CS 378 – Big Data Programming

Lecture 19

Spark Introduction

Assignment 8 – Job Chaining

- Solution review

- Assignment 8 – Directory structure
- After session binning
  - clicker-m-0000x.avro
  - part-m-0000x.avro
  - shower-m-0000x.avro
  - submitter-m-0000x.avro
  - visitor-m-0000x.avro

Review

- After four parallel jobs (collect event subtype stats)
  - clicker-m-0000x.avro
  - **clickerStats /** part-r-00000.avro
  - part-m-0000x.avro
  - shower-m-0000x.avro
  - **showerStats /** part-r-00000.avro
  - submitter-m-0000x.avro
  - **submitterStats /** part-r-00000.avro
  - visitor-m-0000x.avro
  - **visitorStats /** part-r-00000.avro

Review

- After aggregation job (aggregate click subtype stats)
  - **aggregateStats /** part-r-00000.avro
  - clicker-m-0000x.avro
  - **clickerStats /** part-r-00000.avro
  - part-m-0000x.avro
  - shower-m-0000x.avro
  - **showerStats /** part-r-00000.avro
  - submitter-m-0000x.avro
  - **submitterStats /** part-r-00000.avro
  - visitor-m-0000x.avro
  - **visitorStats /** part-r-00000.avro

Issues with MapReduce

- One “template”: map, then reduce
- HDFS is its own file system

- In a data pipeline, each map-reduce step
  - Reads all input data from disk
  - Writes all output data to disk
  - Even if output is just an intermediate result
  - Can use `ChainMapper`, `ChainReducer`

- Addresses failure handling with replicated data
  - Can help performance though

Apache Spark

- Open source project out of AMPLab at UC Berkeley

- A Spark program defines:
  - Transformations and actions on data sets
  - Data flow, or lineage graph among data sets, induced by the transformations

- Data sets in Spark are called RDDs
  - Resilient Distributed Datasets

Spark Features

- Provide domain specific libraries
  - Example: map-reduce library
  - Promotes functional programming model
- Access to multiple data (file) systems
  - Local, HDFS, Cassandra, S3, database tables, …
- Lazy evaluation, and caching for performance
  - Reduce or eliminate disk I/O
- Support multi-stage and iterative apps

Spark RDDs

- Resilient Distributed Dataset
  - One RDD has one or more partition
  - Partitions are distributed across machines
  - Rebuilt from base data on failure (versus replication)
  - Lazy evaluation – created/computed on demand

- RDD types offer various functions
  - map, reduce
  - groupBy, reduceByKey
  - joins (inner, leftOuterJoin, rightOuterJoin)
  - filter, sample

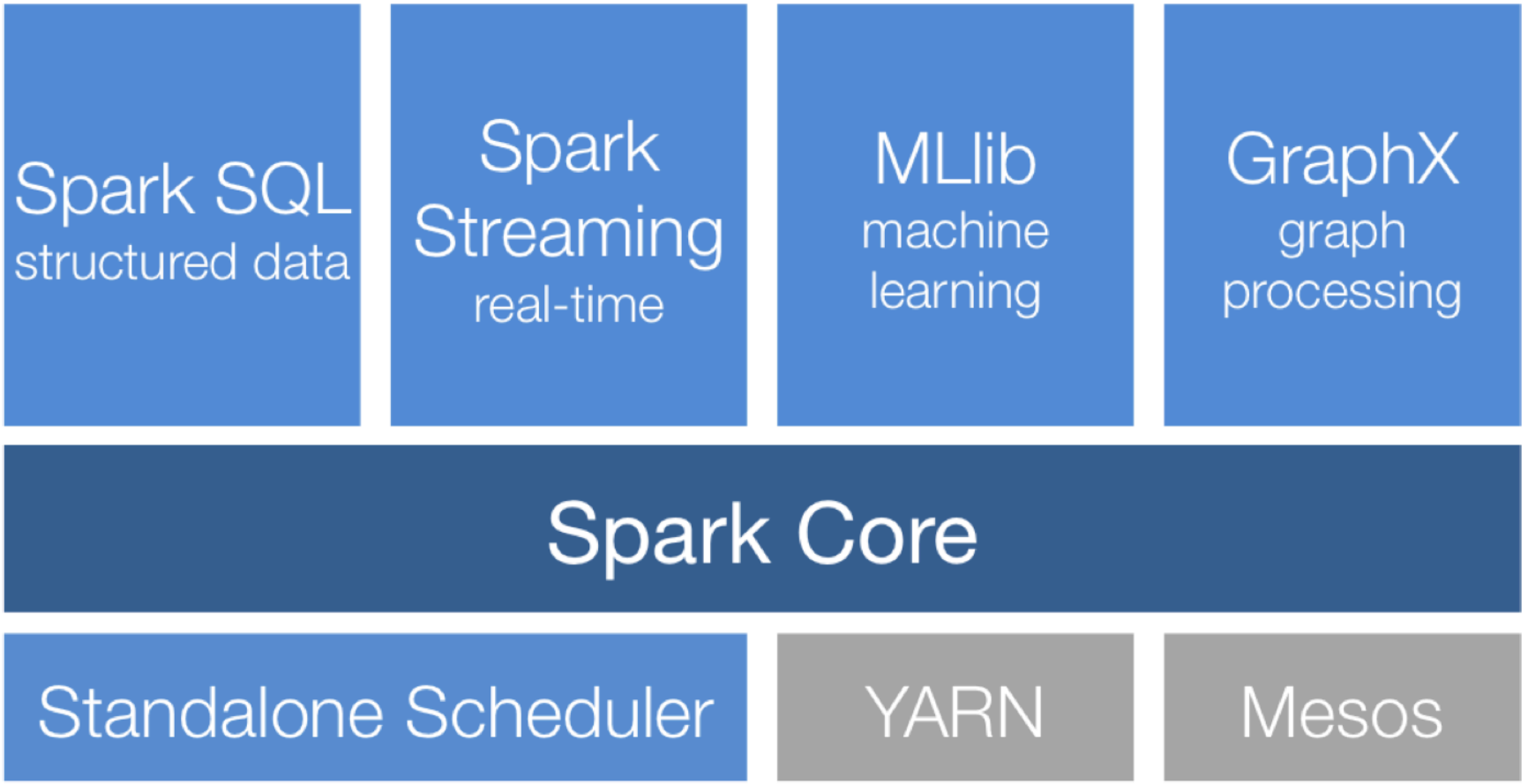
Spark

- Provides a higher level of abstraction for coding
  - Multi-stage map-reduce pipeline in Hadoop …
  - Can be composed functions in Spark

- RDD support and libraries
  - Spark SQL – RDD representing relational table
  - Streaming data – D-Stream, Twitter stream
  - Graph data – GraphX
  - …

Spark Stack

Learning Spark, Figure 1-1.



Spark Programs

- A Spark program defines:
  - RDDs
    - Input from external sources
    - Produced by a transformation
  - Transformations
    - Produce a new RDD from the input RDD(s)
  - Actions
    - Compute something from the input RDD
      - Return non-RDD objects (e.g., number)
    - Write an RDD to external storage

Spark Example

- Interactive
  - Scala
  - Python
- Batch
  - Java

Assignment 9

- Download Spark
- Compile Java WordCount for Spark
- Run Java WordCount using Spark