GENERAL ASSEMBLY

DATA SCIENCE
11 WEEK PART TIME COURSE

Week 8 – Spark
Monday 7th February 2016

1. Tasks from Wednesday
2. Spark
3. Lab
4. Real World Problem
5. Review

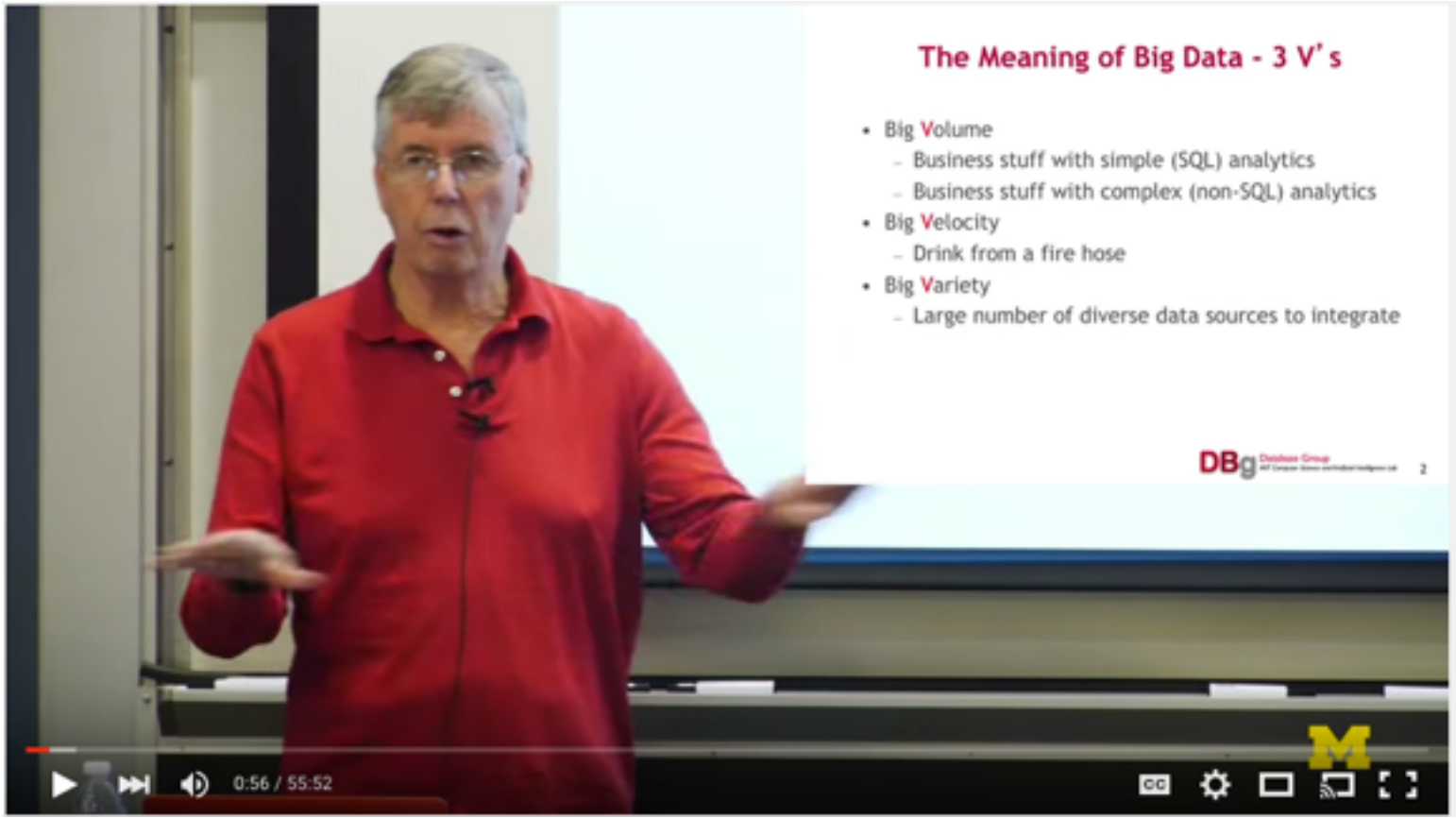
DATA SCIENCE – Week 7 Day 2

Task List

- [ ] Read & Review one chapter from http://www.redbook.io/
- [ ] Watch https://www.youtube.com/watch?v=KRcecxdGxvQ , what are the 5 key points for you?
- [ ] Install Spark on EMR
- [ ] Load data into S3
- [ ] Load data from S3 into your iPython notebook on EC2

1. Background introduced by Michael Stonebraker
2. Traditional RDBMS Systems introduced by Michael Stonebraker
3. Techniques Everyone Should Know introduced by Peter Bailis
4. New DBMS Architectures introduced by Michael Stonebraker
5. Large-Scale Dataflow Engines introduced by Peter Bailis
6. Weak Isolation and Distribution introduced by Peter Bailis
7. Query Optimization introduced by Joe Hellerstein
8. Interactive Analytics introduced by Joe Hellerstein
9. Languages introduced by Joe Hellerstein
10. Web Data introduced by Peter Bailis
11. A Biased Take on a Moving Target: Complex Analytics by Michael Stonebraker
12. A Biased Take on a Moving Target: Data Integration by Michael Stonebraker

