



Home » Big Data Tutorials » Skills Needed to Become a Data Scientist

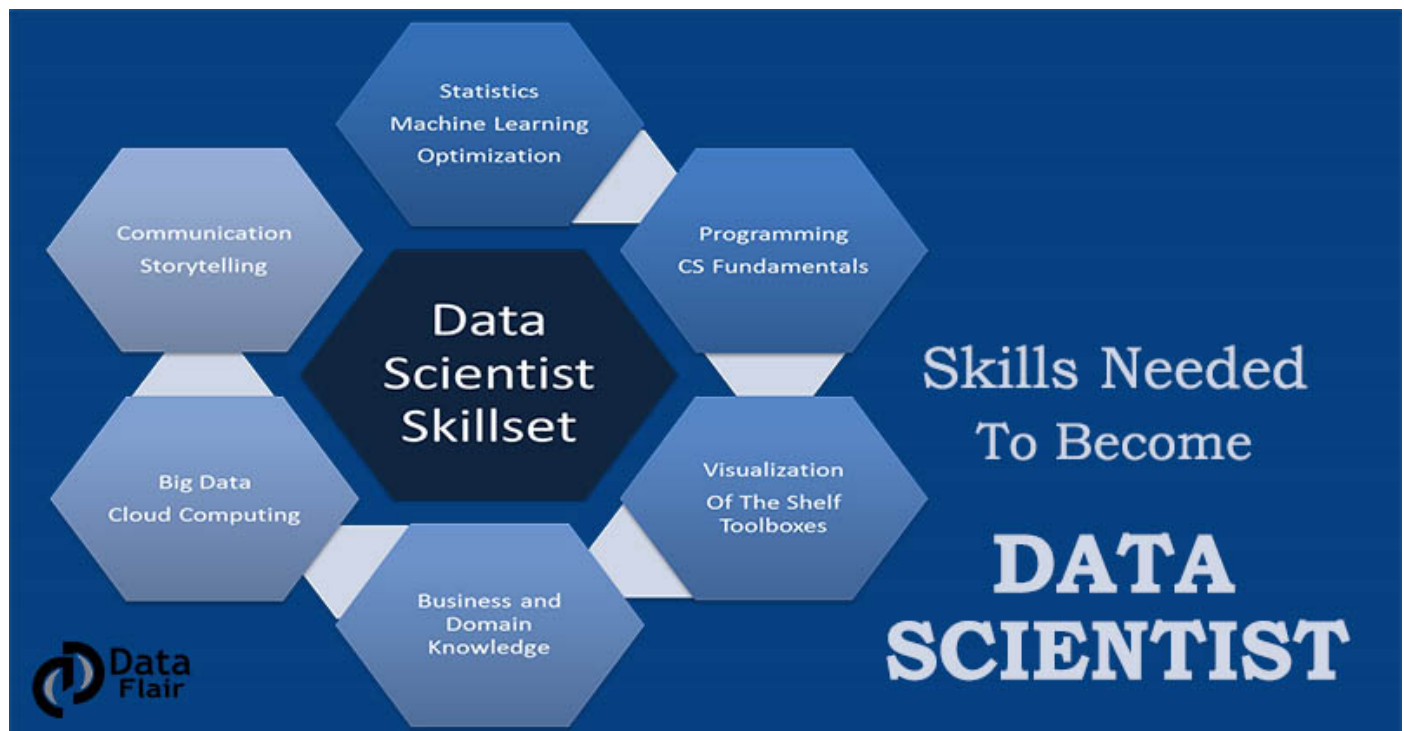
Skills Needed to Become a Data Scientist¹

26 Dec, 2016 in Big Data Tutorials by Support

1. Objective


In this tutorial, we will discuss the Skills Needed to Become a Data Scientist. We will talk about what are the qualifications needed for a data scientist, what are the different data science certification program, what is data scientist's job description.

A data scientist is better statistician than any software engineer and better engineer as compared to any statistician. A data scientist is termed to be the "sexiest job of the 21st century. Let's discuss how to become a data scientist (What are the skills needed).



2. Skills Needed to Become a Data Scientist

Data scientists are big data wranglers. They take a huge amount of messy data points (unstructured and structured) and clean, massage and organize them with their formidable skills in math, statistics, and programming. Then they apply all their analytic powers to uncover hidden solutions to business challenges and present it to the business. In other

 words. Data scientists utilize their knowledge of statistics and modeling to convert data into actionable insights about everything from product development to customer retention to new business opportunities.

