



University of
South Australia

What is Apache Pig

1



- Platform for analyzing large datasets
- Uses its own language called Pig Latin
- Pig Latin is really easy to write!
- Automatically optimizes code execution

Apache Pig

2

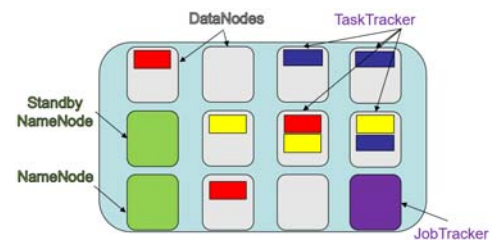
Pig runs on Hadoop. It makes use of HDFS and Yarn.



Hadoop 2.0 uses YARN (Yet Another Resource Negotiator)

This splits up JobTracker's two primary responsibilities, resource management and job scheduling, into two separate daemons; the Resource Manager and Application Master. These handle scheduling and accepting job submissions.

The TaskTracker is effectively replaced by NodeManager, which is responsible for monitoring node resource usage (memory, cpu etc.)



3

Apache pig philosophy

<http://pig.apache.org/philosophy.html>

Pigs Eat Anything

istockphoto.com



Pigs Live Anywhere

shutterstock.com



Pigs Are Domestic Animals

<https://www.nadineathome.com>



Pigs Fly

shutterstock.com



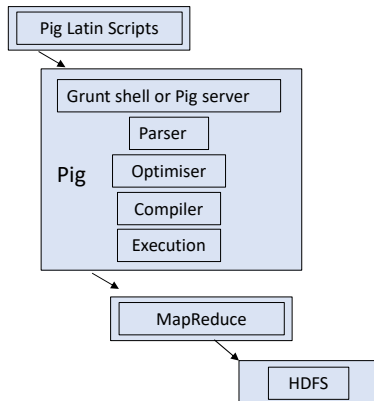
Pigs are ... light?

4

Pig components

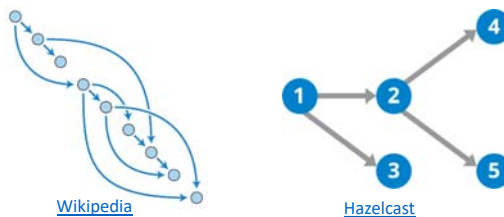
Pig is made up of two main components: the runtime engine and the Pig Latin scripting language

Runtime Engine



Pig Latin

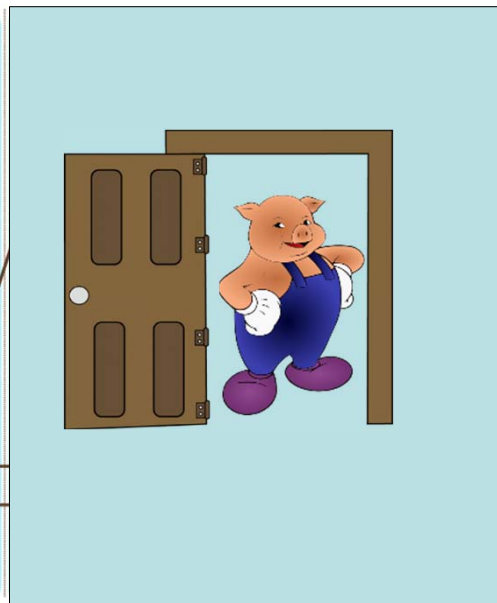
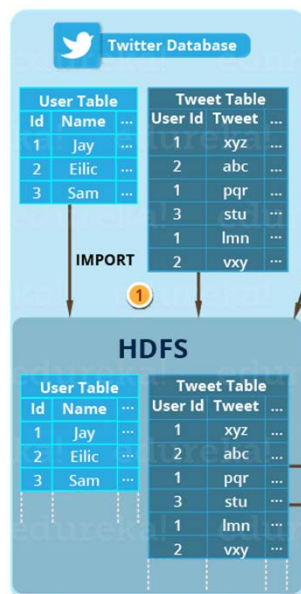
- Dataflow language
- Described through a Directed Acyclic Graph (DAG)



Modes

Local
Vs
MapReduce

Interactive
Vs
Batch



Use case is 
often Twitter

Though Yahoo runs about 40% of its Hadoop jobs in Pig too

Why use
MapReduce at all?

- More control
- Difficult jobs
- Faster

WARNING

This material has been reproduced and communicated to you by or on behalf of the **University of South Australia** in accordance with section 113P of the *Copyright Act 1968* (**Act**).

The material in this communication may be subject to copyright under the Act. Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

Do not remove this notice