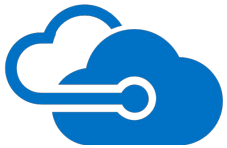
Data Science Tools

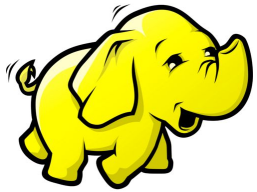
Challenges of Data Engineering

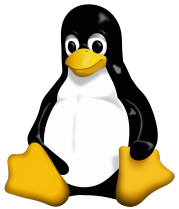
- Efficient retrieval, processing and storage
- High Volume Computing
- Parallel Computing
- Resilience/redundancy

Arsenal of Big Data Tools

- Apache Hadoop
- Apache Spark
- Linux
- Apache Pig
- Hortoworks
- R:
  - parallel
  - doSNOW
- SQL
- Apache Hive







R Gone Wild: parallel

- Package **parallel**
- install.packages(“parallel”)
- library(parallel)
- Easy transition from non-parallel code
- Remember to load variables and packages using
  - clusterExport for variables
  - clusterEvalQ for packages

The Battle of Clouds

1) **AWS**
2) **Microsoft Azure**
3) **Google Cloud**

Offers both high/low-level modules

Allows costs to be more variable

Local cluster(s)

- Easier to control
- Easier to personalize
- Large short-term expense
- May not optimal for cost

High Level vs Low level

High level: UIs, SQL, R, Python, Azure

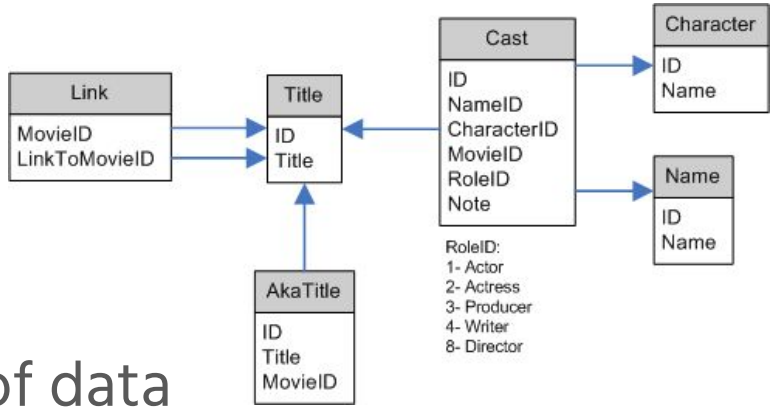
1) Easy to use, easy to understand
2) When things go wrong.. =(

Low Level: Linux, C, C++ (General use of terminals)

1) Much more tedious and complicated
2) More control

Relational Data

Charts, tables (the conventional)

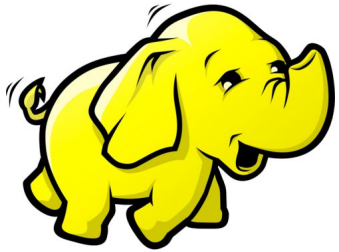


1) Represents traditional form of data
    a) Example: accounting books, information on students
2) SQL traditionally used to handle this data mysql, psql
3) Cannot effectively represent images, text, video

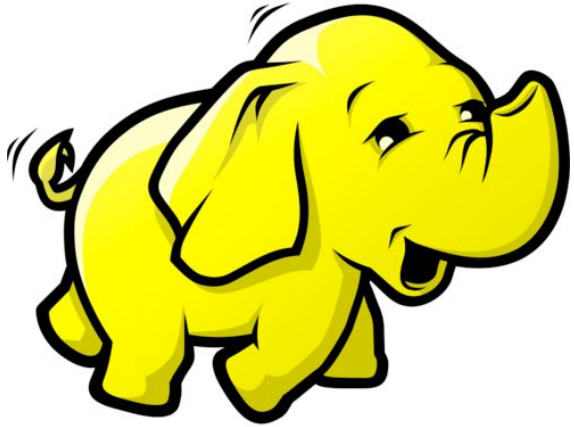
Non - relational Data

Images, music, video, text

1) Sparsity issues if represented by a matrix
2) Large amounts of data in a few files
3) SQL traditionally used to handle this data
4) Examples: images, text, video

