com/hadoop/yarn/

Hive

Data warehouse software that works on top of Hadoop and allows for developers to structure data into schemas.

Benefits of Hive

Schema bound

HiveQL

Extendable

Example HiveQL Script

```
SELECT * FROM tables;
…shows all columns in table
SELECT column1, column2 FROM table;
…displays column1 and column2 in table
SELECT column1 FROM table WHERE colunm2 = “value”;
…displays all column1 in table where column2 is value
```

```
$ Hive
..loading
hive> write query here…
```

Hive Shell

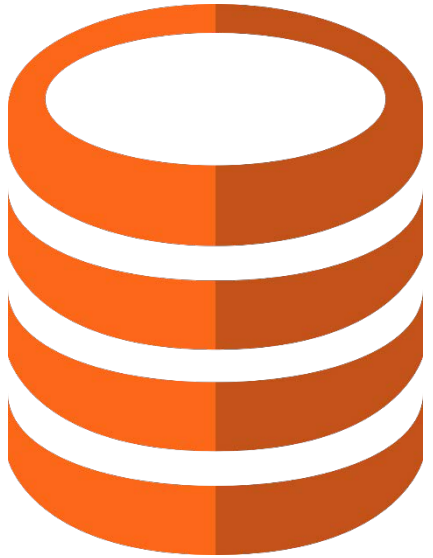
Allows developers to interact with HDFS and write HiveQL queries

Demo Hive

Placeholder

Data

- Apple daily stock quote
- Values
  - Date
  - Open
  - High
  - Close
  - Volume
  - Adj Close



Querying HDFS with HiveQL

Build table in Hive | Write HiveQL query | Extract results

Demo Hive

Placeholder

Apache Pig

Pig

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

Benefits of Pig

Not schema bound

Pig Latin

Unstructured and semi-structured data

Example Pig Latin Script

```
A = LOAD ‘somefile’ USING PigStorage(‘,’)

AS (field1,field2, field3);

results = FOREACH a GENERATE field1;

DUMP results;
```

```
$ Pig –x local
…loading
Pig> interact with shell…
```

Pig Grunt Shell

Allows developers to interact with HDFS, test and debug Pig Latin scripts

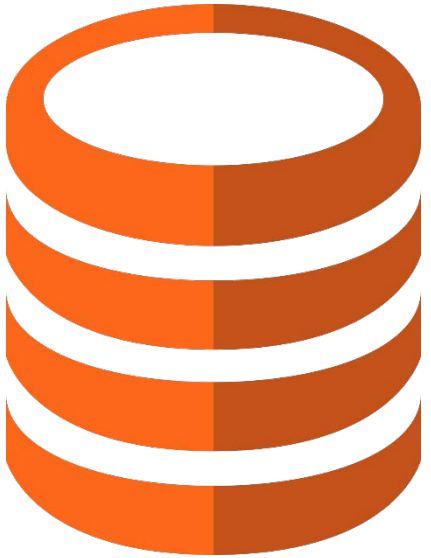
Demo Pig

Command line

Simple script

Data

- Apple daily stock quotes
- NASDAQ daily stock quotes
  - Date
  - Open
  - High
  - Close
  - Volume
  - Adj Close



Querying HDFS with Pig Latin

Load data from HDFS | Write JOIN in Pig Latin | Store results in HDFS

Demo Pig

Use JOIN

Store data in HDFS

Hive vs. Pig

Hive

- Structured data
- Declarative language
- SQL-like

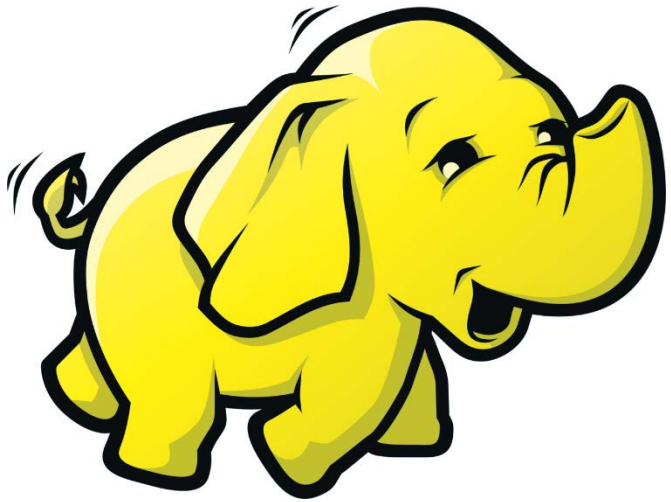
Pig

- Unstructured and semi structured data
- Procedural language
- SQL-like

Hive and Pig Resources

- Official Documentation
  - Hive → hive.apache.org
  - Pig → Pig.apache.org
- Pluralsight Courses
  - SQL on Hadoop – Analyzing Big Data with Hive → Ahmad Alkilani
  - Pig Latin: Getting Started → Thomas Henson

Summary



Learned about Hive and Pig

Examples of querying data in HDFS

Discussed use cases for Hive and Pig

Found resources for Hive and Pig