

Apache Spark Through Email

Markus Dale, medale@asymmetrik.com

Nov 2018

- 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)

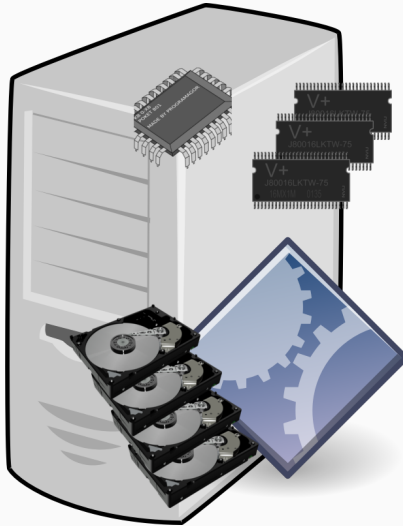


Figure 3: Beefed-up Server

Data Science for Large Datasets (Horizontal Scaling)





Figure 5: HDFS, MapReduce

Hadoop Ecosystem

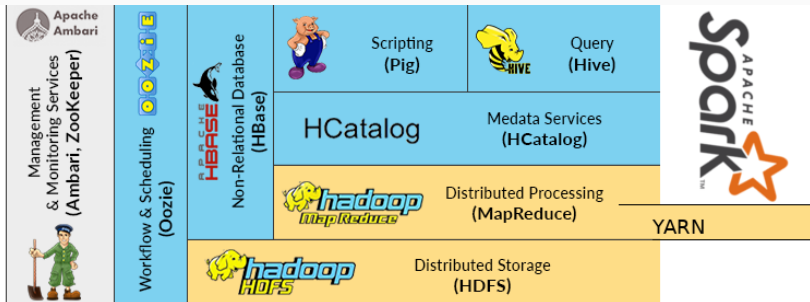


Figure 6: Some Frameworks Around Hadoop

Apache Spark Components

Structured
Streaming

Advanced
Analytics

Libraries &
Ecosystem

Structured APIs

Datasets

DataFrames

SQL

Low-level APIs

RDDs

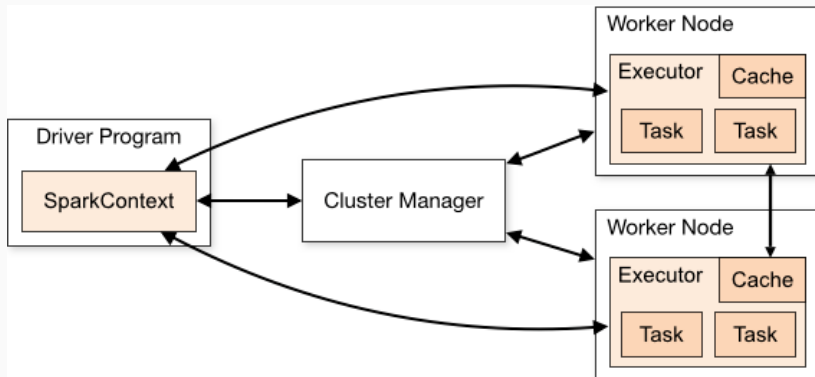
Distributed Variables

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

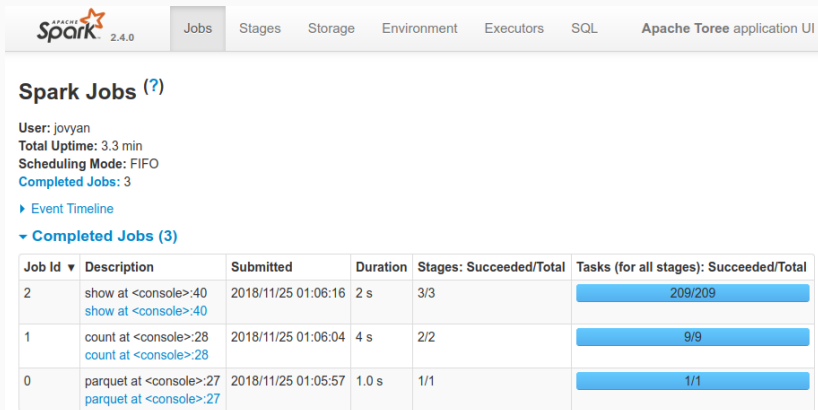


Figure 7: Jobs and Tasks

Stages: Pipeline work per stage - shuffle

▼ DAG Visualization

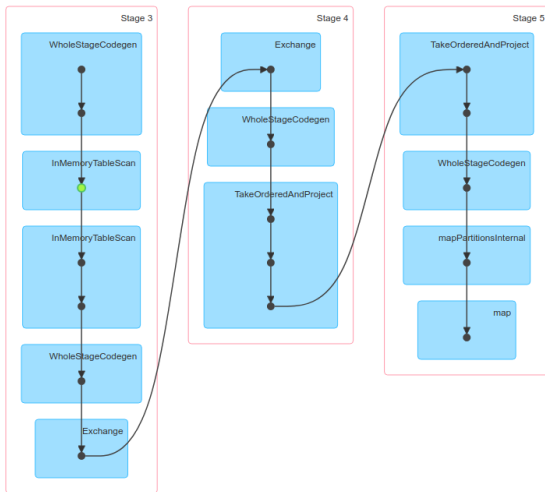


Figure 8: Stages

- 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

- 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?



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