GENERAL ASSEMBLY

DATA SCIENCE

11 WEEK PART TIME COURSE

Week 8 - Cloud Computing
Monday 9th May 2016

AGENDA

1. Guest Speaker
2. Cloud Computing
3. Spark
4. Lab
5. Real World Problem
6. Review

WHAT IS IT AND WHY USE IT?

- EC2
- RDS
- mongoDB
- Redshift
- Spark

AWS - SPARK - JUPYTER

DATA SCIENCE PART TIME COURSE

NO-SQL

SQL

- Traditional rows and columns data
- Strict structure / Primary Keys
- Entire column for each feature
- Industry standard

NoSQL

- No well defined data structure
- Works better for unstructured data
- Cheaper hardware
- Popular among Startups

SQL/NoSQL - Examples

SQL

- MySQL
- Oracle
- Postgres
- SQLite
- SQLServer
- Redshift

NoSQL

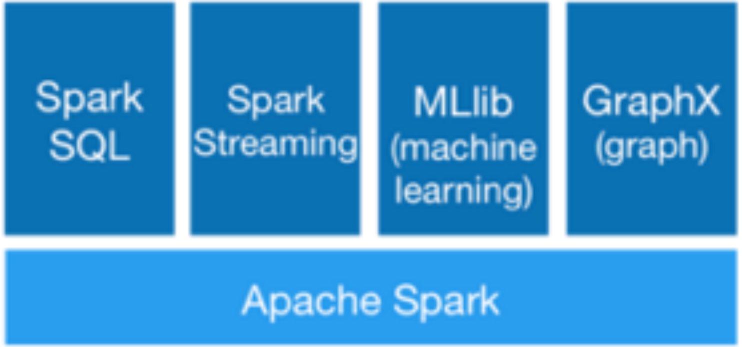
- MongoDB
- CouchDB
- Redis
- Cassandra
- Neo4j
- HBase

DATA SCIENCE PART TIME COURSE

SPARK

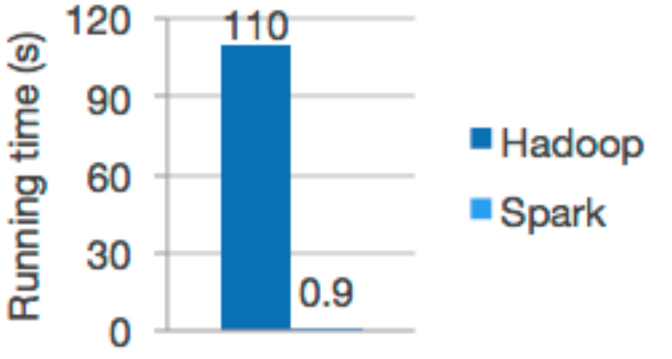
SPARK – WHAT IS IT?

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.



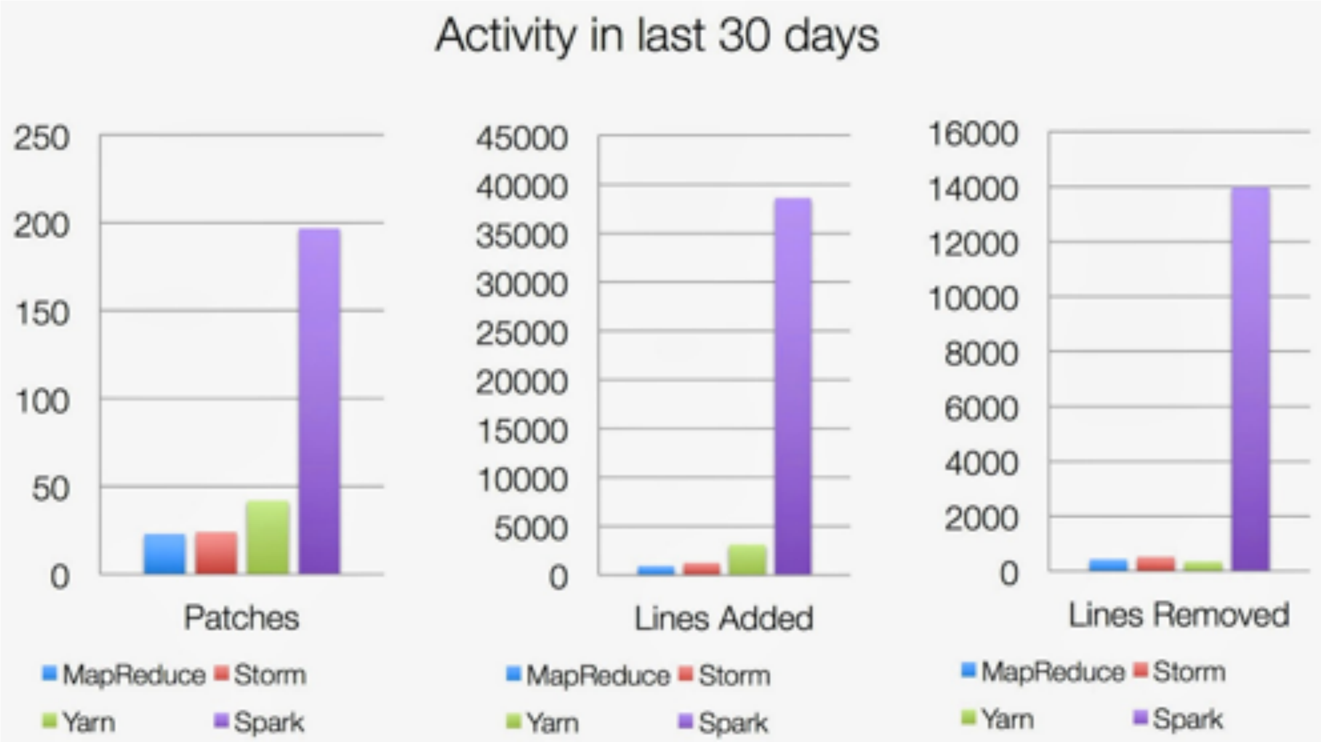
SPARK – HOW DOES IT RELATE TO DATA SCIENCE?

