DATA ENGINEERING 101

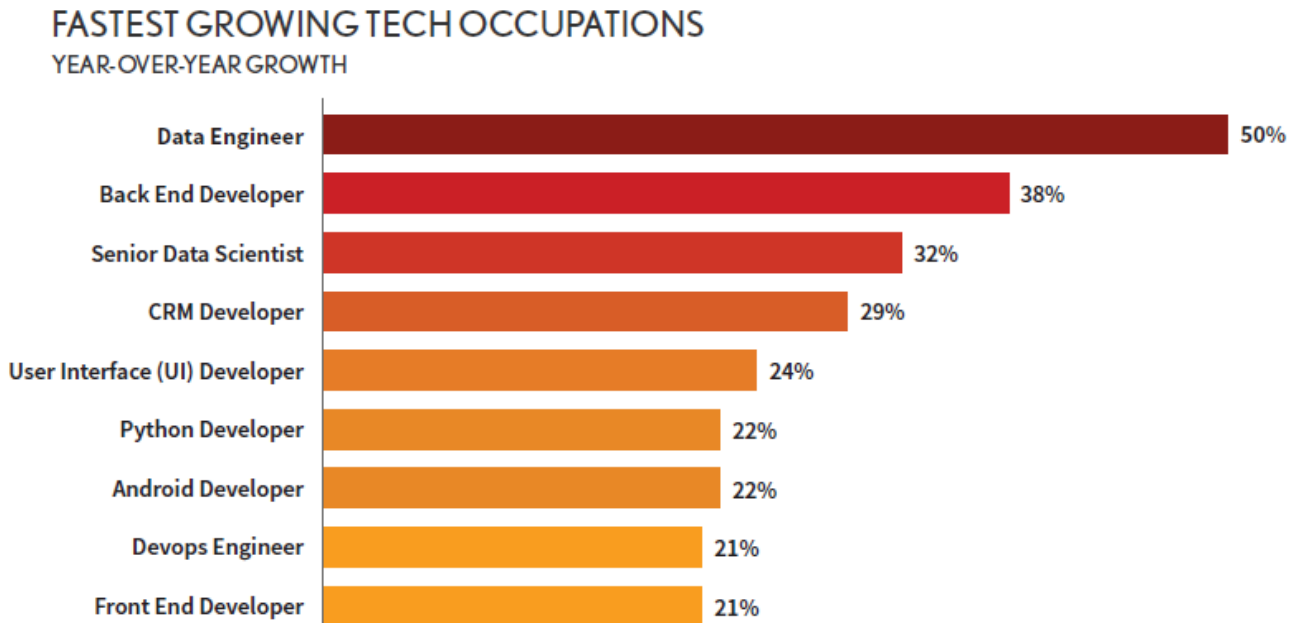
Introduction to Data Engineering

A Q&A for the most frequently asked questions about data engineering.

Xinran Waibel Follow
Feb 23 · 6 min read ★

According to the recently published Dice 2020 Tech Job Report, data engineer was the fastest-growing tech occupation in 2019, with a 50% year-over-year growth in the number of open job positions. As data engineering is a relatively new job category, I often get questions about what I do from people who are interested in pursuing it as a career. In this blog post, I will share my own story of becoming a data engineer and answer some frequently asked questions about data engineering.

> *This article is part of the Ask Me Anything series at Towards Data Science (TDS). This column provides TDS readers with the unqiue opportunity to have their Data Science questions answered by the TDS team.*



Source: The Dice 2020 Tech Job Report

. . .

My journey to data engineering

A couple of years ago, before becoming a data engineer, I mainly worked on database and application development (plus scale and performance testing). I loved working with RDBMS and data so much that I decided to pursue an engineering career that focuses on Big Data. After researching online, I learned about a type of job I had never heard of before but thought might be the "Mr. Right" for me: data engineer. I was lost at the time because I did not know where to start or whether this would work out for me. However, I was lucky enough to have a mentor who showed me the rope and helped me land my first data engineering job. Since then, I have been working full time in the field and am still loving it!