Michael Stonebraker | Big Data is (at least) Four Different Problems



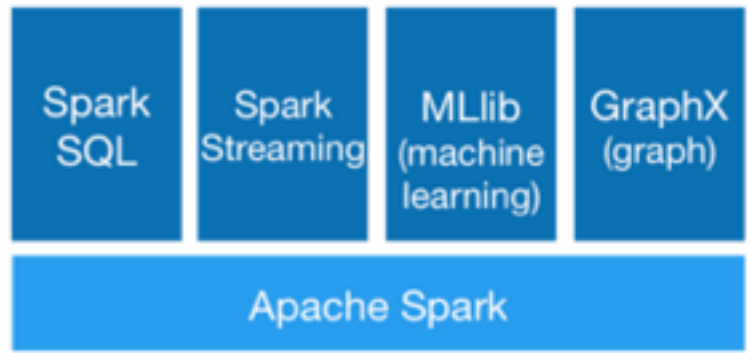
1. Big Volume - Little Analytics
2. Column stores ~ 50x faster
3. Array store for complex analytics
4. SQL is supporting json to perform NoSQL type operations
5. Uncurrated data is a data swamp not a lake

Amazon EC2

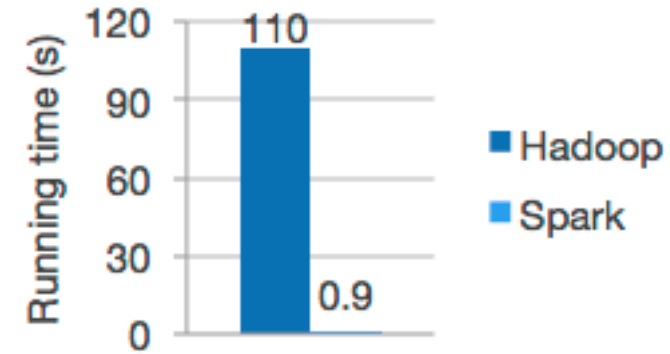
DATA SCIENCE PART TIME COURSE

SPARK

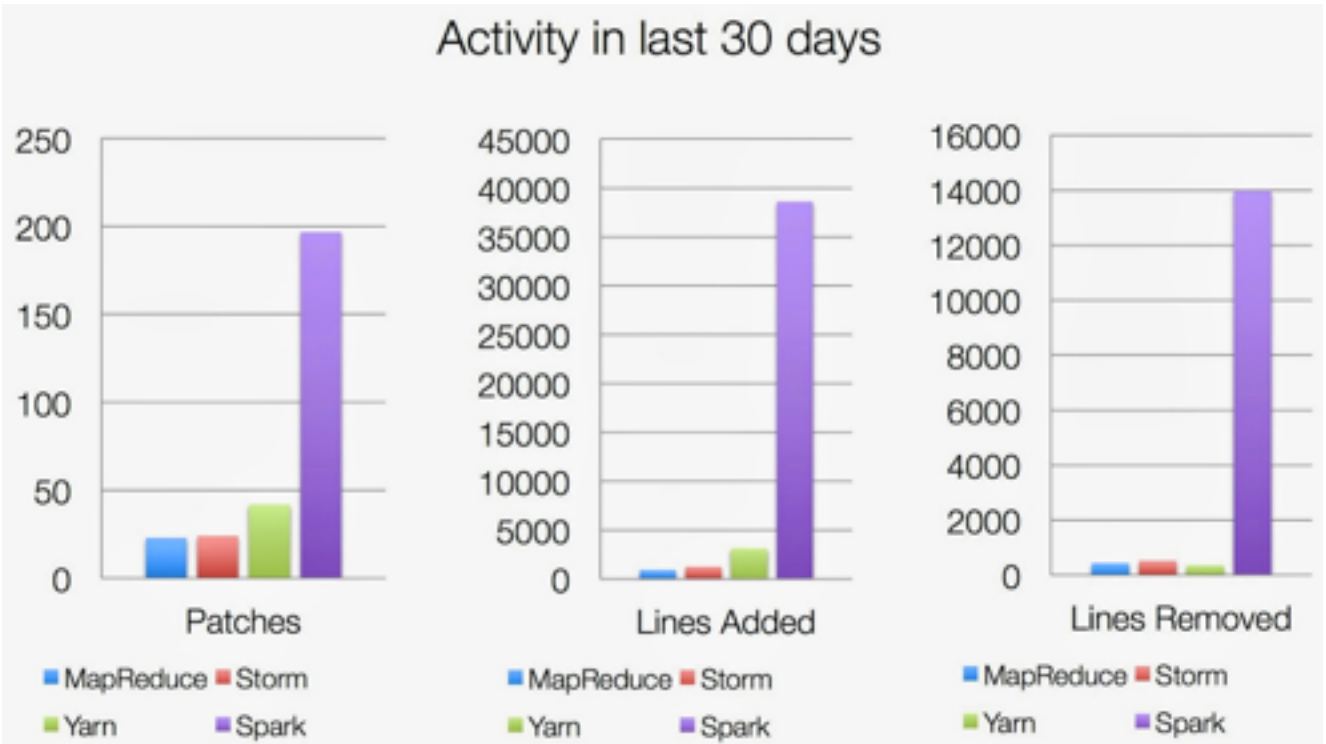
Spark is a fast and general processing engine compatible with Hadoop data. It can process data in HDFS, HBase, Cassandra, Hive, and any Hadoop InputFormat. It is designed to perform both batch processing (similar to MapReduce) and new workloads like streaming, interactive queries, and machine learning.

