



University of
South Australia

What is MapReduce?

1

What is MapReduce? ... through an example



Imagine that the university is interested in doing a kind of census of their wildlife

They make an app that anyone can download and record which animals they have seen

If this were to become a big data problem, how would you process the data?

2

Keys

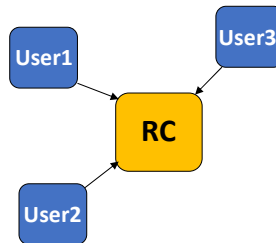
Cat

Fox

Owl

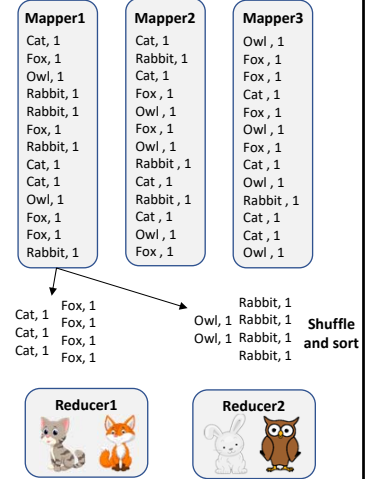
Rabbit

User1	User2	User3
Cat	Cat	Owl
Fox	Rabbit	Fox
Owl	Cat	Fox
Rabbit	Fox	Cat
Rabbit	Owl	Fox
Fox	Fox	Owl
Rabbit	Owl	Fox
Cat	Rabbit	Cat
Cat	Cat	Owl
Owl	Rabbit	Rabbit
Fox	Cat	Cat
Fox	Owl	Cat
Rabbit	Fox	Owl



- Memory issues with large datasets
- Data needs to be transferred
- All 39 entries unsorted and handled iteratively

Better to make use of a distributed file system, like HDFS



Keys

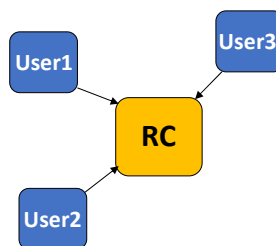
Cat

Fox

Owl

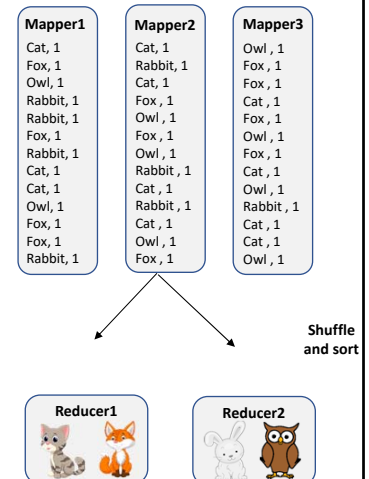
Rabbit

User1	User2	User3
Cat	Cat	Owl
Fox	Rabbit	Fox
Owl	Cat	Fox
Rabbit	Fox	Cat
Rabbit	Owl	Fox
Fox	Fox	Owl
Rabbit	Owl	Fox
Cat	Rabbit	Cat
Cat	Cat	Owl
Owl	Rabbit	Rabbit
Fox	Cat	Cat
Fox	Owl	Cat
Rabbit	Fox	Owl



- Memory issues with large datasets
- Data needs to be transferred
- All 39 entries unsorted and handled iteratively

Better to make use of a distributed file system, like HDFS



Keys

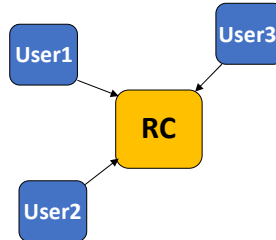
Cat

Fox

Owl

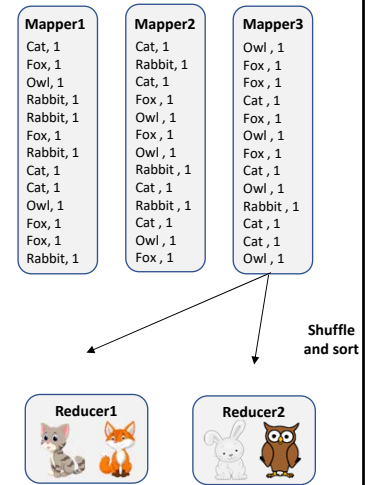
Rabbit

User1	User2	User3
Cat	Cat	Owl
Fox	Rabbit	Fox
Owl	Cat	Fox
Rabbit	Fox	Cat
Rabbit	Owl	Fox
Fox	Fox	Owl
Rabbit	Owl	Fox
Cat	Rabbit	Cat
Cat	Cat	Owl
Owl	Rabbit	Rabbit
Fox	Cat	Cat
Fox	Owl	Cat
Rabbit	Fox	Owl