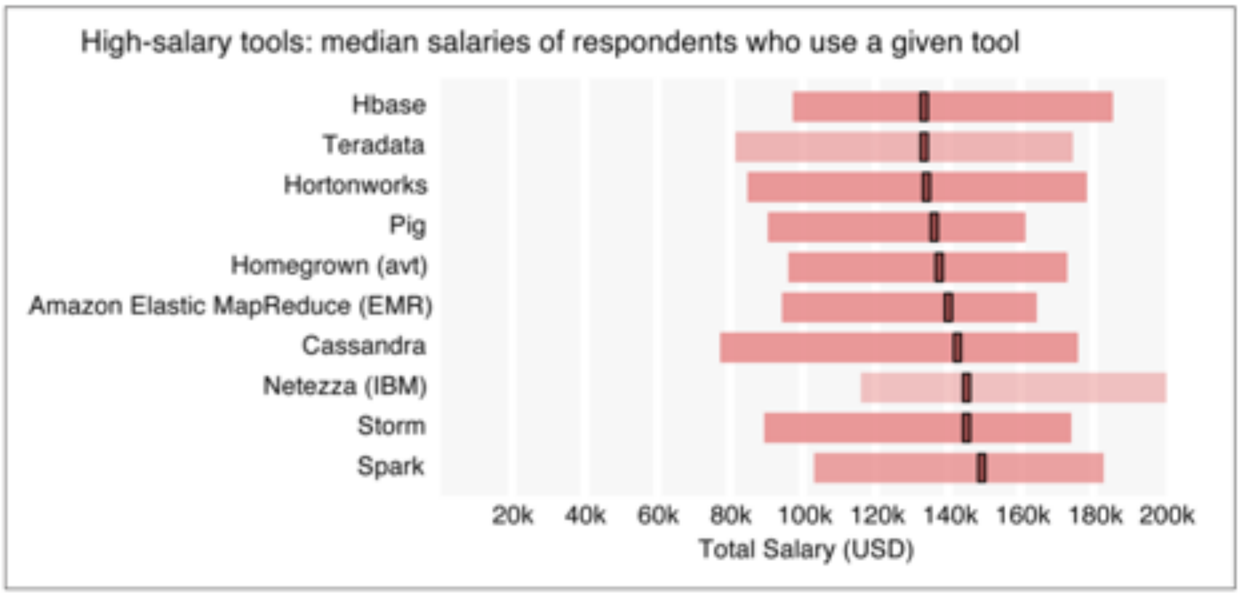
- MLlib is Spark's machine learning library. Its goal is to make practical machine learning scalable and easy. It consists of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as lower-level optimization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



Logistic regression in Hadoop and Spark

SPARK – HOW POPULAR IS IT?





High-salary tools: median salaries of respondents who use a given tool

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

RESILIENT DISTRIBUTED DATASETS (RDDs)

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

SPARK SQL

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.

SPARK DATAFRAMES

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.

SPARK DATASETS

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 & SPARK ML

- 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.