Ask Me Anything

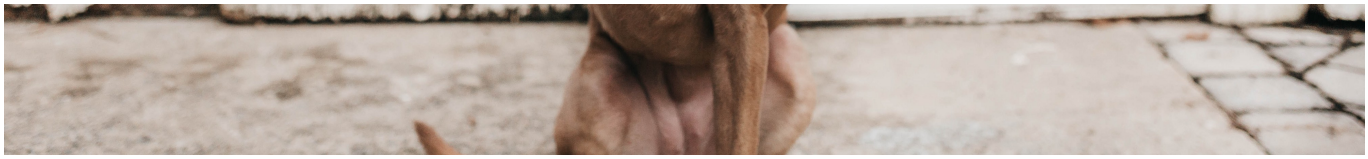


Photo by Camylla Battani on Unsplash

Q: What does a data engineer do?

In short, data engineers are responsible for designing, developing and maintaining the data platform, which includes the data infrastructure, data applications, data warehouse, and data pipelines.

In a big company, data engineers are usually divided into different groups that work with a specific part of the data platform:

- **Data warehouse & pipelines**: Data warehouse engineers build batched and/or real-time data pipelines to integrate data between systems and also support the data warehouse. Since the data warehouse is meant for tackling business problems, data warehouse engineers usually work closely with data analysts, scientists, or business teams that serve a specific business function.
- **Data infrastructure**: Data infrastructure engineers build and maintain the very foundation of data platform: the distributed systems that everything runs on top of. For example, at Target, the data infra team maintains Hadoop clusters used by the whole organization.
- **Data applications**: Data application engineers are software engineers building internal data tools and APIs. Sometimes, a great internal tool may later become an open-source product of the company. For example, one of the data product teams at Lyft built a data discovery tool called Amundsen which was open-sourced in 2019.

Q: What is a data warehouse?

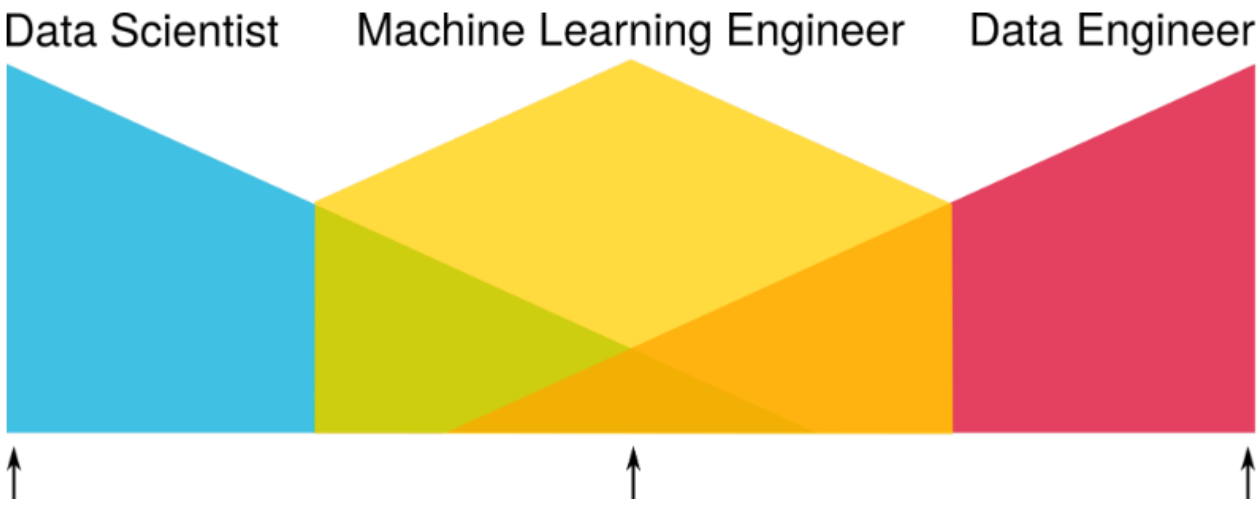
A data warehouse is a data storage system filled with data from various sources and is mainly used for data analysis. A company's data is often stored in different transactional systems (or even worse, as text files) and transactional data is highly normalized and suboptimal for analytics. The main reason for building a data warehouse is to store all types of data in optimized formats in a centralized place so that data scientists can

