Data Engineering

A Brief Overview

Chip Young I320 D

Data Engineering

An Introduction

What Is Data Engineering?

What Do Data Engineers Do?

What Are Their Roles In An Organization?

Why Is Data Engineering Important?

What Do Data Engineers Build?

What is Data Engineering?

And Why Does it Matter?

- Data Engineering is the process of taking raw operational data and extracting it, cleaning it, transforming it, analyzing it, and publishing it to make it meaningful and useful to people and organizations.
- Our customer base as data engineers is, well, *just about everyone*...
- Which means that the veracity, quality, availability, and accessibility of our data is absolutely critical.

Image from http://www.ipdutexas.org/blog/are-you-data-literate

“Operational (Source) Data”
Where does it come from?

- Consumer software systems (Tiktok, Spotify, Amazon)
- Internal systems (Salesforce, CRM, Accounting, HR, etc.)
- Internal business users (Excel spreadsheets)
- IoT devices (solar panels, automobiles, cell phones)
- And everything else you can think of...

“The value of data has become so widely recognized, it likely won’t be long before it’s listed as an asset on a company’s financials.” - NewComp Analytics
“By 2025, the amount of data generated each day will reach 463 exabytes globally.”

What Do Data Engineers Do?

- They build the systems that transform raw source data into meaningful and useful information.
- They do this by building:
  - Pipelines that transform data.
  - Analytics that extract meaning from data.
  - Reports that summarize data.
  - Visualizations that present data.

Are Data Engineers in Demand?

- Data Engineer was the fastest growing tech job in 2019, growing by 50% YoY (Dice Tech Jobs Report).
- “The incentive to invest in these jobs is strong, as there is gold in the data - from more timely and effective data-driven strategy and decision-making to data productization, which opens up new growth and revenue centers”
- $106,000 was median salary for Level 1 data engineers (Information Week) in 2019.
- Dice Tech Jobs report from 2023: Average salary for data engineers is $122,811.
- $145,235 average for data architects (BuiltIn.com).

With a booming growth rate of 30.7 percent, the global big data analytics market is projected to be worth $346.24 billion by 2030, boosting the fortunes of data engineers, business analysts and data analysts.

“.. a search [on LinkedIn] for data engineer jobs produced nearly 230,000 [open positions].”
“As the data industry booms, we are in the middle of a widely publicized data engineer shortage. By some estimates, we are more than 5 years into this striking disparity in talent supply and employer demand.” - https://towardsdatascience.com/a-lack-of-institutionalized-education-fuels-the-data-engineering-shortage-d2d7c5ac7a81

Data Engineering Roles

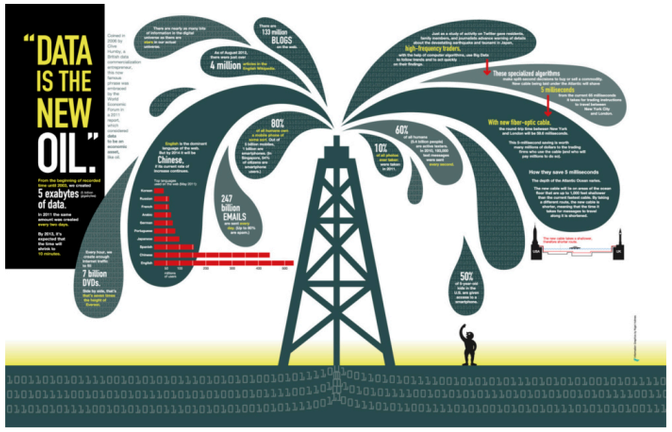
An Ever-Expanding List

- Data Engineer as infrastructure engineer (data engineering + devOps = dataOps)
- Data Engineer as software engineer (“data intensive” systems)
- Data Engineer as data scientist
- Data Engineer as analytics engineer
- Data Engineer as knowledge engineer

“Data is the New Oil”

