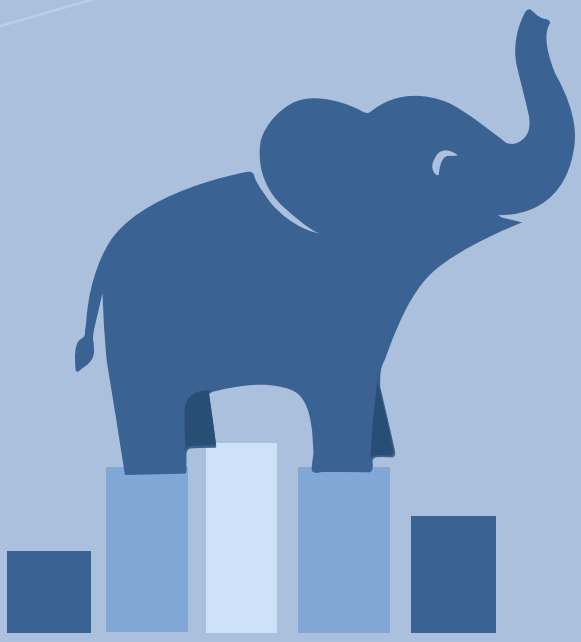
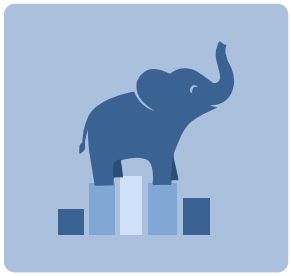


DATA SCIENTIST:

THE NUMBERS GAME DECIPHERED

A STEP-BY-STEP GUIDE

Table of Contents



Big Data is a popular term used for data sets that are so large and complex in nature, that the traditional data processing methods are inadequate for analyzing these. A recent statistics predict that about 2.5 quintillion bytes of data are created every day and 90% of the data in the world is developed in the last two years alone.

However, this entire data is not useful for industries in the raw form. Big Data gives an opportunity to businesses to find new trends in the data that can help in making their processes agile and also assist in better decision making. The main reason behind the wide usage of Big Data is to collect data from all sources, harness the relevant data and analyze it to find answers to key business-related questions related to:

- **Cost Reduction**
- **Time Reduction**
- **New Product Development**
- **Optimized product offerings**
- **Smarter and quicker business decisions**

Data Science – History and Recent Developments

As there is more and more data churned out every day, there is an increased need of procuring this data and making it useful. Data Science refers to collection, preparation, analysis, visualization, management, and preservation of large collection of data.

In simple terms, data science is the extraction of useful information from the available data. The methods that usually deal with Big Data are of particular interest in data science, though the latter is not restricted to such data.

Data Science, the term itself has existed 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) meet, that the term 'data science' was

Hadoop, MapReduce, GridGain, HPCC, Storm are some of the popular Big Data Analysis platforms and tools.

In 1997, C.F Jeff Wu gave an inaugural lecture on “Statistics = Data Science?” at the University of Michigan. In this lecture, the term ‘data science’ was coined and it was advocated that statistics should be renamed data science and statisticians should be renamed data scientist.

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

Data Scientists – What do they do?

Data scientists play an important part in design and implementation of data architecture, acquisition, analysis, and archiving. The overlapping skills of a data scientist are handling Big Data systems like Hadoop, Netezza, knowledge of Python/R and data mining/statistics.

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

Representing the data, transforming it, arranging it in different groups, and linking the data for the analysis part are all the tasks where data scientist is actively involved.

The analysis phase is the one where data scientists are most involved. By analysis, we mean summarizing the input data and drawing key samples from it. These samples need to be carefully studied and conclusions regarding the larger context are drawn from them. Once the conclusions are drawn, it is very important to communicate the same using diagrams, tables, and other visual communication techniques. Else the entire data will be pointless to a data user and all the statistical analysis data will become useless.

Once the data is routed, organized, arranged, analyzed, and communication; the next step is to archive the same. Data curation is an important aspect of data management system. Preservation of data so that it can be re-used is one of the important focus areas for data scientists.

Data Scientists break Big Data into four dimensions: Volume , Velocity, Variety, and Veracity.

In a Nutshell:

- Data Scientist 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 refining data management processes, curation techniques, and scaling the existing processes for achieving better efficiency.

Bridging the Talent Gap

Though the phrase, 'Data Scientist' has been doing rounds since long time, not many 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 data that is available. Also, the fact that about 40,000 exabytes of data will be collected by 2020 makes the talent gap evident.

