RDD

Kristo Raun

Big Data Management 2024

UNIVERSITY OF TARTU
1632

What is RDD?

- An abstraction called resilient distributed datasets.

- Source: Spark: Cluster computing with working sets (2010, Zaharia et al)

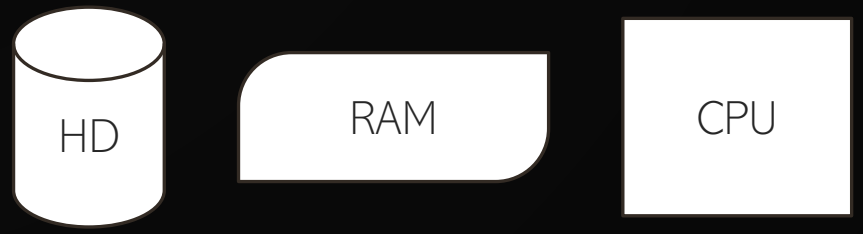
What is RDD?

- An abstraction called resilient distributed datasets.
- A read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost.

- Source: Spark: Cluster computing with working sets (2010, Zaharia et al)

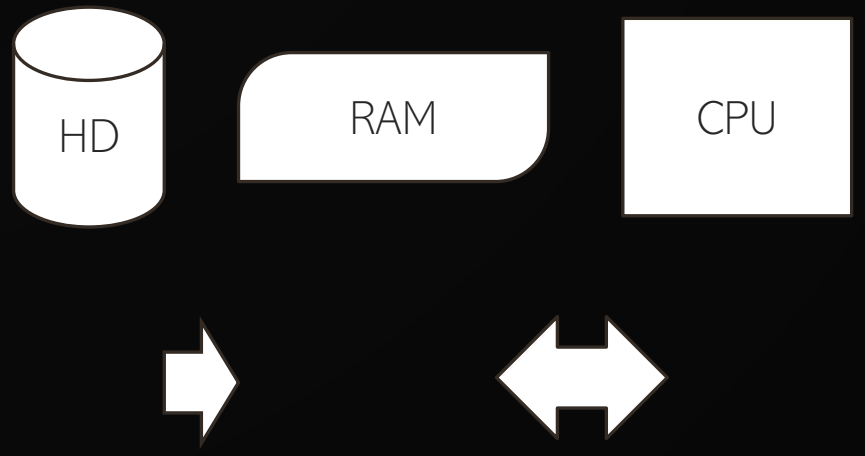
Problem with Hadoop Map/Reduce

Problem with Hadoop Map/Reduce

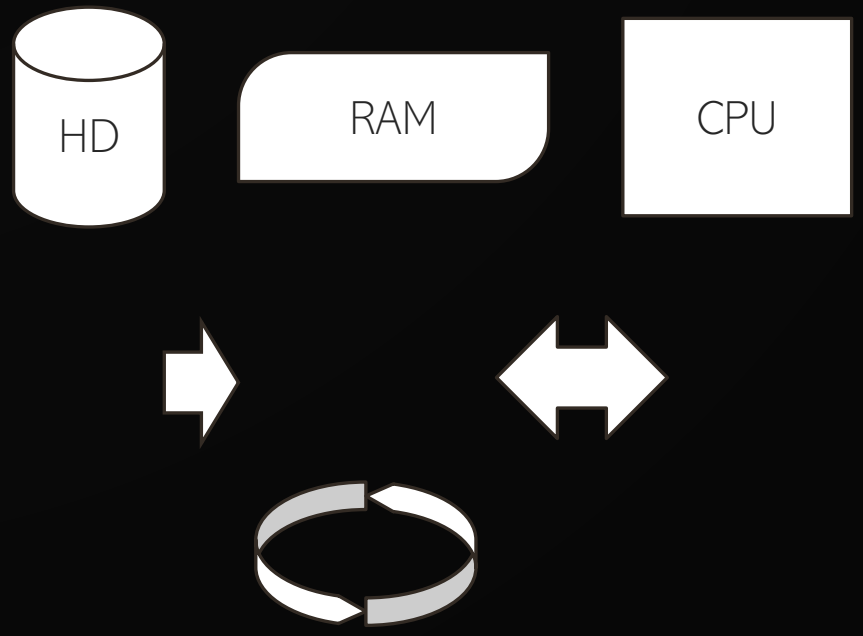


Problem with Hadoop Map/Reduce

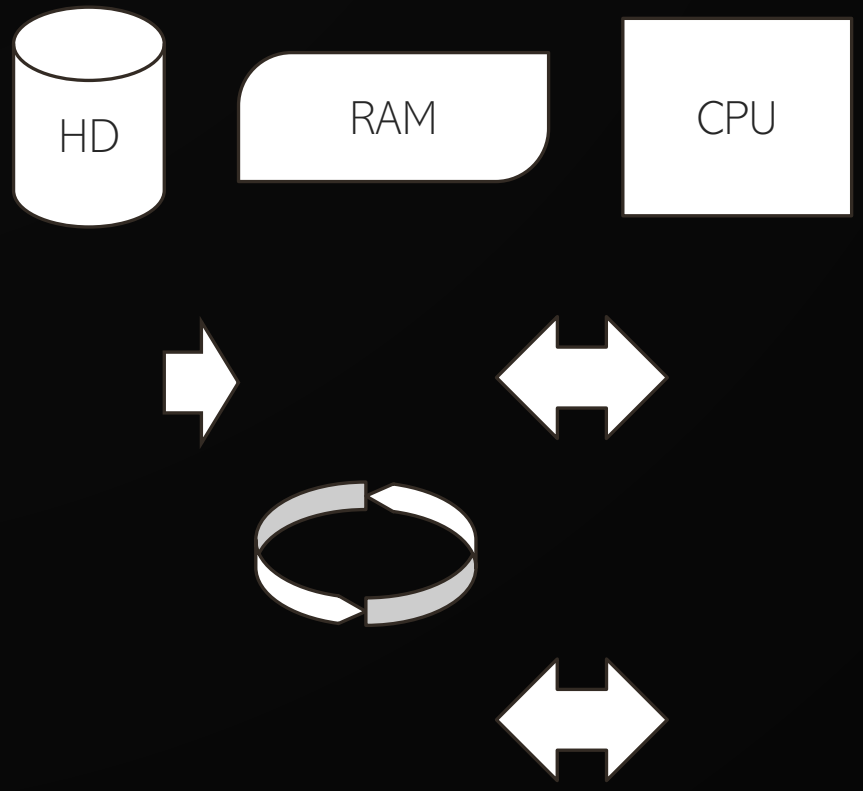
Problem with Hadoop Map/Reduce



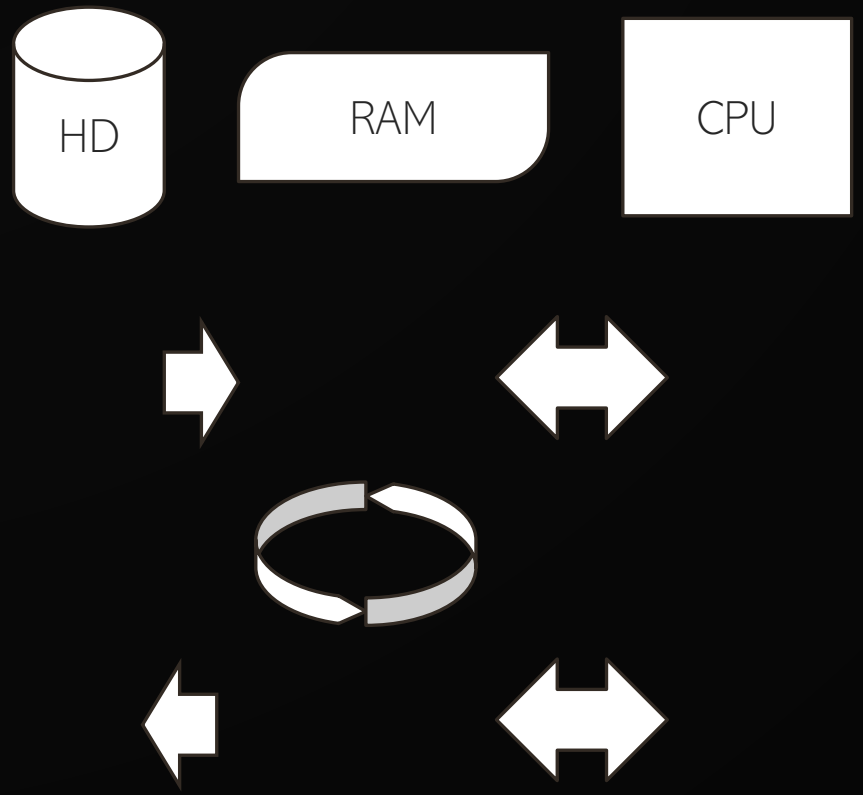
Problem with Hadoop Map/Reduce



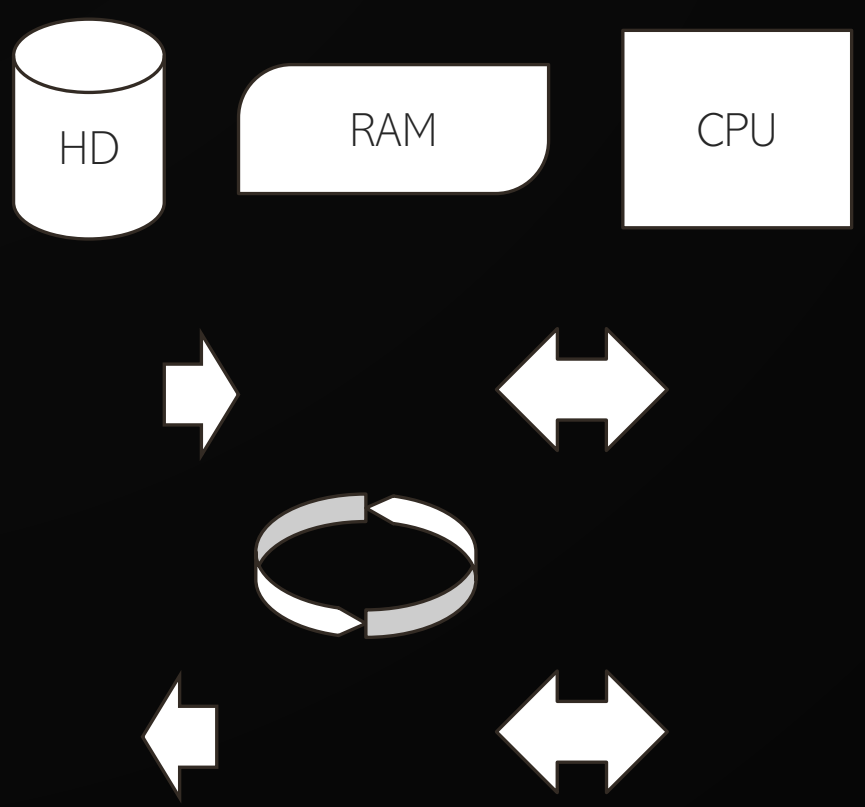
Problem with Hadoop Map/Reduce