SPARK MLLIB

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

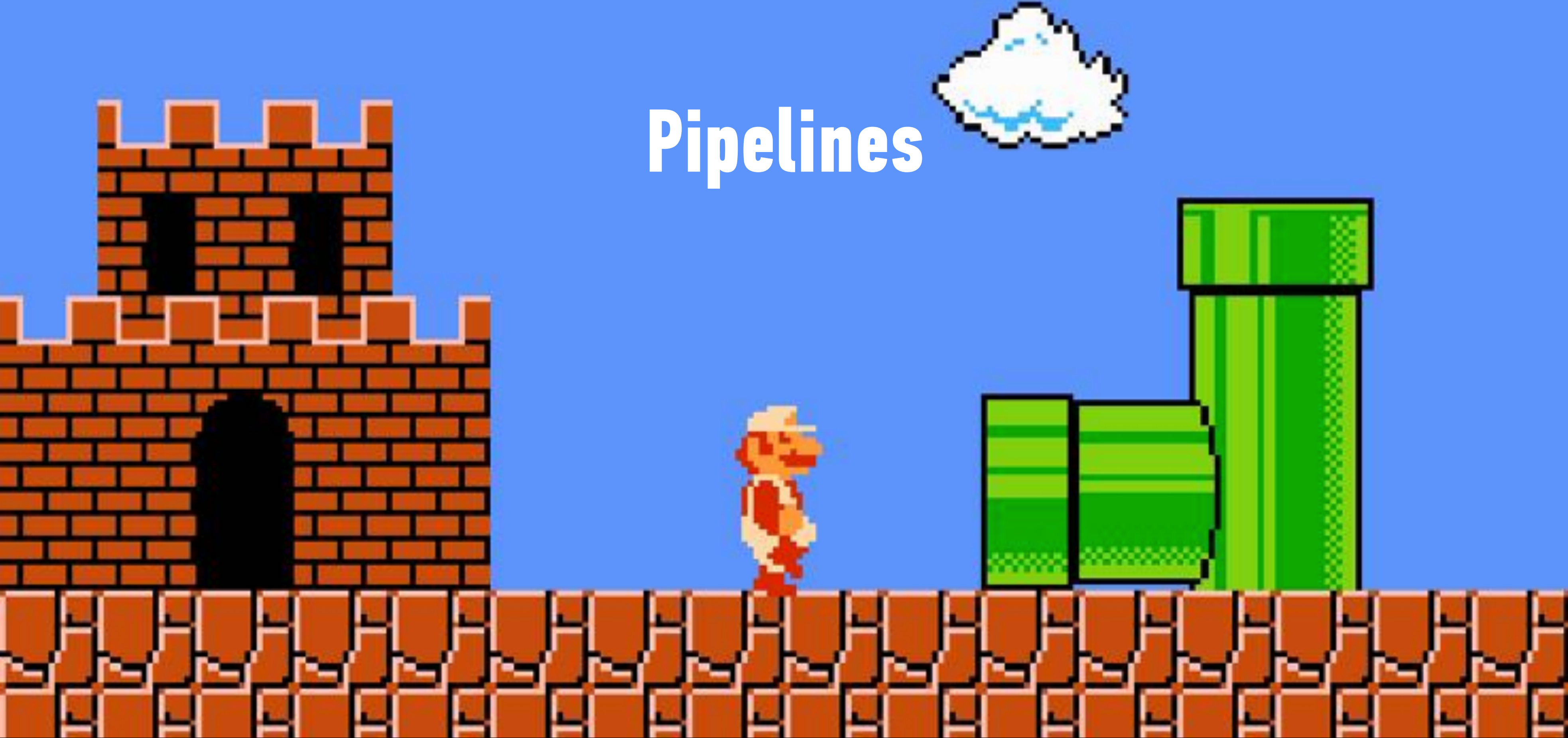
- FP-growth
- association rules
- PrefixSpan

Evaluation metrics

PMML model export

Optimization (developer)

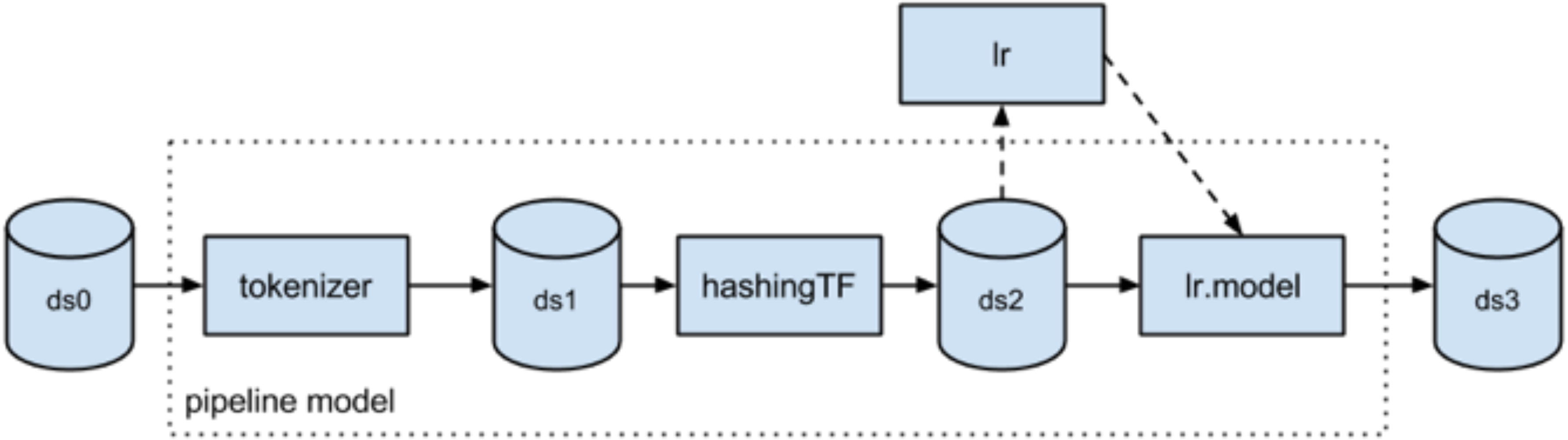
Pipelines



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.

