Introduction to Apache Pig

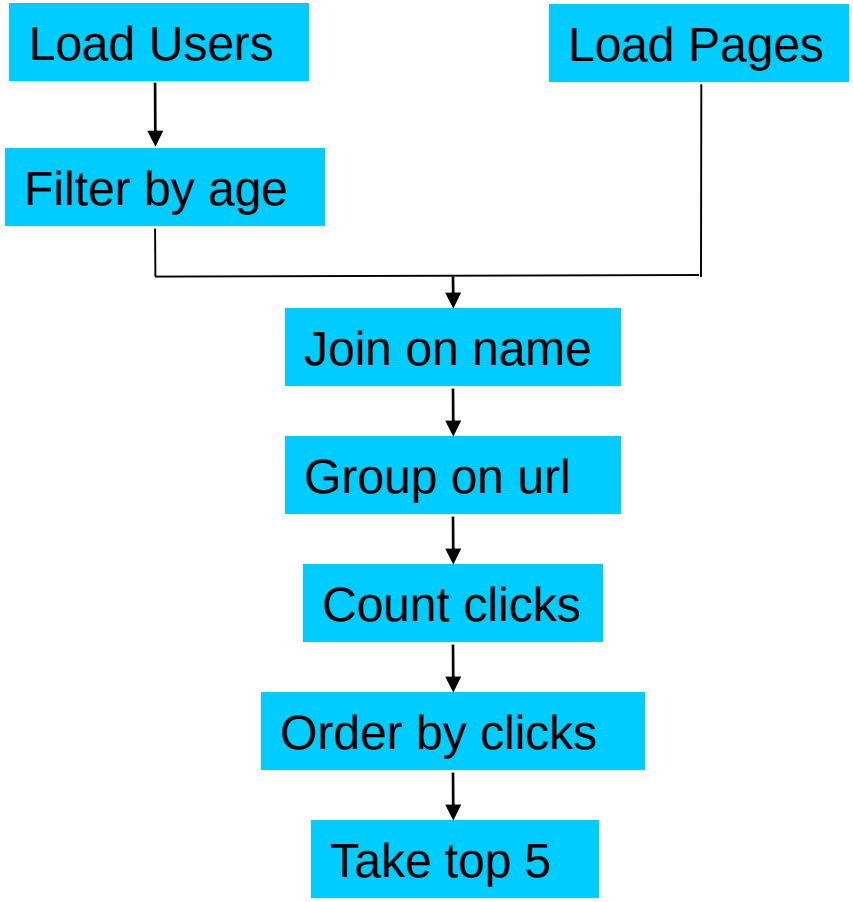
Prof. Dr. Stephan Trahasch
Offenburg University of Applied Sciences

Apache Pig

- Apache Pig is a platform for data analysis and processing on Hadoop
- It offers an alternative to writing MapReduce code directly
- Goals: flexibility, productivity, and maintainability
- Started at Yahoo! Research. Now an open-source Apache project
- Now runs about 30% of Yahoo!'s jobs
- Features
  - Expresses sequences of MapReduce jobs
  - Data model: nested "bags" of items
  - Provides relational (SQL) operators (JOIN, GROUP BY, etc.)
  - Easy to plug in Java functions

An Example Problem

Suppose you have user data in a file, website data in another, and you need to find the top 5 most visited pages by users aged 18-25

