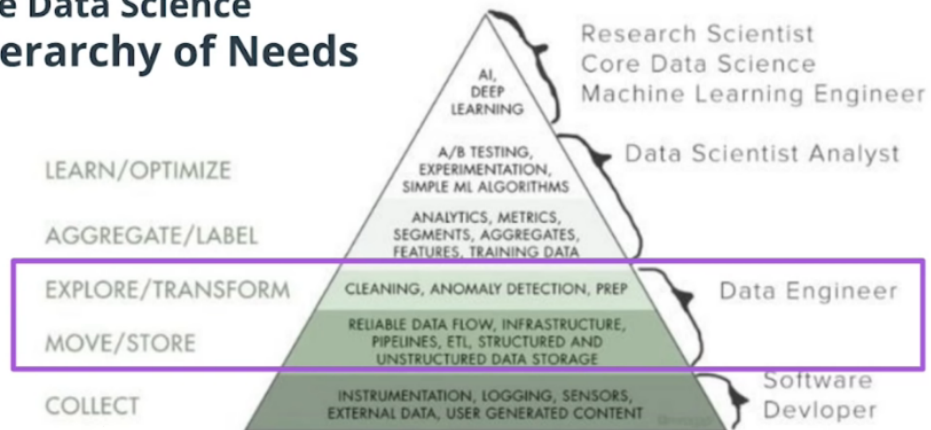


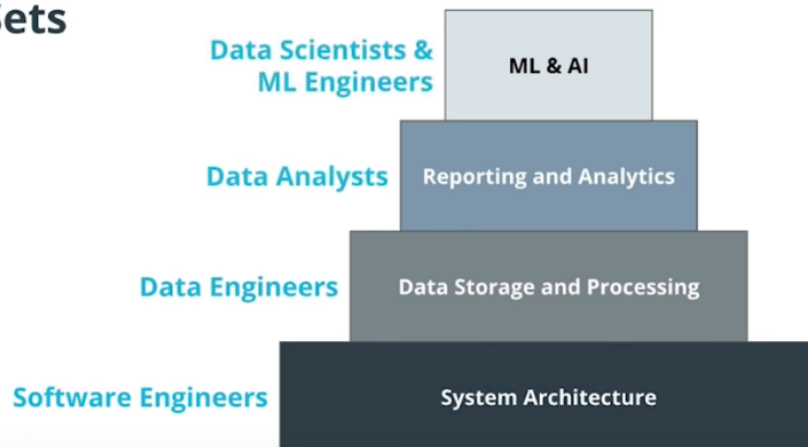
Welcome to the Nanodegree Program

The Data Science Hierarchy of Needs



<https://www.nextacademy.com/blog/how-to-become-a-data-scientist-in-2019/>

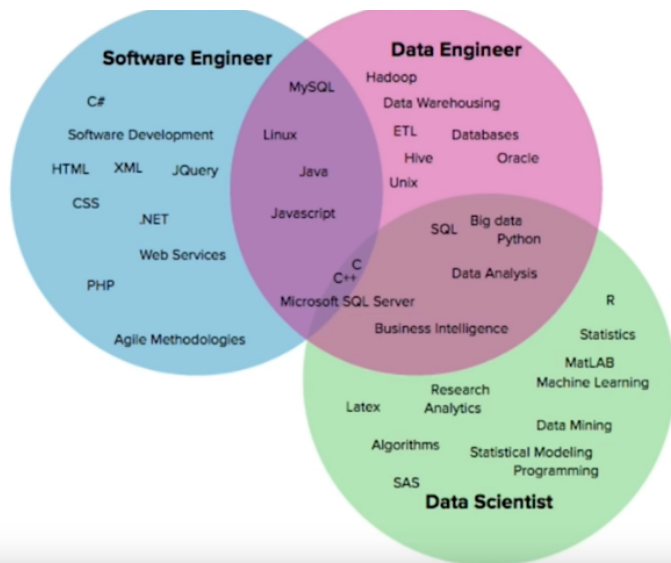
Data Roles and Skill Sets



Common Data Engineering Activities

- Ingest data from a data source
- Build and maintain a data warehouse
- Create a data pipeline
- Create an analytics table for a specific use case
- Migrate data to the cloud
- Schedule and automate pipelines
- Backfill data
- Debug data quality issues
- Optimize queries
- Design a database