PIPELINES


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

SPARK GRAPHX

Useful for graphs and graph parallel processing

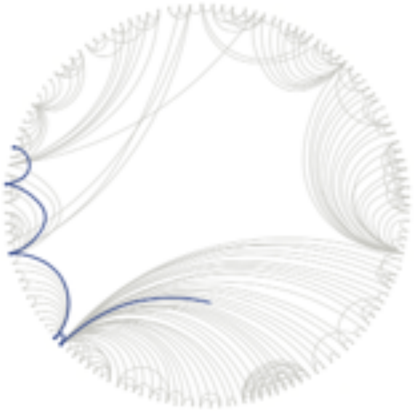
- PageRank
- Label Propagation
- SVD++
- Triangle Counting

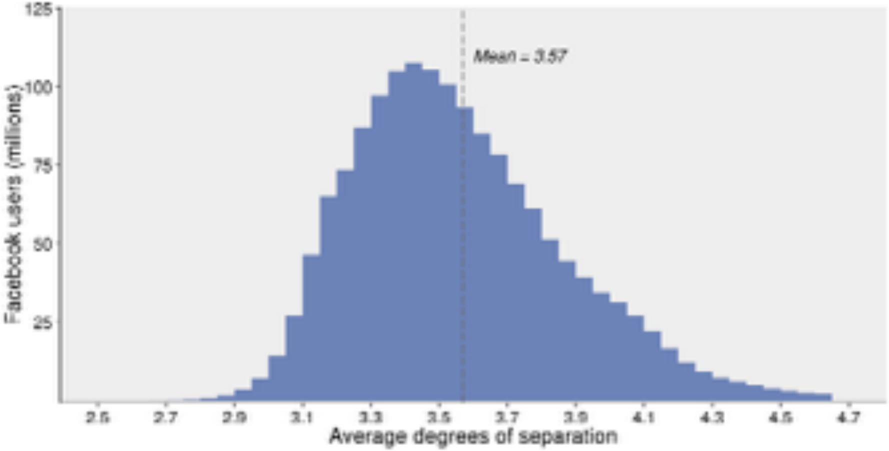
DIGRESSION – GRAPH ANALYSIS – FACEBOOK

How connected is the world?

