# Spark

Spark is a unified computing engine and a set of libraries for parallel data processing on computer clusters

## RDD (Resilient Distributed Dataset):

* Resilient – if data in memory is lost, it can be recreated
* Distributed – stored in memory across the cluster
* Dataset – initial data can come from a file or be created programmatically

RDD are the fundamental unit of data in spark

Most Spark programming consists of performing operations on RDDs

In Spark all work is expressed as either creating new RDDs, transforming existing RDDs, or calling operations on RDDs to compute a result