Introduce to *Spark*

A big data processing tool built with Scala and runs on JVM

ADB 2017

Yen Hao Huang

Big Data

- 4Vs
  - Volume/Variety/Velocity/Veracity

Due to the rise of Big Data, faster tools are required for processing data.

Hadoop

Hadoop

- A platform to store and process large scale data
- Features
  - Scalable
  - Economical : many cheap servers
  - Flexible : schema-less
  - Reliable : replicas

Hadoop MapReduce

- Map
  - Divide job to multiple tiny tasks and distribute to servers
- Reduce
  - Summary the results from those servers

Hadoop MapReduce

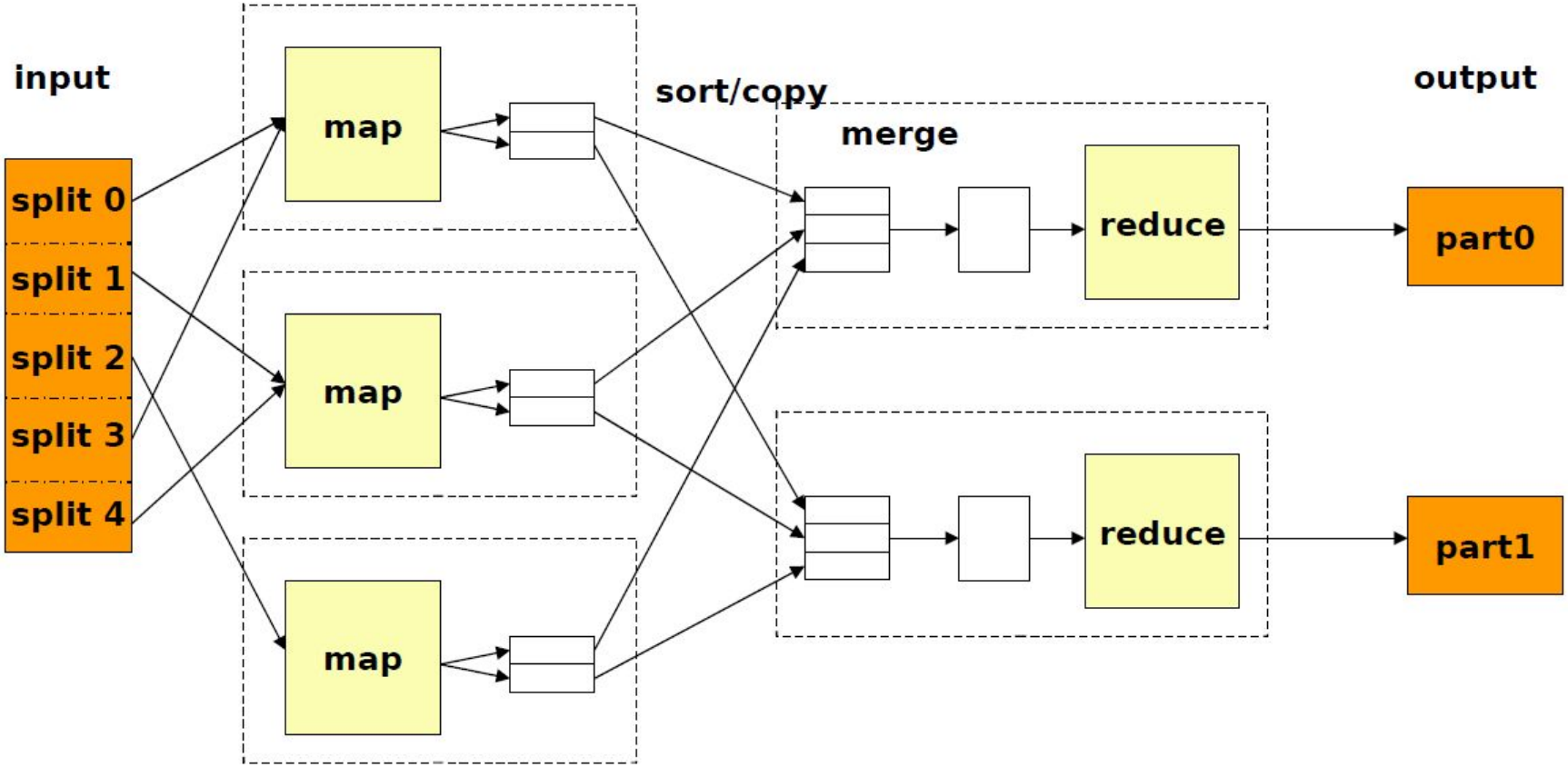
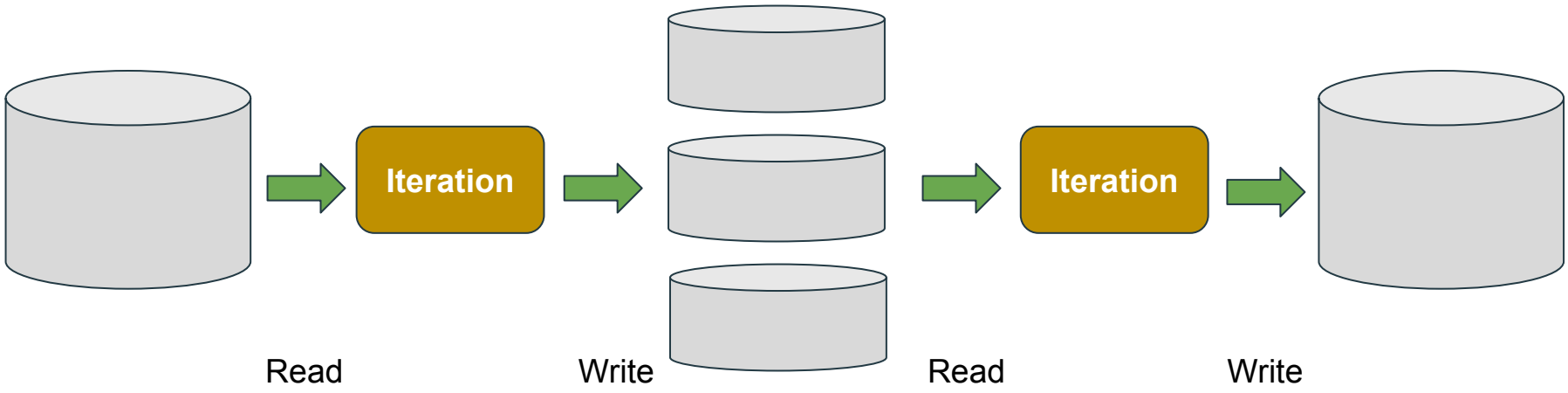


