

A Path To Data Scientist: A Review

Ridhima Dhawan

ridhima.24015@lpu.co.in

*Assistant Professor, department of Computer Applications
Lovely Professional University*

Abstract:

Data science is the forthcoming of artificial intelligence which is going to rule for next one decade. Data science is a notion to combine statistics, data analysis, machine learning and its associated approaches with data to discover hidden designs or patterns from data which outlines the various trends of user behaviour and thus helps in prediction of the user demands.

Nevertheless, nobody has the solid facts and knowledge how to accomplish the data sciences projects. This paper précises the many technologies and tools that one ought to acquire to become the data scientist.

Keywords: *Data Scientist, Data, Tools, Life Cycle of Data Science.*

INTRODUCTION

Data Science is the future of artificial intelligence. Data science is a notion to combine statistics, data analysis, machine learning and its associated approaches with data to discover hidden designs or patterns from data.

Data science is primitively used in making decisions, reaching conclusions and predictions by using various algorithms such as Linear Regression, Logistic Regression, Bayes etc., tools such as SAS, Apache Spark etc., and machine learning.

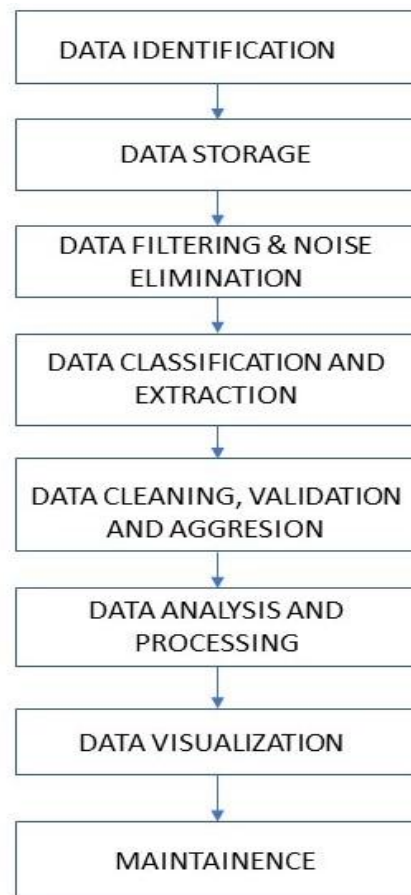
LIFE CYCLE -THE DATA SCIENCE

Now a days, data science and its projects are ruling the market, therefore becoming the backbone of the society. However, nobody has the dense knowledge how to execute the such projects and processes from capturing the data to analysis and outlining the hidden facts and designs in the data.

Following points discuss the lifecycle stages of data science (Chen,2014):

- I. *Data identification and collection-* In this stage large volume of data is collected from multiple resources to discover the buried relationships and associations
- II. *Data storage-* In this stage, the data, structured or unstructured, stored in databases or data warehouse. To accommodate the data, NoSQL databases are required such as Apache, Oracle etc. that permit analytics tools and techniques to fetch and process data from various repositories.
- III. *Data cleaning and redundancy elimination-* In this stage, replicated, corrupt and irrelevant data objects are removed from the gathered information. [3].
- IV. *Data classification and mining-* In this stage, unrelated data is extracted and converted into a common data format that the basic analytics tool can process. By extracting appropriate data, lead to decrease in the volume of data, which is submitted to various analytics software.
- V. *Data cleansing, validation and aggregation-* In this stage various validation rules are applied depending upon the business case to check data relevancy for data analysis. Due to data complexity, Aggregation helps to associate multiple data sets into fewer numbers based on common fields which simplifies further data processing.

- VI. *Data analysis and processing*-In this stage, data mining and analysis is carried out to treasure hidden patterns and design for making business decisions.
- VII. *Data visualization*- Inthis stage involves outlining analysis result into visual or graphical form that makes it easier to understand for the end users.



PROGRAMMING LANGUAGES FOR DATA SCIENCE

PYTHON

Python is a general-purpose construed, collaborating, object-oriented, and elevated programming language. It was developed by Guido van Rossum in 1985. It's an open basis code, obtainable under the GNU General Public License (GPL). It is a high-level language which not only support functional and structured programming methods but also object-oriented programming concepts. It can be easily combined with C, C++, JAVA, CORBA.

Python is one of the famous computer languages used by data scientist for data analysis during their researches. It is used for developing analytical procedure and exploring hidden facts and patterns in the data because it is highly interactive due to availability of the scientific libraries.

It consists of various libraries such as NumPy, pandas, matplotlib, SciPy etc.

NumPy: NumPy is a general-purpose package which process multi-dimensional arrays and also contain various mathematical functions which are useful for linear algebra, Fourier transform, and calculating random number.