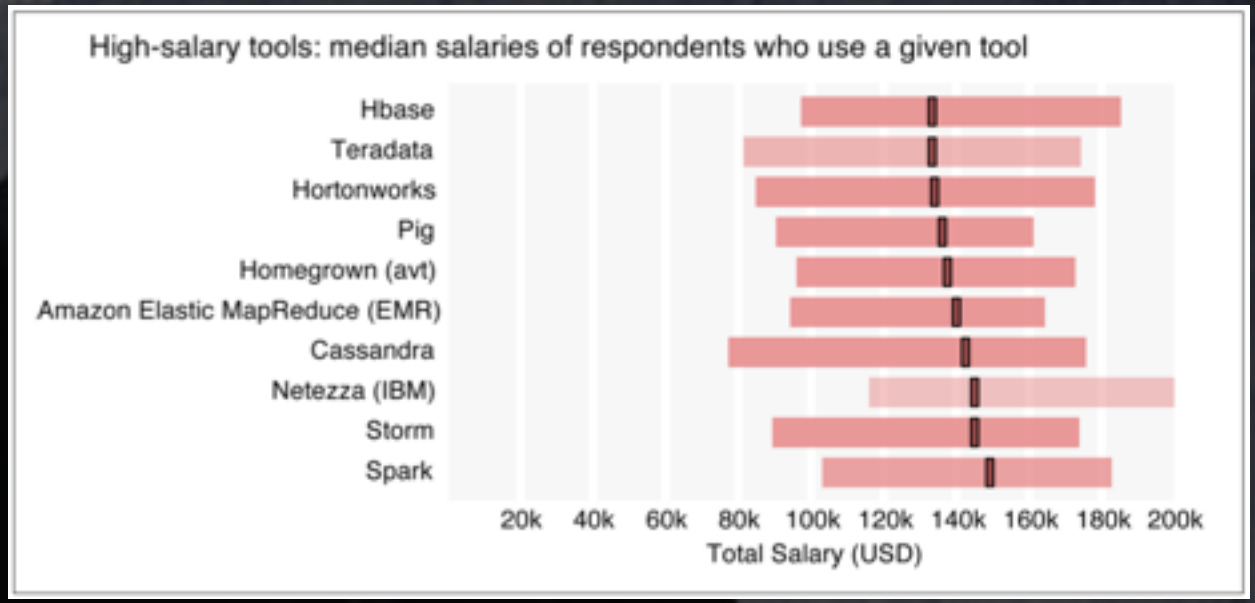


- MLlib is Spark's machine learning library. Its goal is to make practical machine learning scalable and easy. It consists of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as lower-level optimization primitives and higher-level pipeline APIs.
- GraphX in Spark for graphs and graph-parallel computation



Logistic regression in Hadoop and Spark


Activity in last 30 days
250
200
150
100
50
0
Patches
MapReduce
Storm
Yarn
Spark
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
Lines Added
MapReduce
Storm
Yarn
Spark
16000
14000
12000
10000
8000
6000
4000
2000
0
Lines Removed
MapReduce
Storm
Yarn
Spark



‘We can talk, but money talks, so talk more bucks’ - Jay-Z (Izzo - The Blueprint)

Spark revolves around the concept of a resilient distributed dataset (RDD), which is a fault-tolerant collection of elements that can be operated on in parallel.

There are two ways to create RDDs:

1. Parallelizing an existing collection in your driver program
2. Referencing a dataset in an external storage system, such as a shared filesystem, HDFS, HBase, or any data source offering a Hadoop InputFormat

One use of Spark SQL is to execute SQL queries written using either a basic SQL syntax or HiveQL. Spark SQL can also be used to read data from an existing Hive installation.

Spark SQL provide Spark with more information about the structure of both the data and the computation being performed. Internally, Spark SQL uses this extra information to perform extra optimizations. There are several ways to interact with Spark SQL including SQL, the DataFrames API and the Datasets API. When computing a result the same execution engine is used, independent of which API/language you are using to express the computation.

A DataFrame is a distributed collection of data organized into named columns.

It is conceptually equivalent to a table in a relational database or a data frame in R/Python, but with richer optimizations under the hood.