Figure Refence

Hadoop - Bottleneck

- File I/O - write the middle process data to disk



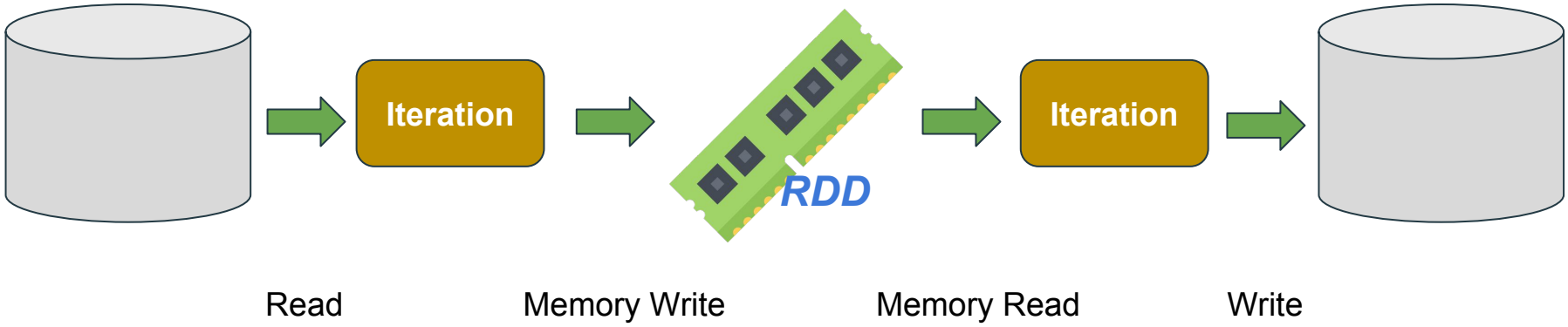
Spark

RDD

In-memory computation framework

RDD (Resilient Distributed Dataset)

