22 Widely Used Data Science and Machine Learning Tools in 2020

ARTIFICIAL INTELLIGENCE | BEGINNER | BIG DATA | BUSINESS ANALYTICS | BUSINESS INTELLIGENCE | DATA SCIENCE | DATABASE | LISTICLE | MACHINE LEARNING

Overview

- There are a plethora of data science tools out there – which one should you pick up?
- Here's a list of over 20 data science tools catering to different stages of the data science lifecycle

Introduction

What are the best tools for performing data science tasks? And which tool should *you* pick up as a newcomer in data science?

I'm sure you've asked (or searched for) these questions at some point in your own data science journey. These are valid questions! There is no shortage of data science tools in the industry. Picking one for your journey and career can be a tricky decision.

Let's face it – data science is a vast spectrum and each of its domains requires handling of data in a unique way that leads many analysts/data scientists into confusion. And if you're a business leader, you would come across crucial questions regarding the tools you and your company choose as it might have a long term impact.

So again, the question is which data science tool should you choose?

In this article, I will be attempting to clear this confusion by listing down widely used tools used in the data science space broken down by their usage and strong points. So let us get started!

And if you're a newcomer to machine learning and/or business analytics, or are just getting started, I encourage you to leverage an incredible initiative by Analytics Vidhya called UnLock 2020. Covering two comprehensive programs – Machine Learning Starter Program and the Business Analytics Starter Program – this initiative is time-bound so you'd need to enroll as soon as you can to give your data science career a massive boost!

Table of Contents

Data Science Tools for Big Data

To truly grasp the meaning behind Big Data, it is important that we understand the basic principles that define the data as big data. These are known as the 3 V's of big data:

- Volume
- Variety
- Velocity

Tools for Handling Volume

As the name suggests, volume refers to the scale and the amount of data. To understand the scale of the data I'm talking about, you need to know that over 90% of the data in the world was created in just the last two years!

Over the decade, with the increase in the amount of data, the technology has also become better. The decrease in computational and storage costs has made collecting and storing huge amounts of data far easier.

The volume of the data defines whether it qualifies as big data or not.

When we have data ranging from 1Gb to around 10Gb, the traditional data science tools tend to work well in these cases. So what are these tools?

- **Microsoft Excel** – Excel prevails as the easiest and most popular tool for handling small amounts of data. The maximum amount of rows it supports is just a shade over 1 million and one sheet can handle only up to 16,380 columns at a time. These numbers are simply not enough when the amount of data is big.

- **Microsoft Access** – It is a popular tool by Microsoft that is used for data storage. Smaller databases up to 2Gb can be handled smoothly with this tool but beyond that, it starts cracking up.

- **SQL** – SQL is one of the most popular data management systems which has been around since the 1970s. It was the primary database solution for a few decades. SQL still remains popular but there's a drawback – It becomes difficult to scale it as the database continues to grow.

We have covered some of the basic tools so far. It is time to unleash the big guns now! If your data is greater than 10Gb all the way up to storage greater than 1Tb+, then you need to implement the tools I've mentioned below:

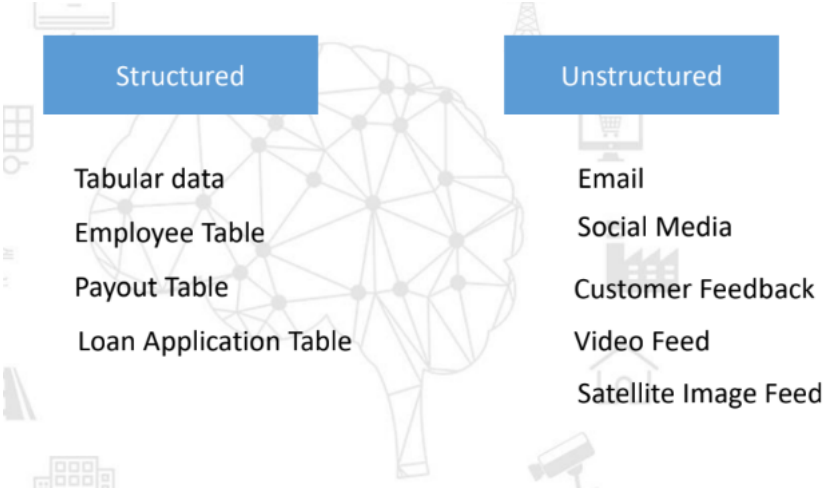
- **Hadoop** – It is an open-source distributed framework that manages data processing and storage for big data. You are likely to come across this tool whenever you build a machine learning project from scratch.

