DSA 8430
Parallel Computing for Data Analytics

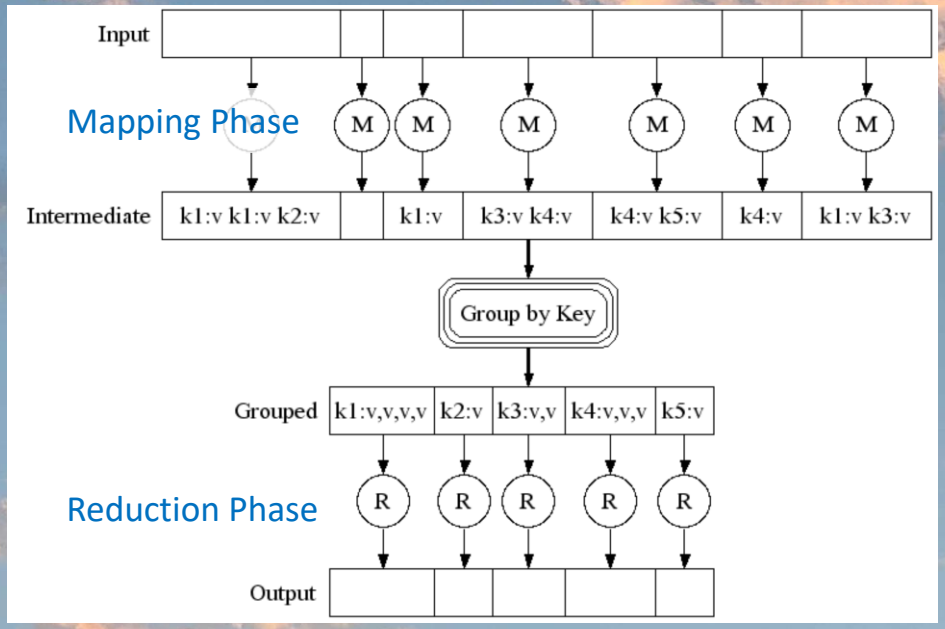
Advanced Distributed Processing

With Spark and Dataproc

Module Topics

- Map-Reduce
- Hadoop
- Spark
- GCP Dataproc

Divide & Conquer: Map - Reduce



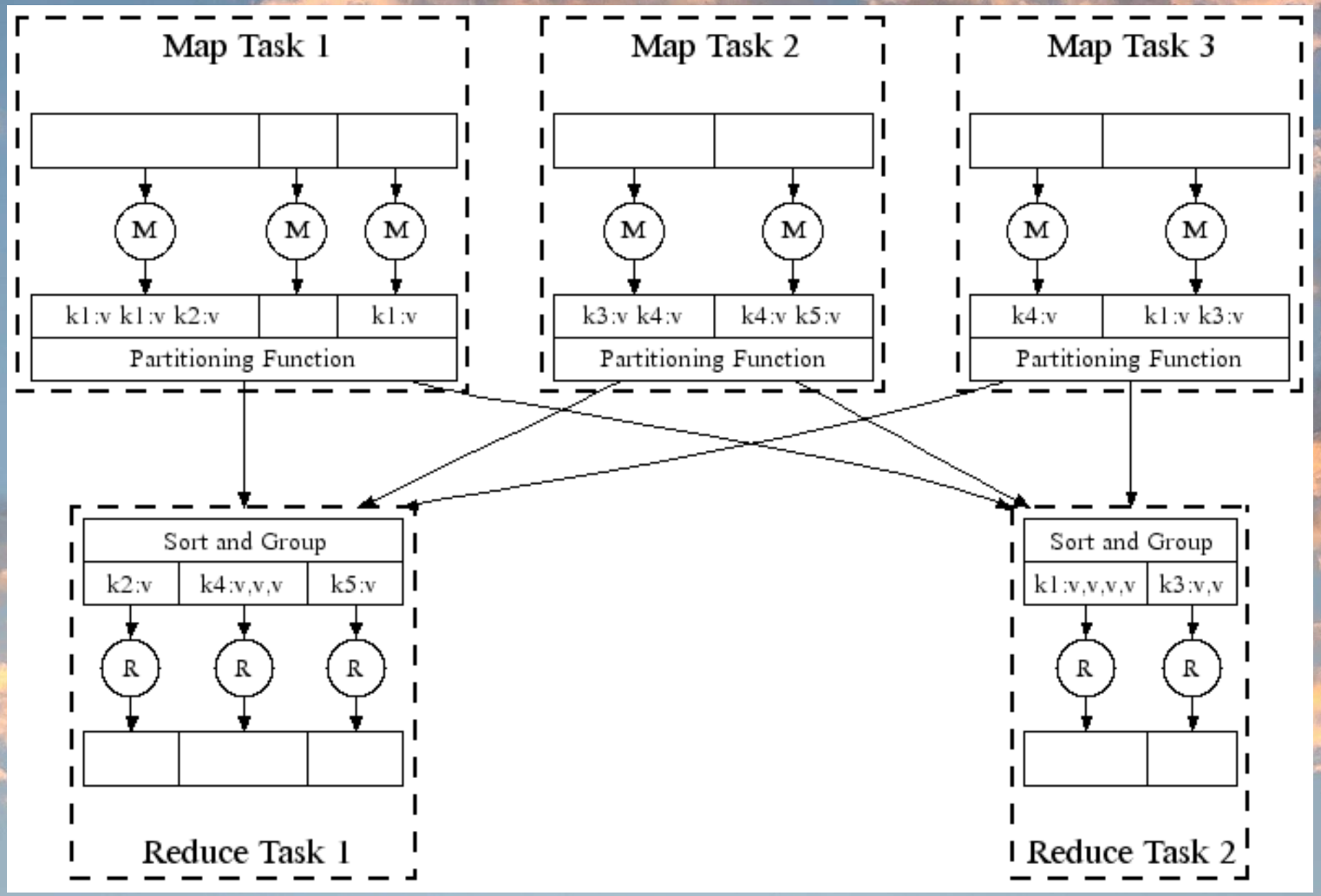
https://research.google.com/archive/mapreduce-osdi04-slides/index-auto-0007.html

