GENERAL ASSEMBLY

DATA SCIENCE
11 WEEK PART TIME COURSE

Week 7 – Technology Evaluation
Wednesday 3rd February 2016

1. Tasks from Monday
2. No-SQL
3. JSON
4. Docker
5. Spark
6. Evaluating Technologies
7. Lab
8. Real World Problem
9. Review

- EC2
- RDS
- mongoDB
- Redshift
- Spark

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

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

NoSQL

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

There are four general types of NoSQL databases, each with their own specific attributes:

Graph database

Based on graph theory, these databases are designed for data whose relations are well represented as a graph and has elements which are interconnected, with an undetermined number of relations between them. Examples include: Neo4j and Titan.

Key-Value store

These databases are designed for storing data in a schema-less way. In a key-value store, all of the data within consists of an indexed key and a value, hence the name. Examples of this type of database include: Cassandra, DyanmoDB, Azure Table Storage (ATS), Riak, BerkeleyDB.

Column store

Instead of storing data in rows, these databases are designed for storing data tables as sections of columns of data, rather than as rows of data. While this simple description sounds like the inverse of a standard database, wide-column stores offer very high performance and a highly scalable architecture. Examples include: HBase, BigTable and HyperTable

Document database

Expands on the basic idea of key-value stores where “documents” contain more complex in that they contain data and each document is assigned a unique key, which is used to retrieve the document. These are designed for storing, retrieving, and managing document-oriented information, also known as semi-structured data. Examples include: MongoDB and CouchDB.