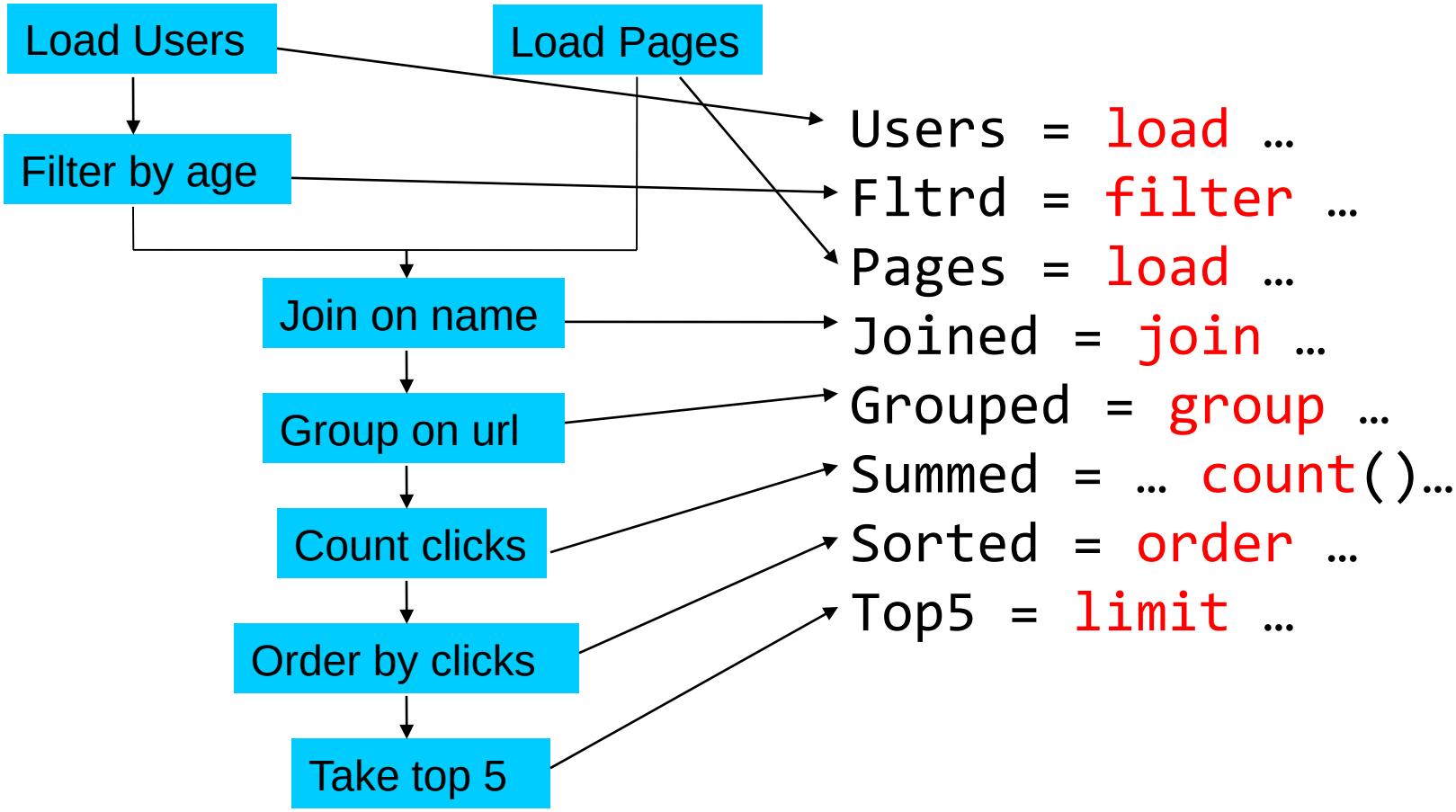


In MapReduce

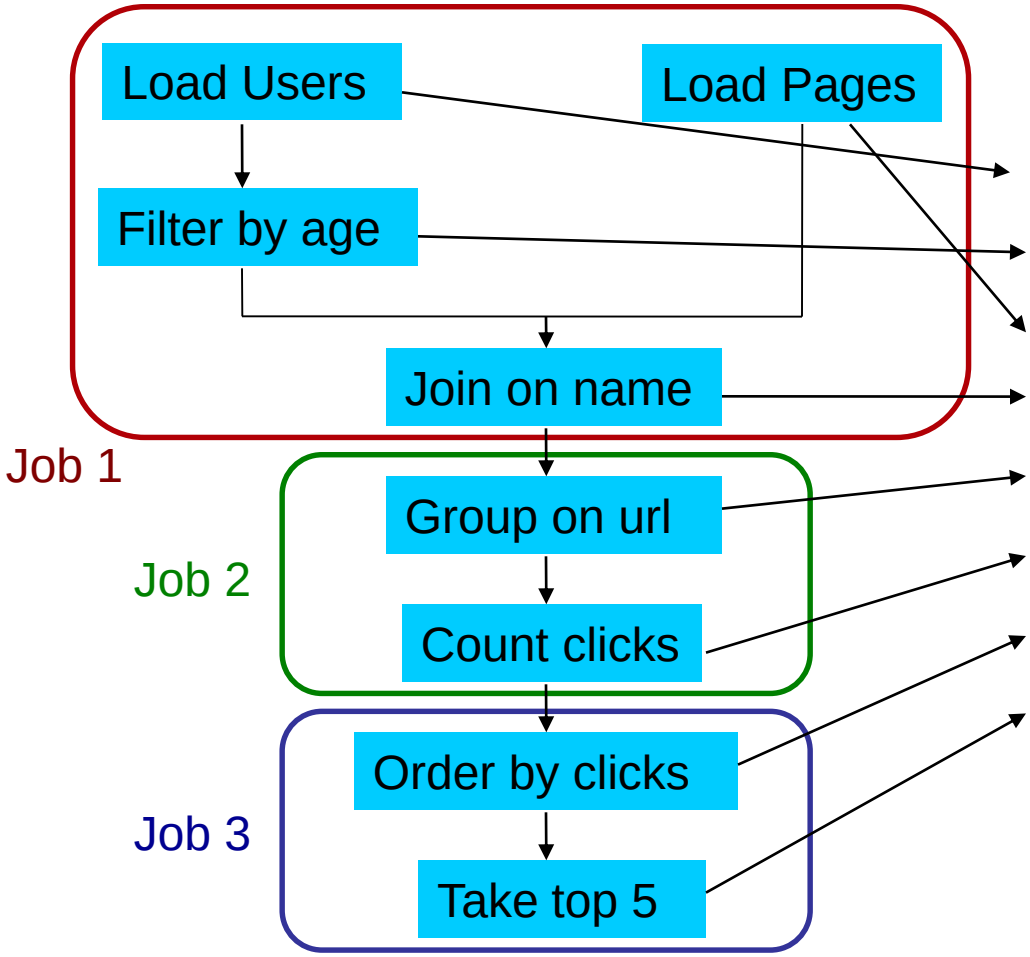
In PigLatin

```
Users = load ‘users’ as (name, age);
Filtered = filter Users by age >= 18 and age <= 25;
Pages = load ‘pages’ as (user, url);
Joined = join Filtered by name, Pages by user;
Grouped = group Joined by url;
Summed = foreach Grouped generate group,
                 count(Joined) as clicks;
Sorted = order Summed by clicks desc;