DataFrames can be constructed from a wide array of sources such as: structured data files, tables in Hive, external databases, or existing RDDs.

A Dataset is a new experimental interface added in Spark 1.6 that tries to provide the benefits of RDDs (strong typing, ability to use powerful lambda functions) with the benefits of Spark SQL's optimized execution engine.

A Dataset can be constructed from JVM objects and then manipulated using functional transformations (map, flatMap, filter, etc.).

The unified Dataset API can be used both in Scala and Java. Python does not yet have support for the Dataset API. Full python support will be added in a future release.

- spark.mllib contains the original API built on top of RDDs.
- spark.ml provides higher-level API built on top of DataFrames for constructing ML pipelines.

Data types

Basic statistics

- summary statistics
- correlations
- stratified sampling
- hypothesis testing
- streaming significance testing
- random data generation

Classification and regression

- linear models (SVMs, logistic regression, linear regression)
- naive Bayes
- decision trees
- ensembles of trees (Random Forests and Gradient-Boosted Trees)
- isotonic regression

Collaborative filtering

- alternating least squares (ALS)

Clustering

- k-means
- Gaussian mixture
- power iteration clustering (PIC)
- latent Dirichlet allocation (LDA)
- bisecting k-means
- streaming k-means

Dimensionality reduction

- singular value decomposition (SVD)
- principal component analysis (PCA)