Hype and Reality

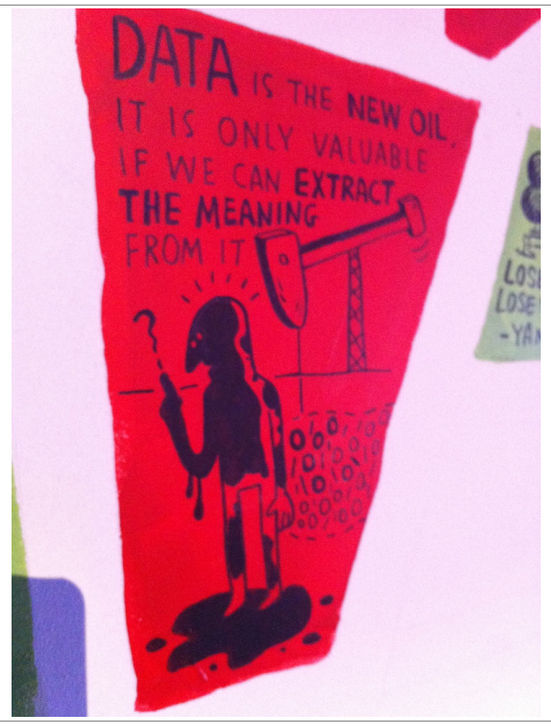
- “Big Data” - Web 2.0
- By 2025, the amount of data generated each day will reach 463 exabytes globally.
- “The value of data has become so widely recognized, it won’t be long before it’s listed as an asset on a company’s financials”
- What do we do with all that raw data?



“The value of data has become so widely recognized, it likely won’t be long before it’s listed as an asset on a company’s financials.” - NewComp Analytics
By 2025, the amount of data generated each day will reach 463 exabytes globally

How is Data Like Oil?

