YARN Applications

School of Information Studies
Syracuse University

YARN Applications

- MapReduce is great, but there's a need for high-level scripting.
- There are also other needs beyond batch capabilities of MapReduce.

Pig

- Platform for analyzing large data sets, performing ETL, data cleanup, etc.
- Write code simpler for MapReduce in “piglatin” instead of Java.

Steps:

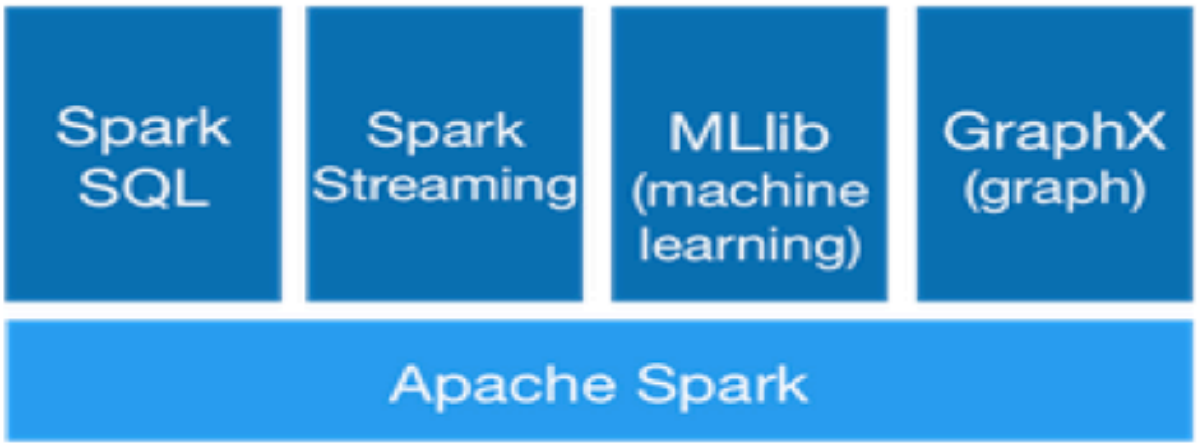
- LOAD
- TRANSFORM
- STORE /DUMP

Hive

- SQL-like syntax over HDFS
- Declarative, not procedural like Pig
- Useful for ad hoc query of HDFS data
- Dimensional models

Spark

- Spark is a cluster computing framework
- Extends the M-R metaphor in memory for large-scale data processing.
- Data Mining and ML in Hadoop



Zeppelin

- Web-based notebook for interactive data analytics
- Interactive and collaborative
- Interpreters for Hive, Pig, Spark, Python, R, and more