Feature extraction and transformation

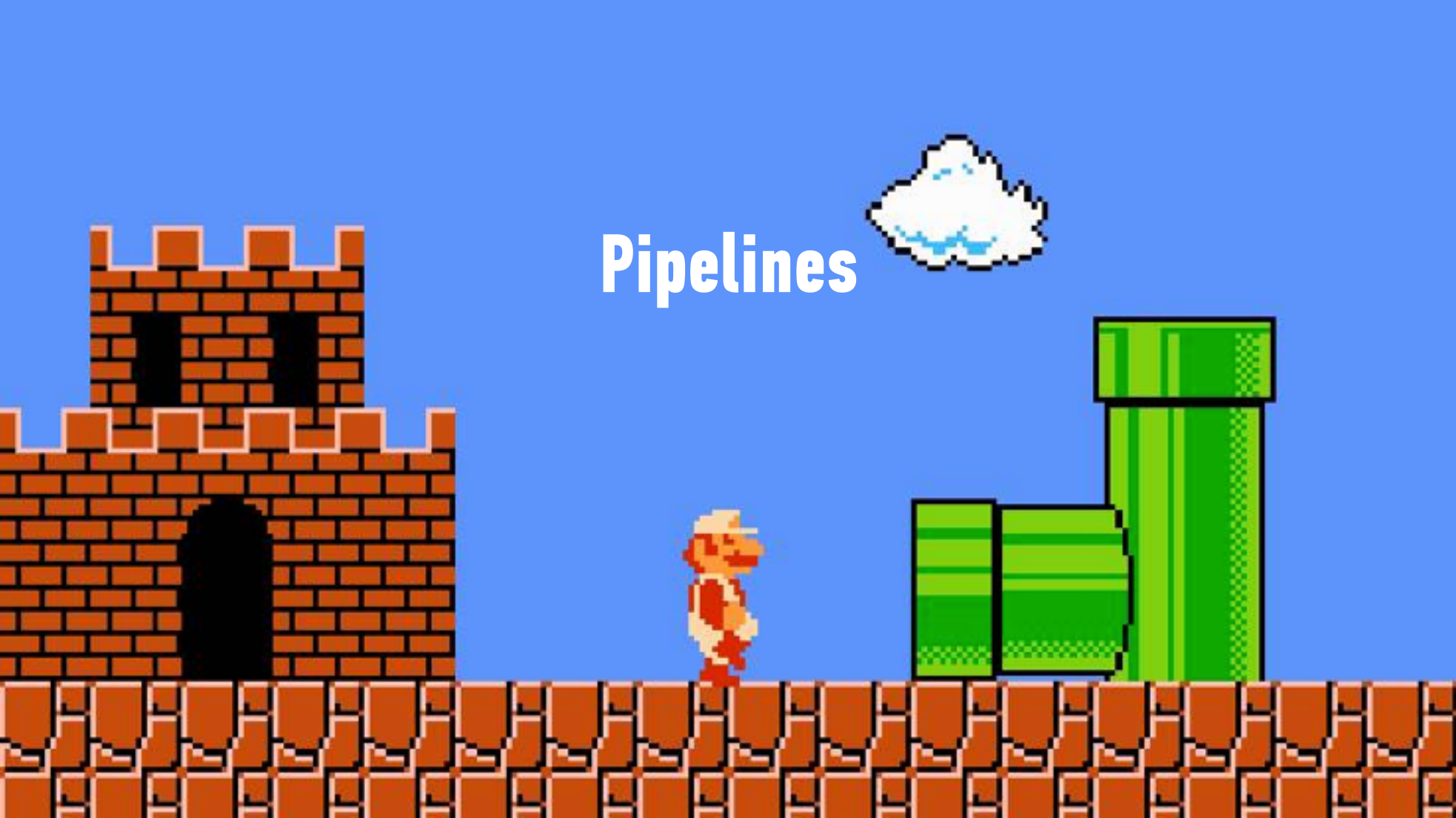
Frequent pattern mining

- FP-growth
- association rules
- PrefixSpan

Evaluation metrics

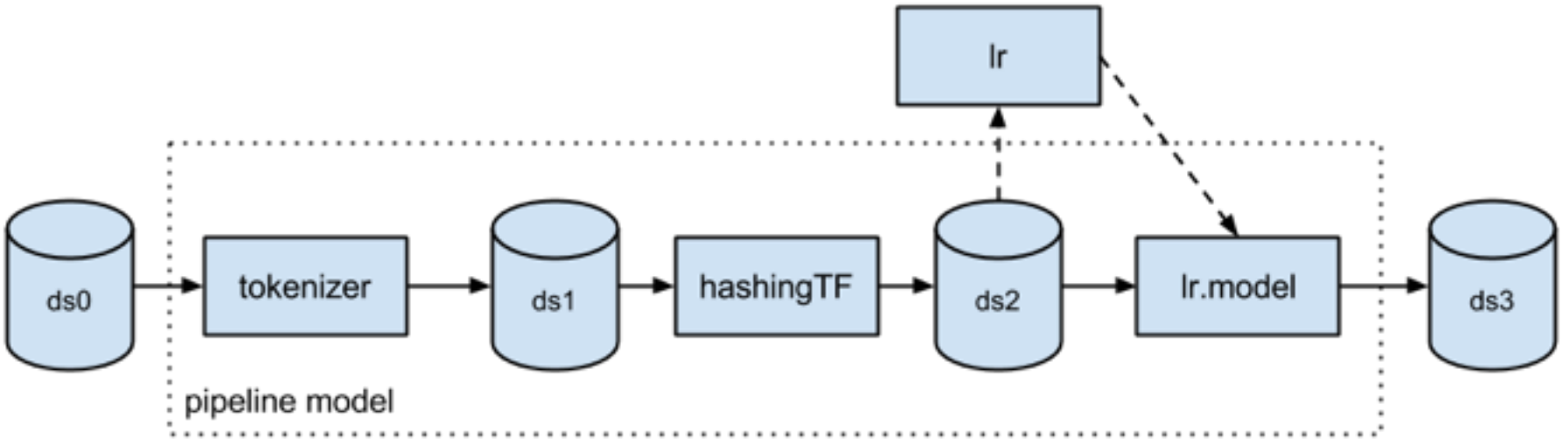
PMML model export

Optimization (developer)


Pipelines

Two types of pipelines

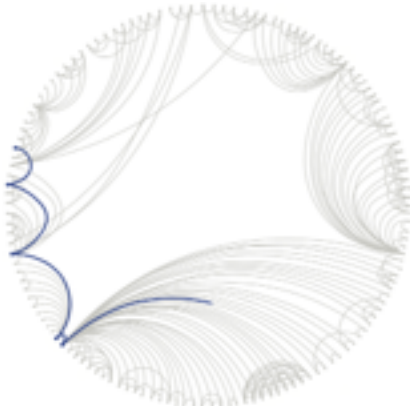
- Transformer - takes a dataset as input and produces an augmented dataset as output. For example, a transformer may read a column (e.g., text), map it into a new column (e.g., feature vectors), and output a new DataFrame with the mapped column appended
- Estimator - basically training a model, it must be first fit on the input dataset to produce a model. For example, a learning algorithm such as LogisticRegression is an Estimator.


lr
ds0
tokenizer
ds1
hashingTF
ds2
lr.model
ds3
pipeline model

Useful for graphs and graph parallel processing

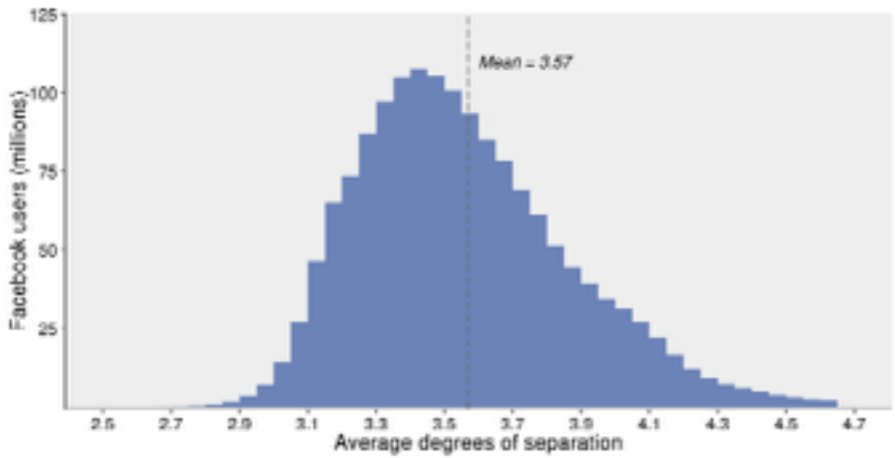
- PageRank
- Label Propagation
- SVD++
- Triangle Counting

How connected is the world?

