Apache Spark Through Email

Markus Dale

Nov 2018

Slides And Code

- Slides: https://github.com/medale/sparkmail/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)

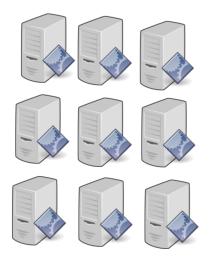


Figure 4: Multiple cooperating Servers

Big Data Framework - Apache Hadoop



Figure 5: HDFS, MapReduce

Hadoop Ecosystem

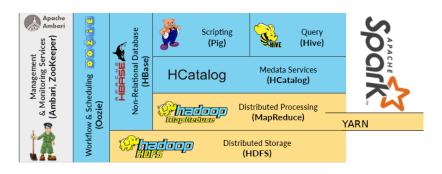


Figure 6: Some Frameworks Around Hadoop

Apache Spark Components

Structured Advanced Libraries & Streaming Ecosystem Analytics Structured APIS Datasets SQL DataFrames Low-level APIS Distributed Variables RDDs

Source: Spark: The Definitive Guide

Hello, Email World!

- Jupyter Notebook with Apache Toree
- ► See ApacheSparkThroughEmail1

DataFrameReader: Built-in Data Formats

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

Figure 7: Jobs and Tasks

Combining work per stage - shuffle

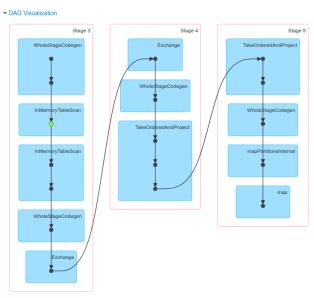


Figure 8: Stages

Colon Cancer



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

Questions?



Figure 9: medale@asymmetrik.com