simplilearn

DATA SCIENTIST

THE NUMBERS GAME DECIPHERED

A Step-By-Step Guide

Table of Contents


BIG DATA

What is Big Data and Data Science?

Big Data is a popular term used to describe data sets that are so large and complex by nature that traditional data processing methods are inadequate for analyzing them. Recent statistics predict that about 2.5 quintillion bytes of data are created every day, and 90 percent of the data in the world was developed in the last two years alone.

However, this data is not useful for industries in its raw form. When properly processed, Big Data allows businesses to find new data trends that can help in making agile processes and assist in better decision making. Big Data's main reason for existing is to provide a means of collecting data from a large number of varied sources, harnessing the relevant data, and analyzing it to find answers to vital business-related questions relating to:

Cost Reduction, Time Reduction

Optimizing New Product Development Product Offerings

Smarter and Quicker Business Decisions

Data Science – History and Recent Developments

Hadoop, MapReduce, GridGain, HPCC, and Storm are some of the most popular Big Data Analysis platforms and tools available today. As there is an increasing amount of data being churned out every day, there is a correspondingly urgent need for procuring this data and making it useful. Data Science refers to the collection, preparation, analysis, visualization, management, and preservation of these large amounts of data.

In simple terms, Data Science is the extraction of useful information from the available data. The methods generally associated with processing Big Data are of particular interest to the field of data science, though the latter deals with all types of data, not just Big Data.

The term "Data Science" has existed for over thirty years and was usually used as a substitute for "computer science". It was only in 1996, at the International Federation of Classification Societies (IFCS) meeting, that the term 'data science' was included in the conference title.

In 1997, C.F. Jeff Wu gave an inaugural lecture on "Statistics = Data Science?" at the University of Michigan. In this lecture, he advocated that statistics should be renamed data science and statisticians should be renamed, data scientists.

In 2008, the term "Data Scientist" was coined by DJ Patil and Jef Hammerbacher to define their jobs at LinkedIn and Facebook, respectively.

What Does a Data Scientist Do?

Data scientists play an essential part in the design and implementation of data architecture, acquisition, analysis, and archiving. The overlapping skills of a data scientist, including knowledge of programming languages and data mining/statistics, are used to handle Big Data systems like Hadoop.

Since data scientists are involved in the design and implementation of data acquisition, they will, in most cases, be partnered with system architects in order to develop a system architecture which will ensure the acquired data is routed and organized for further analysis.

Data scientists are actively involved in representing the data, transforming it, arranging it in different groups, and linking it for analysis. Data scientists as a rule are most involved with the latter task.

In this context, analysis means summarizing the input data and drawing essential samples from it. These samples need to be carefully studied, and conclusions regarding the broader context subsequently drawn from them. Once the conclusions are established, it is imperative to communicate the findings so that non-data scientists can understand them, usually by means of diagrams, tables, and other visual communication techniques. Otherwise, the entirety of the data will be pointless to the average user, and all the statistical analysis data will be rendered useless.

Once the data is routed, organized, arranged, and analyzed, the next step is to archive the information. Data curation is a crucial aspect of the data management system, preserving the data so that it can be reused. This is one of the most critical responsibilities for data scientists.

Data Scientists break Big Data into four Dimensions:

Volume, Velocity, Variety, and Veracity

In a Nutshell, Data Scientists:

- Should have both statistical modeling experience and technical, engineering skills.
- Should have experience in working on granular data, preferably on a Hadoop platform.
- Should have the ability to focus on revenue generation and yield management apart from being only analytics specific.
- Should work with others for the purposes of refining data management processes, curation techniques, and scaling the existing procedures for achieving better efficiency.

Bridging the Talent Gap

Though the phrase 'Data Scientist' has been around for a long time, not enough skilled professionals have entered the field. This talent gap has been very well highlighted in a new report by McKinsey Global Institute (MGI), 'Game changers: Five opportunities for US growth and renewal.' According to the report, Big Data analytics could increase the annual GDP up to $325 billion by 2020 in retail and manufacturing.

According to the same report, there is a shortage of 190,000 skilled data scientists and 1.5 million managers and analysts who can draw useful conclusions from the available data. The report also highlights the fact that about 40,000 Exabytes of data will be collected by 2020, adding further proof that a talent gap exists.

As most companies (except the A-listers in Silicon Valley) find it hard to get skilled data scientists on-board, they have had to get creative by assembling teams of people to fill the role of a data scientist. To that end, these teams have data crunchers, statisticians, computer scientists, analysts, and managers who collectively put up the data in a usable form.

Though this system works on paper, in reality, it's nothing but a stop-gap arrangement for most companies. With this vast scarcity in the market for skilled data scientists, this becomes a lucrative certification option for most professionals. After all, the best job opportunities come from the fields where the demands are higher.

In 2015, the Big Data market was about $23.0 billion and expected to hit $118.52 billion by 2022.