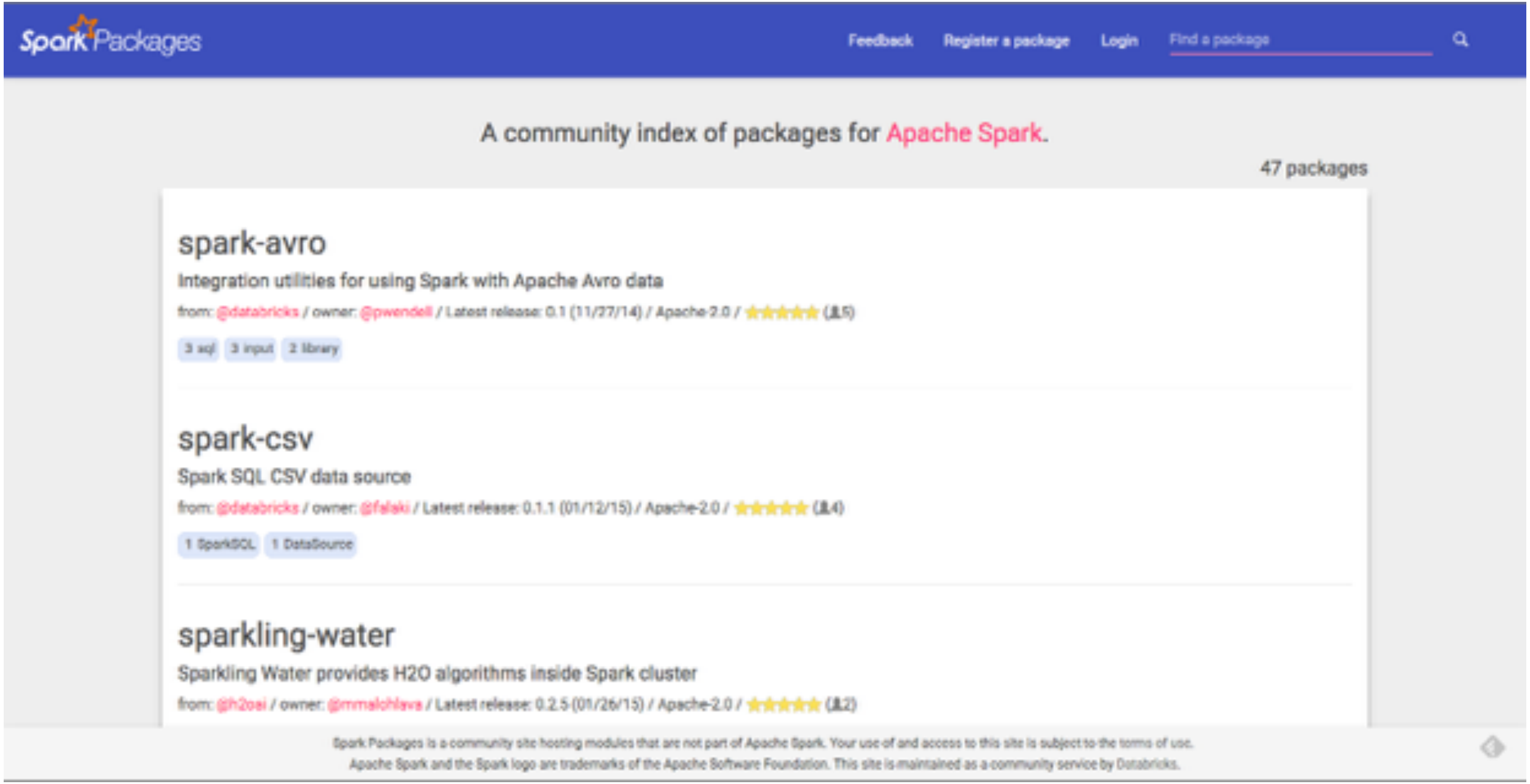
Each person in the world (at least among the 1.59 billion people active on Facebook) is connected to every other person by an average of three and a half other people.

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





SPARK PACKAGES



A community index of packages for Apache Spark.

47 packages

spark-avro

Integration utilities for using Spark with Apache Avro data

from: @databricks / owner: @pwendell / Latest release: 0.1 (11/27/14) / Apache-2.0 / ★★★★★ (5)

3 sql 3 input 2 library

spark-csv

Spark SQL CSV data source

from: @databricks / owner: @falaki / Latest release: 0.1.1 (01/12/15) / Apache-2.0 / ★★★★★ (4)

