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/

Data Science for Small Dataset

Figure 1: Laptop

Data Science for Larger Dataset

Figure 2: Standalone Server

Data Science for Larger Dataset (Vertical Scaling)

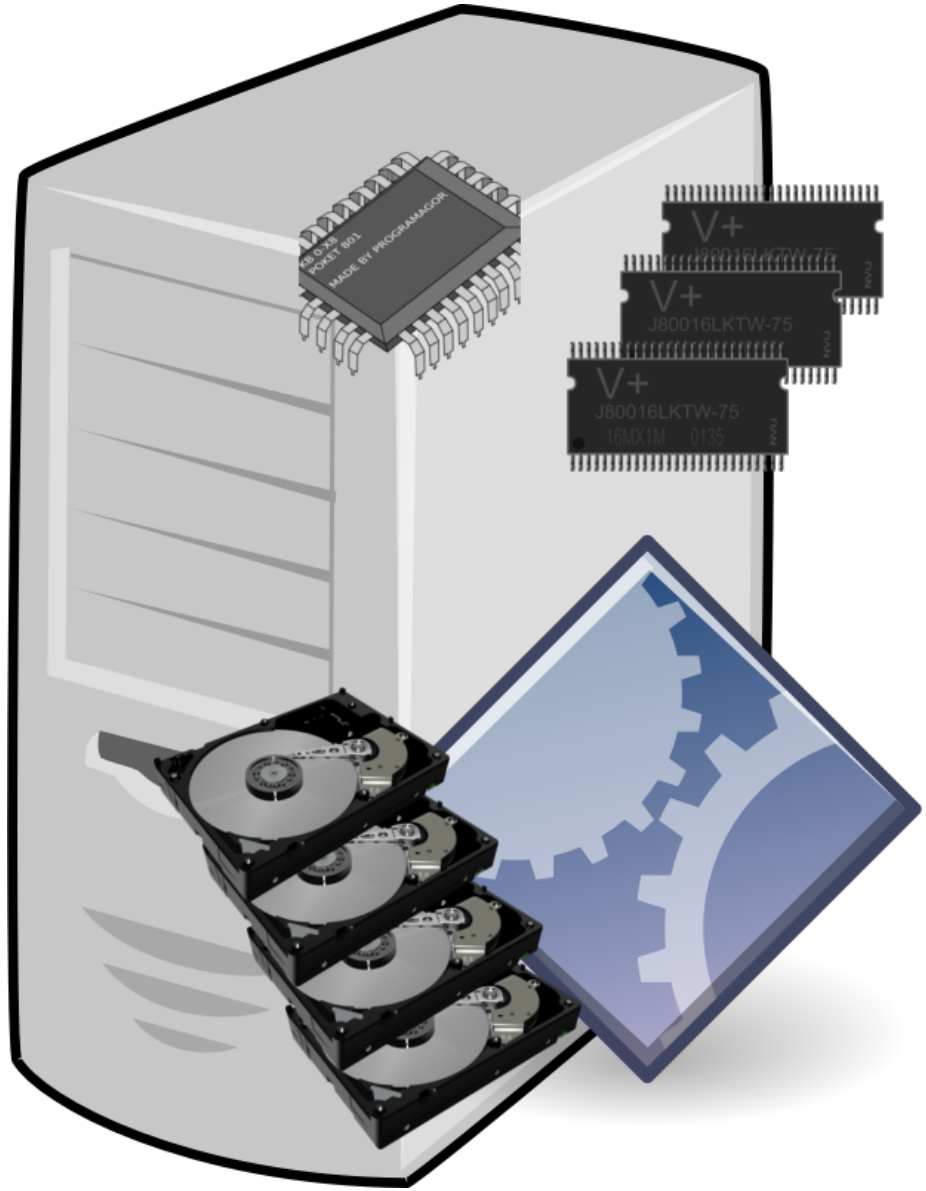


Figure 3: Beefed-up Server

Data Science for Large Datasets (Horizontal Scaling)