As most companies (Except the A-listers in Silicon Valley) find it hard to get skilled data scientists on-board, they create a team of people who fit-in the shoes of a data scientist, not individually but as a group. So, this team will 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 this is nothing but a stop-gap arrangement for most companies. With this huge scarcity in the market for skilled data scientists, this becomes a lucrative certification option for most professionals.

In 2010, Big Data market was about $3.2 billion while in 2015 it has grown to $16.9 billion.

Pre-requisites for Becoming a Data Scientist

Though there are no defined pre-requisites for taking up certification training, it is important to brush up a few skills like **Multivariable Calculus, Linear Algebra, and Statistics (basics)**. Multivariable Calculus is necessary in some parts of machine learning and also in probability calculations. Similarly, linear/matrix algebra holds a lot of importance in machine learning concepts.

For becoming a data scientist having the basic hands-on knowledge of statistics, is necessary. Though there is a lot of debate doing rounds in the data science circle regarding statistics being out-dated and stodgy, statistical modeling is still an important part of the job profile of a data scientist. Thus, candidates need to have the basic knowledge of stats, so that they can apply this logic in R or other languages.

The other important pre-requisite for taking up data scientist certification training, is coding. Computer Science fundamentals knowledge is essential for taking up any training. As it is widely known, Data Scientists need to write codes, 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.

Preferred Educational Qualifications

The debate regarding the best qualification in order to learn Data Science is still pretty much on. With a few experts claiming that a Bachelor's degree with good practical skills can do the trick, some other believe a Master's degree or a PhD is needed to prove your skills in this field. However, as aptly mentioned in Data Science: A Introduction/ A Mash-up of Disciplines, this field is a mash-up of several disciplines like Math, Science, Advanced Computing, Visualization, Data Engineering, Hacker Mindset, and Domain expertise. So, it's difficult to single out a particular field which can act as a pre-requisite for learning data science.

In general, one will find minimum 80% of the data scientists with a Master's degree and about 40% having a PhD qualification. The most common fields of study are Mathematics and Statistics, Economics, Computer Science, and Engineering, amongst others.

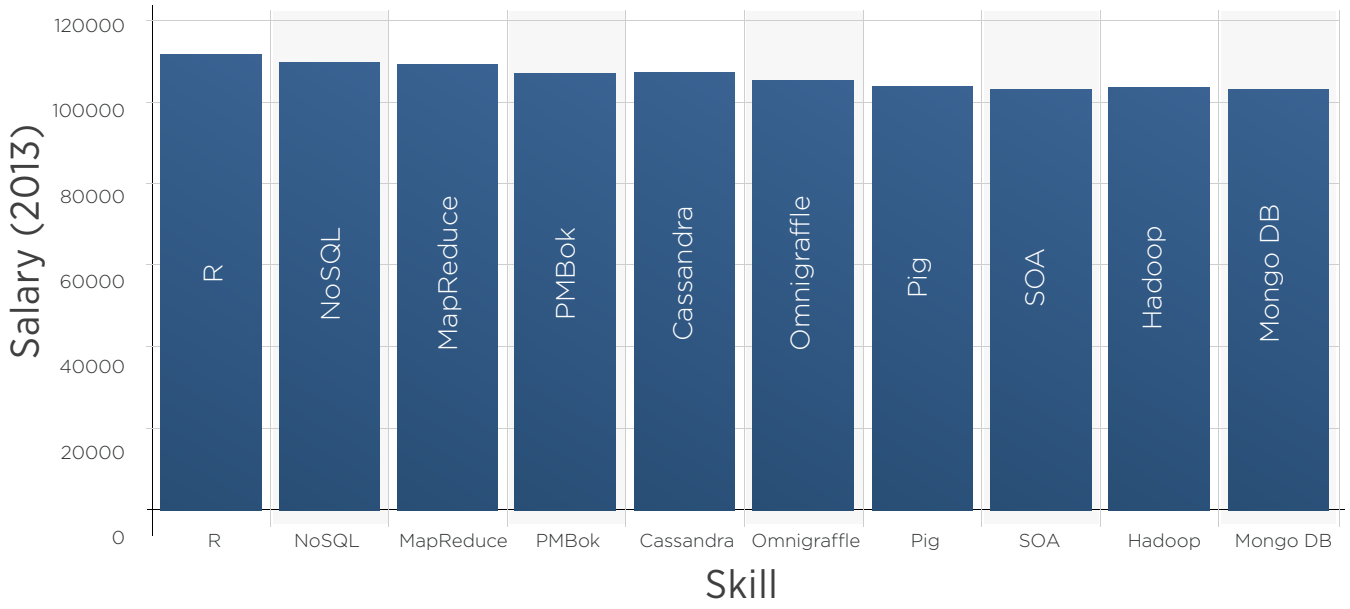
Though a few institutes are planning to start a Bachelor's degree program which will be in line with the Computer Science programs; in general trainings are provided for Master's degree programs. Apart from these programs, a number of institutes offer certification trainings in online, live-virtual classroom, and classroom learning modes, to facilitate students across the globe.