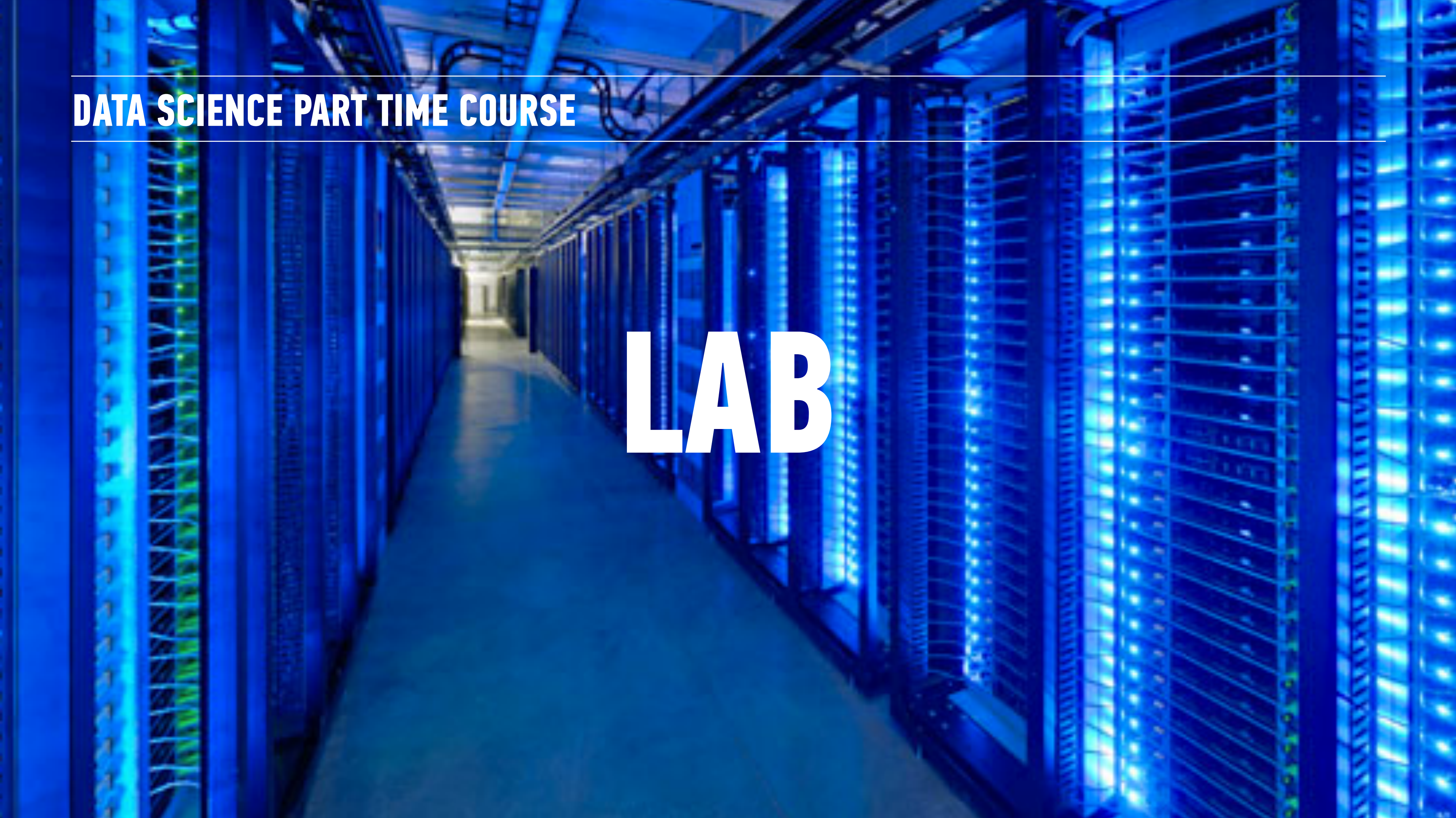
1 SparkSQL 1 DataSource

sparkling-water

Sparkling Water provides H2O algorithms inside Spark cluster

from: @h2oai / owner: @mmalohlava / Latest release: 0.2.5 (01/26/15) / Apache-2.0 / ★★★★★ (2)

Spark Packages is a community site hosting modules that are not part of Apache Spark. Your use of and access to this site is subject to the terms of use.
Apache Spark and the Spark logo are trademarks of the Apache Software Foundation. This site is maintained as a community service by Databricks.



DATA SCIENCE PART TIME COURSE

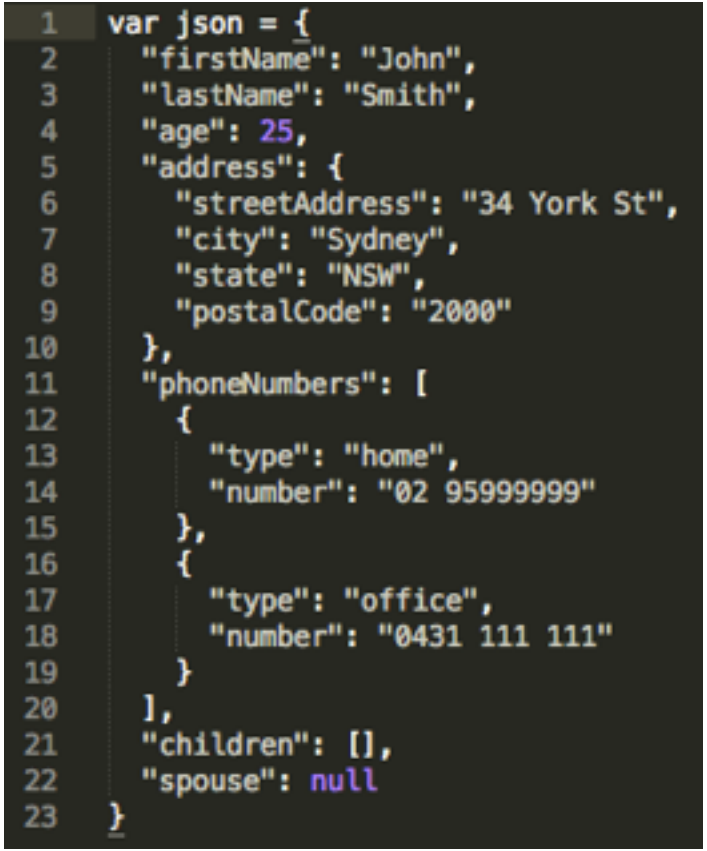
LAB

{JSON}

JSON – WHAT IS IT?