{JSON}

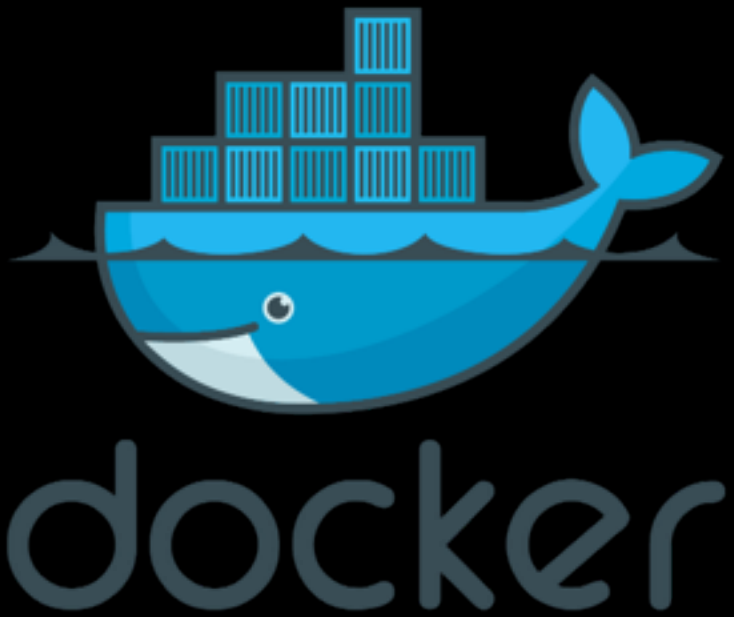
JSON - JavaScript Object Notation

- 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

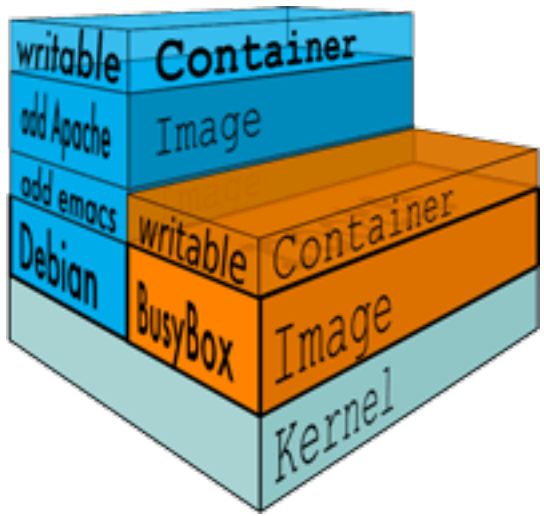
```
var json = {
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "34 York St",
    "city": "Sydney",
    "state": "NSW",
    "postalCode": "2000"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "02 95999999"
    },
    {
      "type":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

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

DATA SCIENCE PART TIME COURSE



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.

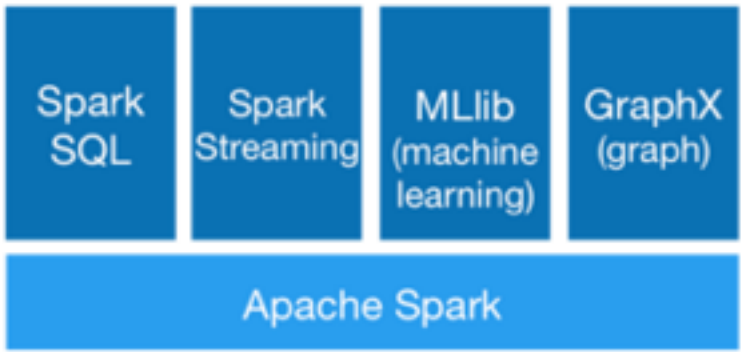


- Lightweight
- Open
- Secure

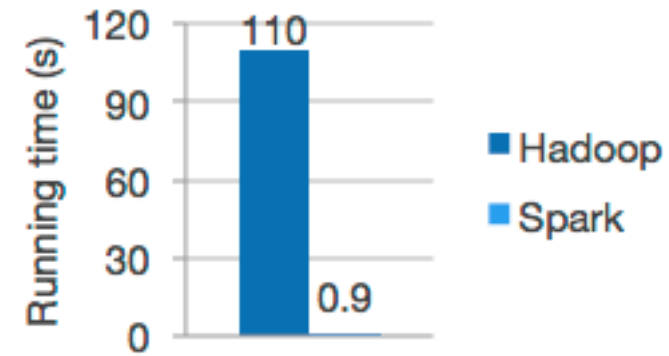
- Installing data science software can be a pain because of software dependencies and different OS environments. Docker helps solve this problem