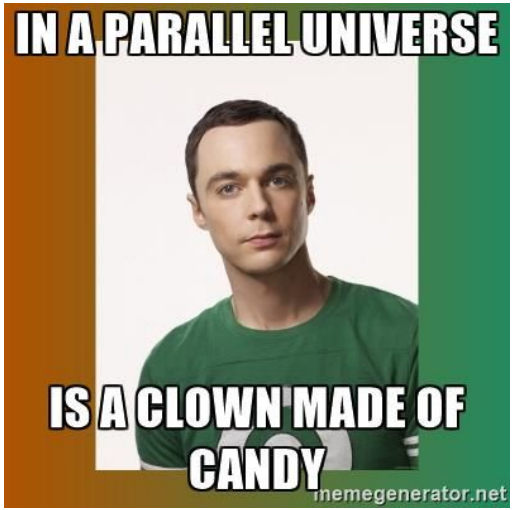


Hadoop Ecosystem



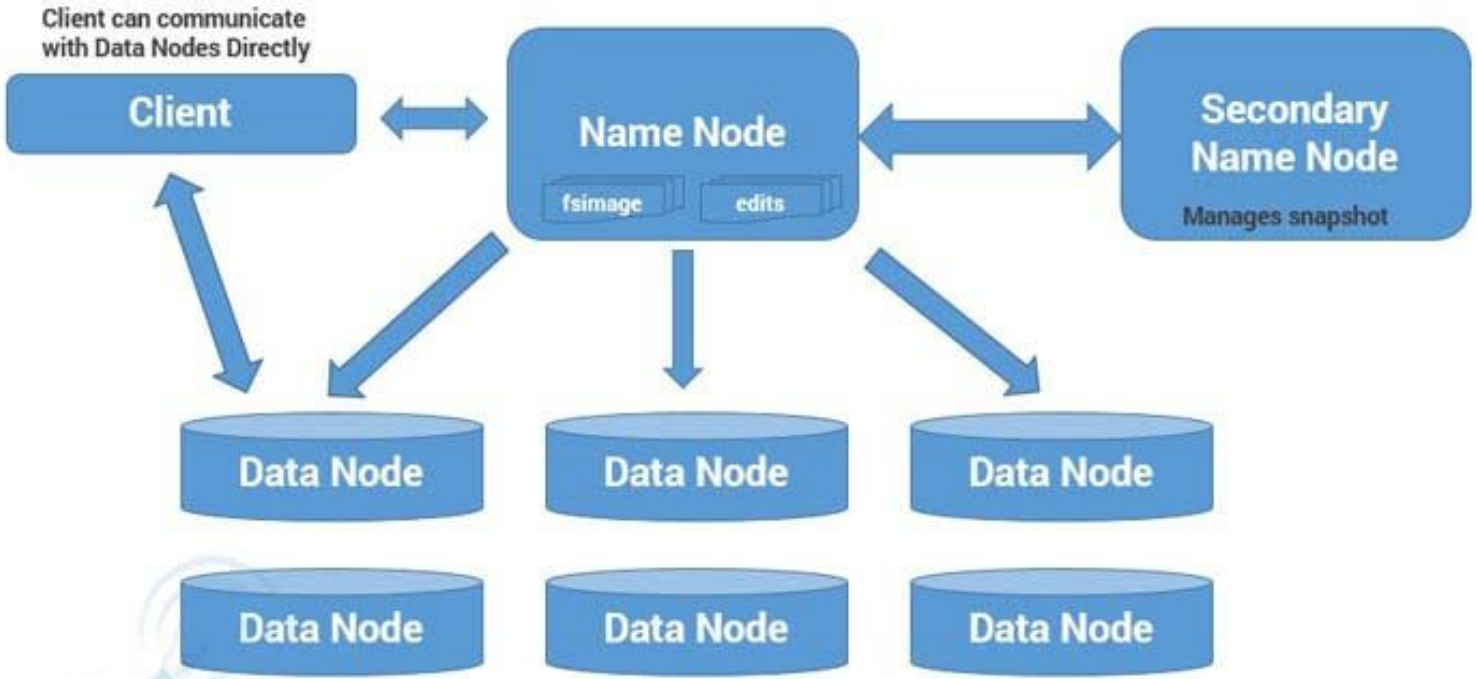
- A non-relational file system
- Very scalable, parallel architecture
- A high-latency, high throughput system
- Built-in resilience and redundancies
- Uses large file blocks to maximize throughput

Parallel Computing Terminology



- Core: processing unit, usually a cpu
- Chip: a chip that contain cpus
- Socket: Physical connector to a chip
- Node: A single unit that can store, send, and receive information (servers)
- Thread: a single process that can be run concurrently on a core

Hadoop Components: HDFS



http://backtobazics.com/wp-content/themes/twentyfourteen/images/hadoop/hdfs-1.x-architecture.jpg

Hadoop Components #2:YARN

https://cdn.edureka.co/blog/wp-content/uploads/2014/05/YARN-431x300.jpg

Hadoop Ecosystem

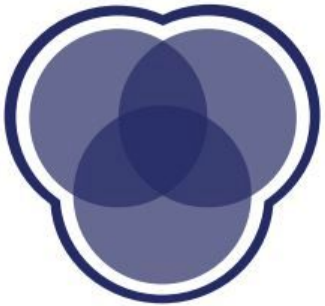
- MapReduce
- Spark
- Hive: SQL platform
- Pig: High-level hadoop language
- Tez

Coming Up

Your assignment: Project 3 and survey

Next week: How to be lit in the Summer

See you then!



CDS

Cornell Data Science