Pandas: Pandas is a python library whose code is either written in C or Python. It provides the data structure which are quite useful in data analysis and data is analysed with series and data frames.

SciPy: SciPy stands for scientific python, data structures are reflected from NumPy. It is basically used for image processing, optimization, linear algebra, integration and various other mathematical functions.

R LANGUAGE

R is a language and environment for statistical calculating and graphics. R make available a variety of statistical techniques such as linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering etc., and graphical techniques, and is highly extensible. It has following features:

- syntax is quite easy which quicken our work on information.
- It has a variant format for loading and storing data for both local and over internet tasks.
- By using various consistent and easily understandable syntax, it has an ability to perform tasks in memory
- Since it has an open source, it has a long list of tools which are quite useful for analysis of data
- Users don't require to load again their data every time because the system keeps the data amid the sessions, and save the history of the commands.
- GUI, for R such as: RStudio, R commander, Rattle, JGR Java GUI for R (Siddiqui, T., Alkadri, M., & Khan, N. A. (2017))

COMPARISON BETWEEN PYTHON & R

R and python are the two programming languages which are extensively use for data sciences. Both the languages are quite flexible and open source languages. Following table compare the characteristics of the python and R.

Comparison between python and R¹

PYTHON	R
Object oriented language	Procedural language
Support all kind of data formats such as .CSV, JASON. Also, SQL tables can be imported directly.	Import data from excel, CSV and from text files into R. Also support SPSS format.
Pandas library is use for data analysis.	R has built functions and packages for statistical and numerical analysis
NumPy is used for numerical modelling analysis.	Some modelling analysis packages are not available such as Poisson distribution and mixture of probability laws

ARTIFICIAL INTELLIGENCE(AI)

AI program is called as an intelligent agent which interrelates the environment with agents in order to identify the state of an environment by using sensors and yielding output through actuators.

¹https://medium.com/@data_driven/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197

The critical issue of AI is to convert and transfer the inputs obtain by using sensors on the actuators. This is ended by a function inside the AI agent, which map the sensors to the actuators.

The final task of AI is to machines which has intelligence like humans. Nevertheless, such task can be completed by learning algorithms which attempt to imitate brains of humans work and learn (Das, Dey, A., Pal, A., & Roy, N(2015)).

Why is Artificial Intelligence Important?

Artificial intelligence systems play vital role for organizations that extract value from data by automation and optimization process, thus producing practical insights. Machine learning is a backbone of an artificial intelligence, enable companies to influence huge amounts of data yielding designs and patterns in the data which is impossible for a user to identify and predict.

- AI automates repetitive knowledge and sighting through data
- AI enhances intelligence
- AI familiarizes through advanced learning code driven approach which construct the programmes. AI finds patterns and symmetries in data so that the algorithm obtains ability: such code driven approaches develops a classifying algorithm or a prediction algorithm
- AI examines supplementary and bottomless data by means of neural networks that have numerous unseen layers.
- AI attains unbelievable accurateness
- AI grows the utmost out of data².

MACHINE LEARNING

Machine learning is one of the utmost electrifying fresh technologies in Artificial Intelligence. It is an ability of a machine to draw output on the basis of experiences without providing explicit instructions.

Tom Mitchell, an American computer scientist, has given the more prescribed explanation, said in respect to some task(T) an experience(E) and performance(P) must be measured, and if performance(P) improves the experience(E) then it is called as a program of a machine learning.

Types of the Machine learning Algorithms:

- **Supervised learning:** This algorithm compares the expected output with calculated output and if there is an error, then error is attuned in order to reach the expected output.
- **Unsupervised learning:** this algorithm finds the hidden pattern and design on the bases of the input pattern.
- **Reinforcement learning:** In this algorithm, software agents must yield actions in an atmosphere in order to make the most of output.
- **Recommender system:** this algorithm, online user, filters the information in order to predict the rating and preferences of online customer (Das, Dey, A., Pal, A., & Roy, N (2015)).
- **Regression algorithm:** Linear regression is a machine learning algorithm which predict the value on the basis of a dependent variable. Thus, outlining the relationship between the variables and the forecasted values by performing regression task.
- **Classification:** Classification is a technique of classifying the data into multiple classes i.e. discrete values by discovering or finding the functions. If-then rules are

²https://www.sas.com/en_in/insights/analytics/what-is-artificial-intelligence.html

applied on the derived mapping functions. The classification process deal with the problems where the data can be divided into binary or multiple discrete labels³.

DATA PROCESSING & ANALYSING

In this stage, data mining and analysis is carried out to treasure hidden patterns and design for making business decisions. Following table outline the characteristics and comparison of various tools used for data analysing processing (Narang, A. (2017))⁴.