Data Roles and Skill Sets



Histoire

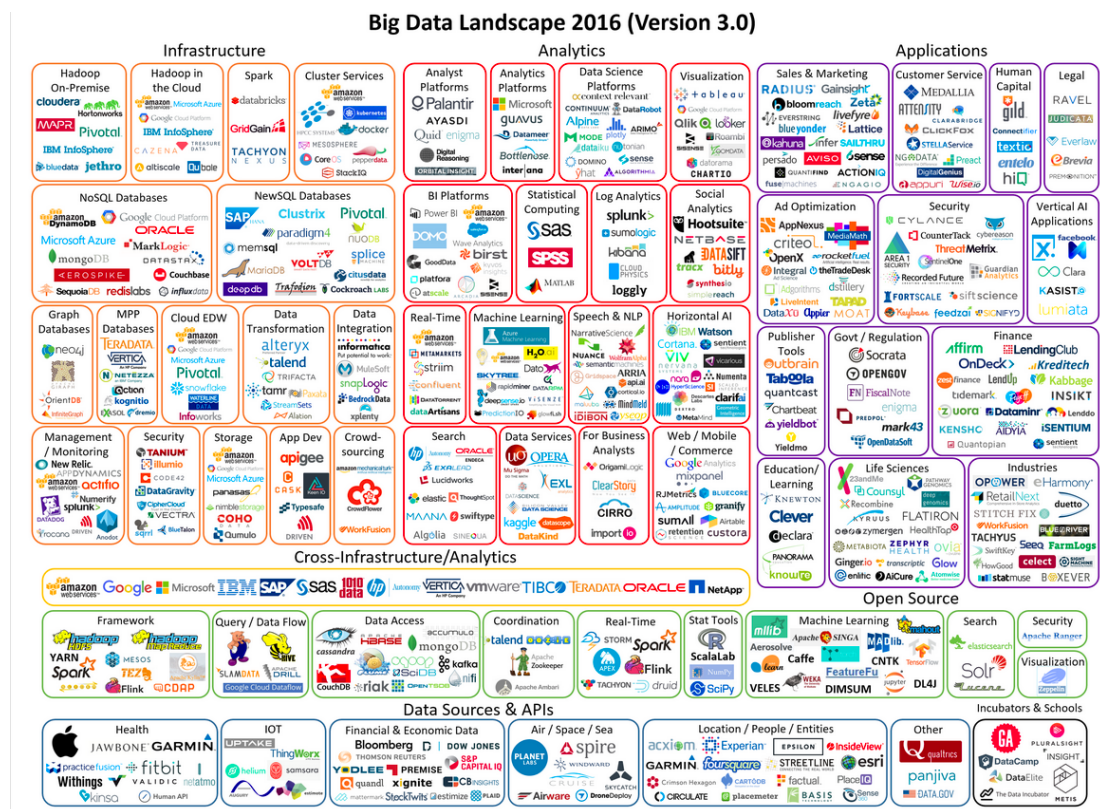
[ON the evolution of Data Engineering \(https://medium.com/analytics-and-data/on-the-evolution-of-data-engineering-c5e56d273e37\)](https://medium.com/analytics-and-data/on-the-evolution-of-data-engineering-c5e56d273e37)

- Big data, growing number of technologies
- 2006, with Hadoop, more easier and cheaper to store large amount of data
- Data Orchestration Engines, python ecosystem in the PyData
- Machine Learning: improvement in SKLearn (2007), produce ready workflow for moderately sized datasets
- Spark and Real-time: 2014, democratized machine learning computation on Big Data, a way to easily process streaming data, window toward real-time processing
- Cloud development and Serverless. AWS launched in 2006. construit sur Hadoop, Elastic Map Reduce launch in 2009, making easier spin up and scale Hadoop clusters
- move to the cloud : storage and compute was infinite one can pay for it

[Data Engineering Introduction and Epochs \(https://learn.panoply.io/hubfs/Data%20Engineering%20-%20Introduction%20and%20Epochs.pdf\)](https://learn.panoply.io/hubfs/Data%20Engineering%20-%20Introduction%20and%20Epochs.pdf)

- 60s, computers more popular, development of language and scripting language, so programming possible for more people. Cobol, Fortran, Algol
- 70s & 80s built the framing for the data information engineering structure
 - by understanding how data is be used, appropriate struct can be created ... Naur
 - Data processing, driven information engineering. Planning Analysis, Design Construction. travail pour comprendre les business process pour que la data soit bien utilise
 - business centric and no more history driven. create an information system for where the business was going.
- 90s digital storage became more cost-effective than paper. so digital data became more accessible, more manipulate, cheaper
 - people view data
 - the use of data to boost top-line revenue
- 00s emergence of Mainstreaming of big data
 - Volume, key of big data, petabytes, exabytes, zettabytes, require specialized tools
 - Velocity, the amount of data captured increases, new way to store data, which impact storage et networks. need to find ways to process et analyze data far more quickly
 - Variety. more data more quickly, different type such as the internet of Things
 - Variability: data today no have the same meaning tomorrow. Ensuring that you're able to assign consistent meaning to that data, even as underlying conditions change, is increasingly necessary
 - Veracity, for big data systems to be reliable and usable, you need to be able to ensure that the data is also accurate.
 - Visualization, turning raw data elements into beautiful images that the viewer can immediately grasp. charts, infographics, spreadsheets, make underlying information visually accessible
 - Value. objectif de generer de la valeur pour l'entreprise, reduire les couts, accroitre les revenus, rationaliser les processus

Data Engineering Tools



Dataflog (<https://dataflog.com/big-data-open-source-tools/os-home/>)

Joviam (<https://joviam.com/this-infographic-of-big-data-tools-will-blow-your-mind-infographic/>)

sites et Livres (<https://analyticsindiamag.com/data-engineering-101-top-tools-and-framework-resources/>)

Burtch works (<https://www.burtchworks.com/2018/09/10/the-rise-of-data-engineering-common-skills-and-tools/>)

Entrée []: