DATA ENGINEER VS DATA SCIENTIST: WHAT'S THE DIFFERENCE?

Workplace job titles are often far from accurate or precise. Many employees are quick to point out that their job titles don't align with the work they actually do. Some companies even choose to forego job titles altogether, instead embracing the theory that everyone knows their rule – and sometimes to underscore the idea that hierarchies aren't the best way towards innovation.

In technology in particular, things are little different. It might seem that anyone who works in tech is a programmer, or at least has some programming skills, but with big data on the rise, two jobs are in high demand: data engineers and data scientists.

The positions may sound the same – and companies may think they're the same, with similar job descriptions or candidates. But, they're very different, with less overlap than the names may imply.

As big data integrates into all types and sizes of companies, the positions of data engineers and data scientists are increasingly vital. Let's explore what these job titles mean and how they support two different, both necessary, parts of big data.

Big data growth, big data jobs

In the last two years, the world has generated 90 percent of all collected data. Two years! That means two things: data is huge and data is just getting started. As such, companies are seeking employees who can help them understand, wrangle, and put to use the potential of big data. Data engineers and data scientists are increasingly vital to this effort.

A simple distinction, though not complete or always accurate, is that a data scientist is more math-oriented while a data engineer is more IT-minded. This correlates to necessary job skills: while data scientists and data engineers both possess some analytics and programming skills, the scientist has more advanced analytics skills and the engineer has higher programming capabilities.

But it may be the *way* these skills play out in the workplace that is the key difference. In order for a data scientist to perform data science, a data engineer must first create the structure and provide the data for the analysis. Data pipelines are a key part of data analysis – the infrastructures that gather, clean, test, and ensure trustworthy data. Depending on the business, data pipelines can vary widely: this is the data engineer's specialty.

But once the data infrastructure is built, the data must be analyzed. Enter the data scientist.

Who is a data scientist?

Like scientists who tend to work in universities or R&D environments, data scientists often come from a more academic background. They may have degrees in math, statistics, physics, or a similar type of applies math, and they want to focus on analytics – the discovery, understanding, and communication of data patterns. The results of a data scientist's work may be developing new algorithms or features, extracting data patterns, and visualizing data. At the farther end, a data scientist may build a machine learning model or a form of artificial intelligence.

But data scientists that are employed by companies don't exist in a theory vacuum. Instead, they must understand how their analytical work can inform vital business decisions. As such, a data scientist must be able to translate data findings and then communicate it to business-minded folk.

Programming may be another tool in the data scientist's toolkit, but it's not the sharpest one. While a data scientist may have chosen to pursue a degree in applied math, picking up some basic programming skills was likely more a matter of necessity, in order to complete their analyses.

Day-to-day, a data scientist works with systems like MatLab and Rstudio and languages like Python and R in order to model and analyze data through statistics and algorithms.

Who is a data engineer?

Data engineers likely followed a different route to their profession, with a chosen background in programming, like Java or Python. A data engineer may not necessarily have an advanced degree, depending on his hands-on experience, but he likely has moved far beyond building apps and simple systems to specialize in distributed systems and big data. In that way, a data engineer uses advanced to expert programming skills as well as system creation skills in order to create software solutions for big data.

Because big data is so, well, big, the systems that rein data in for analysis are necessarily complicated and unwieldy. Data pipelines that incorporate massive data sets likely make use of multiple types, sometimes dozens, of big data systems.

Therefore, a data engineer understands not just how to build a data pipeline, but how to combine a variety of technologies and frameworks in order to create the best solution for the business.

A data engineer's toolkit may include systems like Oracle, Hadoop, MySQL as well as languages including Java, Linux, SQL, and JavaScript. The goal of an engineer? To accomplish tasks related to data warehousing, ETL, databases, and business intelligence. (Unlike most programmers, however, a data engineer may not need access to the production-side systems.)

Don’t confuse engineers and scientists

Knowing the difference between your engineers and your scientists will keep everyone happy. For the sake of (seeming) efficiency, cost savings, or a whole host of other reasons, a company or team may seek to fulfill only one position – an engineer or a scientist. Beware: a team who expects a scientist to build the data pipeline or an engineer to develop and understand algorithms will be disappointed and frustrated. In fact, some experts say it’s quite common for a data scientist to be forced into engineer work, but less so vice-versa.

This may be due to the way data engineers tend to develop in technology companies. Data scientists, however, have tended to exist in academic environments or mega-size tech firms until very recently, as smaller companies adapt big data practices. A good rule of thumb, at least for now, is one data scientist for every two or three data engineers.

These roles, in fact, complement each other: data engineers use their programming and system creation skills to design, build, and maintain big data pipelines.