Apache Spark Through Email

Markus Dale

Nov 2018

Intro, Slides And Code

- Slides: https://github.com/medale/spark-mail/blob/master/presentation/ApacheSparkThroughEmail.pdf
- Spark Code Examples: https://github.com/medale/spark-mail/
  - README.md describes how to get and parse Enron email dataset

Data Science for Small Dataset

Figure 1: Laptop

Data Science for Larger Dataset

Figure 2: Standalone Server

Data Science for Larger Dataset (Vertical Scaling)

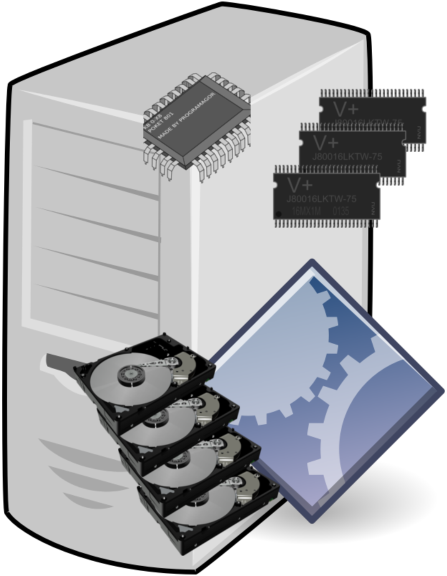


Figure 3: Beefed-up Server

Data Science for Large Datasets (Horizontal Scaling)

Figure 4: Multiple cooperating Servers

Big Data Framework - Apache Hadoop

Figure 5: HDFS, MapReduce

Hadoop Ecosystem

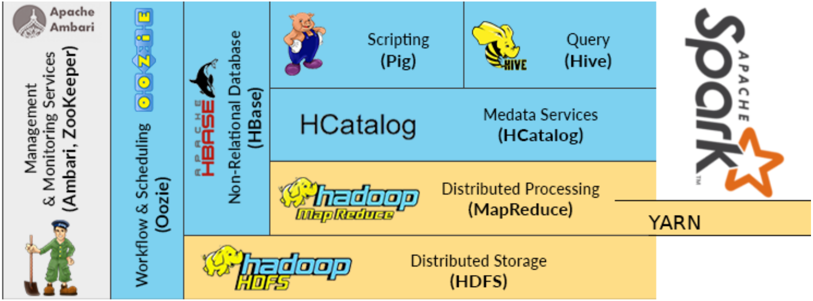
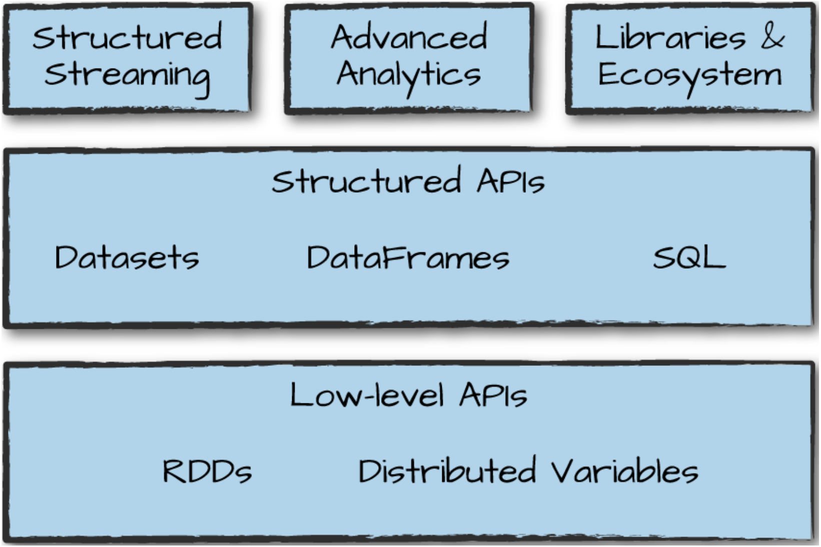

Apache Ambari
Management & Monitoring Services (Ambari, ZooKeeper)
Oozie
Workflow & Scheduling (Oozie)
Apache HBase
Non-Relational Database (HBase)
Scripting (Pig)
HIVE
Query (Hive)
HCatalog
Medata Services (HCatalog)
hadoop MapReduce
Distributed Processing (MapReduce)
hadoop HDFS
Distributed Storage (HDFS)
Apache Spark
YARN

Figure 6: Some Frameworks Around Hadoop

Apache Spark Components

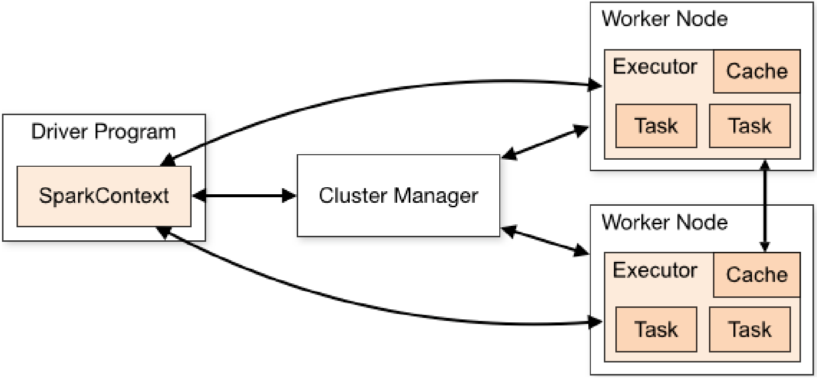


Source: Spark: The Definitive Guide

Hello, Spark Email World!