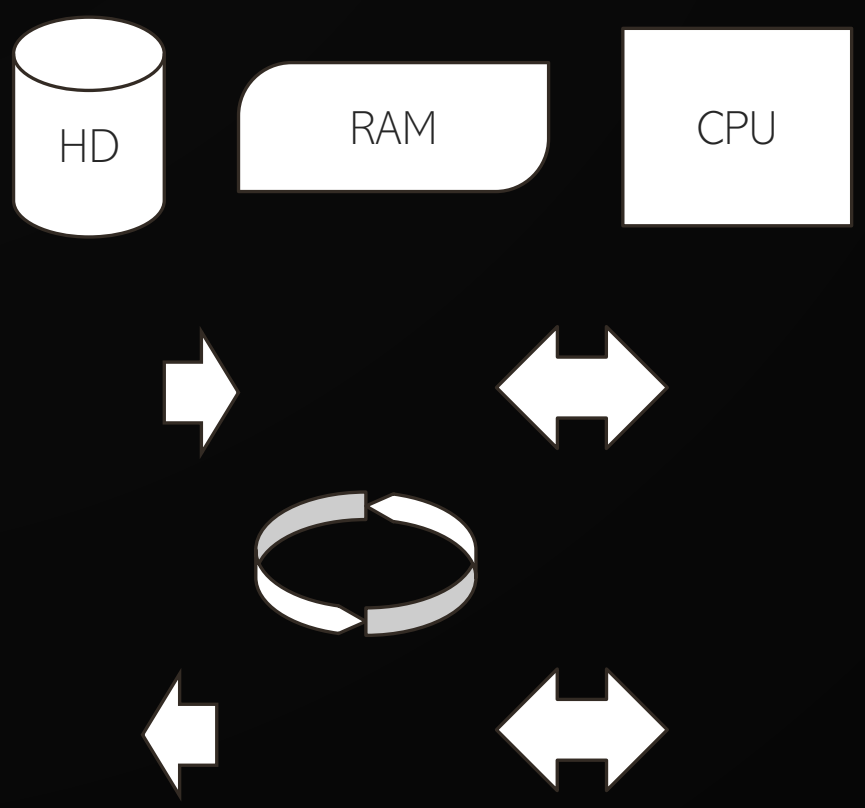
Problem with Hadoop Map/Reduce

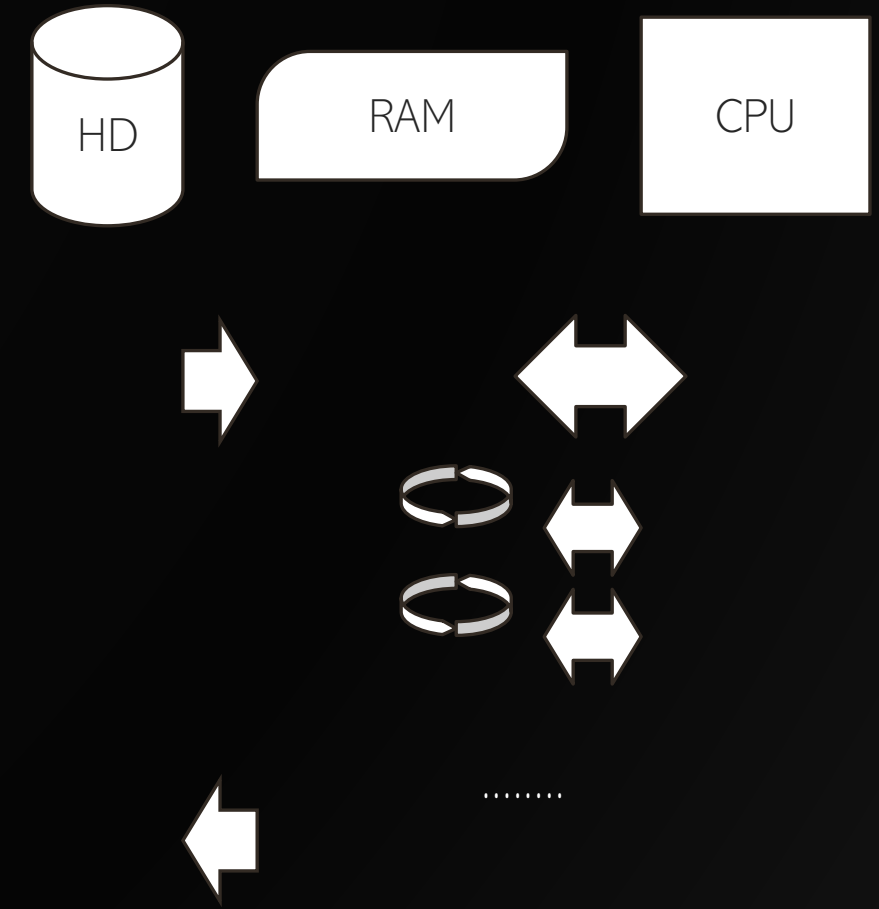


Problem with Hadoop Map/Reduce



Problem with Hadoop Map/Reduce





Spark RDDs

Spark RDDs

- Solved reuse of data within transformations

Spark RDDs

- Solved reuse of data within transformations
- Use cases:
  - Interactive data analysis
  - Iterative ML
  - Graph processing
  - Streaming

Things to note

Things to note

- Spark RDDs were introduced in ~2010

Things to note

- Spark RDDs were introduced in ~2010
- By now, almost everyone uses DataFrames
  - Higher level
  - Usually more performant
  - More user-friendly

Things to note

- Spark RDDs were introduced in ~2010
- By now, almost everyone uses DataFrames
  - Higher level
  - Usually more performant
  - More user-friendly
- RDDs are *theoretically* still relevant if you want more control on your dataset