- See Kaggle Scripts

DATA SCIENCE PART TIME COURSE

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 (ML) 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

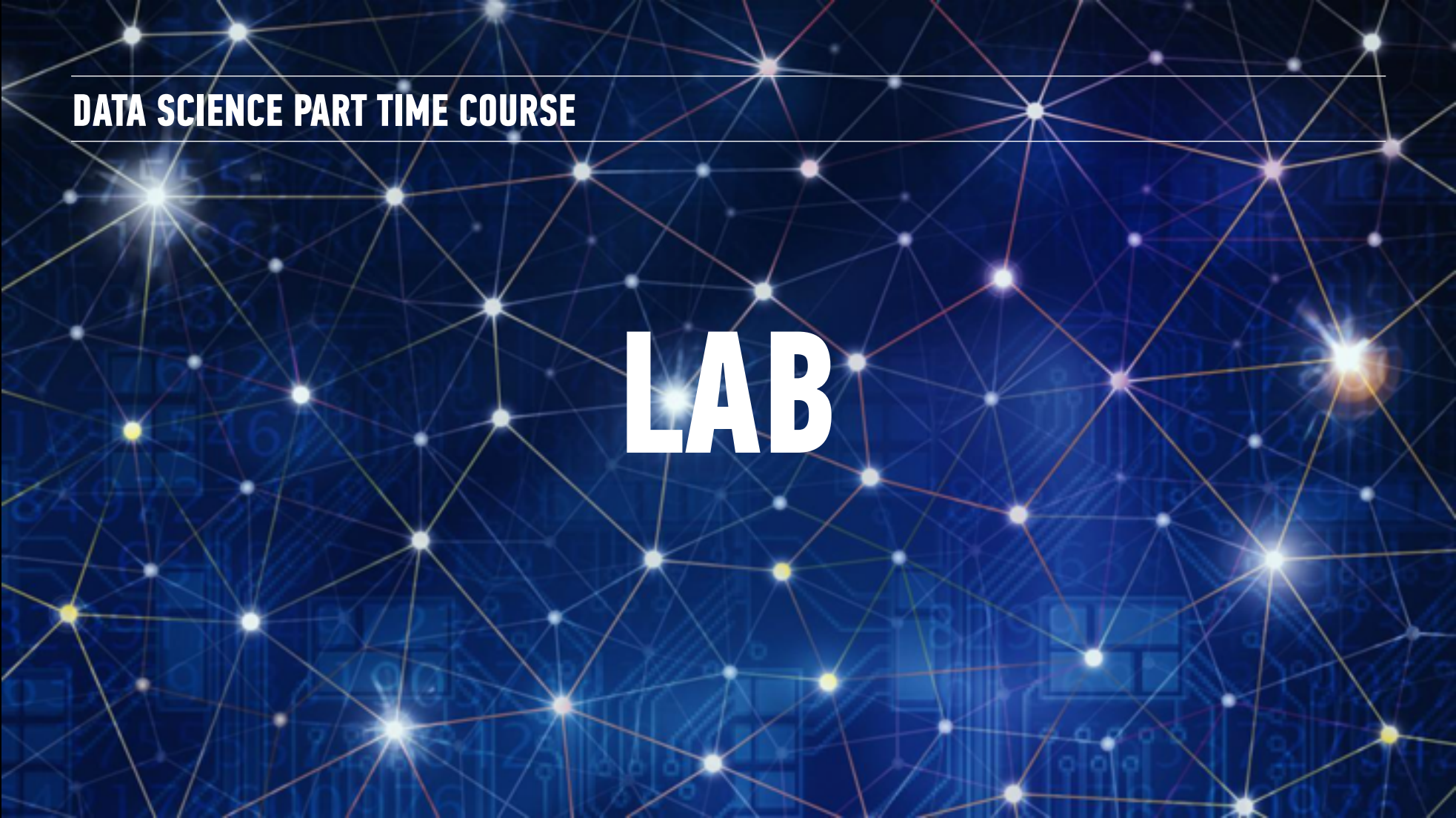


Logistic regression in Hadoop and Spark

DATA SCIENCE PART TIME COURSE

EVALUATING TECHNOLOGIES

DATA SCIENCE PART TIME COURSE



LAB

- Read in the Yelp.json file with Pandas
- Setup a Linux ec2 instance
- Load data into your RDS instance
- Start a Spark cluster with EMR

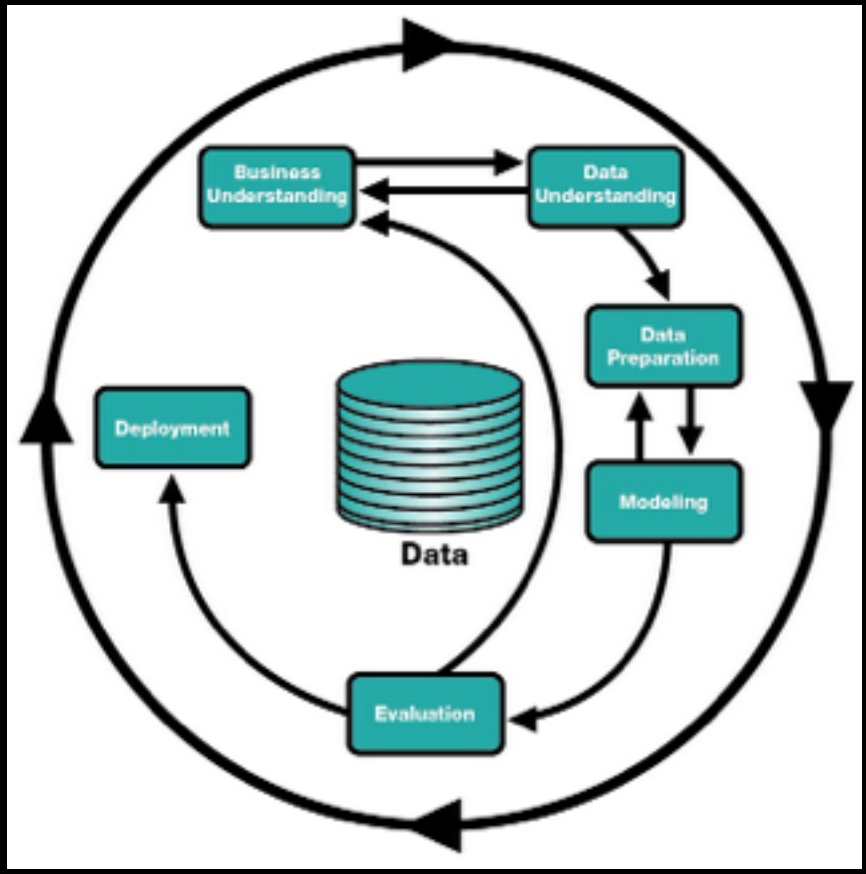
DISCUSSION TIME

- Talk through a real problem
- Questions
- Task List

DATA SCIENCE – Week 7 Day 2

REAL PROBLEMS

DATA SCIENCE – Week 7 Day 2




EVERYBODY
HAS A
PLAN
UNTIL THEY GET
PUNCHED
IN THE
MOUTH
— Mike Tyson

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