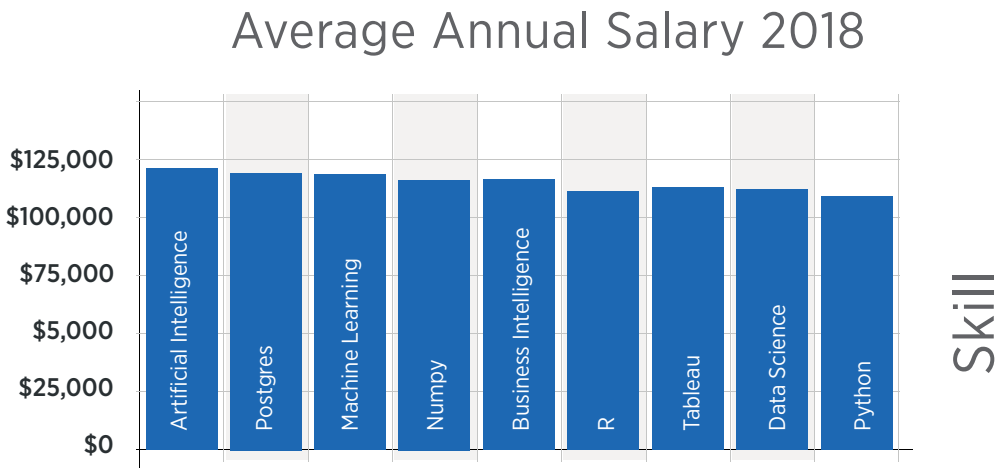
Prerequisites for Becoming a Data Scientist

Though there are no defined prerequisites for taking up certification training, it is essential to brush up on some appropriate skills such as Multivariable Calculus, Linear Algebra, and Statistics. Multivariable Calculus is necessary for various stages of machine learning and probability calculations. Similarly, linear/matrix algebra often shows up in machine learning concepts.

A data scientist must have the basic hands-on knowledge of statistics in order to do their job successfully. While there is a lot of debate in data science circles regarding statistics being outdated and stodgy, statistical modeling is still an important part of the job profile. Thus, candidates need to have the basic knowledge of stats so that they can apply this logic in R or other languages.

The final vital data scientist certification training prerequisite is coding, a Computer Science fundamental. It's common knowledge that data scientists need to write code for the simple reason that if one can't use R or similar languages, they cannot work on real-world data. One need not be an expert in coding, but basic knowledge is always helpful.

It's hard not to talk about coding without bringing up programming, considering how the two disciplines often blur together. This in turn brings the subject of programming languages to the forefront. With that in mind, it is essential for data scientists to work with languages like R and Python if they want to work with real data. Before we jump into the skill sets involved with each of these languages, let us take a look at the impact of various popular IT skills on the salary structures.



Source: Dice.com 2019 Tech Salary Survey

It is essential for any aspiring data scientist to learn the Python language. When compared with other tools for the purposes of data processing, Python emerges as the best. This is due to Python's simplicity and the availability of ready to use machine learning tools like scikit-learn, Orange, etc. Python is a vital ingredient in the 'data processing' toolbox and is a great starting language.

The next logical step is to take R language training. Employers look for candidates who are skilled in R language because it facilitates data analysis and helps data scientists get an idea of what works best.

According to a 2018 survey of most-used statistics/programming languages, SAS is one of the most popular languages in the Data Science community. Based on this, it's easy to conclude that the next likely destination for a data scientist aspirants will be a SAS Base Programmer.

And finally, though not a requirement, industry experts conclude Hadoop platform knowledge is essential for dealing with real data sets. Hadoop makes it easier to process Big Data, a plus for any data scientist. Additionally, employers are always on the lookout for data scientists with Hive or Pig experience alongside familiarity with cloud tools like Amazon S3.

Preferred Educational Qualifications

There is still an ongoing debate regarding the best qualifications for learning Data Science. Although some experts claim that a Bachelor's degree with excellent practical skills is enough, others believe a Master's degree or a Ph.D. is needed to do justice to the profession.

However, as previously mentioned, Data Science is a fusion of several disciplines including Math, Science, Advanced Computing, Visualization, Data Engineering, Hacker Mindset, and Domain expertise. Therefore, it's difficult to single out one particular field as a prerequisite for learning data science.

According to several studies, 80 percent of data scientists have a Master's degree, and about 40 percent have a Ph.D. The most common fields of study are Mathematics and Statistics, Economics, Computer Science, and Engineering.

Though a few institutes are planning to start a Bachelor's degree program which will be in line with the Computer Science programs, training is typically focused on Master's degree programs. Apart from these programs, several institutes offer certification training online, live-virtual classroom, and classroom learning modes, in order to reach out to and address the needs of students across the globe.

Miscellaneous Non-technical Skills

Intellectual Curiosity

As mentioned in a post on Burtch Works, the primary motivating factor for data scientists is the curiosity associated with making meaningful inferences from the available data sets. Aspirants can initiate data science projects on their own and draw conclusions from them, thereby enhancing their analytics skills.

Industry experience

As most of the data that is being analyzed is related to critical business decisions, it is essential that the data scientist should have adequate knowledge about the industry that she is working in and must understand the problems that the company is trying to solve. Thus, she must be able to ascertain which business problems are best solved by the application of data science.

