LinkedIn version : https://www.linkedin.com/pulse/data-analyst-vs-scientist-engineer-nicolas-hubert/

Data Analyst vs. Scientist vs. Engineer

Nicolas Hubert

Obviously, Data Science is now a buzzword that refers to a vast array of jobs, tasks and skills that sometimes differ from one another. And if they do, that is mainly because there is no consensus on a clear and widely accepted definition of what Data Science really is. Actually, it is no wonder that experts have so much trouble defining the scope of this new field, as it emerged no sooner than in the late 1980's. One of the yet unaddressed challenges remains to differentiate a Data Analyst from a Data Scientist and from a Data Engineer. That is where we come in !

DATA ANALYST

Overview : Basically, Data analysts collect, process and perform statistical analyses of data. They may not have the same mathematical or programming background as Data Scientists, but the very purpose of their job is definitely the same : to leverage existing data and to make them insightful. Only then can the Data Analysts formulate relevant recommendations to their superiors or other teams.

That is why a big part of their job consists in presenting the conclusions of their analysis to their teammates. Whether it is a deliverable, a dashboard, or a PowerPoint presentation, any Data Analyst should have a good industry-sector knowledge as well as strong interpersonal and communication skills. Add to that some basic knowledge of descriptive statistics, data query and data visualization tools, and you may have found your guy !

Skills : Data Analysts are at the very heart of the company. The better ones are able to easily navigate through the organization, so that they can have a grasp of the different challenges the other units are trying to solve. Then, Data Analysts are expected to fit into many roles and teams, and help

teammates make better data-driven decisions. Here are some of common tasks for any Data Analyst :

- Acquiring, cleaning and organizing data ;
- Analyzing data in order to identify trends and gain insight ;
- Creating visualizations and dashboards (using BI tools such as Tableau, QlikView, Microsoft Power BI, Looker, etc.) to help the company interpret and make wise decisions ;
- Defining the data-driven strategy of the company ;
- Elaborating segmentation criteria (especially when it is a marketing-related position) ;
- Presenting the results of a technical analysis to business clients or internal teams.

Education : Whereas a Master from a top-tier University or Engineering school – with a focus on such subjects as Applied Mathematics, Computer Science or Data Science – is a prerequisite to secure a job as Data Scientist/Engineer, Data Analysts come from a wider range of educational backgrounds. Of course, some of them studied Statistics or Computer Science at University, although there is a growing need for graduates from Business Schools. A specialization in Digital Marketing or Information Technology is a plus, but anyway many companies are fond of students in Economics or Business Schools, as they have been taught how to effectively communicate and present their work.

PS : Beware of placing too much emphasis on the job title : depending on the company, the data analyst can go by different titles such as : Operations Analyst, Business Intelligence (BI) Analyst, or even just Business Analyst.

DATA SCIENTIST

Some companies indistinctly post job offers of 'Data Analyst' and 'Data Scientist', considering these job titles are just synonymous. But considering the different range of skills and tasks required in these two jobs, a clear distinction must be drawn.

Even if Data Scientists and Data Analysts are motivated by the same goal – to gain insight from the huge amount of datasets – a Data Scientist's work requires an extensive knowledge of both applied mathematics and computer science in order to tackle larger bunches of data.

Overview : The main difference between a Data Scientist and a Data Analyst lies in the type of data to analyze : while a Data Analyst works with already formatted data, any Data Scientist has to cope with raw data which often come in large amount (hence the term 'Big Data'). Moreover, Data Scientists are used to perform a kind of upstream work : they carry out the preliminary work so that Data Analysts can move forward. In a way, the Data Scientist substitutes for the Data Analyst whenever data analysis gets more complex and requires a good command of more advanced techniques and technologies. For instance, when the problem to solve implies to develop an accurate model – in which case the project is akin to R&D – then the Data Scientist has a role to play.

Here are some of the major challenges Data Scientists are currently involved in :

- **Recommendation algorithms** which objective is to provide customers with personalized services and content. This kind of algorithm is instrumental in the way you can be alerted by mail when a job offer similar to those you applied for has just been posted online;

- **Natural language processing (NLP) algorithms**, thanks to which some companies are now able to extract relevant information from a vast array of different documents, whether it is an e-mail, a resume or a job offer. Machine translation (i.e the automatic translation from one language to another) is a common example : did you realize that when a Facebook post is written in a foreign language, a translation into your native language is automatically performed ? Many thanks, machine translation !