- **Hive** – It is a data warehouse built on top of Hadoop. Hive provides a SQL-like interface to query the data stored in various databases and file systems that integrate with Hadoop.

Tools for Handling Variety

Variety refers to the different types of data that are out there. The data type may be one of these – Structured and Unstructured data.

Let us go through the examples falling under the umbrella of these different data types:



Take a moment to observe these examples and correlate them with your real-world data.

As you might have observed in the case of **Structured data**, there is a certain order and structure to these data types whereas in the case of **unstructured data**, the examples do not follow any trend or pattern. For example, customer feedback may vary in length, sentiments, and other factors. Moreover, these types of data are huge and diverse.

It can be very challenging to tackle this type of data, so what are the different data science tools available in the market for managing and handling these different data types?

The two most common databases are **SQL** and **NoSQL**. SQL has been the market-dominant players for a number of years before NoSQL emerged.

Some examples for SQL are Oracle, MySQL, SQLite, whereas NoSQL consists of popular databases like MongoDB, Cassandra, etc. **These NoSQL databases are seeing huge adoption numbers because of their ability to scale and handle dynamic data.**

Tools for Handling Velocity

The third and final V represents the velocity. This is the speed at which the data is captured. This includes both real-time and non-real-time data. We'll be talking mainly about the real-time data here.

We have a lot of examples around us that capture and process real-time data. The most complex one is the sensor data collected by self-driving cars. Imagine being in a self-driving car – the car has to dynamically collect and process data regarding its lane, distance from other vehicles, etc. all at the same time!

Some other examples of real-time data being collected are:

- CCTV
- Stock trading
- Fraud detection for credit card transaction
- Network data – social media (Facebook, Twitter, etc.)

> Did you know?
>
> More than 1Tb of data is generated during each trade session at the New York stock exchange!

Now, let's head on to some of the commonly used data science tools to handle real-time data:

- **Apache Kafka** – Kafka is an open-source tool by Apache. It is used for building real-time data pipelines. Some of the advantages of Kafka are – It is fault-tolerant, really quick, and used in production by a large number of organizations.

- **Apache Storm** – This tool by Apache can be used with almost all the programming languages. It can process up to 1 Million tuples per second and it is highly scalable. It is a good tool to consider for high data velocity.

- **Amazon Kinesis** – This tool by Amazon is similar to Kafka but it comes with a subscription cost. However, it is offered as an out-of-the-box solution which makes it a very powerful option for organizations.

- **Apache Flink** – Flink is yet another tool by Apache that we can use for real-time data. Some of the advantages of Flink are high performance, fault tolerance, and efficient memory management.

Now that we have a solid grasp on the different tools commonly being used for working with Big Data, let's move to the segment where you can take advantage of the data by applying advanced machine learning techniques and algorithms.

