Apache Spark RDD

Ladle Patel

Definition

- Resilient Distributed Dataset (RDD) is the primary data abstraction in Apache Spark and the core of Spark ("Spark Core").

- Resilient :Fault-tolerant with the help of RDD lineage graph and so able to recompute missing or damaged partitions due to node failures.

- Distributed:Data residing on multiple nodes in a cluster.

- Dataset :Collection of partitioned data .

Definition Cont..

- RDD is fault-tolerant ,In-Memory, Immutable,Lazy evaluated, collection of distributed elements which are partitioned across the nodes of the cluster can be operated on in parallel.

- Partitioned collections of objects spread across a cluster, stored in memory or on disk .

Definition Cont..

- In-Memory:Data inside RDD is stored in memory as much (size) and long (time) as possible.
- Immutable or Read-Only : It does not change once created and can only be transformed using transformations to new RDDs.
- Lazy evaluated :The data inside RDD is not available or transformed until an action is executed that triggers the execution.
- Cacheable :You can hold all the data in a persistent "storage" like memory (default and the most preferred) or disk (the least preferred due to access speed).
- Parallel :Process data in parallel.

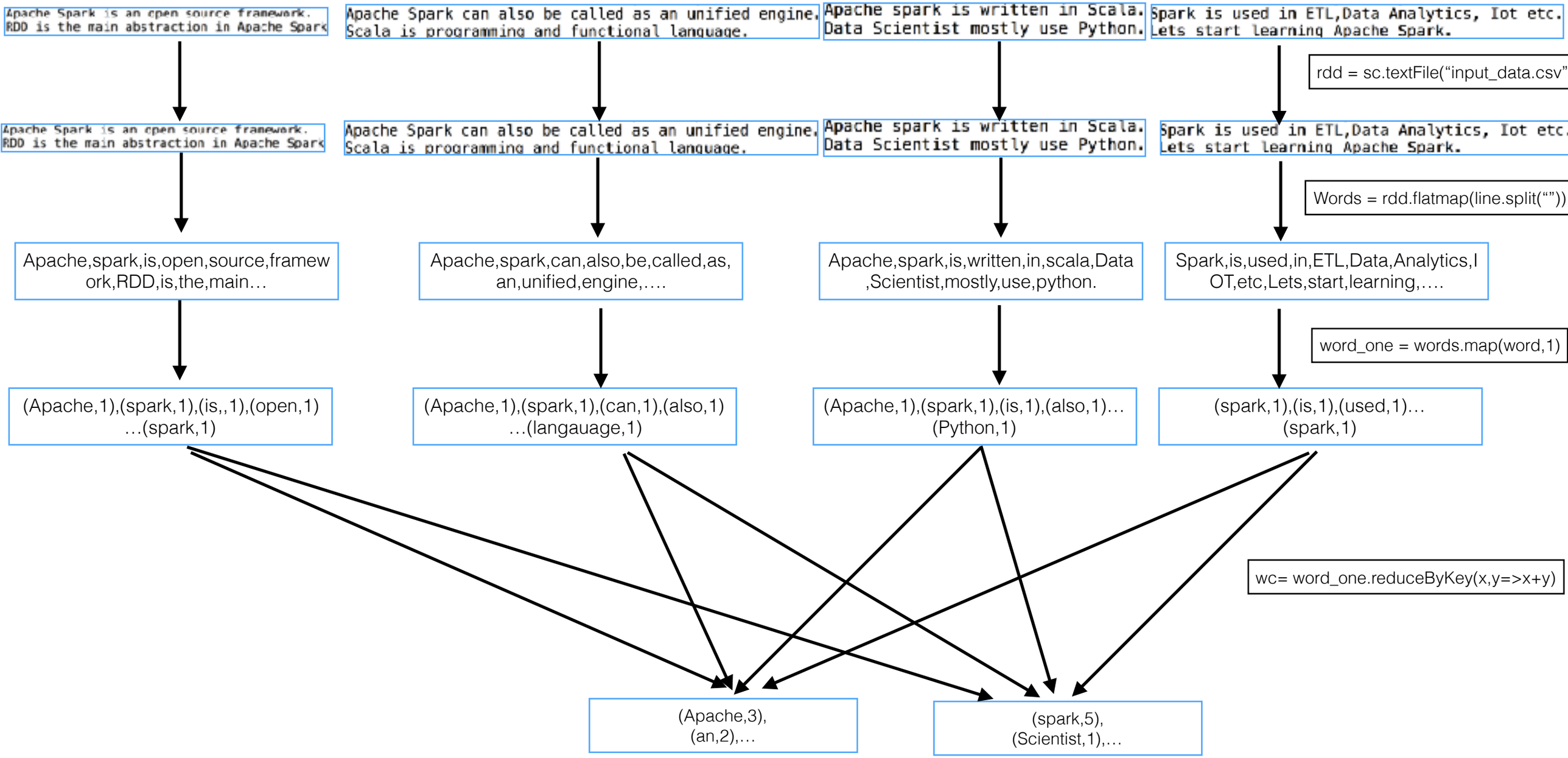
Definition Cont..

- Typed : RDD records have types, e.g. Long in RDD[Long] or (Int, String) in RDD[(Int, String)].

- Partitioned — records are partitioned (split into logical partitions) and distributed across nodes in a cluster.

- RDDs built and manipulated through a diverse set of parallel transformations (map, filter, join) and actions (count, collect, save)

Word Count Flow



Thank You.