- Jupyter Notebook with Apache Toree
- See Notebook ../notebooks/html/ApacheSparkThroughEmail1.html

Cluster Manager, Driver, Executors, Tasks



Source: Apache Spark website

SparkSession: Entry to cluster

- spark: spark.sql.SparkSession

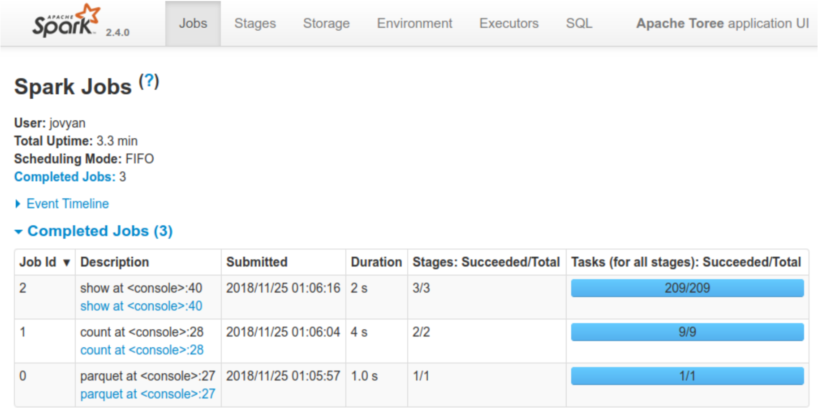
```
//SparkSession provided by notebook as spark
val records = spark.read.
   parquet("/datasets/enron/enron-small.parquet")

//In regular code for spark-submit
//com.uebercomputing.spark.dataset.TopNEmailMessageSenders
val spark = SparkSession.builder().
   appName("TopNEmailMessageSenders").
   master("local[2]").getOrCreate()
```

DataFrameReader: Input for structured data

- `spark.read`: spark.sql.DataFrameReader
  - jdbc
  - json
  - parquet
  - text...
  - Also: https://spark-packages.org - Avro, Redshift, MongoDB...

Scaling Behind the Scenes



Spark Jobs (?)

User: jovyan
Total Uptime: 3.3 min
Scheduling Mode: FIFO
Completed Jobs: 3

▸ Event Timeline

▾ **Completed Jobs (3)**

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	show at <console>:40 show at <console>:40	2018/11/25 01:06:16	2 s	3/3	209/209
1	count at <console>:28 count at <console>:28	2018/11/25 01:06:04	4 s	2/2	9/9
0	parquet at <console>:27 parquet at <console>:27	2018/11/25 01:05:57	1.0 s	1/1	1/1

Figure 7: Jobs and Tasks

Stages: Pipeline work per stage - shuffle

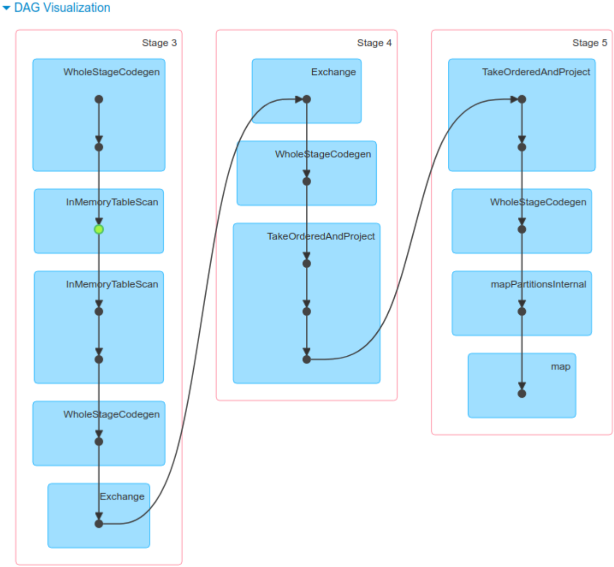


Figure 8: Stages

Where clause, Column methods, Built-in functions

- See Notebook
  ../notebooks/html/ApacheSparkThroughEmail2.html

Parallelism and Partitioning

- Goldilocks - not too many, not too few
- Initial parallelism - number of input "blocks"
- Shuffle - `spark.sql.shuffle.partitions` configuration

Explode, Shuffle Partitions, UDF, Parquet partition

- See Notebook ../notebooks/html/ApacheSparkThroughEmail3.html

And now for something completely different: Colon Cancer

- Screening saves lives!
  - Colonoscopy - talk to your doc
- Colorectal Cancer Alliance

Questions?

Figure 9: medale@asymmetrik.com

- Baltimore Scala Meetup https://www.meetup.com/Baltimore-Scala/
- Spark Mail repo https://github.com/medale/spark-mail/