

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)

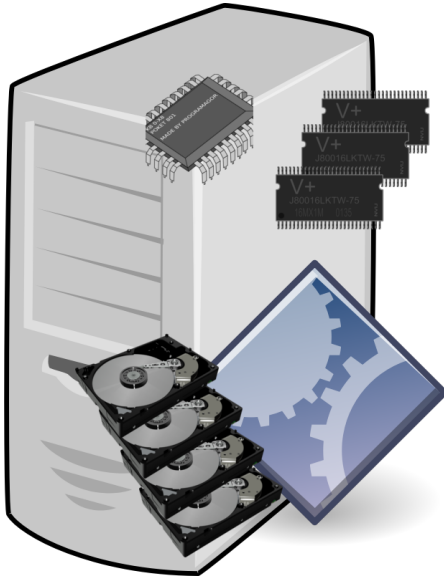


Figure 3: Beefed-up Server

Data Science for Large Datasets (Horizontal Scaling)



Figure 4: Multiple cooperating Servers

Big Data Framework - Apache Hadoop



Figure 5: HDFS, MapReduce

Hadoop Ecosystem

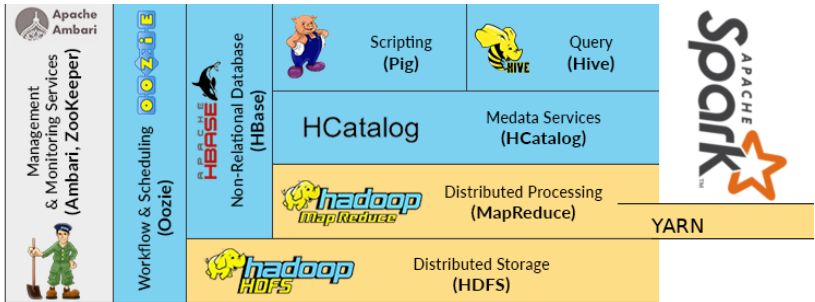


Figure 6: Some Frameworks Around Hadoop

Apache Spark Components

Structured
Streaming

Advanced
Analytics

Libraries &
Ecosystem

Structured APIs

Datasets

DataFrames

SQL

Low-level APIs

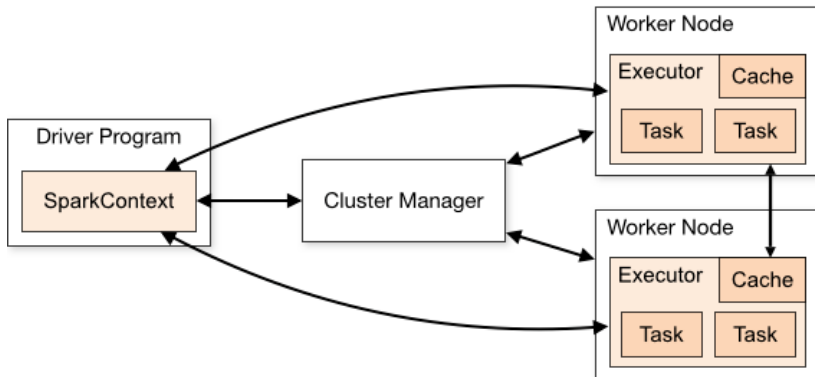
RDDs

Distributed Variables

Hello, Spark Email World!

- ▶ Jupyter Notebook with Apache Toree
- ▶ See `ApacheSparkThroughEmail1`

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

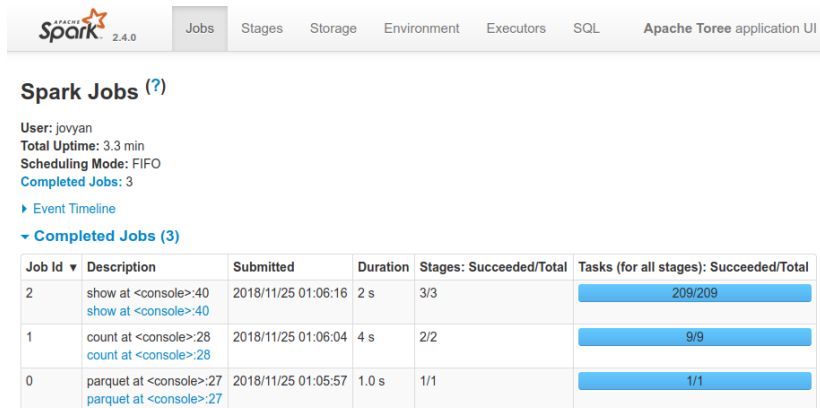


Figure 7: Jobs and Tasks

Stages: Pipeline work per stage - shuffle

▼ DAG Visualization

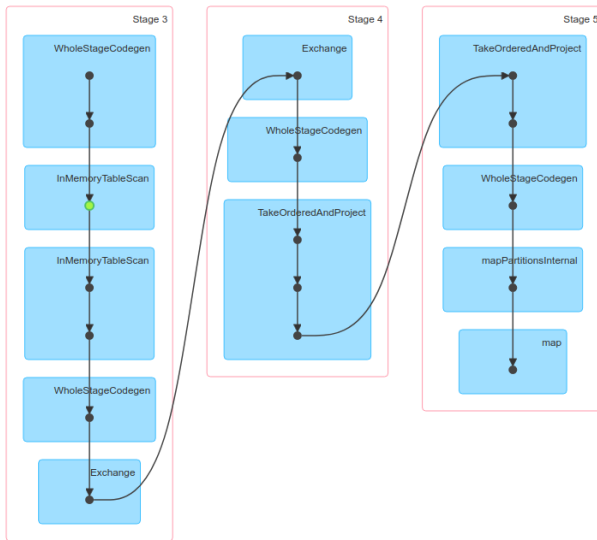


Figure 8: Stages

Where clause, Column methods, Built-in functions

- ▶ See Apache Spark Through Email Notebook 2

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 Apache Spark Through Email Notebook 3

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
- ▶ Spark Mail repo <https://github.com/medale/spark-mail/>