Data Engineering with Apache Spark

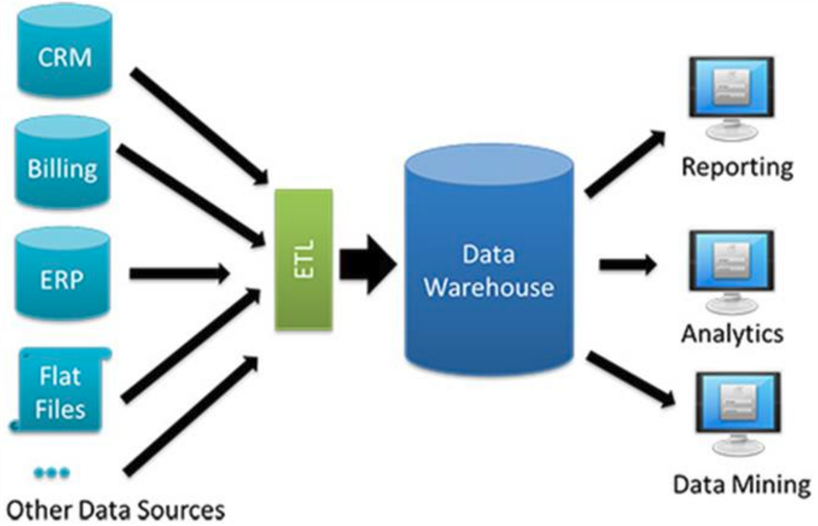
Markus Dale, medale@asymmetrik.com

May 2019

Intro, Slides And Code

- Slides: https://github.com/medale/
- Scala Spark Code Examples: https://github.com/medale/

Data engineering



Data engineering on small dataset

Figure 1: Laptop

Data engineering for larger dataset (Vertical Scaling)

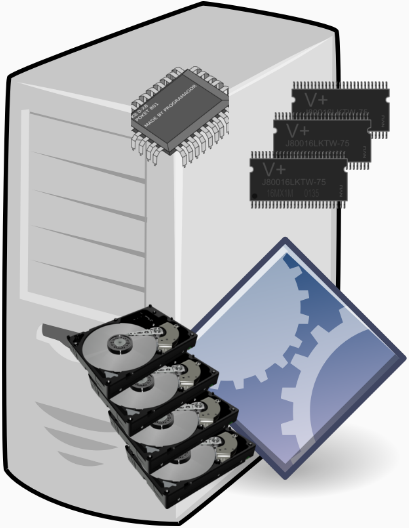
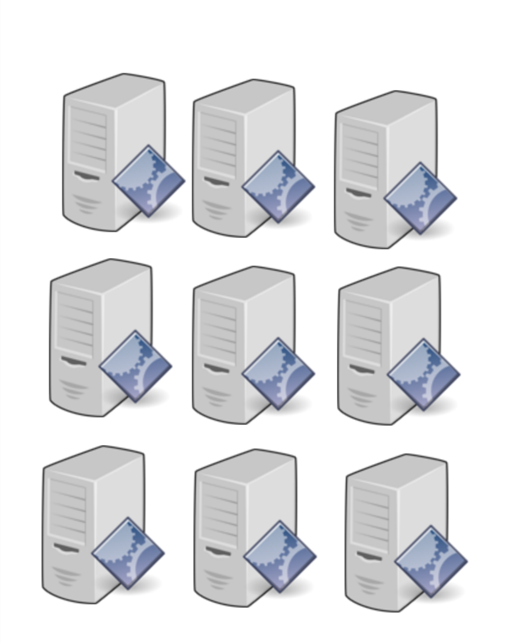
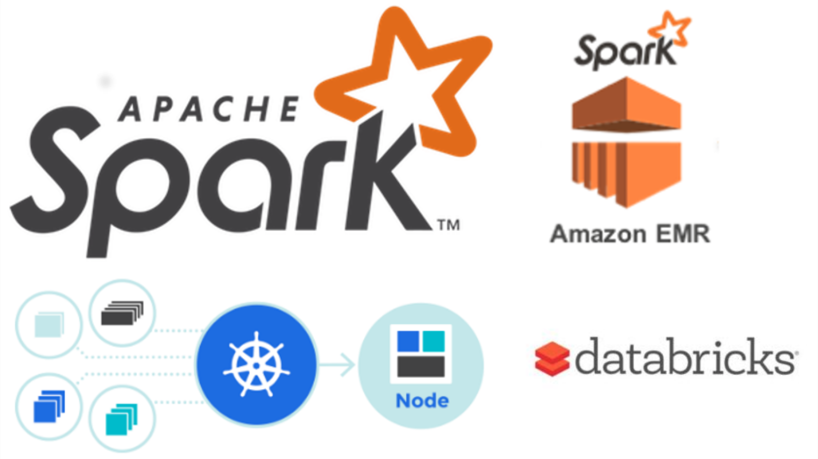


Figure 2: Beefed-up Server

Data engineering for large datasets (Horizontal Scaling)



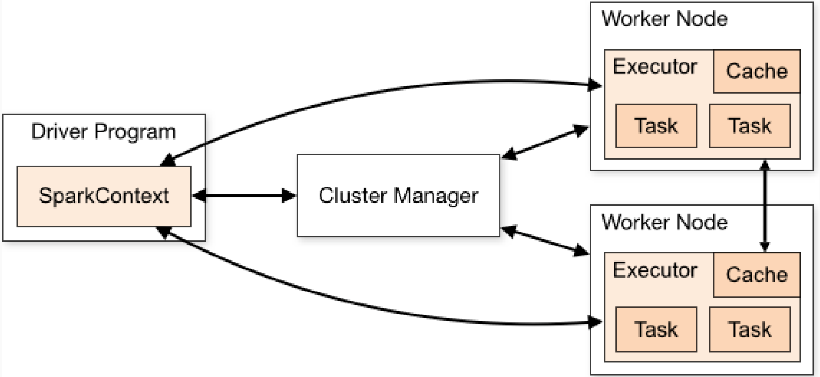
Cluster Manager



Layers of Spark

- TODO: distributed resource management
- Spark Standalone
- Kubernetes
- Hadoop YARN
- Mesos
- Databricks
- Spark application: Driver, executors
- tasks, partitions (distributed file system, commonly accessible data store)

Cluster Manager, Driver, Executors, Tasks



Source: Apache Spark website

And now for something completely different: Colon Cancer

- Screening saves lives!
    - Colonoscopy - talk to your doc
    - Dave Barry: A journey into my colon — and yours
- Colorectal Cancer Alliance

Questions?

- medale@asymmetrik.com
- Infrequent blog/past presentations http://uebercomputing.com/