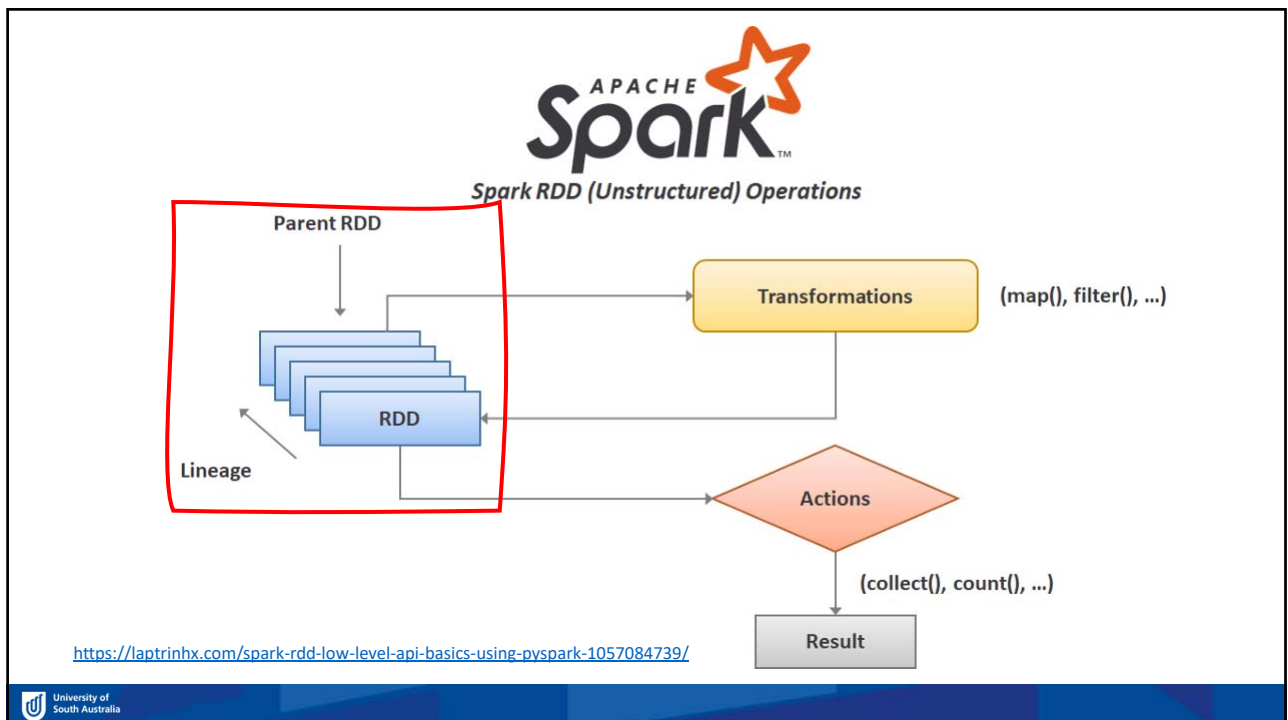




University of
South Australia

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

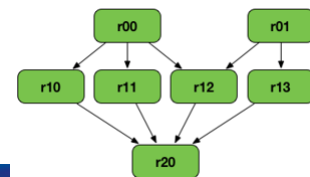
Processed in memory

Grants Spark substantial
increases in speed

Lazy evaluation

Just like Pig

Fault tolerance ... without the redundancy



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