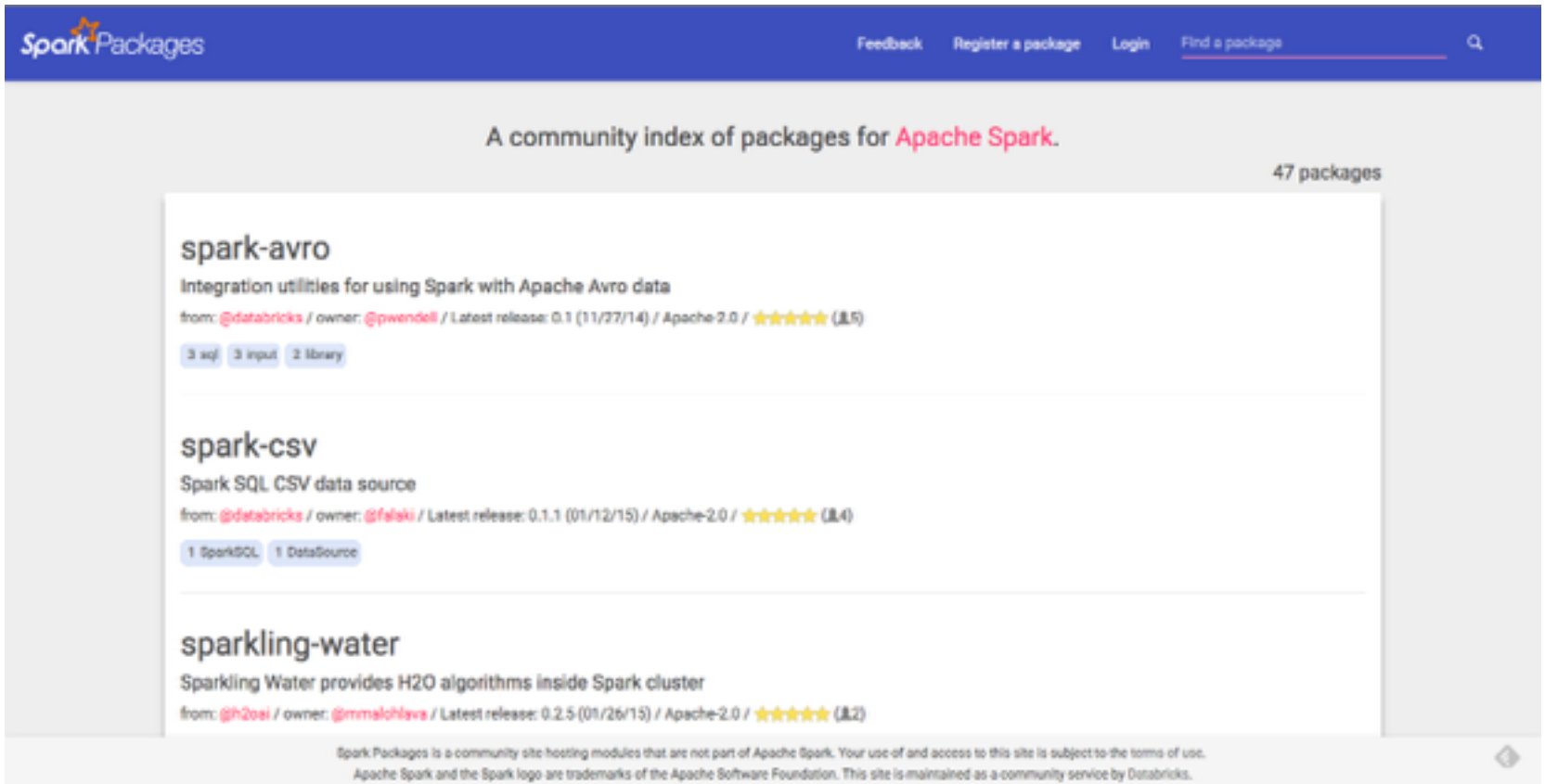


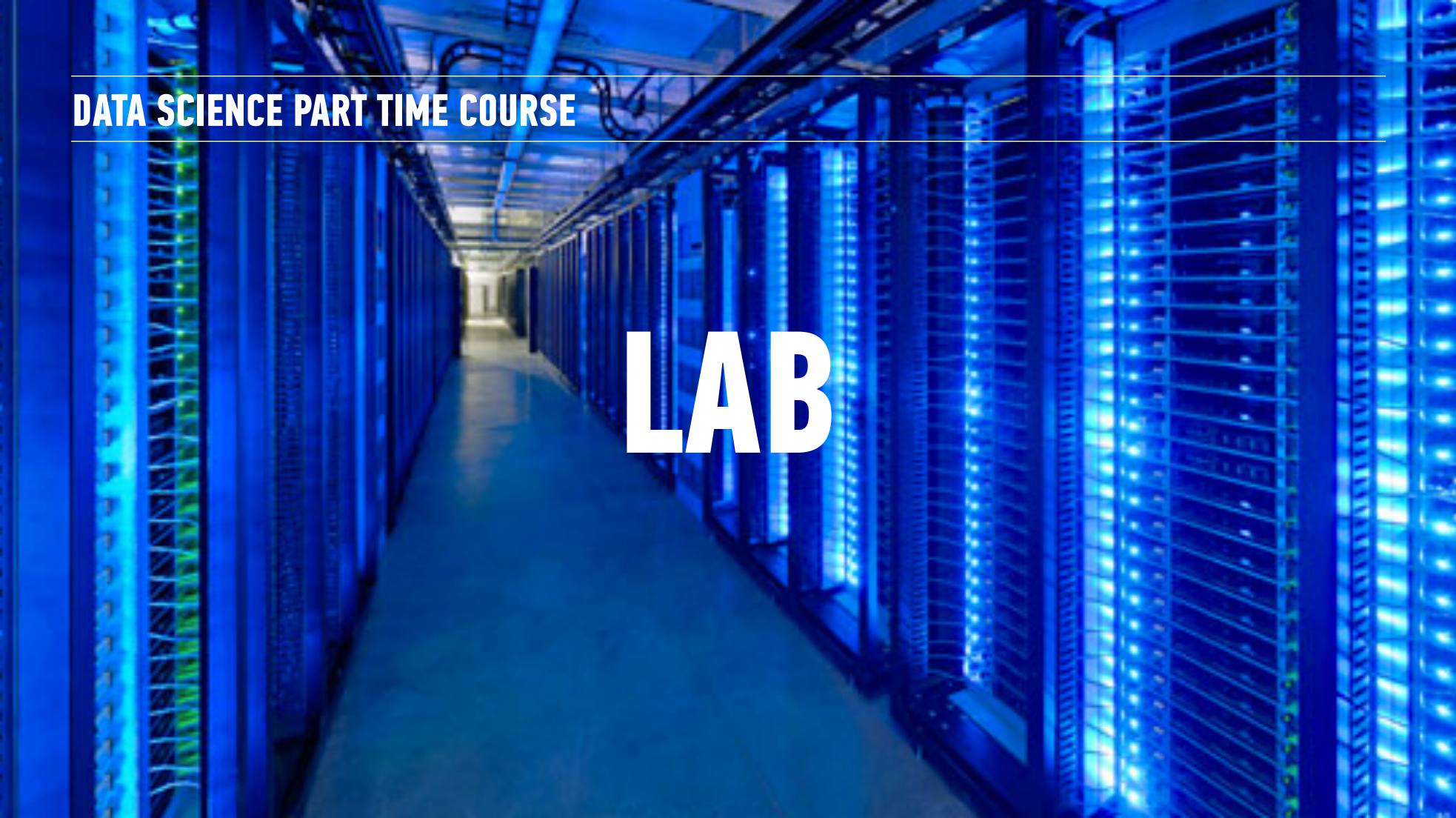
Each person in the world (at least among the 1.59 billion people active on Facebook) is connected to every other person by an average of three and a half other people.

Rather than calculate it exactly, they estimate distances with statistical algorithms



SPARK PACKAGES





LAB

LAB

- Start a Spark cluster with EMR
- Run a notebook in Zepplin and connect to it
- Load and analyse data in Spark

