Speaker Notes: Data Engineering with Apache Spark

Markus Dale, medale@asymmetrik.com

May 2019

Setup

- Open Spark API:
  https://spark.apache.org/docs/latest/api/scala/index.html

Intro, Slides And Code

- Bio:
  - mostly Java, big data with Hadoop
  - big data with Spark, Databricks, Scala
  - Now Asymmetrik - Scala, Spark, Elasticsearch, Akka...
  - Data Engineer
- Slides: https://github.com/medale/
- Scala Spark Code Examples: https://github.com/medale/
- Also https://github.com/medale/spark-mail

Goals

Data Engineering

- dataquest.io: “transform data into a useful format for analysis”

Narrow vs. wide transformations

https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-rdd-transformations.html

Splittable data formats

Shuffle partitions

Schemas

Parquet

https://www.gharchive.org/ wget

http://data.gharchive.org/2019-04-28-0.json.gz wget

http://data.gharchive.org/2019-04-28-1.json.gz wget

http://data.gharchive.org/2019-04-28-13.json.gz

store under data directory run spark-shell from parent of data directory (gz of .json file with one json per line)

```
val records = spark.read.json("data")
//slow - needs to figure out JSON schema
records.cache
records.count
//235728


//huge
records.printSchema
```