Data Engineering

Our Story

What is the difference between a *data engineer* and a *data scientist*?

Data engineering is a broad field encompassing everything from developing, constructing, testing and maintaining architectures to deploying large scale processing systems and databases. Although the definition of the role varies, the principal responsibility of a data engineer is to develop, test, and maintain a software ecosystem. This ecosystem is the environment that data scientists and software engineers work in. When comparing a data engineer and a data scientist the tools of trade and the skills have some amount of overlap, but the roles are nevertheless quite distinct. Data scientists can only devote a percentage of time to maintaining data ingestion pipelines, deployment tools and other necessary infrastructure. They typically focus most of their work on modeling, training and analyzing data to drive business outcomes.

At AAVAIL, you soon discover that you are going to be doing a lot of data engineering as well as data science. Even in large enterprises like AAVAIL, that is a common situation. The goal of this unit is to help you formalize the process of automating data ingestion. Data ingestion in this context is the *collection of data followed by the readying of that data for use in the AI workflow.*

Overlap of data science and data engineering

At the intersection of the engineer and scientist roles is the data itself. It is generally unavoidable that data scientists must dedicate a portion of time to data wrangling. The data wrangling process can change with the choice of modeling solution—so the data scientist must be intimately aware of the process by which the data was obtained. A data engineer is often tasked with readying data in a way that is easy to consume by data scientists and it is here that the distinction between the two roles becomes less clear. The *readying* of data it turns out has historically been referred to as Extract, Transform, Load or ETL. ETL is a component of the unit, but with the modern landscape of tools the term no longer adequately describes all aspects of this part of the AI workflow.

Besides the data ingestion part of the workflow the other important area where the duties of data scientists and data engineers overlap is at the level of model deployment and this will be discussed during module 6. There is a trade-off between spending time on infrastructure and focusing on the AI workflow itself. Often smaller sized companies do not have data engineers and data scientists perform some of the engineering tasks. It is a delicate balance that needs to be maintained—generally as a company grows there will inevitably be a point where infrastructure needs, like accommodating scale, become so important that a dedicated data engineer will be necessary to maintain efficiency.