Optum Day 2

Data engineering Lecture 1

What is DE - hacking skills, math & stats, Domain expertise.

kinda like data science + software engineer

science - enigneer: what's the difference:

Why do we need this? - nexus of tech: tech is cheap, storage is available, computing power is near unlimited , and the data exists.

Free resources such as software, data, and courses, open source etc

high scale, low latency expectations

Tools

Apache Hadoop - what is it?

distributed large scale proccesing computer for data

Apache Kafka

large scale streaming

Apache Airflow

large scale machine learning pipelines

AWS

large scale cheap cloud storage infrastructure

Apache Hive

SQL like grammar based on Hadoop

PyTorch

automatic differentiation engine

Principles

Highly extensible ie more than GUI, highly configurable, pipelining systems, specification of dependencies, consistent grammar, parallel + distributed processing

Do we need to learn all of these to be a DE?

we can just learn, unix, sql, tmux, and Make etc.

Unix CLI - approx 3 weeks

Navigation

"pwd" - print working dir

"cd -" go back to prev dir

" cd .. " go back one dir

"ls" list out contents

"ls -all" list out all contents including hidden and extras in long format

"ls - ." list out files and fir in home and wor

"ls -l" long formatting

"ls * .[file type]" list all things with this file type

"ls -a " all files in short format

“man” Manuel
“man [command” command instructions and more
“cp [source] [target]” copy
“mv [source] [target]” move
“ls /[dir]” list out what’s in dir whilst being in another dir
“touch” to create/modify a new file
“mkdir” make dir
“mkdir -p” iterateive dir
“cat [file]” concatenate files
“more [file]” view a file in a Line by line manner
“wc [file]” word count
“wc -l” line count
“head” & “tail” beginning and ending of a file