**A Few Terms**

* Resilient Distributed Dataset (RDD) - a distributed collection of items
* DataFrame, DataSet - abstractions on top of RDD
* Operations:
  + actions (required to emit values)
  + transformations (return pointers to new RDD)
* RDD transformations list ([link](https://spark.apache.org/docs/latest/rdd-programming-guide.html#transformations))
* Application - custome driver program + executors
* SparkSession - access to Spark (interactive or created in an app)
* Job - parallel computation with tasks spawned by Spark action (e.g., save(), collect())
* Stage - set of tasks
* Task - unit of work sent to Spark executor
* Cluster manager (built-in, YARN, Mesos, or Kubernetes)
* Spark executor (usually one per worker node)
* Deployment modes (local, standalone, YARN, etc.)