Must-have Skill sets

Technical Skills:

For working on real data, it is essential for data scientists to work with languages like R, Python, SAS, Hadoop, and more. Before we jump into the skill sets involved with each of these languages, let us take a look at the implications of these IT skills on the salary structures.

**Source: Dice.com 2014 Salary Survey*



For an aspiring data scientist, Python will be most important language to learn. When compared with other skills, Python is better at data processing. The simplicity of Python and the availability of ready to use machine learning tools like scikit learn, orange, etc. Python is a very important ingredient in the 'data processing' toolbox. It is a great starting language.

The next logical step would be to take R language training. Analyzing data and getting an idea (offline) of what works best, is what makes R language most-sought after skill by employers.

A 2014 survey of most-used statistics/programing language by data scientists have indicated that SAS user base has increased from the previous year. This means, the next destination for Data scientist aspirants will be SAS Base Programmer.

Though not a requirement, industry experts conclude Hadoop platform knowledge is essential for dealing with real data sets. Employers are always on a lookout on for data scientists with Hive or Pig experience alongside familiarity with cloud tools like Amazon S3 and similar ones.

Non-Technical Skills:

Apart from the technical skills, an aspiring data scientist must also focus on inculcating key non-technical, business-related skills.

Intellectual Curiosity

As mentioned in a post on Burtch Works, the main motivation factor for data scientists is the curiosity of making meaning inferences from the available data sets. Aspirants can initiate data science projects on their own, and draw inferences from them, in order to enhance their analytics skills.

Business Implications

As most of the data that is being analyzed is related to key business decisions, it is important that the data scientist should have adequate knowledge about the industry that he is working in and must understand the problem that the company is trying to solve. Thus, he must be able to discern business problems that are to be solved with data science.

Communication Skills

Employers prefer to hire data scientists who can easily translate their technical findings to a non-technical team. Thus, communication skills are very important skillsets for datascientists. Also, they need to understand the non-technical needs of data analysis and present quantified insights to 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 created below that will help understand the nuances of the topic.

4.4 million data scientists needed by 2015.

Useful Resources

Apart from enrolling for training with a renowned institute, it is also important to keep yourself well-informed about the basics and nuances of the field. This will be achieved only when you spend time in reading books, watching key videos, and going through some of the best articles on the subject.

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

- Big Data: A Revolution That Will Transform How We Live, Work, and Think by Viktor Mayer-Schonberger and Kenneth Cukier
- Big Data at Work: Dispelling the Myths, Uncovering the Opportunities by Thomas H. Davenport
- Data Science for Business: What you need to know about data mining and data-analytic thinking by Foster Provost and Tom Fawcett
- Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die by Eric Siegel
- Big Data Analytics with R and Hadoop by Prajapati

Apart from this, you can also visit sites like KDnuggets, R-blogger, DataTau, for keeping yourself updated with the latest trends in the field.

Data Science in Future

With the increasing use of data science across all industries, employers are now looking for skilled and certified professional sin the field. A recent report by Gartner predicts that more than 4.4 million IT jobs will be created by 2015 to support Big Data.

Data Science is expected to mature, consolidate, become the mainstream career option, and even surprise us with the advancements in the field. The field is expected to mature over time and slowly shift to cloud environment. Data science practitioners should be able to build predictive models in temporary cloud environments in order to increase their performance requirements. Unlike the current trend of single algorithm or tool being used to solve most of the data-related problems, the future is predicted to break this jinx. Data scientists are building new data algorithm to suit their needs, which are predicted to take advantage of parallel data processing to improve efficiency.

Simplilearn's Certified Analytics Professional course helps you in understanding the nuances of the field.

Classroom Sessions

16 hrs of High-Quality e-Learning content

10+ Case studies and tools videos

Course completion certificate

8 Business Analytics simulation exams

Learn R-studio

4 Real-life industry projects

Course aligned with SAS Global Certification Program