JSON - JavaScript Object Notation

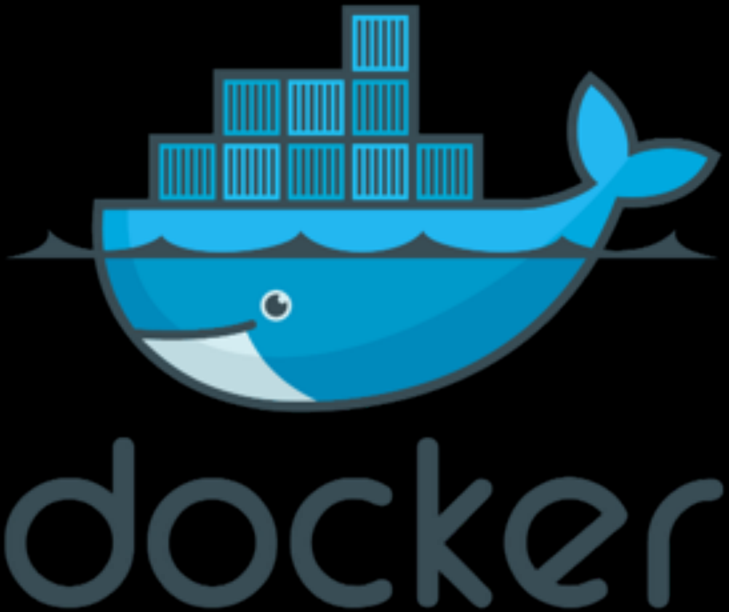
- Human readable data with attribute-value pairs.
- What is inside the curly brackets is an object
- In the object we declare variables with ‘attribute’ : ‘value’ pairs



```
var json = {
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "34 York St",
    "city": "Sydney",
    "state": "NSW",
    "postalCode": "2000"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "02 95999999"
    },
    {
      "type": "office",
      "number": "0431 111 111"
    }
  ],
  "children": [],
  "spouse": null
}
```

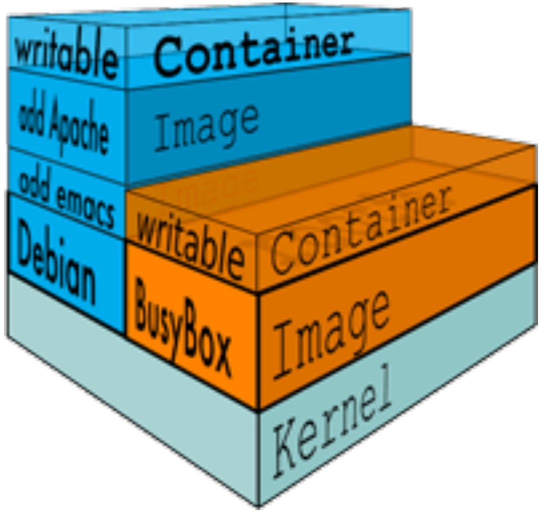
JSON – HOW DOES IT RELATE TO DATA SCIENCE?

- Webservices provide application programming interfaces (APIs) are now usually transferring data via JSON
- Underlying document databases like MongoDB
- Increasingly common data format


docker

DOCKER – WHAT IS IT?

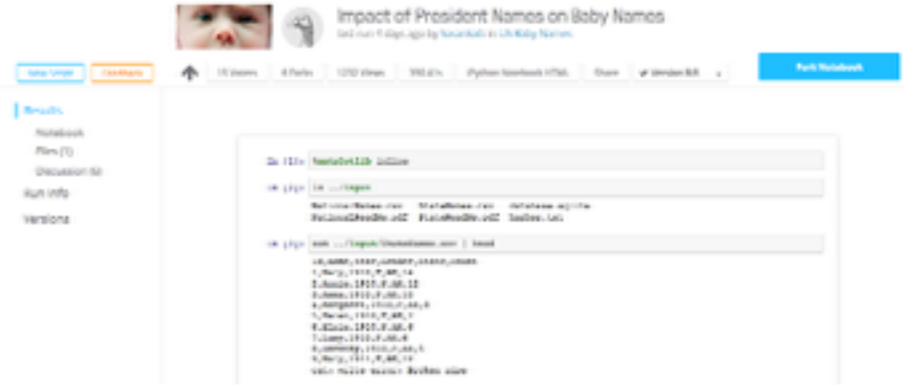
Docker containers wrap up a piece of software in a complete filesystem that contains everything it needs to run: code, runtime, system tools, system libraries – anything you can install on a server. This guarantees that it will always run the same, regardless of the environment it is running in.



- Lightweight
- Open
- Secure

DOCKER – HOW DOES IT RELATE TO DATA SCIENCE?

- Installing data science software can be a pain because of software dependencies and different OS environments. Docker helps solve this problem
- See Kaggle Scripts



