What is Spark?

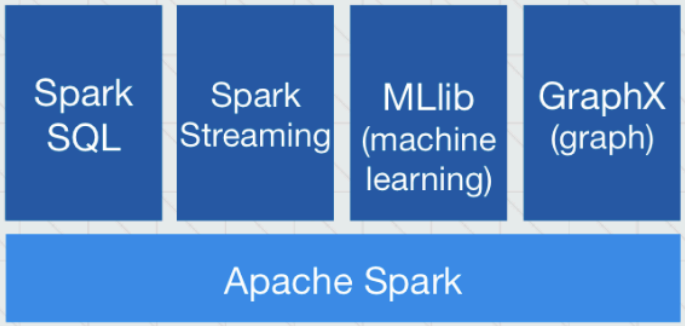

Image courtesy of https://spark.apache.org/

- Way to distribute work across a cluster
- Abstraction on top of HDFS, JSON, SQL, Parquet...
- Framework for data preparation and Machine Learning
- SIMD implementation that supports Map-Reduce and more

Spark Execution

- Driver defines the data preparation and processing
- Spark forms data partitions and serializes the code
- Distribution of data and code implements SIMD
- Serialization has implications on performance

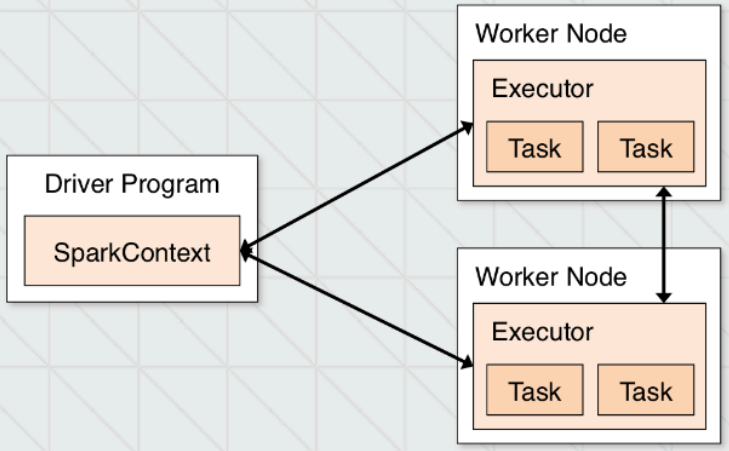

Image courtesy of *Learning Spark – Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia*

RDD: Define data and processing

What is it?

Resilient Distributed Dataset (RDD), the basic abstraction in Spark. Represents an immutable, partitioned collection of elements that can be operated on in parallel.

- Abstract class with subclasses: HadoopRDD, SchemaRDD, JdbcRDD...
- Resilient means that if a partition of the RDD (the data the Worker sees) drops or is evicted by caching, the system can recompute it easily to complete the work
- Distributed via a default or custom partitioning scheme

Laziness and Lineages

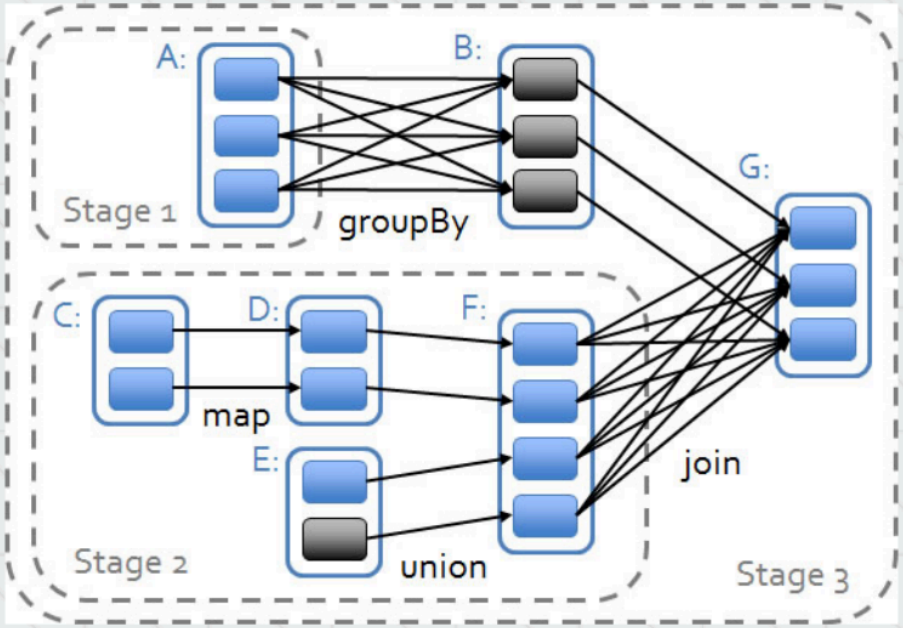

Image courtesy of http://www.cs.duke.edu

- Spark Context used to load data
- Spark SQL Context gets access to DBMS sources
- RDD contents are defined by:
  - Code using collections
  - Loading data via Contexts
- RDD inherits a series of Transformations on loaded data called a **lineage**
- Lineage is computed on demand when an action is performed
- Most transformations are performed as iterative composed functions - a row at a time
- Persisting an RDD will cache it and prevent recomputation

Persistence & Serialization

- Persistence of RDD is related to caching
  - Memory, Disk or Spillover
  - Serialized or not
  - Some experimental features using Tachyon for Off Heap memory (instead of files)
  - Removes costs of subsequent Actions recomputing lineage
- Serialization concerns
  - variables in closure scopes are serialized - deref
  - instance methods of objects drag entire object into serialization - use functions vs. methods
  - Output size and processing performance

What Are Your Next Steps?

- https://spark.apache.org/docs/latest/quick-start.html
- http://www.cs.berkeley.edu/~matei/papers/2012/nsdi_spark.pdf
- http://ampcamp.berkeley.edu/wp-content/uploads/2013/02/Machine-Learning-on-Spark-Shivaram-Venkataraman-Strata-2013.pptx
- http://www.safaribooksonline.com/library/view/learning-spark/9781449359034/