Data Scientist needs to have both technical and non-technical skills to perform their job in an effective manner. Technical skills are involved at 3 stages in Data Science. They include:

1. Data Capture & pre-processing
2. Data Analysis & pattern recognition
3. Presentation & visualization

For performing above 3 stages, 3 categories of tools are needed – tools for pulling data, tools for analyzing the data, and tools for presenting the results. Here are the different tools available for performing the same:

2.1. Tools for data pulling & pre-processing

a. SQL

This is a must skill for all data scientists, regardless of whether you are using structured or unstructured data. Companies are using latest SQL engines like Apache Hive, Spark-SQL, Flink-SQL, Impala, etc.

b. Big Data Technologies

This is the must out of the Skills Needed to Become a Data Scientist. The data scientist needs to know about different big data technologies – 1st Gen technologies like Apache Hadoop & its ecosystem (hive, pig, flume, etc.), Next Gen like – Apache Spark and Apache Flink (Apache Flink is replacing Apache Spark quickly as Flink is a general purpose Big data engine, which can handle real-time stream as well, for more details about Flink follow this comprehensive tutorial).

c. UNIX

As most raw data is stored on a UNIX or Linux server before it's put in a data-store so it's nice to be able to access the raw data without the dependency of a database. So Unix knowledge is good for Data Scientists. Follow this command guide to practice Linux commands.

d. Python

Python is a most popular language for the data scientist. Python is an interpreted, object-oriented programming language with dynamic semantics. It is a high-level language with dynamic binding and typing.

2.2. Tools for Data Analysis & pattern matching

This depends on your level of statistical knowledge. Some tools are used for more advanced statistics and some for more basic statistics.

a. SAS

Lots of companies use SAS, so some basic SAS understanding is good. You can manipulate equations easily.



R is most popular in the statistical world. R is an open-source tool and language that is object oriented, so you can use that anywhere. It is the first choice of any data scientist as most things are implemented in R. To get the comparison between top data analytics tool, follow this comparison guide between R vs SAS vs SPSS.

c. Machine Learning

Machine learning is the most demanding and most useful tool the data scientists must have. Machine learning algorithms are used for advanced data analytics, predictive analytics, advanced pattern matching. There are lots of machine learning tools are available in the market like weka, nltk, etc. but machine learning tools on top of big data technologies are grabbing industry attention like Mahout (on top of Hadoop), MLlib (on top of Spark), FlinkML (on top of Flink).

2.3. Tools for Visualization

a. Tableau

It is a popular tool, especially in Silicon Valley.

b. JMP (SAS subsidiary)

JMP has some nice visualization.

c. R

R also has great visualization support such as ggplot2, lattice, rCharts, google charts, shiny for web apps for presentations, etc.

Apart from above-mentioned tools following tools are also popular – JasperSoft, SAP BI, QlikView, MicroStrategy, etc.

2.4. Non-Technical Skills

a. Business Acumen

One needs to have a solid understanding of the industry he is working in, to know the issues faced by the organization. The data scientist should be able to determine which problems are critical and which aren't, for identifying new ways to which the data can be used as a leverage.

b. Communication Skills

Companies are searching for data scientists who can clearly and confidently translate their insights on the data to other teammates. A data scientist arms them with quantified insights.

c. Analytical Problem-Solving

Analytical problem-solving skill is highly demanding for Data Scientist so that the right approach can be used to get maximum output in available time and resources.

3. Various Certifications for Data Scientist



Once you have learned the above Skills Needed to Become a Data Scientist, you can opt for Data Scientist certification. Here are few Data Scientist certifications that focus on useful skills:

a. Cloudera Certified Professional: Data Scientist (CCP: DS)

CCP: DS is aimed at data scientists to demonstrate advanced skills in working with big data. Candidates are drilled in 3 exams – Descriptive and Inferential Statistics, Unsupervised Machine Learning and Supervised Machine Learning – and must prove their skill set by developing a production-ready data science solution under real-world conditions. To learn more about Cloudera Data Scientist Certification follow this Tutorial.

b. Certified Analytics Professional (CAP)

This certification was created in 2013 by the Institute for Operations Research and the Management Sciences (INFORMS) and is targeted towards data scientists. Candidates need to demonstrate their expertise of the end-to-end analytics process. This certification includes the framing of business and analytics problems, data, and methodology, model building, deployment and lifecycle management.

c. EMC: Data Science Associate (EMCDSA)

The EMCDSA certification tests the ability to apply common techniques and tools required for big data analytics. Candidates are judged on their technical expertise in tools such as “R”, Hadoop, and Postgres, etc and their business acumen.

Leave a comment

Your email address will not be published. Required fields are marked *

Comment

Name *

Email *

