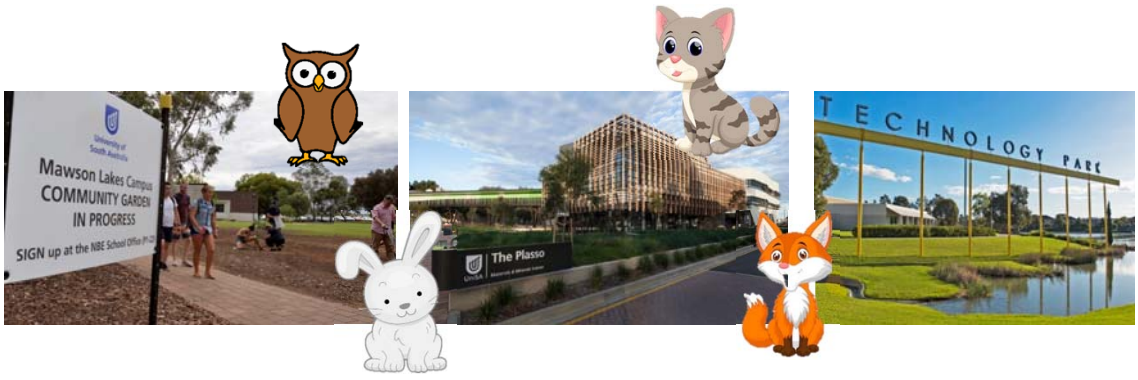
University of South Australia

What is MapReduce?

1

What is MapReduce? … through an example



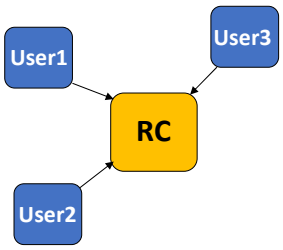
Imagine that the university is interested in doing a kind of census of their wildlife

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?

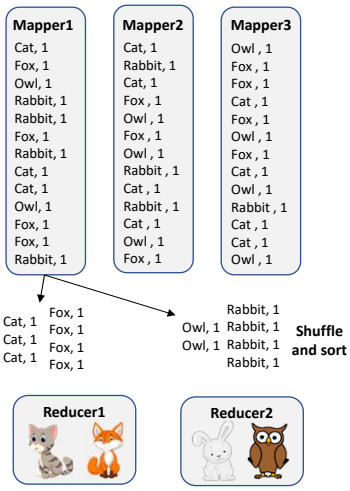
2

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



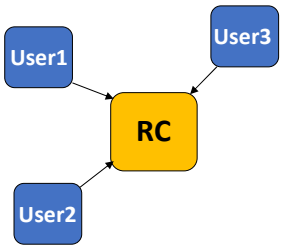
- Memory issues with large datasets
- Data needs to be transferred
- All 39 entries unsorted and handled iteratively

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



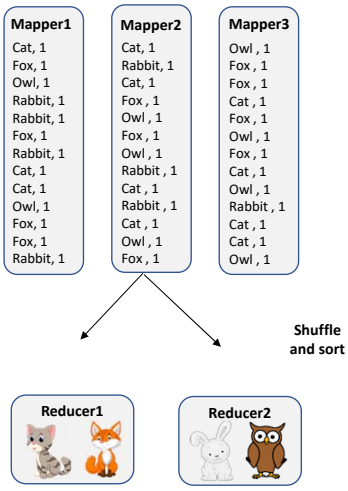
3

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



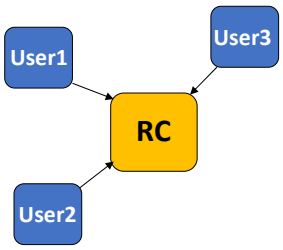
- Memory issues with large datasets
- Data needs to be transferred
- All 39 entries unsorted and handled iteratively

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



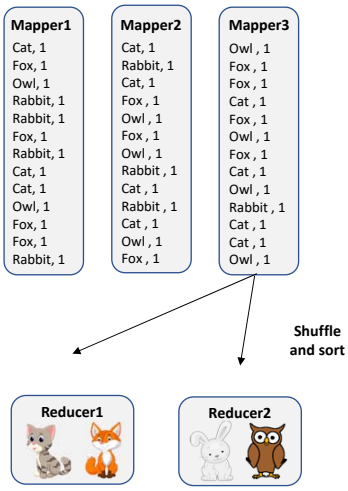
4

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



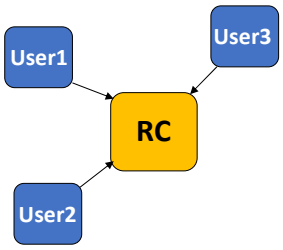
- Memory issues with large datasets
- Data needs to be transferred
- All 39 entries unsorted and handled iteratively

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



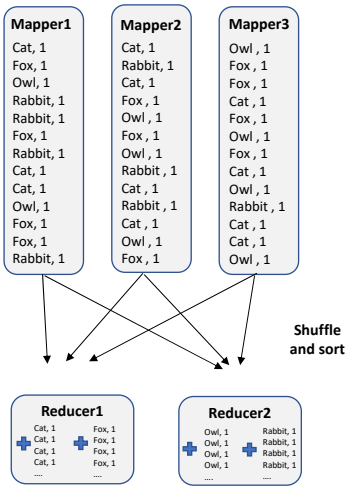
5

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



Cat, X Fox, Y Owl, Z Rabbit, W

6

Keys

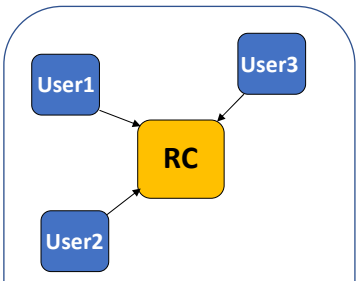
Cat

Fox

Owl

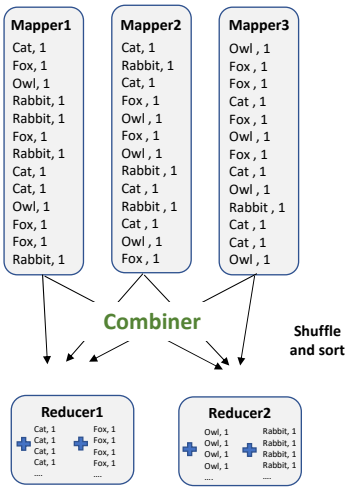
Rabbit

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



Cat, X Fox, Y Owl, Z Rabbit, W

7

8