analyze this data altogether. There are many databases that serve well as a data warehouse, such as Apache Hive, BigQuery (GCP), and RedShift (AWS).

Q: What is a data pipeline?

A data pipeline is a series of data processes that extract, process and load data between different systems. There are two main types of data pipelines: batch-driven and real-time:

- **Batch-driven**: Batch data pipelines only process data at a certain frequency and are often scheduled by a data orchestration tool, such as Airflow, Oozie, or Cron. They usually process a large batch of historical data all at once, therefore taking a long time to finish and inducing more data delay at the end system. For example, a batch-based data pipeline downloads the previous day's data from an API at 12 AM every day, transforms the data, and then loads it into a data warehouse.
- **Real-time**: Real-time data pipelines process new data as soon as it is available and there is almost no delay between the source and end system. The architecture for real-time data processing is very different from that of batched pipelines because data is treated as a stream of events instead of chunks of records. For example, to rebuild the pipeline mentioned above into a real-time pipeline, an event streaming tool like Kafka is needed: a Kafka Connector will stream data from the API into a Kafka topic, and a Kafka Streams (or Kafka Producer) process will perform the transformation on the raw data from the Kafka topic and load the transformed data into another Kafka topic. The delay between the source API and the destination Kafka topic may be within a second!



Research ML/AI	Operationalizing ML	Adv. Programming
Adv. Analytics	Optimizing ML	Distributed Sys.

Source: Data Engineer vs. Data Scientists

Q: How is a data engineer different from a data scientist?

Data engineers build the data platform which enable data scientists to analyze data and train Machine Learning (ML) models. Sometimes data engineers also need to do data analysis and help data scientists integrate ML models into data pipelines. In some data teams, you might find data scientists do data engineering work. There are several overlapped skills between data engineers and data scientists: programming, data pipelining and data analysis.

There is an emerging role called ML engineer that builds a bridge between both worlds. An ML engineer wields strong skills in both engineering and machine learning and is responsible for optimizing and productionalizing ML models.

Photo by Christopher Gower on Unsplash

Q: To become a data engineer, what programming languages should I learn?

The short answer is: Python, Java/Scala, and SQL.

For the data application track, you also need to learn common programming languages for full-stack development, e.g. HTML, CSS, and JavaScript.

Q: What skills and tools should I learn?

Here is a list of core skills along with one of the popular frameworks:

- Distributed systems: Hadoop
- Databases: MySQL
- Data processing: Spark
- Real-time data ecosystem: Kafka
- Data orchestration: Airflow
- Data science and ML: pandas (Python library)
- Full-stack development: React

Since one-size-fits-all solutions no longer exist in the big data world, each company is leveraging different tools for its data platform. Therefore, I recommend you first learn the foundation knowledge for each core skill, then pick a popular tool to learn in-depth and understand the trade-offs between what you picked and the other tools out there. To get to the next level, you also need to know how all the tools work together in a data architecture.

(Interested in ways to efficiently learn a tech stack? Check out the Systematic Learning Method.)

. . .

Today, about 1.7 MBs of new data are generated per second per human being on the planet, and this data contains huge values that cannot be harvested without data engineering. Data engineers help people make smarter decisions, which why I love my job so much. :)

Feel free to let me know if you have any questions related to data engineering or data science!

(Want to learn more about Data Engineering? Check out my Data Engineering 101 column on Towards Data Science!)