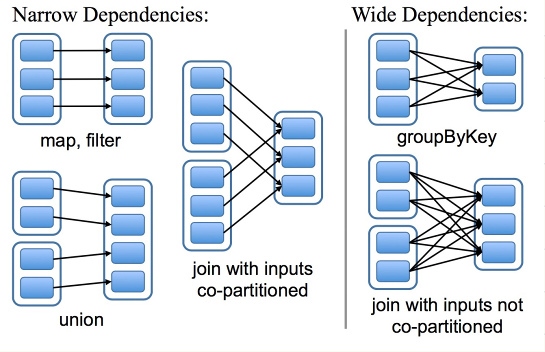
SPARK


Narrow Dependencies:
Wide Dependencies:
map, filter
union
join with inputs co-partitioned
groupByKey
join with inputs not co-partitioned

RDD:

- list of partitions
- function computing split
- list of dependencies
- partitioner (optional for key-value)
- list of preferred locations (optional)

Partitioning

- Default partitioning - split in equally sized partitions

Pair RDDs only:

- Range partitioning - split according to natural order of keys
- Hash partitioning - split according to key hash

Dependencies:

- Narrow - each partition of source used by at most 1 target
- Wide - multiple partitions in target depend on single in source

Persistence

data stored as:

- Java objects
- serialised data
- file system

storage levels:

- MEMORY_ONLY - deserialised Java objects in JVM
- MEMORY_AND_DISK - deserialised Java objects in JVM
- MEMORY_ONLY_SER - serialised Java objects
- MEMORY_AND_DISK_SER - serialised Java objects
- DISK_ONLY - RDD partitions only on disk

if memory only and doesn't fit, compute on fly instead of write on disk

SPARK Architecture

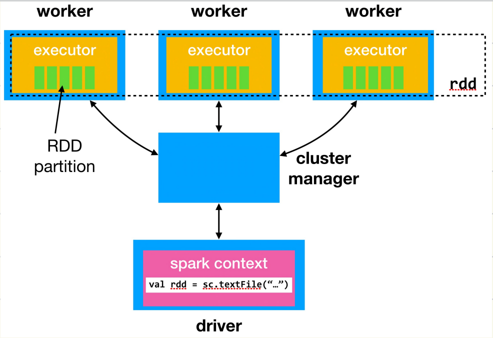
Executor - actual processing

Worker - can contain multiple executors

Driver - accepts user programs;

returns processing results

Cluster manager - resource allocation



job - action requested; graph worked back

stage - job with wide dependencies

task - minimum unit of execution