Communication Skills

Employers prefer to hire a data scientist who can easily translate technical findings to a non-technical team. Thus, communication skills are fundamental. Also, a good data scientist needs to understand the non-technical needs of data analysis and present quantified insights into the non-technical teams.

Study Plan

Once you have decided to take up the data scientist path, the next step is to excel in all the key areas in the subject. A detailed study plan is presented below to help you understand the nuances of data science.

Learning Path

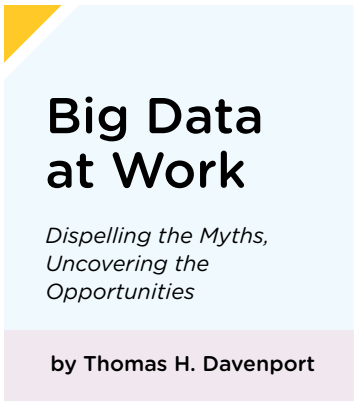
Useful Resources

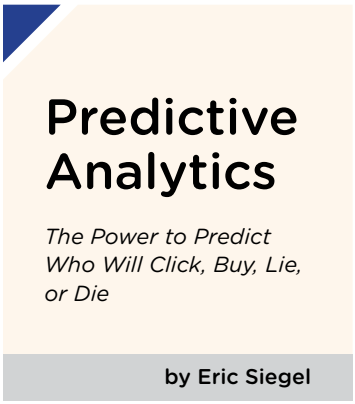
Apart from enrolling for training with an accredited institute, it is also important to keep yourself well-informed about new developments and changes in the field. You can accomplish this by spending time reading books, watching key videos, and going through some of the best articles on the subject.

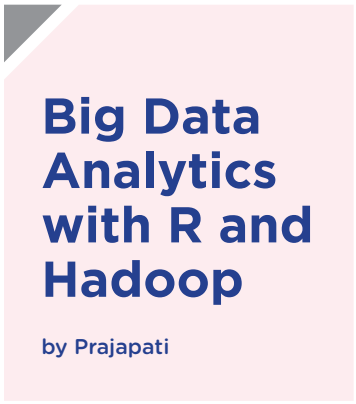
If videos are your preferred means of gathering information, Simplilearn's YouTube channel has everything you need to know, conveniently assembled in one place. You will find the subject of Big Data well-represented.

Simplilearn also offers a useful collection of articles covering topics such as Big Data and Analytics. For those who want to take learning to a level beyond just online reading material, consider taking a Simplilearn course such as Data Analyst, Big Data Architect, or Data Engineer.

Ever since the term data scientist was coined, hundreds of books have been written on the subject. We have put together a list of some of the most useful guides on the subject.







Additionally, you can also visit sites such as; ***KDnuggets, R-blogger, DataTau,*** for keeping up with the latest trends in the field.

The Future of Data Science

With the increasing use of data science across all types of industries, employers are now looking for skilled and certified professionals in the field. For instance, a recent report from LinkedIn names data scientist as the number one most promising job in America in 2019. Forbes echoes these positive trends, predicting a 12 percent increase in Big Data/tech-related positions through 2024, compared to the 6.5 percent increase predicted for other jobs.

Data Science is expected to mature, consolidate, become the mainstream career option, and even surprise us with new advancements in the field. These changes will come over time and happen concurrently with a gradual shift to the cloud environment. Data science practitioners should be able to build predictive models in temporary cloud environments to increase their performance requirements. Currently, most data-related problems are solved by employing a single algorithm or tool, but this is expected to change. Data scientists are building new data algorithms to suit their needs, which are expected to take advantage of parallel data processing to improve efficiency.

Simplilearn's Data Scientist Master's Program can walk you through every nuance of becoming a successful data professional.

4% increase in Big Data/tech-related positions through 2024.

Key Features

- Industry recognized certifications from IBM and Simplilearn for this unique co-developed program
- Portfolio worthy capstone demonstrating mastered concepts
- 15+ Real-life projects providing hands-on industry training
- 30+ In-demand skills
- Lifetime access to self-paced learning and class recordings

More Information

https://www.simplilearn.com/big-data-and-analytics/senior-data-scientist-masters-program-training

Founded in 2009, Simplilearn is one of the world's leading providers of online training for Digital Marketing, Cloud Computing, Project Management, Data Science, IT Service Management, Software Development and many other emerging technologies. Based in Bangalore, India, San Francisco, California, and Raleigh, North Carolina, Simplilearn partners with companies and individuals to address their unique needs, providing training and coaching to help working professionals meet their career goals. Simplilearn has enabled over 1 million professionals and companies across 150+ countries train, certify and upskill their employees.

Simplilearn's 400+ training courses are designed and updated by world-class industry experts. Their blended learning approach combines e-learning classes, instructor-led live virtual classrooms, applied learning projects, and 24/7 teaching assistance. More than 40 global training organizations have recognized Simplilearn as an official provider of certification training. The company has been named the 8th most influential education brand in the world by LinkedIn.

For more information, visit www.simplilearn.com.