- 2004 Google paper

https://research.google/pubs/pub62/

MapReduce: Simplified Data Processing on Large Clusters

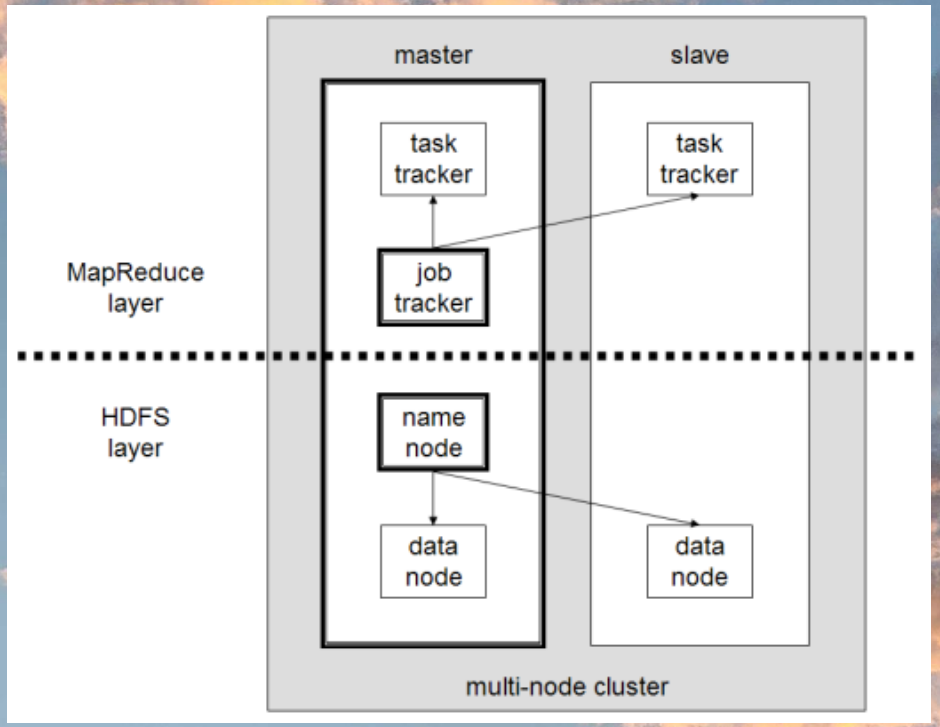
Divide & Conquer: Map - Reduce



Hadoop – Map Reduce for All

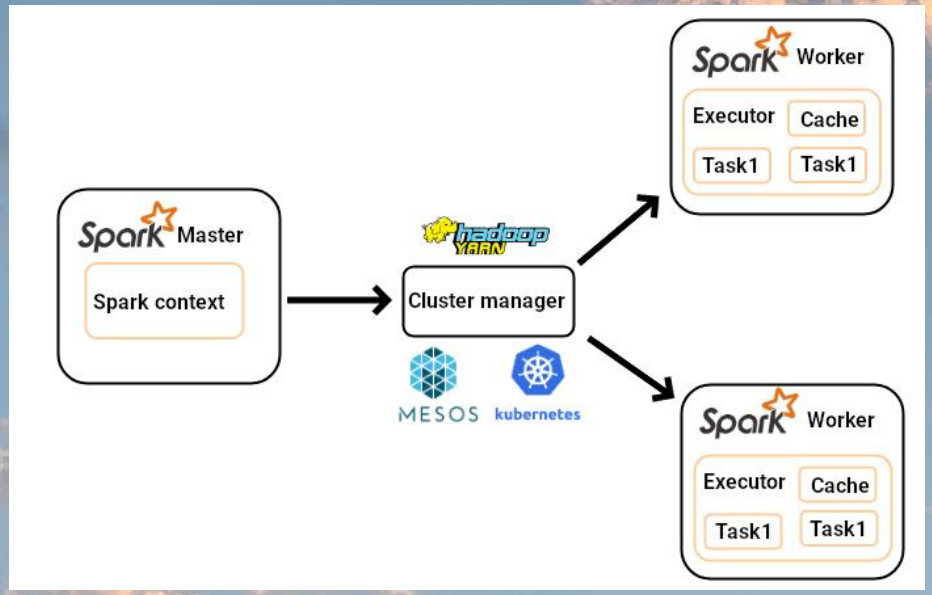
Software Ecosystem

- Hadoop Commons
- HDFS
- Yarn
- MapReduce



https://en.wikipedia.org/wiki/Apache_Hadoop#/media/File:Hadoop_1.png

Apache Spark – Hadoop Evolved



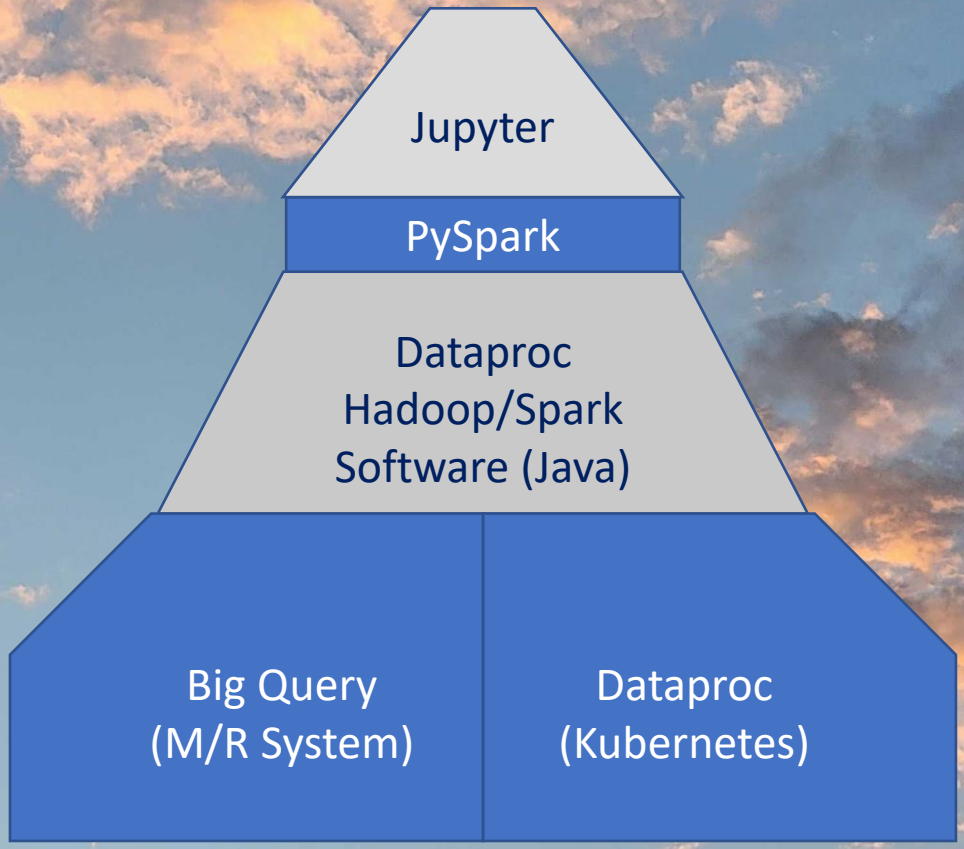
https://medium.com/southworks/movie-data-statistics-with-apache-spark-58c2ef8fe452

Evolution of distributed computing ideas from Hadoop

- In memory computing
- Data instantiated into RDD
  - resilient distributed dataset
  - ReadOnly Distributed RAM Objects

GCP Dataproc

- Google Cloud-based
- Kubernetes underneath
- Ties into data sources from all of GCP
- Supports Hadoop and Spark workloads
- Allows use of libraries built upon Spark (in-memory distributed computing)



Pay Attention to Detail!

Start early, work methodically!

Take notes as you work through things!