It has to be refined to be useful

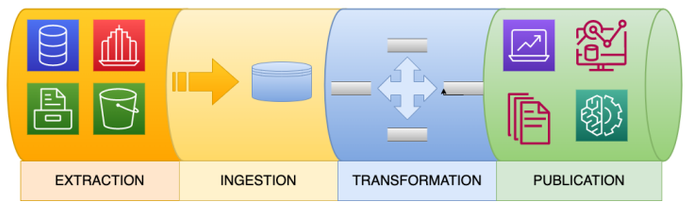


Data Pipelines

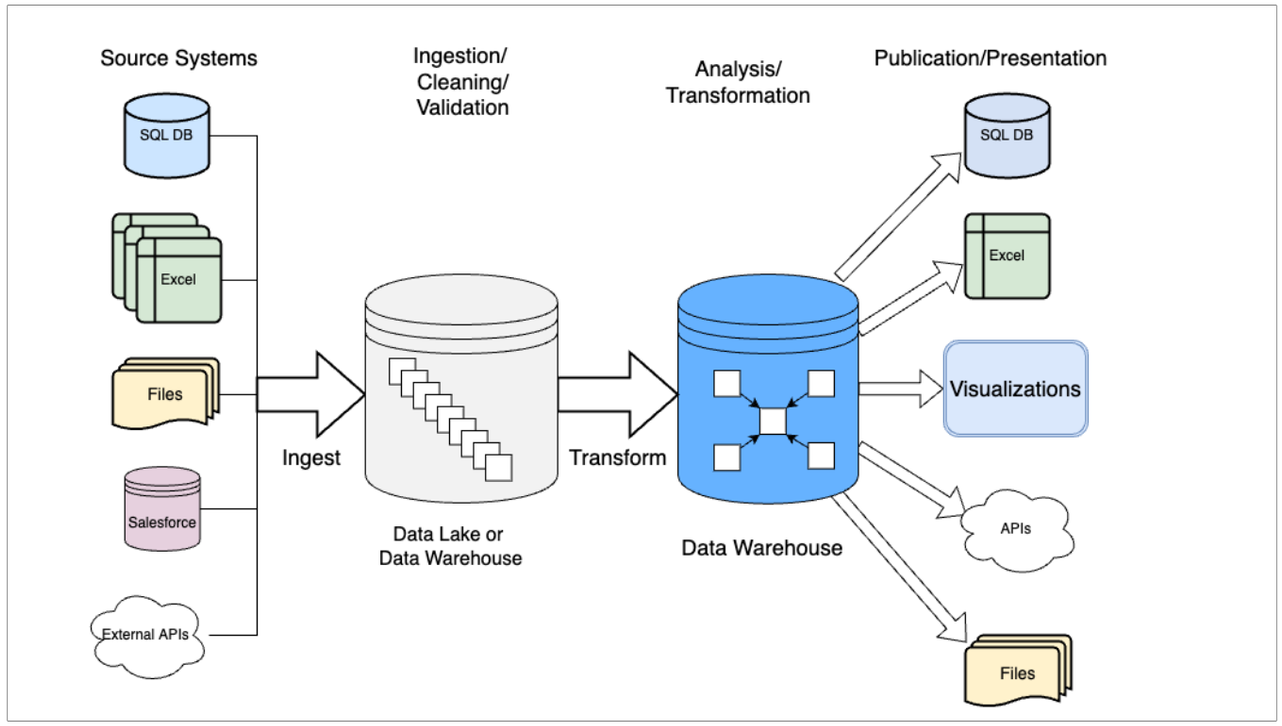
Transform Data into Information

- Pipelines are sequences of tasks that combine raw source data and transform it into useful information
- Pipelines have the following stages:
  - Extraction
  - Ingestion
  - Transformation
  - Publication/Presentation

Conceptual Data Pipeline

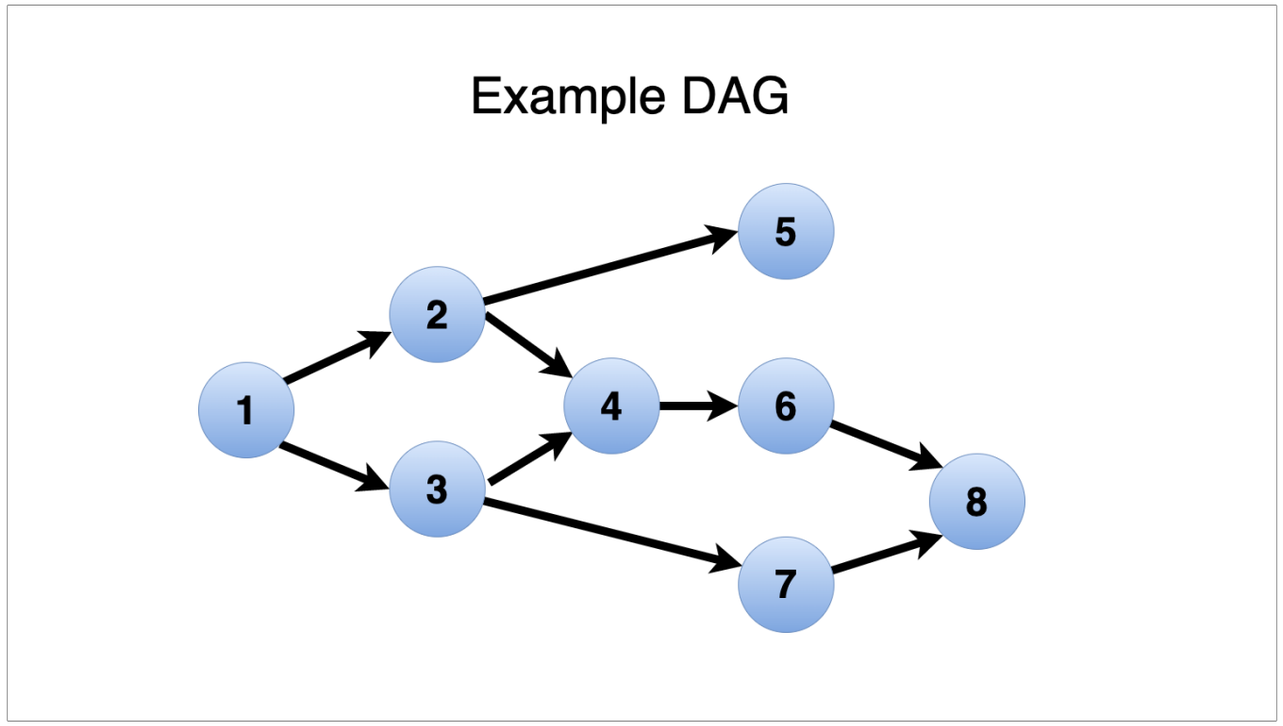


How do we refine raw data and turn it into useful forms of information?
We process it in *pipelines*


Source Systems
Ingestion/
Cleaning/
Validation
Analysis/
Transformation
Publication/Presentation
SQL DB
Excel
Files
Salesforce
External APIs
Ingest
Data Lake or
Data Warehouse
Transform
Data Warehouse
SQL DB
Excel
Visualizations
APIs
Files

How is Data Transformed?