Top5 = limit Sorted 5;
store Top5 into ‘top5sites’;
```

Ease of Translation



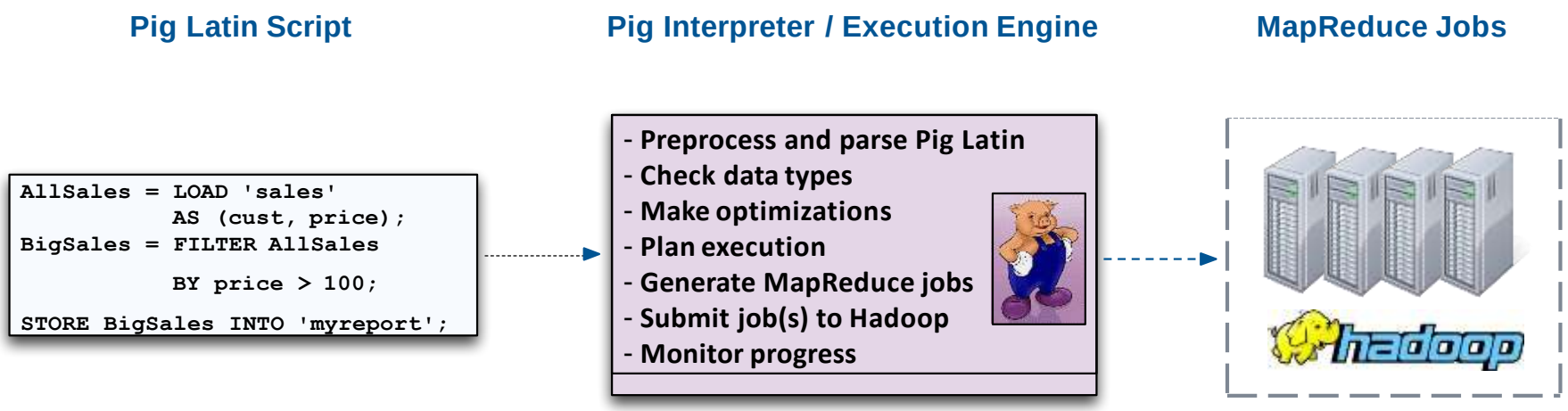
Ease of Translation




```
Users = load …
Fltrd = filter …
Pages = load …
Joined = join …
Grouped = group …
Summed = … count()…
Sorted = order …
Top5 = limit …
```

The Anatomy of Pig

- Pig Latin: The data fow language
- Grunt: interactive shell where you can type Pig Latin statements
- The Pig interpreter and execution engine