- **Clustering algorithms** that basically aim at grouping data points in different groups, in which every element shares the same features. But data points are not *just* data points. Instead, they may represent customers (with, for example : age on the X-axis and amout spent on the Y-axis), or other things you just want to categorize. Do not rack your brain trying to implement your own clustering algorithm ! If it is sometimes required, though, in most cases you can just use existing algorithms and adapt them to your own specifications. That often pays off ! The purpose of this article is not to go over all the details, but here are the names of some of the most popular cluster algorithms : K-Means Clustering, Gaussian Mixture Model Clustering, Density based Clustering, etc.

Skills : In short, a Data Scientist must possess (but not limited to) the competencies listed below :

- **Mathematics** : the reason why some tests are sometimes inappropriate or irrelevant is due to a bad use of Statistics. So, you should know how to properly use such statistical tools as : distributions, maximum likelihood estimators, acceptance-reject methods, confidence intervals, etc. On top of that, it may be good to have some knowledge of multivariable calculus and linear algebra, as for some companies, algorithm optimization can be significantly time saving ;

- **Programming Skills** : statistical programming language (R and/or Python) and a database querying language such as SQL ;

- **Machine Learning Concepts** : K-nearest neighbors, random forests, logistic regression, etc. All these machine learning techniques are directly implemented in Python and R libraries, but you should know how and when to use them, and above all understand the main idea that lies beneath these techniques ;

- **Data Wrangling** : Often, the data you are analyzing is awfully messy. That is why you should be able to deal with imperfections in data. Some examples include missing values ('N/A' in a certain cell) or inconsistent string formatting (e.g.,'France' vs. 'FRANCE' vs. 'FR') ;

- **Data Visualization & Communication** : as Data Scientists are often seen as people who help their company make better data-driven decisions, a good command of visualization tools as well as strong communication skills are really helpful when it comes to describing your findings, or the conclusion you draw from your data analysis.

Education : Data scientists typically hold a Master or PhD in a quantitative field, such as Computer Science, Statistics, or Applied Mathematics. A few of them have a Bachelor in a related field. However, any good Data Scientist must be eager to learn beyond formal education, and keep honing his own skills. In order to do that, Kaggle is definitely the website to register ! This is basically a website that provides loads of datasets on real-world cases and fosters competitions among Data Scientists and Machine Learning enthusiasts. In short, Kaggle is an ecosystem for learning by doing and sharing about Data Science.

Needless to say, a good Kaggle ranking really matters and you can easily leverage this experience during an HR interview. Want to know what it looks like? Just follow this link : https://www.kaggle.com/.

DATA ENGINEER

Overview : One must acknowledge that Data Scientists' efficiency strongly relies on the easy access to data. Most companies store data in different formats. This is where Data Engineers come in : they are the ones responsible for preparing 'Big Data' (or even better, 'Smart Data') infrastructure on which Data Scientists will be able to work. That's why they are mainly software engineers : they design, build, integrate data from various ressources. Then, they make sure that data is easily accessible. Their final purpose is to improve the performance of their company's big data ecosystem. Data Engineers are as important as Data Scientists, but tend to be less visible because they are further from the end product of the analysis.

Data Engineers might create big data warehouses that can be used for reporting or analysis. A typical day for any Data Engineer would be like : 50% software development, 20% studies and specifications, 20% IT support/debugging and 10% technology monitoring. As you know understand, Data Engineers focus more on the design and architecture. They do not need to know any machine learning algorithms and the required background in mathematics does not matter as much as for Data Scientists.

Skills : Below are some of the competencies a company can expect from a Data Engineer. We won't focus too much on mentioning IT tools because it really depends on the company. Nevertheless, mastering tools such as Hadoop, Python, Scala and Spark is a must.

- Designing, installing, testing and maintaining data management systems ;
- Integrating new data management technologies and software engineering tools into existing structures ;
- Creating custom software components and analytics applications ;
- Using a variety of languages and tools to make systems work together ;
- Recommending ways to improve data reliability, efficiency and quality ;
- Collaborating with data science teams and building the right solutions for them.

If you REALLY want to know more about the tools they are using everyday, here is a non-exhaustive list :

- Database architectures ;
- Data warehousing solutions ;
- Data modeling tools (e.g. ERWin, Enterprise Architect and Visio) ;
- Hadoop-based technologies (e.g. MapReduce, Hive and Pig) ;
- SQL-based technologies (e.g. PostgreSQL and MySQL) ;
- NoSQL technologies (e.g. Cassandra and MongoDB).

Data Analyst, Data Scientist, Data Engineer… for some people it is just gibberish that relates to the same thing. Now you know it is not. And even better, you know what the main features are for these 3 positions.

Sources :

Lebigdata.fr

Lemondeinformatique.fr

Blogdumoderateur.com

Urbanlinker.com

Blog.udacity.com

Quora.com

Dataquest.io

Mastersindatascience.org