- In a pipeline, we execute a series of *tasks:*
  - Tasks are programs that read data from a *source*, transform it in a *process*, and write it to a *sink*.
    - *Sources* and *Sinks* are some kind of permanent storage (a database, a file, a log, etc.).
    - *Processes* are written in some language and execute in memory.
  - The series of tasks follows an specific sequence of execution.
  - The sequence can be represented as a *graph,* specifically a *directed acyclic graph* - commonly referred to as a *DAG*.

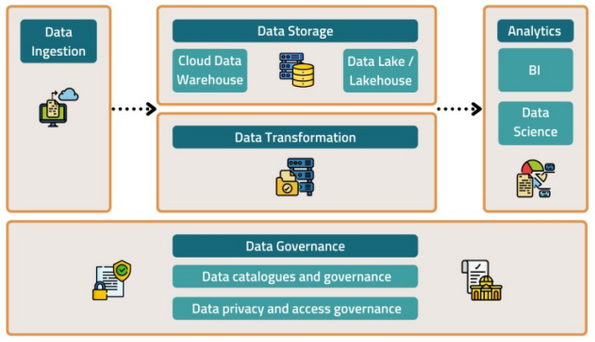


A DAG is a graph consisting of vertices
(nodes) and edges (lines) where the edges
connect the vertices.
- starting from a single vertice
- edges join each of the vertices (a graph)
- in one single direction (directed)
with no cycling back to previously joined
vertices (acylic)
This represents the order of execution of a series of tasks
The order implies *dependencies* - 8 cannot execute until 6 and 7 successfully complete, 6 cannot execute until 4 completes, 7 until 3, 4 until 2 and 3, 5 until 2, etc.

The Modern Data Stack

This Year's Model

- The MDS is a collection of platforms, tools and technologies for delivering, managing, and analyzing data.
- We will use some of those tools in this course, like DBT and Superset.