Use Case: Web Log Session

…

Web Server Log Data

10.174.57.241 - - [03/May/2013:17:57:41 -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/2013:17:57:43 -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/2013:17:58:03 -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/2013:17:58:04 -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/2013:17:58:25 -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/2013:17:58:26 -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/2013:17:59:36 -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/2013:17:59:47 -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

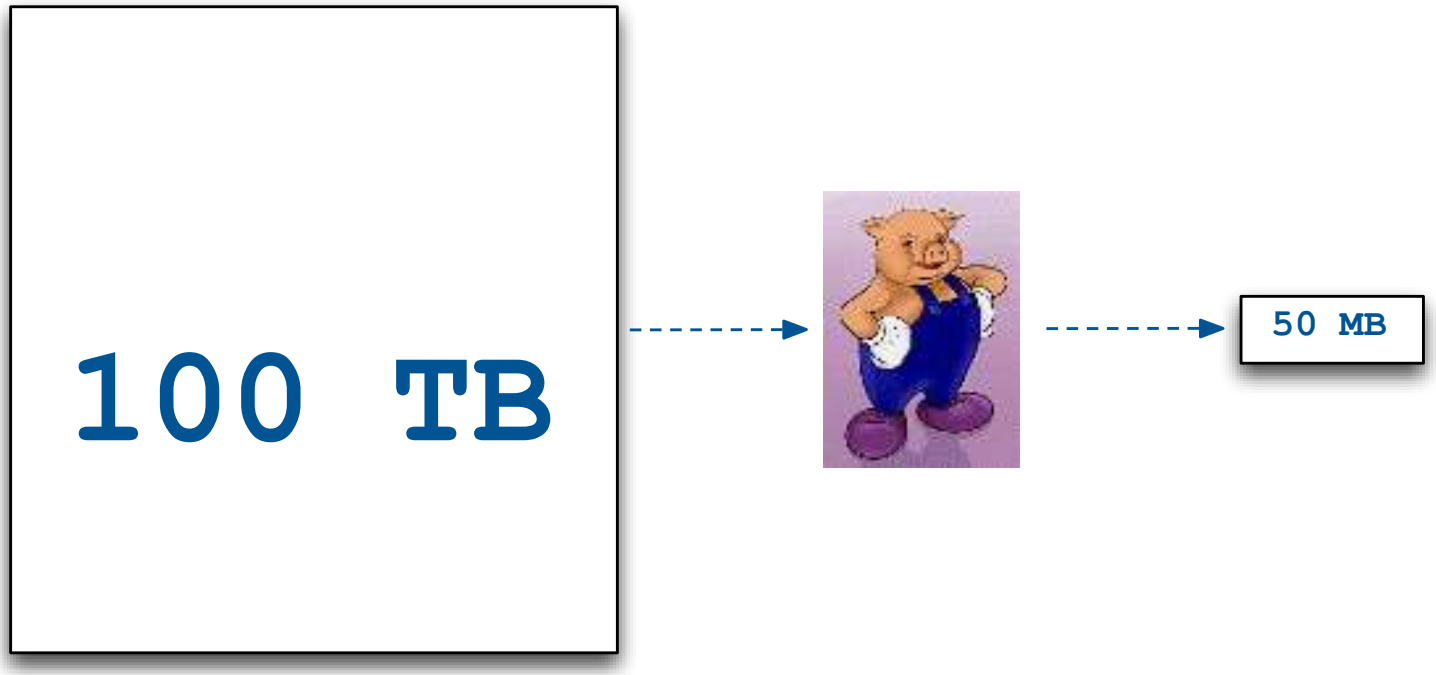
May 3, 2013	May 12, 2013
Search for 'Widget'	Track Order
Widget Results	Contact Us
Details for Widget X	Send Complaint
Order Widget X	

Use Case: Data Sampling

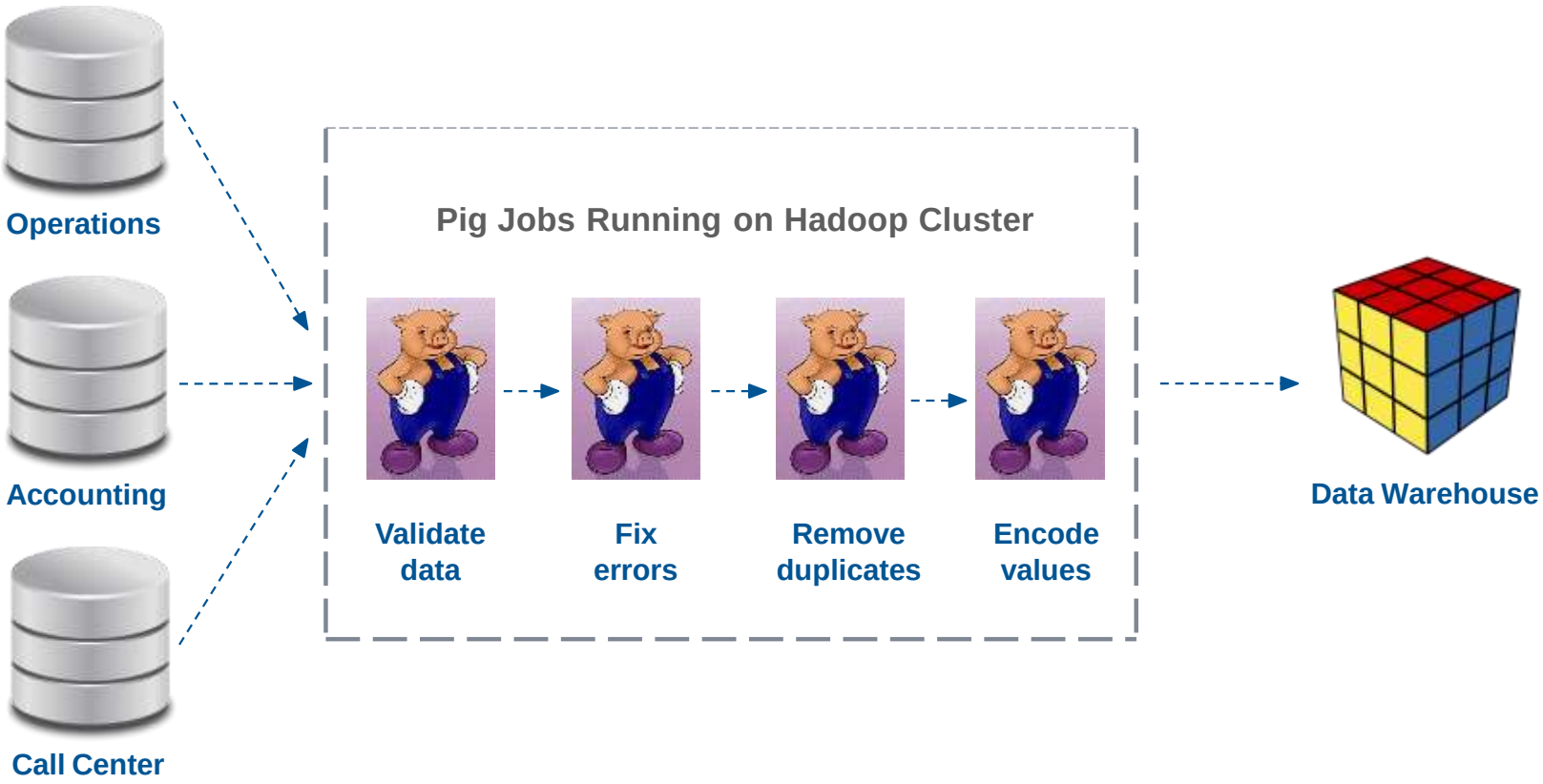
Sampling can help you explore a representative portion of a large data set

Allows you to examine this portion with tools that do not scale well

Supports faster iterations during development of analysis jobs



Use Case: ETL Processing



Pig Features

Pig is an alternative to writing low-level MapReduce code

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

Using Pig Interactively

You can use Pig interactively, via 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 > 100;
grunt> STORE bigsales INTO 'myreport';
grunt> quit;
```

You can also execute a Pig Latin statement from the UNIX shell via the -e option

You can manipulate HDFS with Pig, via 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 > 100;
grunt> STORE bigsales INTO 'myreport';
grunt> fs -getmerge myreport/ bigsales.txt;
```

The sh command lets you run UNIX programs from Pig

```
grunt> sh date;
Fri May 10 13:05:31 PDT 2013
grunt> fs -ls; -- lists HDFS files
grunt> sh ls; -- lists local files
```

Running a Pig Script

A Pig script is simply Pig Latin code stored in a text ?le

By convention, these files have the .pig extension

You can run a Pig script from within the Grunt shell via the run command.

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

- 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 MapReduce 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
```

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

- Pig offers an alternative to writing MapReduce code directly
- Pig interprets Pig Latin code in order to create MapReduce jobs
- It then submits these MapReduce jobs to the Hadoop cluster
- You can execute Pig Latin code interactively through Grunt
- 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