13/02/24 **Lecture 4: Apache Spark and RDD**

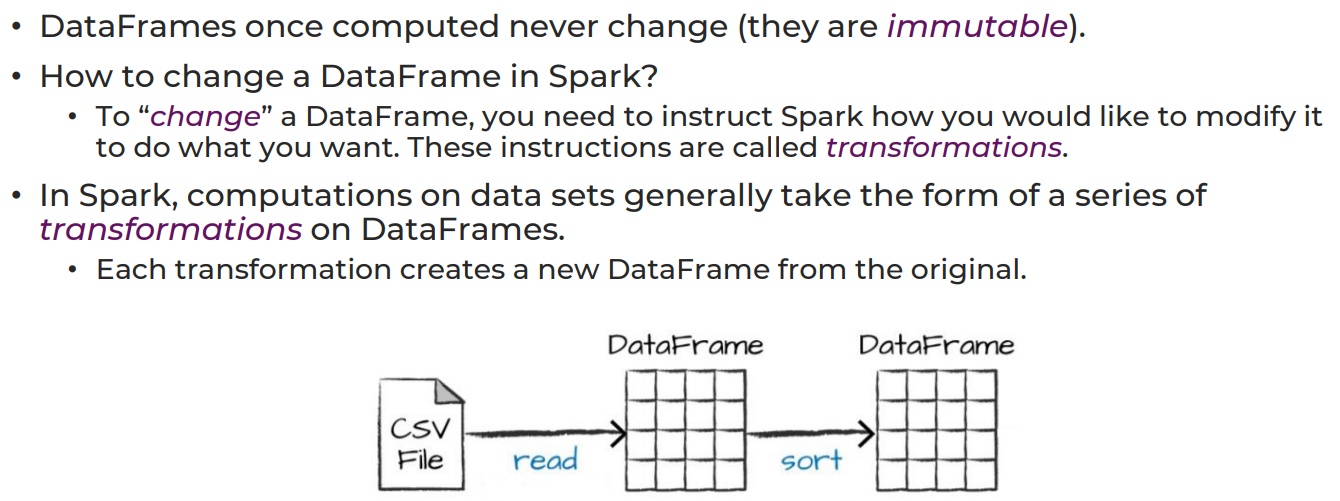
Apache Spark

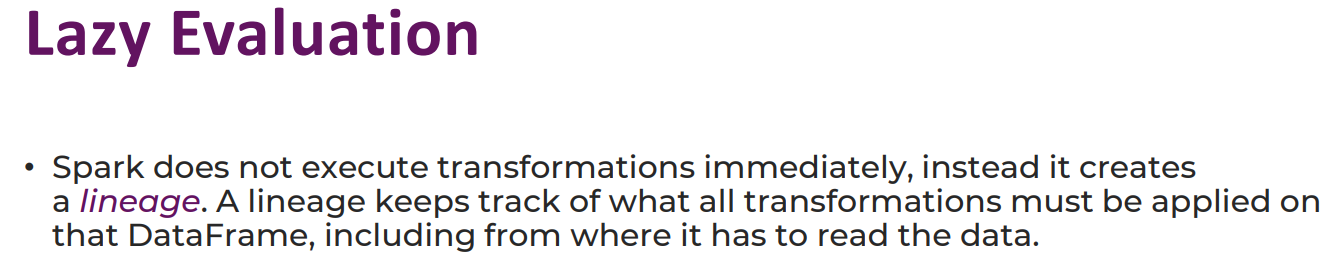
* Unify computing engine for parallel processing with sets of library
* Supports many languages (Python, Scala, Java, etc)
* In memory computation

Resilient Distributed Dataset (RDD)

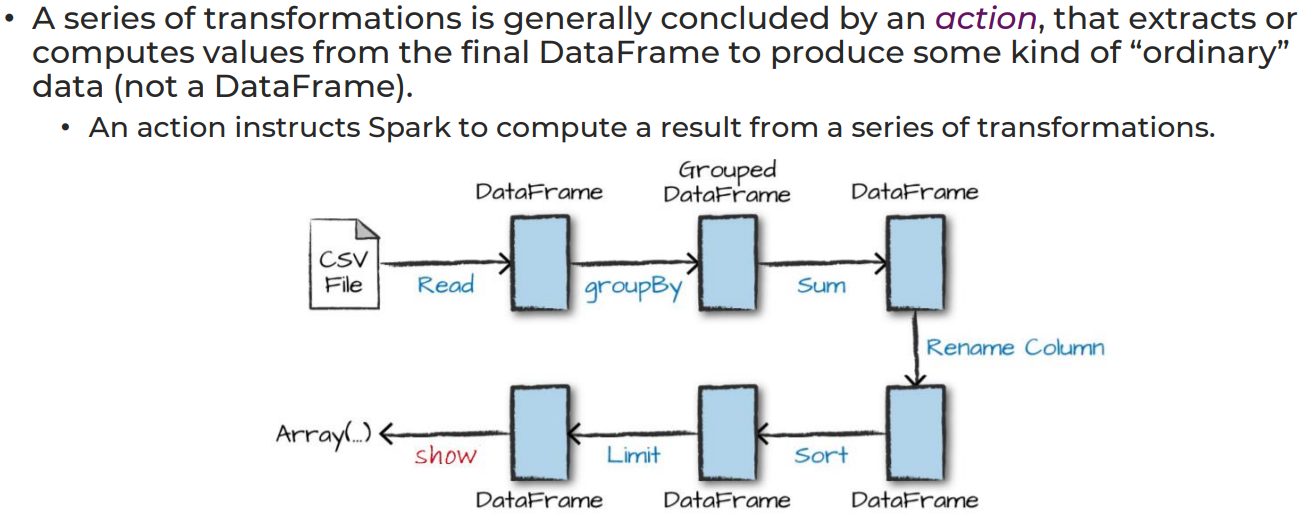
* Read-only distributed collection of elements of data, partitioned across nodes in a cluster (Non-mutable)
* In-memory computations
* Fault tolerance similar to MapReduce()
* Utilisation of Dataframe

Transformation





Action



Spark vs Hadoop - Spark (RAM), Hadoop (Disk) - Spark is better for iterative