Apache Spark Through Email

Markus Dale

Nov 2018

Slides And Code

- Slides: https://github.com/medale/spark-mail/blob/master/presentation/ApacheSparkThroughEmail.pdf
- Spark Code Examples: https://github.com/medale/spark-mail/

Data Science for Small Dataset

- Laptop
- Explore subset, develop approaches, find features

Data Science for Larger Dataset

- Standalone server - more memory, faster CPU, more storage

Data Science for Larger Dataset (Vertical Scaling)

- Big iron - lots of cores, memory, disk/SSDs, GPUs

Data Science for Large Datasets (Horizontal Scaling)

- Parallelize, coordinate compute among many "commodity" machines
- Deal with failure

Big Data Framework - Apache Hadoop

- Google GFS (2003), Google MapReduce (2004)
- Hadoop (Nutch - open source web crawler/Lucene) - Doug Cutting, Mike Cafarella
  - Yahoo, Cloudera, Hortonworks, MapR

Hadoop Ecosystem

- HDFS, YARN, MapReduce (Spark replaces MR)
- HBase (Google BigTable), Cassandra, Accumulo
- Pig, Hive - MR scripting DSL/SQL

Apache Spark Components

- Foundation: Resilient Distributed Datasets (RDD)
  - Broadcast variables, accumulators
  - Java objects, should use Kryo serialization
- Structured APIs - Datasets, DataFrames, SQL
  - Spark manages object layout in memory, schemas, code generation
- Streaming, MLlib (Advanced analytics)
- Scala, Java, Python, R + library ecosystems
- Submit (Batch/Stream) or Shell/Notebooks (e.g. Zeppelin, Jupyter)