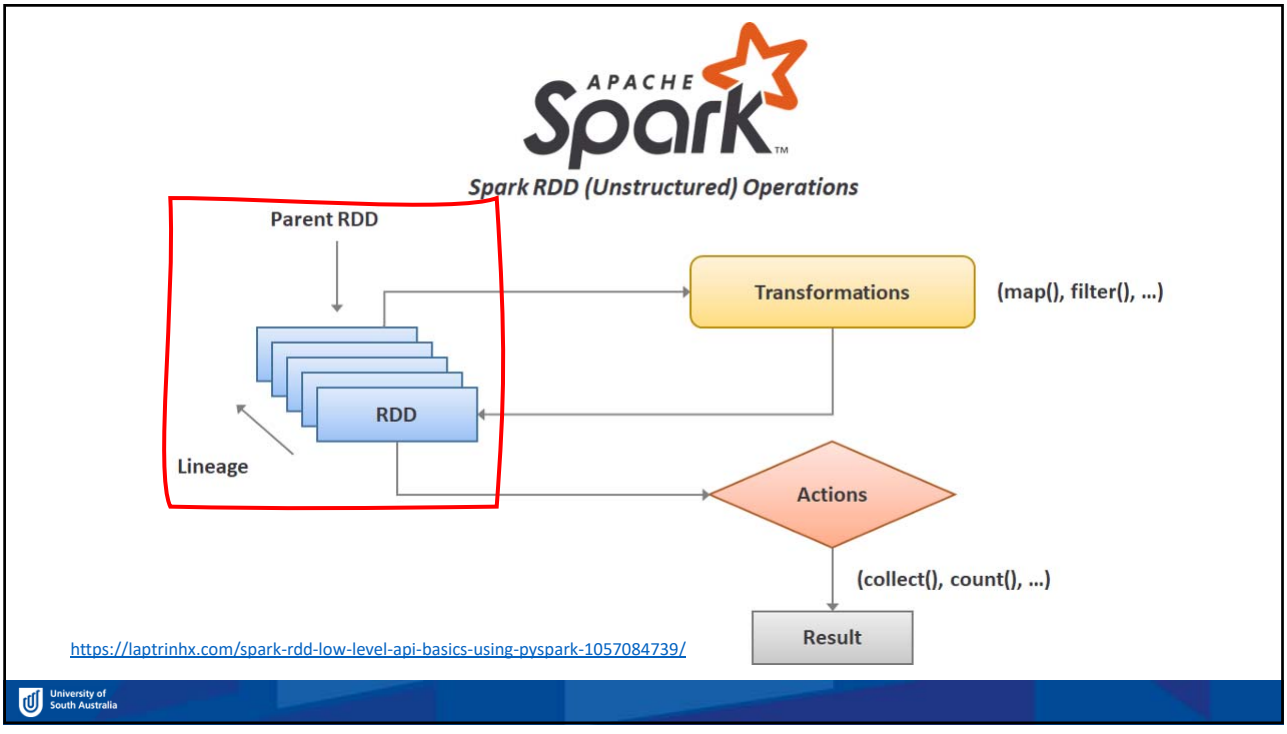
University of South Australia

# What is an RDD

1




2

# Resilient Distributed Datasets (RDD)

The main abstraction Spark provides is a *resilient distributed dataset* (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. RDDs are created by starting with a file in the Hadoop file system (or any other Hadoop-supported file system), and transforming it.

## RDD Properties

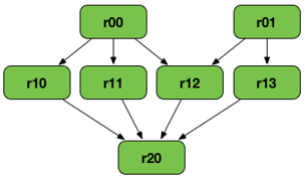
**Processed in memory**

Grants Spark substantial increases in speed

**Lazy evaluation**

Just like Pig

**Fault tolerance … without the redundancy**






3





4