DATA SCIENCE – Week 8 Day 1

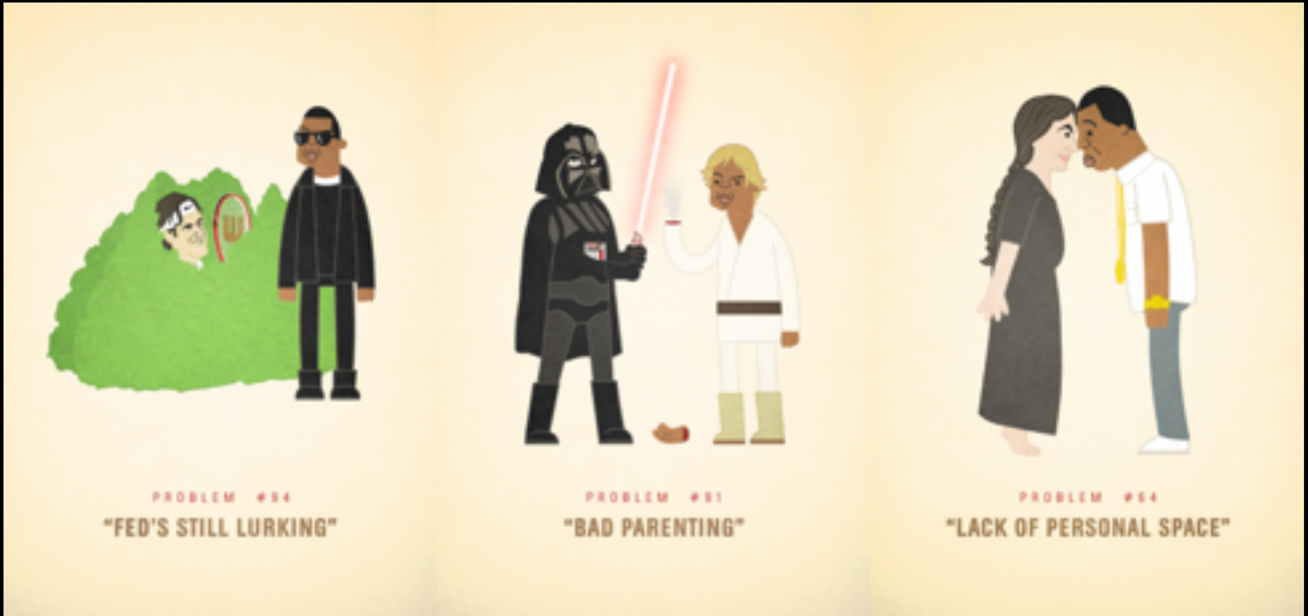
LAB

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

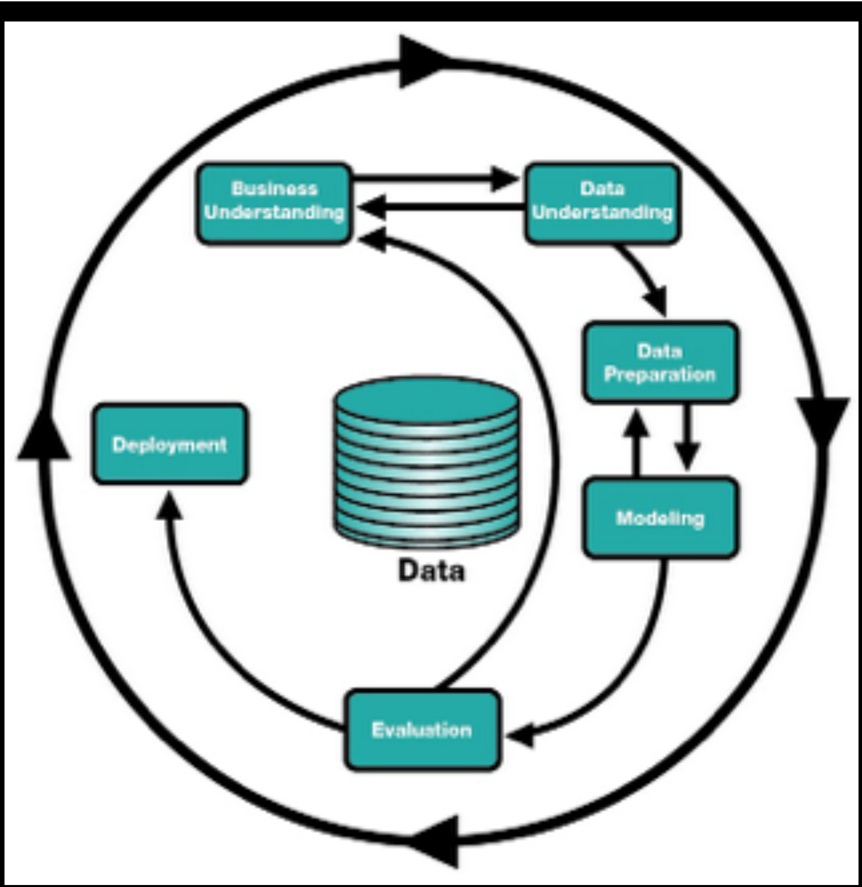
DISCUSSION TIME

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

REAL PROBLEMS



DATA SCIENCE – Week 8 Day 1



DATA SCIENCE – Week 8

REVIEW

DATA SCIENCE - Week 8 Day 1

Task List (25 mins)

□ Read Natural Language Processing website - http://www.nltk.org/ (5 mins)

□ Read and be able to explain one use case of the Alchemy API (10 mins)

□ Download and install NLTK for Python (10mins)