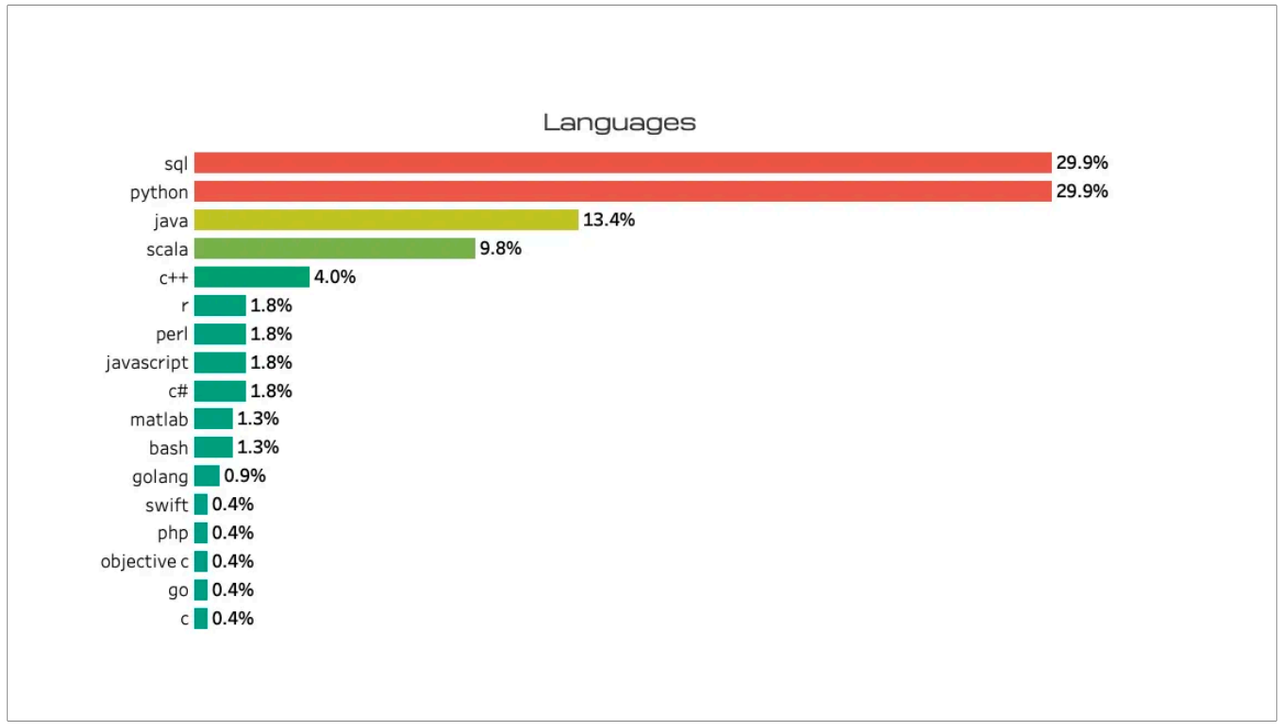


Data Engineering
An Introduction

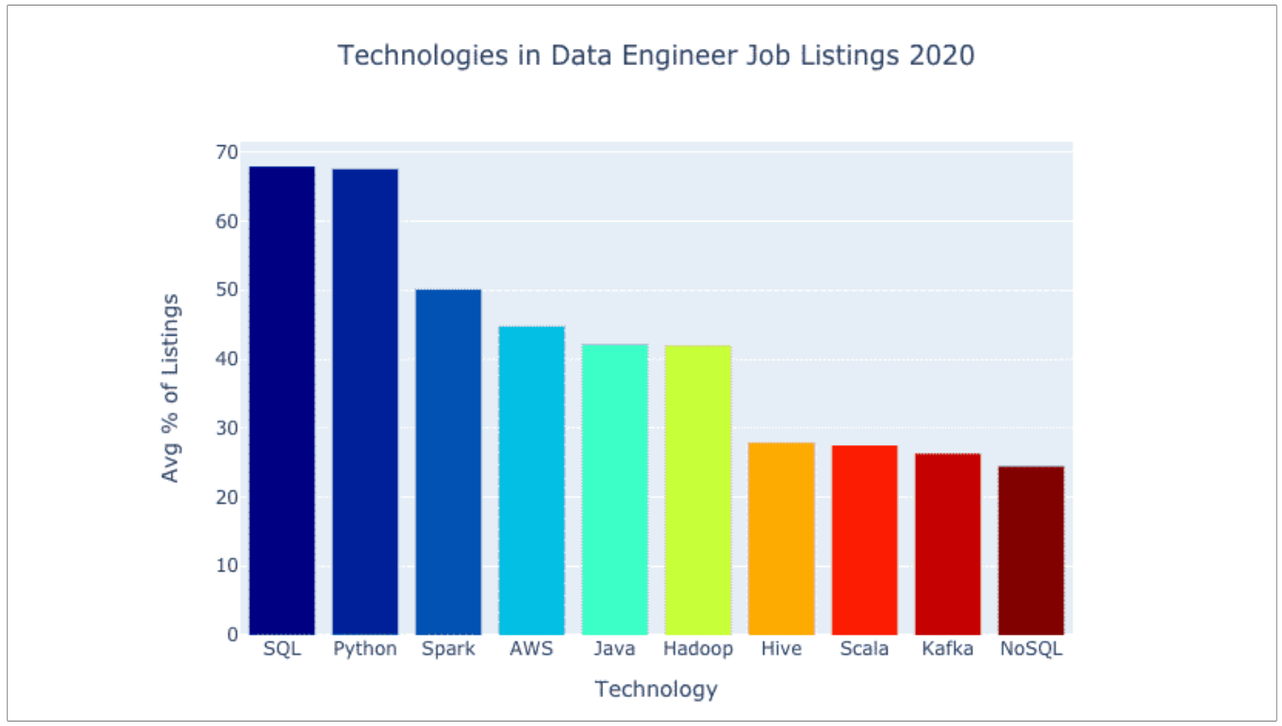
- We discussed:
  - What Data Engineering is.
  - What Data Engineers do.
  - What roles Data Engineers have in an organization.
  - How much demand there is for Data Engineers.

Data Engineering
An Introduction

- We discussed the analogy of “Data is the new Oil”
  - Like oil, data has to be refined to be useful.
  - We build “pipelines” to refine our data
  - These pipelines have several distinct stages:
    - Ingestion from source systems
    - Cleaning and validation
    - Analysis and transformation
    - Publication and presentation
  - We build our pipelines using platforms, tools, and technologies of the Modern Data Stack



https://blog.devgenius.io/become-a-data-engineer-in-2022-analysis-of-over-1-000-faang-job-postings-38784fa727a8


Technologies in Data Engineer Job Listings 2020
Avg % of Listings
0
10
20
30
40
50
60
70
SQL
Python
Spark
AWS
Java
Hadoop
Hive
Scala
Kafka
NoSQL
Technology