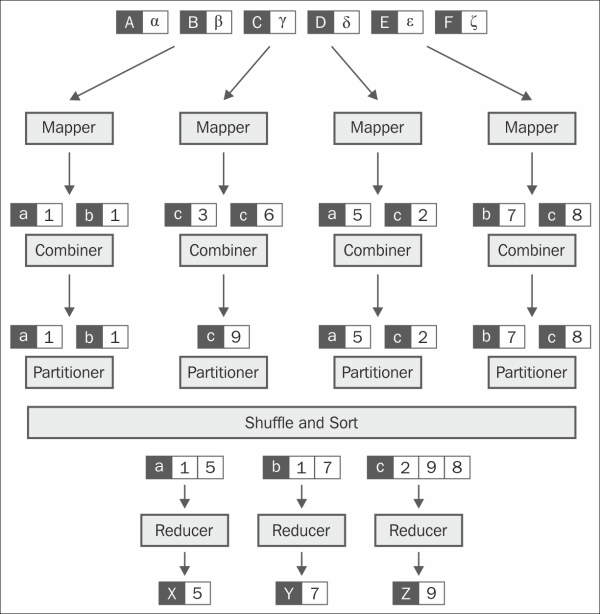
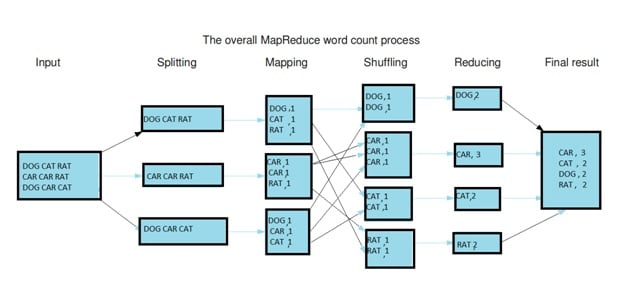
**Large Scale Data Analysis with Spark**

**Map Reduce**



****

**From Hadoop to Spark**

* Disk based operations
* Shuffle-sort between map and reduce process
* Lacks flexibilities for micro-grained parallelism
* Best for batch simple disk based processing

Apache Spark is a cluster computing platform designed to be fast and general purpose.

- Speed:

\* Run computation in memory.

\* Faster than MapReduce on disk.

- Generality:

\* Supporting batch,interactive and streaming.

- Accessibility:

\* Simple APIs in Python, Java, Scala and SQL, and built-in libraries

- Usability:

\* Sparck programming model is more than MapReduce model.

A core concept in Spark: **RDD**

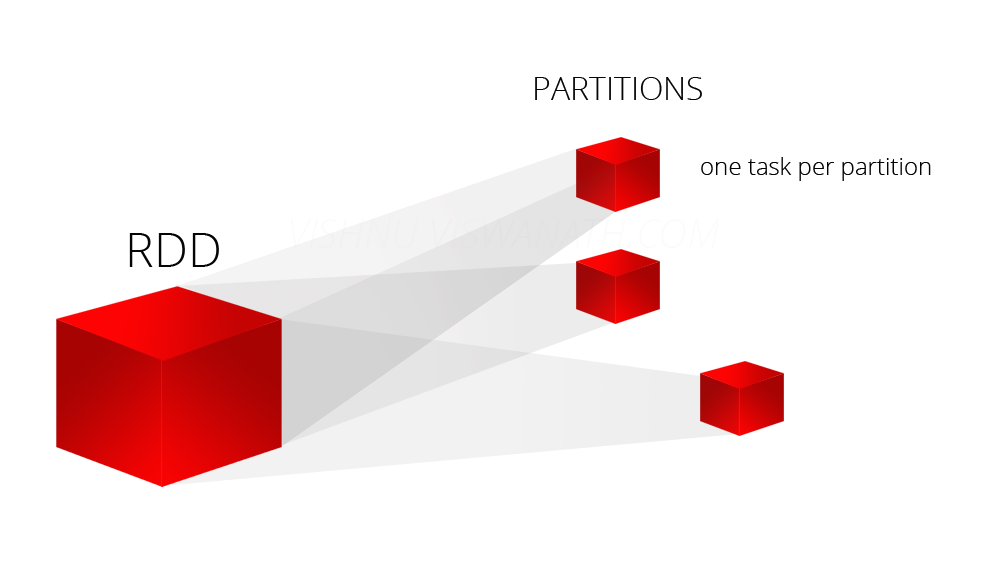
* A resilient Distributed Dataset (RDD)
* The basic abstraction in Spark.
* Represents an immutable, partitioned collection.
* An RDD contains a set of partitions, elements that can be operated on in parallel.
* An RDD is immutable, value cannot be changed.

Why RDD is immutable?

* Safe to share across multiple process
* Easy to move around among resources, cache
* Simplifies development

RDD can be recreated at any time.

* List of dependencies
* Function to compute a partition given its parent.
* Think RDD as a deterministic function rather than a data object.



Fault-tolerance: if the data on once computer gets lost, the system should be able to recover it.

**Introduction on Zepelin**

* A web-based notebook that enables interactive data analytics and visualization.
* Multiple Languages Backend.