Widely Used Data Science Tools

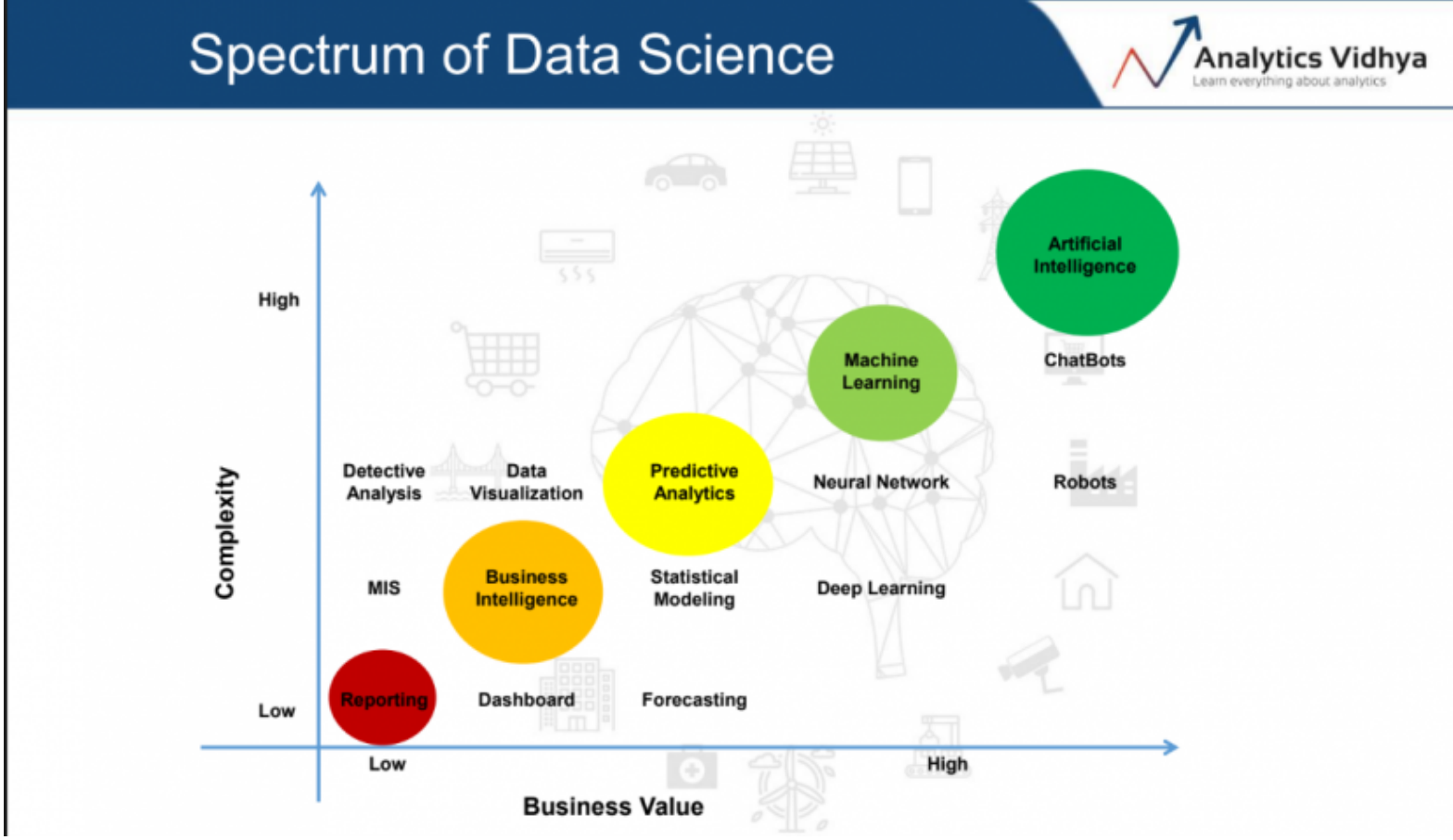
If you're setting up a brand new data science project, you'll have a ton of questions in mind. This is true regardless of your level – whether you're a data scientist, a data analyst, a project manager, or a senior data science executive.

Some of the questions you'll face are:

- Which tools should you use in different domains of data science?
- Should I buy licenses for the tools or opt for an open-source one?, and so on

In this section, we will be discussing some of the popular data science tools used in the industry according to different domains.

Data Science is a broad term in itself and it consists of a variety of different domains and each domain has its own business importance and complexity which is beautifully captured in the below image:



The data science spectrum consists of various domains and these domains are represented by their relative complexity and the business value that they provide. Let us take up each one of the points I've shown in the above spectrum.

Reporting and Business Intelligence

Let's begin with the lower end of the spectrum. It enables an organization to identify trends and patterns so as to make crucial strategic decisions. The types of analysis range from MIS, data analytics, all the way over to dashboarding.

The commonly used tools in these domains are:

- **Excel** – It gives a diverse range of options including Pivot tables and charts that let you do analysis in double-quick time. This is, in short, the Swiss Army Knife of data science/analytics tools
- **QlikView** – It lets you consolidate, search, visualize, and analyze all your data sources with just a few clicks. It is an easy and intuitive tool to learn which makes it so popular.

- Tableau – It is amongst the most popular data visualization tools in the market today. It is capable of handling large amounts of data and even offers Excel-like calculation functions and parameters. Tableau is well-liked because of its neat dashboard and story interface.