- Memory issues with large datasets
- Data needs to be transferred
- All 39 entries unsorted and handled iteratively

Better to make use of a distributed file system, like HDFS



Keys

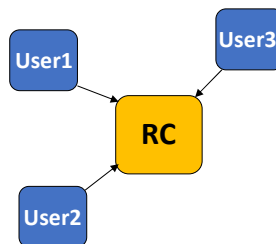
Cat

Fox

Owl

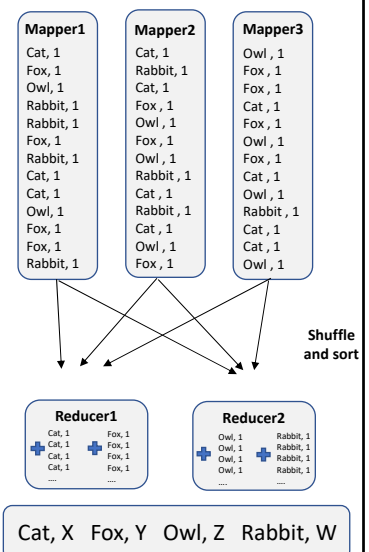
Rabbit

User1	User2	User3
Cat	Cat	Owl
Fox	Rabbit	Fox
Owl	Cat	Fox
Rabbit	Fox	Cat
Rabbit	Owl	Fox
Fox	Fox	Owl
Rabbit	Owl	Fox
Cat	Rabbit	Cat
Cat	Cat	Owl
Owl	Rabbit	Rabbit
Fox	Cat	Cat
Fox	Owl	Cat
Rabbit	Fox	Owl



- Memory issues with large datasets
- Data needs to be transferred
- All 39 entries unsorted and handled iteratively

Better to make use of a distributed file system, like HDFS



Keys

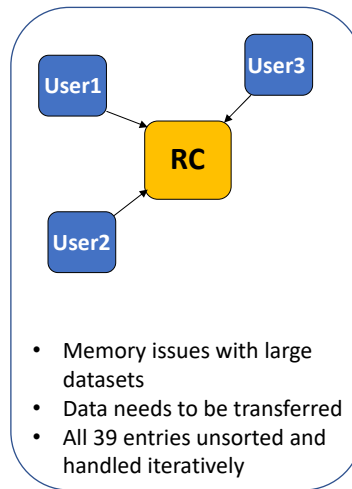
Cat

Fox

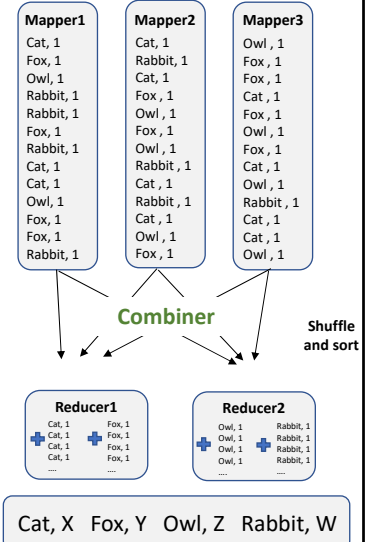
Owl

Rabbit

User1	User2	User3
Cat	Cat	Owl
Fox	Rabbit	Fox
Owl	Cat	Fox
Rabbit	Fox	Cat
Rabbit	Owl	Fox
Fox	Fox	Owl
Rabbit	Owl	Fox
Cat	Rabbit	Cat
Cat	Cat	Owl
Owl	Rabbit	Rabbit
Fox	Cat	Cat
Fox	Owl	Cat
Rabbit	Fox	Owl



Better to make use of a distributed file system, like HDFS



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