Comparison of data processing and analysing tools

TOOLS	CHARACTERISTICS OF TOOL		
	ANALYSING MODEL	PROGRAMMING LANGUAGE	LATENCY
Hive	Streaming	SQL-like	high
Apache Spark	Mini/ micro batches, streaming	Scala, Java, Python	seconds
Apache Storm	A record at a time	Any	milli-seconds
MapReduce	Parallel Processing	Java, Ruby, Python, C++	More (seconds)
Qubole	Stram processing, ad-hoc queries	Python, Scala, R, Go	seconds
Flink	Batch and stram processing	Scala, Java, Python	seconds

DATA VISUALIZATION

Data visualization is a technique of representing the information and data through graph which helps in easier understanding for trends, designs present in the data. Data visualization include graphs, charts, maps, tables, dashboards, infographics. Following table defines and compares the various data visualization tool (Kaushal, S., & Bajwa, J. K. (2012)) (Ali, S, Gupta, N, Nayak, G. K., & Lenka, R. K. (2016, December)):

Comparison of data visualization tools

TOOLS	CHARACTERISTICS		
	SOURCE OF DATA	CODE DRIVEN APPROACH	OUTPUT
Data Wrapper	CSV, PDF, Excel, CMS	Ready-to-use codes	Bar chart, line chart, map, graphs
Tableau	Database, API	Not required	Maps, Bar charts, Scatter plots
Orange	files, SQL tables, and data tables or can paint random	Not required	scatter plots, bar charts, trees, dendrograms, networks and heat maps

³<https://www.geeksforgeeks.org/>

⁴<https://data-flair.training/blogs/apache-storm-vs-spark-streaming/>

	data		
Qlik	Database, spreadsheet, website	Programming language and SQL	Dashboard, Apps
Google Fusion tables	Comma-separated value file formats	Not required	pie charts, bar charts, lineplots, scatterplots, timelines

TABLEAU

Tableau is the most famous data visualization tool, used for business intelligence for outlining various hidden patterns and designs in the form of graphs and charts. It can join to various files, relational and big data sources to capture and process data. It not only blends the data, but also provides the real time collaboration, thus making it unique. It is mainly used for visual data analysis by various organization for analysing trade prediction, research scholars and government administrations.

Tableau join and pull the data stored in various repositories and clouds either connected live or extracted by tableau's data engine, on tableau desktop where the data analyst, data engineer work with the abstracted data and hence creating data visualizations by analysing it. From simple to complex databases, tableau is used for extracting data from the databases⁵.

COMPARISON STUDY BETWEEN R AND TABLEAU

Following table discuss the comparison study between R besides tableau (Rajeswari,Basu, & Maurya, N. (2017, November)).

Comparison of R language and tableau

R LANGUAGE	TABLEAU
R is the utmost systematic quantifiable examination package existing	Tableau has an outstanding user interface.
The graphical facilities of R are odd, giving an entirely code driven strategy dialect that beats most other computable and graphical packages.	Since it has an integration characteristic, therefore it can integrate with big-data software's such as Hadoop.
R has an open source programming license which allow anyone to use and change it.	Tableau supports in mobile gadget which optimize the tableau dashboard routinely.
Since R is an open source software, therefore it allows to run anywhere and anytime.	It can be easily upgraded and also consumes less memory.

CONCLUSION

⁵<https://www.guru99.com/what-is-tableau.html>

Data science is rapidly developing and is becoming one of the hottest areas in the technology industry. With swift developments in computational performance and using and learning the above-mentioned tools, can permit for the analysis of huge datasets, discovering various hidden patterns and designs in the data which manifest the various trends prevailing, helps in meeting and predicting the online customer demands.

By learning fore-mentioned tools and technologies, one can become data scientist which is becoming high ranking profession for next one decade in the era of big data yielding pack of countless careers.

REFERENCES

- 1) Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, 19(2), 171-209.
- 2) Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. *International Journal of Computer Applications*, 115(9).
- 3) Rajeswari, C., Basu, D., & Maurya, N. (2017, November). Comparative Study of Big data Analytics Tools: R and Tableau. In *Materials Science and Engineering Conference Series* (Vol. 263, No. 4, p. 042052).
- 4) Siddiqui, T., Alkadri, M., & Khan, N. A. (2017). Review of Programming Languages and Tools for Big Data Analytics. *International Journal of Advanced Research in Computer Science*, 8(5).
- 5) Kaushal, S., & Bajwa, J. K. (2012). Analytical review of user perceived testing techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(10).
- 6) Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016, December). Big data visualization: Tools and challenges. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 656-660). IEEE.
- 7) Narang, A. (2017). A review-Cloud and cloud security. *International journal of Computer Science and mobile Computing*, 6(1), 178-181.