Figure 4: Multiple cooperating Servers

Big Data Framework - Apache Hadoop

Figure 5: HDFS, MapReduce

Hadoop Ecosystem

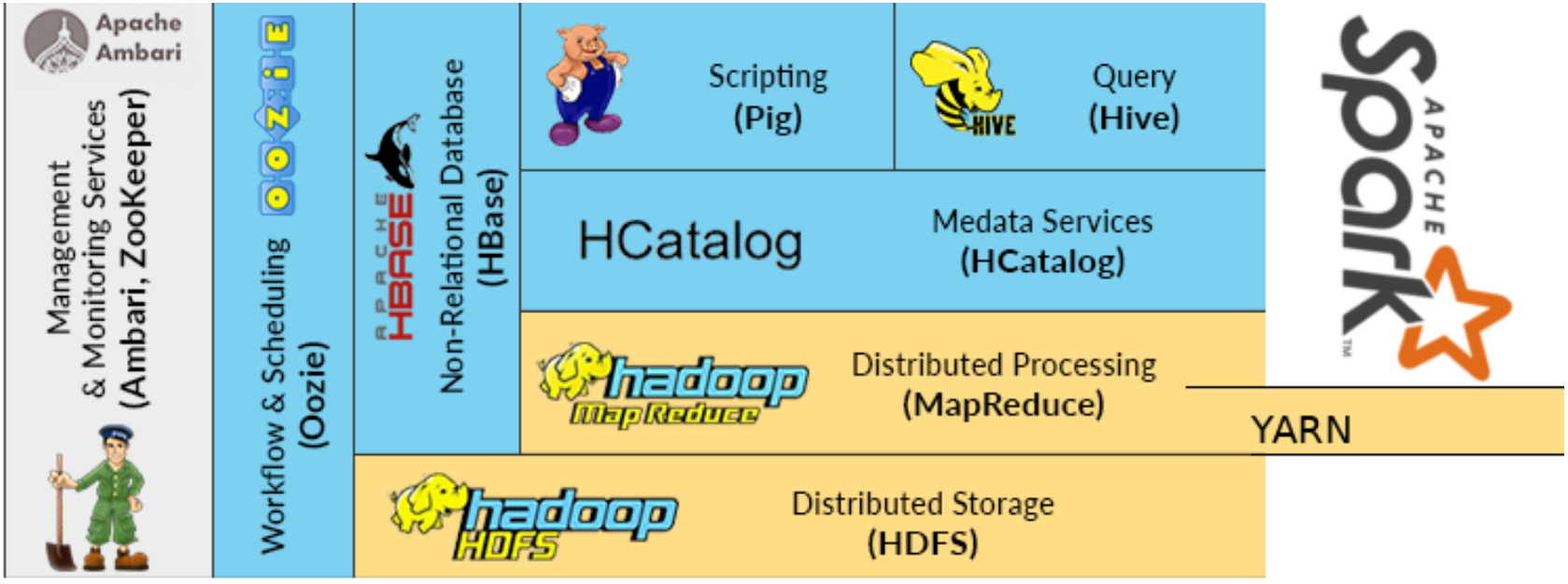
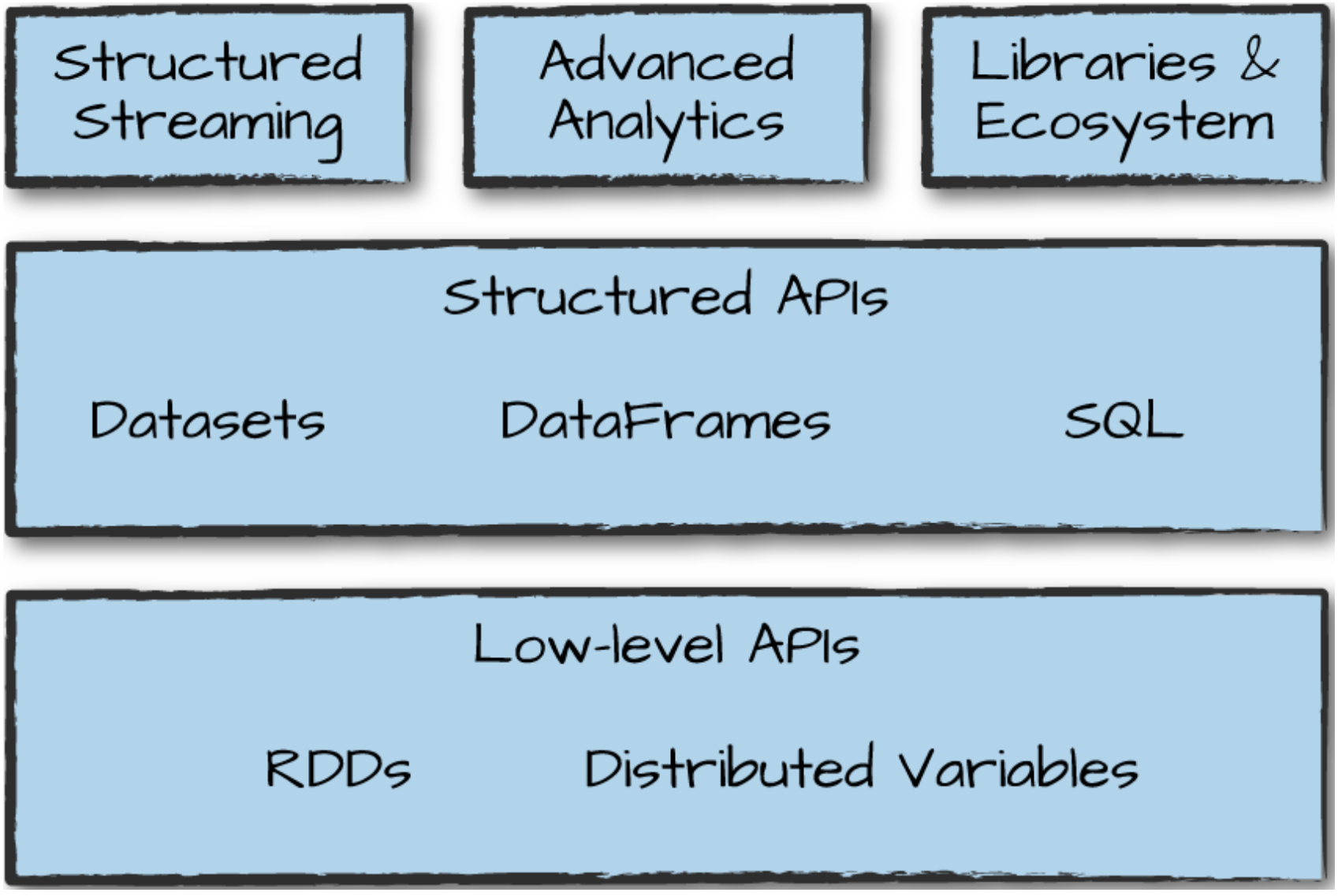