- Write the middle process data to memory
- 10 - 100 times faster than hadoop

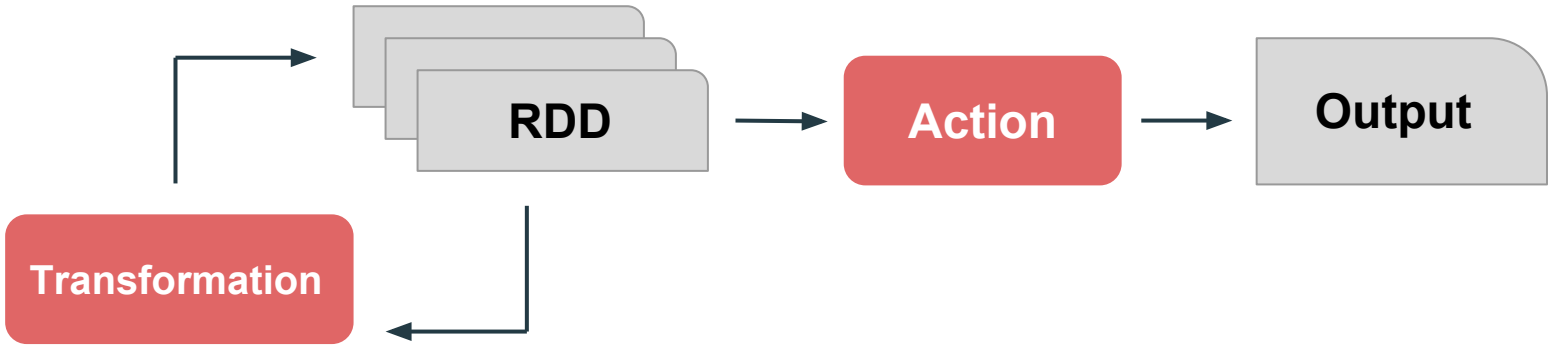


Spark

- Features
  - Speed
  - Ease of use : Scala、Python、Java、R
  - Supports hadoop : HDFS、MapReduce
  - Accessibility : runs on many platforms

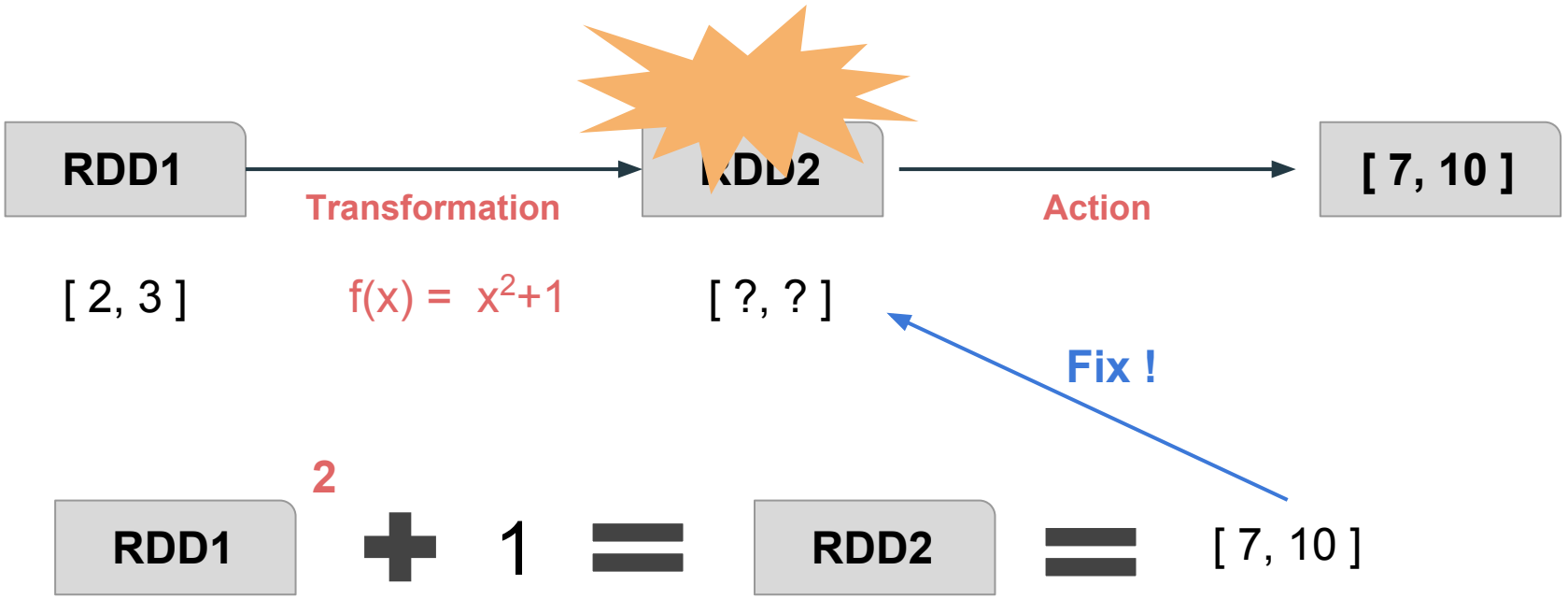
RDD Features

- Computations
  - Transformation - Lazy compute
  - Action - Execute the computations
  - Persistence - Keep RDD in ram/ disk



RDD Lineage

- Error Fixing



Spark Functionality