By Lei Zhang

My dream job after graduation would be a data engineer. The work environment requires a data engineer to coordinate with data analysts, data scientists, business leaders, and system architects to understand their specific requirements.

In a broad sense, a data engineer is responsible for building databases that enable the company to sort through massive quantities of data. Gathering the required data is the preliminary task. Before starting any work on the database, a data engineer needs to gather data from the right sources. After adopting some dataset standards, the data engineer stores the upgraded data.

Creating a data model is the next step. Data engineers utilize the data collection's bound data model to separate recorded knowledge points. In addition, they build predictive models where they understand future developments using anticipatory strategies and experiences. A data engineer has to ensure data security and organization by using unified security controls to encode data, and investigate inductions into data. Data engineers also process data for clearing prerequisites. They use tools to input data from different sources, change and upgrade data, aggregate data and store data in a quota system. Data engineers use assorted data to improve models for the specific use of the data, for example, social media feeds, NoSQL feeds.

The data engineer job title requires strong developer skills. When data engineers use data pipelines, they need to write codes and scripts like data scientists. Highly analytical skills and a knowledge of data visualization are required. SQL is the most important tool for data engineers to communicate data. SQL has always been the lingua franca of data. A data engineer is required to have the technical ability to express any correlated subquery and function complexity in SQL.

I would prefer to work as a data engineer in a company like coursera that combines technology with education. Their mission fits my belief that education can change life, and we can also change education. They provide a highly reliable platform to support concurrent online access by large-scale users. A data engineer then uses data mining to analyze student behavior to conduct personalized course learning and improve course satisfaction, identify homework and establish discussion groups through machine learning. To this end, a data engineer use technology to allow people to access educational services equally and conveniently.

I believe the field of data science is going through its adolescence, data engineering is affirming and defining itself, and at the same time it is going through something similar like its "brother" of data science. What I find unique about the job is that while data engineers learn from data scientists, they also define themselves from the opposite of data science and finds their identity. One of Gordon Lindsay Glegg’s famous quotes would wrap up the job’s merits perfectly — A scientist can discover a new star but he cannot make one. He would have to ask an engineer to do it for him.

It is precisely because there are no strict rules and regulations that restrict how a data engineer should carry out their work, the job title allows more freedom. From technology selection to theoretical research to business scenarios, there is room for a data engineer to play. In collaboration with the colleagues, it is possible for a data engineer to discuss business issues at the same time as product and data analysis. Personally, I could enjoy this work rhythm.

Although some would say data engineers know a little about everything, and not proficient at everything. However, it is not. Once the discussion is set on which technology tools should apply, a data engineer could immediately invest his entire energy to study it thoroughly, not only at the usage level, but sometimes even to analyze its source code, and then modify it according to actual needs, or make plug-ins development to transform open source technologies into proprietary data products suitable for corporate use. From there, data engineers play the role of architects, and can flexibly combine and even transform various open source technologies to exert greater value. In this process, data engineers learn to become proficient in technology.

The rapid development of the Internet has allowed many companies to have thousands of user data, and everyone wants to tap into this gold mine, which extends data application in different scenarios of their own business: e-commerce such as Amazon uses user portraits to make personalized recommendations; internet finance companies such as PayPal implement risk control by identifying the characteristics of high-risk behaviors; travel and transport businesses such as uber analyze data for real-time pricing to maximize their profits. The challenge faced by data engineers is not limited to a specific business, but to how to convert complex business logic into algorithm and models, so as to use massive data to solve this problem. These types of problems do not require engineers to explore algorithms deep enough, but require breadth and cross-skills. They need to understand common machine learning algorithms and know the pros and cons of each. At the same time, data engineers must also have the ability to quickly understand the business, know the source, destination and processing of the data, and have a high degree of sensitivity to the data.

My dream job in an ultimate universe would be data scientist. There is an overlap between data scientist and data engineer in terms of analytic skills. However, overlap always occurs at the irregular edges of each individual's abilities. The analytical skill of a data scientist exceeds that of a data engineer. Data engineers use their programming and system building skills to create big data pipelines. Data scientists leverage their programming skills and use their advanced math skills to create advanced data products using already existing data pipelines. Data scientists are more like researchers.

To a certain extent, data engineers at the moment spend long hours in tasks like data cleaning, data processing, model or feature selection and testing, which takes up to 80 percent of their time. In 2047, I believe, companies will mostly move beyond this curve and choose to embrace the automation of machine learning. The adoption of cloud computing or third-party software solutions will give the company a clear strategic advantage. Potential replacements for data engineers at that time include machine learning engineers, AI communicators, and AI architects. One of the ways to capitalize on this trend is to not only invest in data science and machine learning skills, but also take another approach to consider tasks that cannot be easily automated. Over time, the roles and responsibilities of data engineers are changing. In 2047, data engineers are increasingly geared to be responsible for managing data performance and reliability.