Figure 6: Some Frameworks Around Hadoop

Apache Spark Components

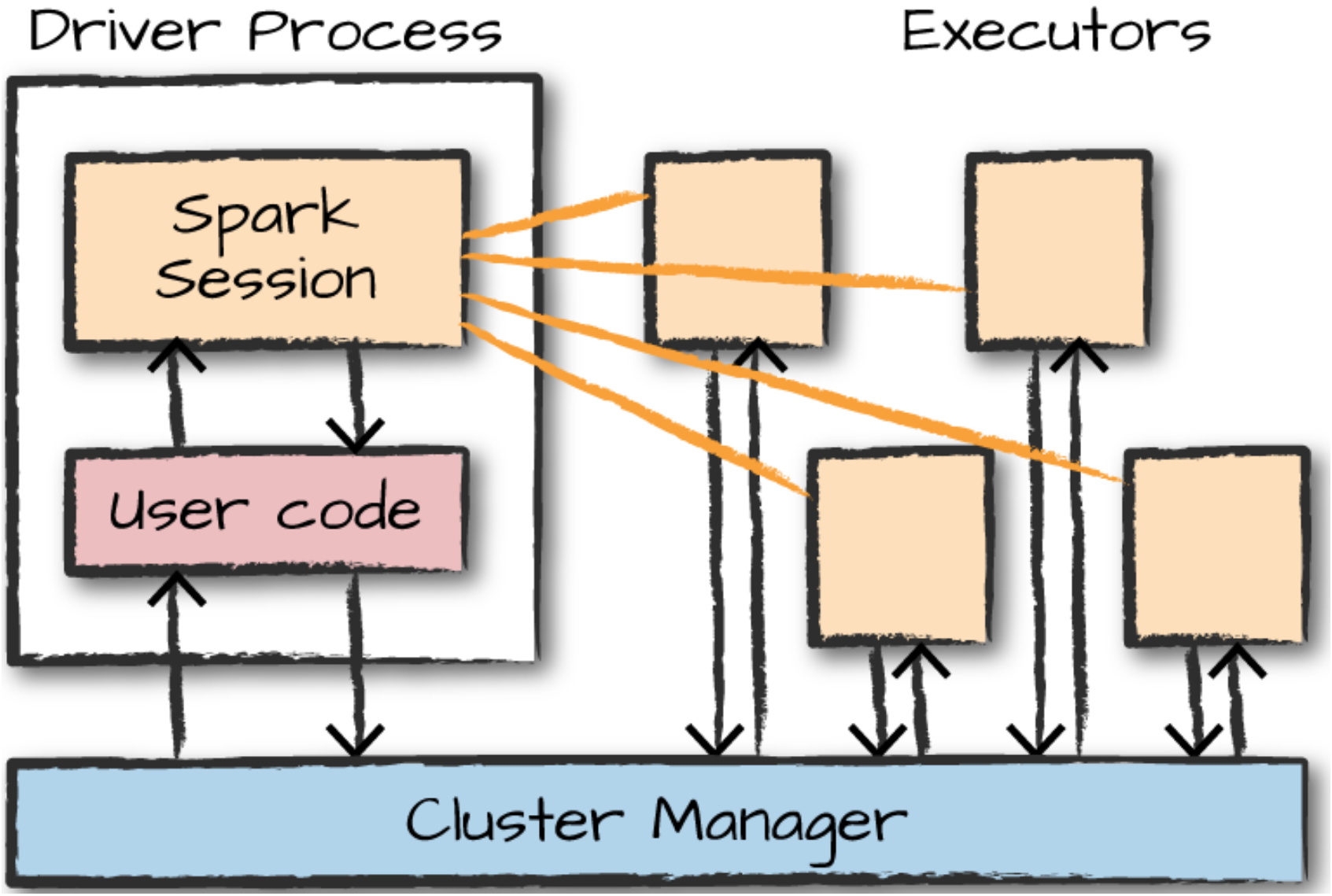


Source: Spark: The Definitive Guide

Hello, Email World!

- Jupyter Notebook with Apache Toree
- See ApacheSparkThroughEmail1

A Spark Application - Driver, Executors, Tasks, Cluster Managers



Source: Spark: The Definitive Guide

DataFrameReader: Built-in Data Formats

`spark.read`:

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

Scaling Behind the Scenes

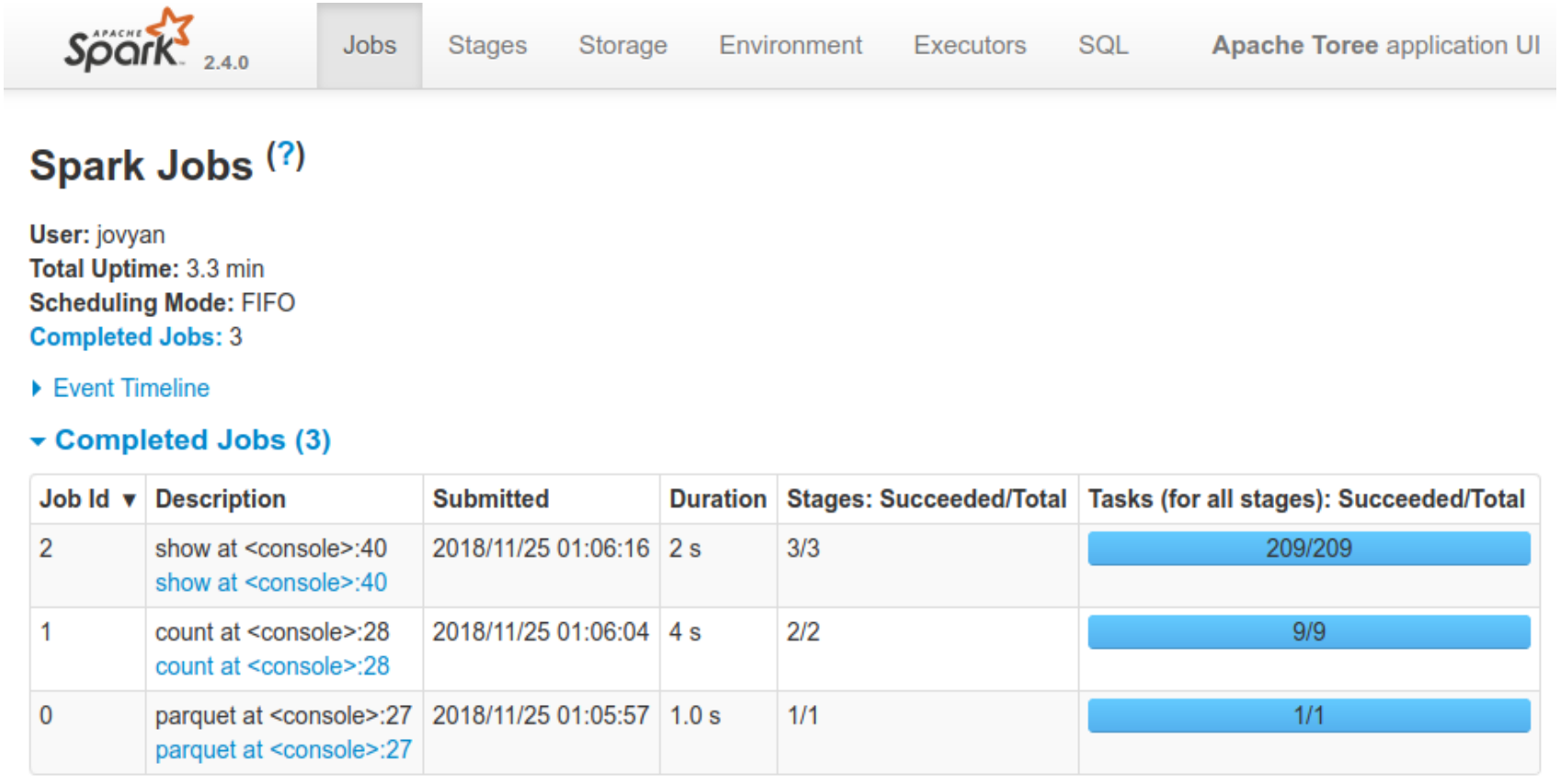


Figure 7: Jobs and Tasks

Combining work per stage - shuffle

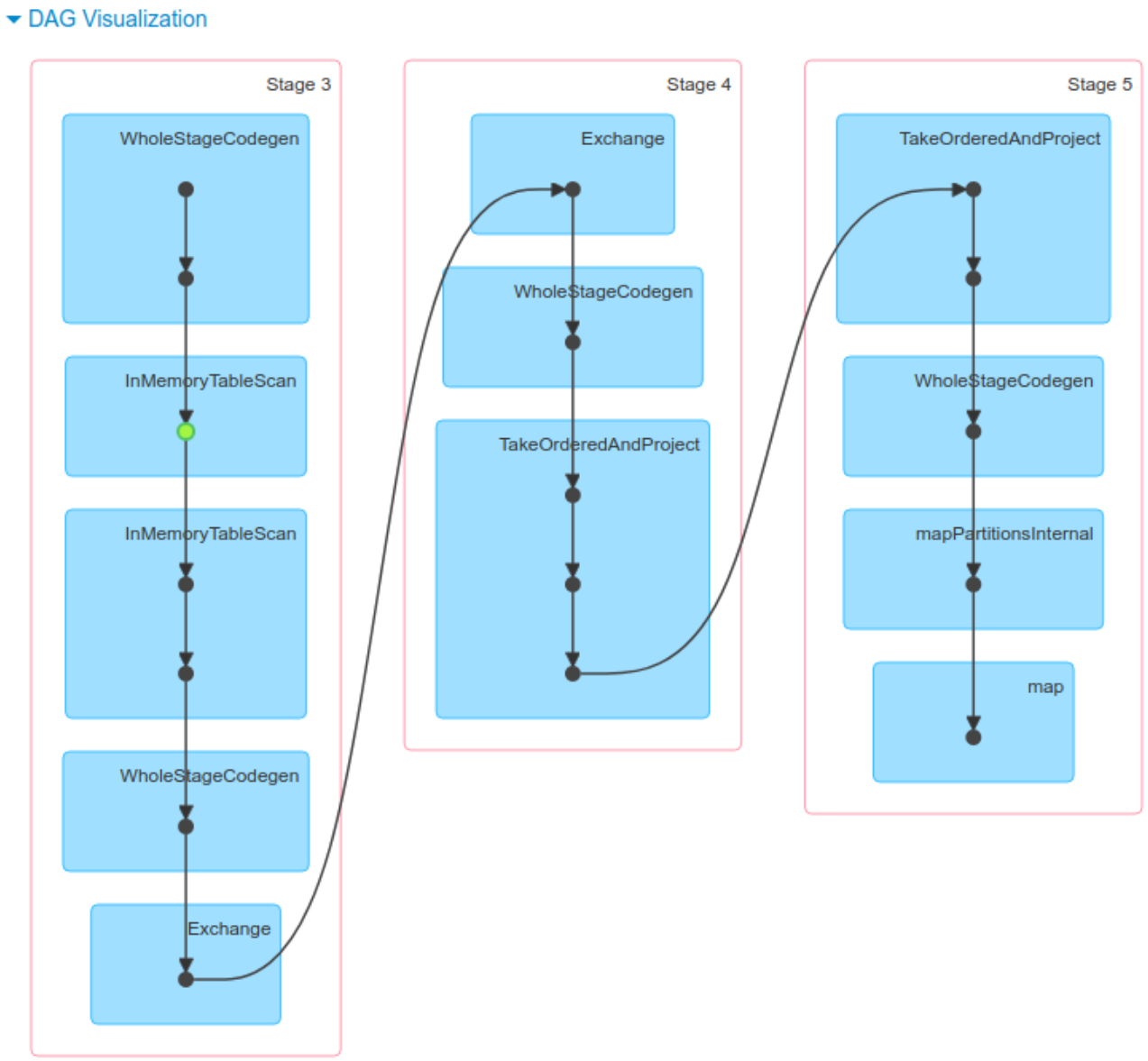


Figure 8: Stages

Where clause, Column methods, Built-in functions

- See Apache Spark Through Email Notebook 2

Parallelism and Partitioning

- Initial parallelism - number of input "blocks"
- Shuffle - `spark.`

Colon Cancer

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

Questions?

Figure 9: medale@asymmetrik.com