- **Microstrategy** – It is yet another BI tool that supports dashboards, automated distributions, and other key data analytics tasks.

- **PowerBI** – It is a Microsoft offering in the Business Intelligence (BI) space. PowerBI was built to integrate with Microsoft technologies. So if your organization has a Sharepoint or SQL database user, you and your team will love working on this tool.

- **Google Analytics** – Wondering how did Google Analytics make it to this list? Well, digital marketing plays a major role in transforming businesses and there's no better tool than this to analyze your digital efforts.

Predictive Analytics and Machine Learning Tools

Moving further up the ladder, the stakes just got high in terms of complexity as well as the business value! This is the domain where the bread and butter of most data scientists come from. Some of the types of problems you'll solve are statistical modeling, forecasting, neural networks, and deep learning.

Let us understand the commonly used tools in this domain:

- **Python** – This is one of the most dominant languages for data science in the industry today because of its ease, flexibility, open-source nature. It has gained rapid popularity and acceptance in the ML community.

- **R** – It is another very commonly used and respected language in data science. R has a thriving and incredibly supportive community and it comes with a plethora of packages and libraries that support most machine learning tasks.

- **Apache Spark** – Spark was open-sourced by UC Berkley in 2010 and has since become one of the largest communities in big data. It is known as the swiss army knife of big data analytics as it offers multiple advantages such as flexibility, speed, computational power, etc.

- **Julia** – It is an upcoming language and is being touted as the successor of Python. It's still in its nascent stage and it will be interesting to see how it performs in the future.

- **Jupyter Notebooks** – These notebooks are widely used for coding in Python. While it is predominantly used for Python, it also supports other languages such as Julia, R, etc.

The tools we have discussed so far are true open-source tools. You don't require to pay for them or buy any extra licenses. They have thriving and active communities that maintain and release updates on a regular basis.

Now, we will check out some premium tools that are recognized as industry leaders:

- **SAS** – It is a very popular and powerful tool. It's prevalently and commonly used in the banking and financial sectors. It has a very high share in private organizations like American Express, JP Morgan, Mu Sigma, Royal Bank of Scotland, etc.

- **SPSS** – Short for Statistical Package for Social Sciences, SPSS was acquired by IBM in 2009. It offers advanced statistical analysis, a vast library of machine learning algorithms, text analysis, and much more.

- **Matlab** – Matlab is really underrated in the organizational landscape but it is widely used in academia and research divisions. It has lost a lot of ground in recent times to the likes of Python, R, and SAS but universities, especially in the US, still teach a lot of undergraduate courses using Matlab.

Common Frameworks for Deep Learning

Deep Learning requires high computational resources and needs special frameworks to utilize those resources effectively. Due to this, you would most likely require a GPU or a TPU.

Let us look at some of the frameworks used for Deep Learning in this section.

- **TensorFlow** – It is easily the most widely used tool in the industry today. Google might have something to do with that!
- **PyTorch** – This super flexible deep learning framework is giving major competition to TensorFlow. PyTorch has recently come into the limelight and was developed by researchers at Facebook
- **Keras** and **Caffe** are other frameworks used extensively for building deep learning applications

Artificial Intelligence Tools

The era of AutoML is here. If you haven't heard of these tools, then it is a good time to educate yourself! This could well be what you as a data scientist will be working with in the near future.

Some of the most popular AutoML tools are **AutoKeras, Google Cloud AutoML, IBM Watson, DataRobot, H20's Driverless AI,** and **Amazon's Lex.** AutoML is expected to be the next big thing in the AI/ML community. It aims to eliminate or reduce the technical side of things so that business leaders can use it to make strategic decisions.

These tools will be able to automate the complete pipeline!

End Notes

We have discussed the data collection engine and the tools required to accomplish the pipeline for retrieval, processing, and storage of data. Data Science consists of a large spectrum of domain and each domain has its own set of tools and frameworks.

Picking your data science tool will often come down to your personal choice, your domain or project, and of course, your organization.

Let me know in the comments about your favorite data science tool or framework that you love to work with!

Article Url - https://www.analyticsvidhya.com/blog/2020/06/22-tools-data-science-machine-learning/

Ram Dewani

Product Growth Analyst at Analytics Vidhya. I'm always curious to deep dive into data, process it, polish it so as to create value. My interest lies in the field of marketing analytics.