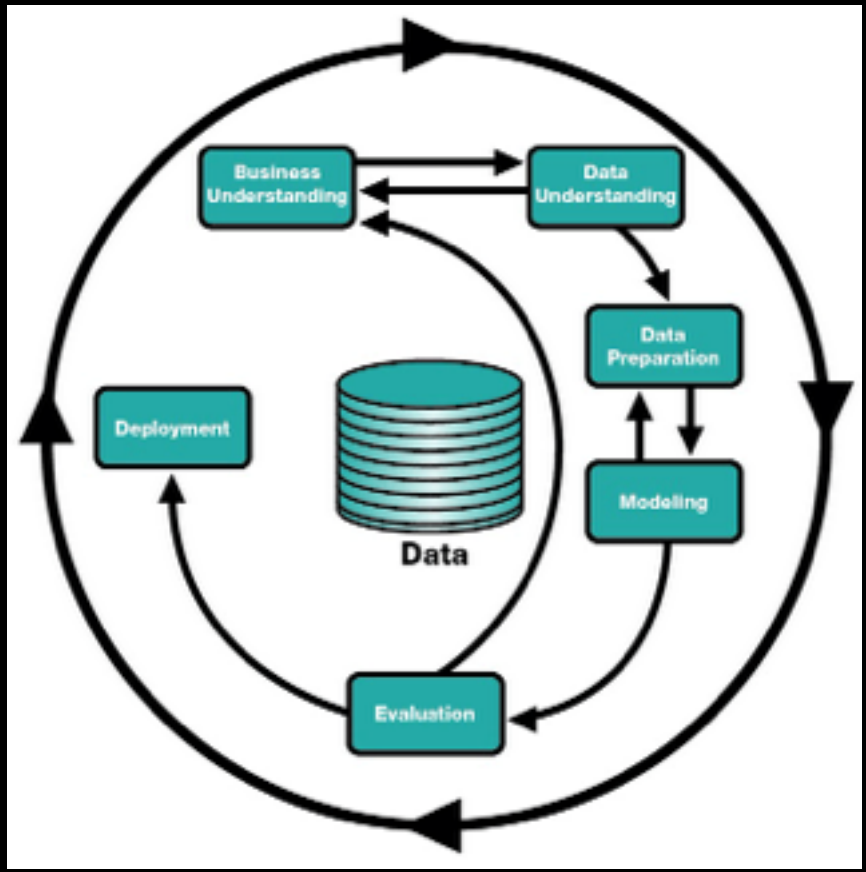
DISCUSSION TIME

- Talk through a real problem
- Review last week
- Questions
- Task List

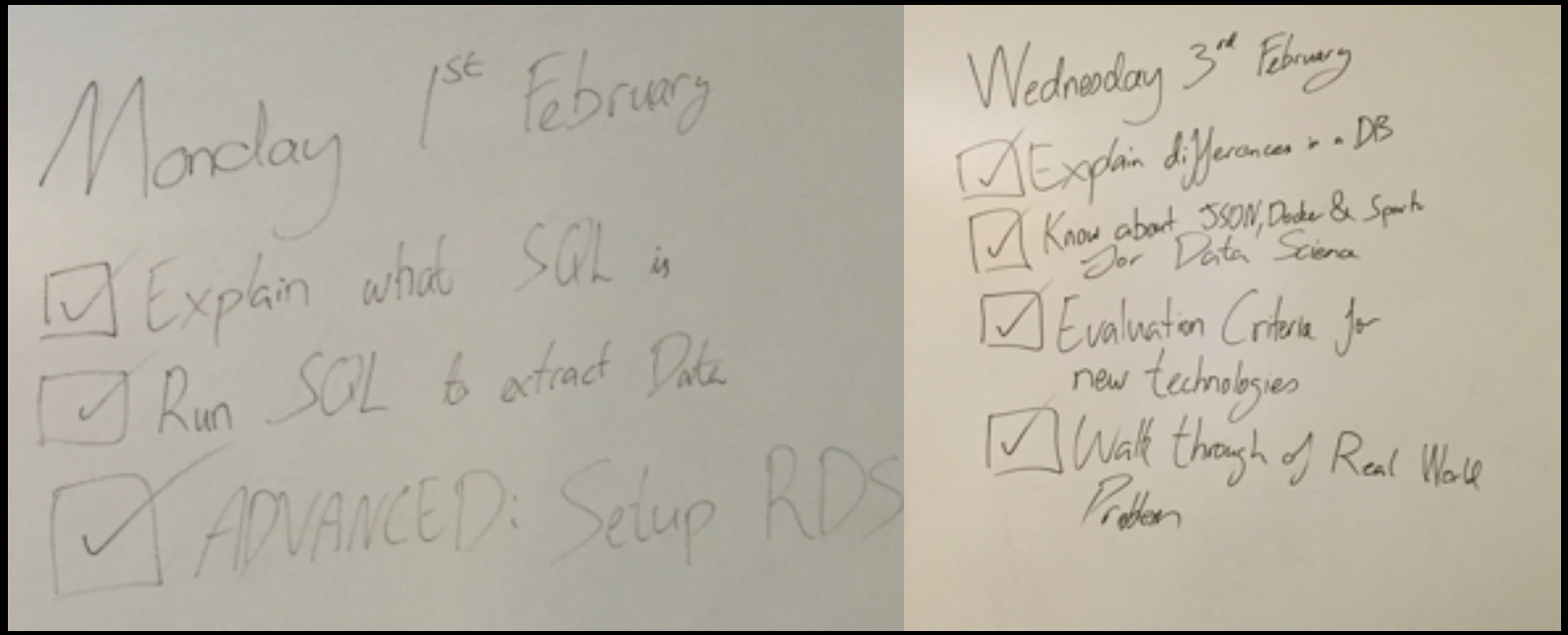
DATA SCIENCE – Week 8 Day 1

REAL PROBLEMS

DATA SCIENCE – Week 8 Day 1



REVIEW


Monday 1st February
Explain what SQL is
Run SQL to extract Data
ADVANCED: Setup RDS
Wednesday 3rd February
Explain differences in a DB
Know about JSON, Docker & Spark for Data Science
Evaluation Criteria for new technologies
Walk through of Real World Problem

DATA SCIENCE – Week 8 Day 1

Task List (30 mins)

- [ ] Read first 2 chapters of Forecasting Principles and Practice https://www.otexts.org/fpp (15 mins)
- [ ] Download and Install R (10 mins)
- [ ] Download and Install RStudio (3 mins)
- [ ] Download the forecast package in R (2 mins)