Data Scientist, Data Engineer & Other

Data Careers

1. Data Architect

The data architect focuses on engineering and managing data stores and the data that reside within them.

Data Architect

The data architect is concerned with managing data and engineering the infrastructure which stores and supports this data. There is generally little to no data analysis needing to take place in such a role (beyond data store analysis for performance tuning), and the use of languages such as Python and R is likely not necessary. An expert level knowledge of relational and non-relational databases, however, will undoubtedly be necessary for such a role. Selecting data stores for the appropriate types of data being stored, as well as transforming and loading the data, will be necessary. Databases, data warehouses, and data lakes; these are among the storage landscapes that will be in the data architect's wheelhouse. This role is likely the one which will have the greatest understanding of and closest relationship with hardware, primarily that related to storage, and will probably have the best understanding of cloud computing architectures of anyone else in this article as well.

Data Architect

SQL and other data query languages — such as Jaql, Hive, Pig, etc. — will be invaluable, and will likely be some of the main tools of an ongoing data architect's daily work after a data infrastructure has been designed and implemented. Verifying the consistency of this data as well as optimizing access to it are also important tasks for this role. A data architect will have the know-how to maintain appropriate data access rights, ensure the infrastructure's stability, and guarantee the availability of the housed data.

This is differentiated from the data engineer role by focus: while a data engineer is concerned with building and maintaining data pipelines, the data architect is focused on the data itself. There may be overlap between the 2 roles, however: ETL; any task which could transform or move data, especially from one store to another; starting data on a journey down a pipeline.

Data Architect

Like other roles in this article, you might not necessarily see a "data architect" role advertised as such, and might instead see related job titles, such as:

- **Database Administrator**
- **Spark Administrator**
- **Big Data Administrator**
- **Database Engineer**
- **Data Manager**

2. Data Engineer

The data engineer focuses on engineering and managing the infrastructure which supports the data and data pipelines.

Data Engineer

What is the data infrastructure? It's the collection of software and storage solutions that allow for the retrieval of data from a data store, the processing of data in some specified manner (or series of manners), the movement of data between tasks (as well as the tasks themselves), as data is on its way to analysis or modeling, as well as the tasks which come after this analysis or modeling. It's the pathway that the data takes as it moves along its journey from its home to its ultimate location of usefulness, and beyond. The data engineer is certainly familiar with DataOps and its integration into the data lifecycle.

Data Engineer

From where does the data infrastructure come? Well, it needs to be designed and implemented, and the data engineer does this. If the data architect is the automobile mechanic, keeping the car running optimally, then data engineering can be thought of as designing the roadway and service centers that the automobile requires to both get around and to make the changes needed to continue on the next section of its journey. The pair of these roles are crucial to both the functioning and movement of your automobile, and are of equal importance when you are driving from point A to point B.

Data Engineer

Truth be told, some the technologies and skills required for data engineering and data management are similar; however, the practitioners of these disciplines use and understand these concepts at different levels. The data engineer may have a foundational knowledge of securing data access in a relational database, while the data architect has expert level knowledge; the data architect may have some understanding of the transformation process that an organization requires its stored data to undergo prior to a data scientist performing modeling with that data, while a data engineer knows this transformation process intimately. These roles speak their own languages, but these languages are more or less mutually intelligible.

You might find related job titles advertised for such as:

- **Big Data Engineer**
- **Data Infrastructure Engineer**

3. Data Analyst

The data analyst focuses on the analysis and presentation of data.

Data Analyst

I'm using data analyst in this context to refer to roles related strictly to the descriptive statistical analysis and presentation of data. This includes the preparation of reporting, dashboards, KPIs, business performance metrics, as well as encompassing anything referred to as "business intelligence." The role often requires interaction with (or querying of) databases, both relational and non-relational, as well as with other data frameworks.

While the previous pair of roles were related to designing the infrastructure to manage and facilitate the movement of the data, as well managing the data itself, data analysts are chiefly concerned with pulling from the data and working with it as it currently exists. This can be contrasted with the following 2 roles, machine learning engineers and data scientists, both of which focus on eliciting insights from data above and beyond what it already tells us at face value. If we can draw parallels between data scientists and inferential statisticians, then data analysts are descriptive statisticians; here is the current data, here is what it looks like, and here is what we know from it.

