University of South Australia

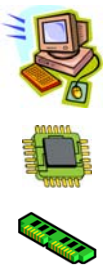
A brief history of Hadoop

1

Big data analysis

When trying to analyse big data we can try two things; scale up or scale out.

- More powerful computer
- More powerful processor
- More memory



Scale up

- Distributes over more computers
- Scales better
- Presents new problems

Scale out

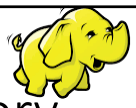
2

Hadoop History

- In 1997 Doug Cutting started writing the first version of Lucene
- Lucene was a full text search library. The same type of library used by Google.
- In 2001 Doug was joined by Mike Cafarella, and a new Lucene subproject called Apache Nutch was born

Hadoop History

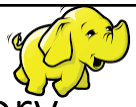
- Nutch was a web crawler
- Nutch used Lucene to index every page it visited
- The pair's first effort was on a single $3000 machine
- They couldn't index the whole web on this machine, so they tried four of them



Hadoop History

- Without any cluster management tool all data transfer and space allocation had to be handled manually.
- This meant extra time and effort programming, constant oversight, increased risk of data loss etc.

☑ Structureless ☑ Durable ☑ Robust ☑ Autonomous



Hadoop History

- By the end of 2004 Doug and Mike had created the 'Nutch Distributed File System'
- Later (2006), when the project grew large enough, it was removed from Nutch, renamed to Hadoop (after Doug's son's toy elephant) and this cluster management tool came to be known as the Hadoop Distributed File System (HDFS).

• • •