Use all of the data to avoid missing out

Disk read & write time

Want to spend maximum amount of time processing data (not copying)

Send the processing to the data (stored locally)

NameNodes

Secondary NameNode used for memory intensive things

It won’t cover the main one if that fails

So important it’s good

DataNode

Fault tolerance

DataNode lets NameNode know who’s available all the time

Admin

How set up cluster

Distributed across virtual machines

# Pig

Useful for taking small samples to run on local machine

LOAD assumes PigStorage & tab-separated columns (USING gives other options)

Use regular expression for loading multiple files

As sets the column name, otherwise refer to columns by $#

Pig assumes null if data is unexpected

A field is a specific element

A tuple is an ordered set of related elements

A bag is a collection of unordered tuples

A relation is a named bag

DUMP prints to console

STORE sends to disk (HDFS)

DESCRIBE gives variable types

FOREACH & GENERATE allows you to extract columns

DISTINCT eliminates duplicate records

ORDER BY & LIMIT is great for selecting things

There are lots of nice functions for manipulating strings

GROUP BY & DESCRIBE is great for breaking up the data

COGROUP creates a nested data structure

Outer joins do not require the key to be in both data sets

Left & right returns all left & right

CROSS will end up insane

## Pig Sampling

1. Everything = LOAD data
2. Subset = SAMPLE everything 0.05
3. Something

Or you could use ILLUSTRATE

# ETL

Processing types

* Extract
* Transform
* Load

Useful for validating data, fixing errors & removing duplicates