Data Analyst

Data analysts require a unique set of skills among the roles presented. Data analysts need to have an understanding of a variety of different technologies, including SQL & relational databases, NoSQL databases, data warehousing, and commercial and open-source reporting and dashboard packages. Along with having an understanding of some of the aforementioned technologies, just as important is an understanding of the limitations of these technologies. Given that a data analyst's reporting can often be ad hoc in nature, knowing what can and cannot be done without spending an ordination amount of time on a task prior to coming to this determination is important. If an analyst knows how data is stored, and how it can be accessed, they can also know what kinds of requests — often from people with absolutely no understanding of this — are and are not serviceable, and can suggest ways in which data can be pulled in a useful manner. Knowing how to quickly adapt can be key for a data analyst, and can separate the good from the great.

Related job titles include:

- **Business Analyst**
- **BI Analyst**

4. Machine Learning Engineer

The machine learning engineer develops and optimizes machine learning algorithms, and implements and manages (near) production level machine learning models.

Machine Learning Engineer

Machine learning engineers are those crafting and using the predictive and correlative tools used to leverage data. Machine learning algorithms allow for the application of statistical analysis at high speeds, and those who wield these algorithms are not content with letting the data speak for itself in its current form. Interrogation of the data is the modus operandi of the machine learning engineer, but with enough of a statistical understanding to know when one has pushed too far, and when the answers provided are not to be trusted.

Machine Learning Engineer

Statistics and programming are some of the biggest assets to the machine learning researcher and practitioner. Maths such as linear algebra and intermediate calculus are useful for those employing more complex algorithms and techniques, such as neural networks, or working in computer vision, while an understanding of learning theory is also useful. And, of course, a machine learning engineer must have an understanding of the inner workings of an arsenal of machine learning algorithms (the more algorithms the better, and the deeper the understanding the better!).

Once a machine learning model is good enough for production, a machine learning engineer may also be required to take it to production. Those machine learning engineers looking to do so will need to have knowledge of MLOps, a formalized approach for dealing with the issues arising in productionizing machine learning models.

Machine Learning Engineer

Related job titles:

- **Machine Learning Scientist**
- **Machine Learning Practitioner**
- **<specific machine learning technology> Engineer, e.g. Natural Language Processing Engineer, Computer Vision Engineer, etc.**

5. Data Scientist

The data scientist is concerned primarily with the data, the insights which can be extracted from it, and the stories that it can tell.

Data Scientist

The data architect and data engineer are concerned with the infrastructure which houses and transports the data. The data analyst is concerned with pulling descriptive facts from the data as it exists. The machine learning engineer is concerned with advancing and employing the tools available to leverage data for predictive and correlative capabilities, as well as making the resulting models widely-available. The data scientist is concerned primarily with the data, the insights which can be extracted from it, and the stories that it can tell, regardless of what technologies or tools are needed to carry out that task.

Data Scientist

The data scientist may use any of the technologies listed in any of the roles above, depending on their exact role. And this is one of the biggest problems related to "data science"; the term means nothing specific, but everything in general. This role is the Jack Of All Trades of the data world, knowing (perhaps) how to get a Spark ecosystem up and running; how to execute queries against the data stored within; how to extract data and house in a non-relational database; how to take that non-relational data and extract it to a flat file; how to wrangle that data in R or Python; how to engineer features after some initial exploratory descriptive analysis; how to select an appropriate machine learning algorithm to perform some predictive analytics on the data; how to statistically analyze the results of said predictive task; how to visualize the results for easy consumption by non-technical folks; and how to tell a compelling story to executives with the end result of the data processing pipeline just described.

Data Scientist

And this is but one possible set of skills a data scientist may possess. Regardless, however, the emphasis in this role is on the data, and what can be gleaned from it. Domain knowledge is often a very large component of such a role as well, which is obviously not something that can be taught here. Key technologies and skills for a data scientist to focus on are statistics (!!!), programming languages (particularly Python, R, and SQL), data visualization, and communication skills — along with everything else noted in the above archetypes.

There can be a lot of overlap between the data scientist and the machine learning engineer, at least in the realm of data modeling (read: the process of employing predictive analytics) and everything that comes along with that. However, there is often confusion as to what the differences are between these roles as well. For a very solid discussion of the relationship between data engineers and data scientists, a pair of roles which also can also have significant overlap.

Stay in touch !

@DataCleanic

www.DataCleanic.ml