What is Data?

Data is a collection of information gathered by observations, measurements, research or analysis. They are stored in the formats of tables, text, images, video, audio, and more.

If data availability is less than 2.5 quintillion bytes referred simply as “Data”.

If data availability is greater than 2.5 quintillion bytes referred as “Big Data”. It has more volume, velocity, and variety of data.

In real time applications the data are generated is very big or Large.

For example, Twitter Monthly data, if we have lots of data - analysis, understanding and making organizational decision-making becomes hard.

To solve this, we use different Technologies like Data Science, artificial intelligence (AI) and machine learning. All these technologies combinedly enable analysts and business professionals to use a company’s big data to improve organizational decision-making, optimize processes, and give an optimal solution to the query.

1.What is Data Science?

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to “extract meaningful insights from data”.

Data science is all about using data to solve problems. The problem could be **decision making** such as

- identifying which email is spam and which is not.
- a product recommendation such as which movie to watch?
- predicting the outcome such as who will be the next CEO

So, the core job of a data scientist is to **understand the data, extract useful information** out of it and apply this in solving the problems.

Data science practitioners apply machine learning algorithms to tables, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into tangible business value.

What is data science used for?

Data science is used to study data in four main ways:

Descriptive Analysis - “What happened?” – Visualizations.

Diagnostic Analysis - “Why did this happen?” – data operations and Transformations.

Predictive Analysis - “What might happen in the future?” – forecasting the occurrence.

Prescriptive Analysis - “What should we do next?” - suggests an optimum response.

1. Descriptive analysis

Descriptive analysis examines data to gain insights into what happened or what is happening in the data environment. It is characterized by data visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives.

For example, a flight booking service may record data like the number of tickets booked each day. Descriptive analysis will reveal booking spikes, booking slumps, and high-performing months for this service.

2. Diagnostic analysis

Diagnostic analysis is a deep-dive or detailed data examination to understand why something happened. It is characterized by techniques such as drill-down, data discovery, data mining, and correlations. Multiple data operations and transformations may be performed on a given data set to discover unique patterns in each of these techniques.

For example, the flight service might drill down on a particularly high-performing month to better understand the booking spike. This may lead to the discovery that many customers visit a particular city to attend a monthly sporting event.

3. Predictive analysis

Predictive analysis uses historical data to make accurate forecasts about data patterns that may occur in the future. It is characterized by techniques such as machine learning, forecasting, pattern matching, and predictive modeling. In each of these techniques, computers are trained to reverse engineer causality connections in the data.

For example, the flight service team might use data science to predict flight booking patterns for the coming year at the start of each year. The computer program or algorithm may look at past data and predict booking spikes for certain destinations in May. Having anticipated their customer’s future travel requirements, the company could start targeted advertising for those cities from February.

4. Prescriptive analysis

Prescriptive analytics takes predictive data to the next level. It not only predicts what is likely to happen but also suggests an optimum response to that outcome. It can analyze the potential implications of different choices and recommend the best course of action. It uses graph analysis, simulation, complex event processing, neural networks, and recommendation engines from machine learning.

For example, flight booking, prescriptive analysis could look at historical marketing campaigns to maximize the advantage of the upcoming booking spike. A data scientist could project booking outcomes for different levels of marketing spend on various marketing channels. These data forecasts would give the flight booking company greater confidence in their marketing decisions.

2. Who is Data Scientist

A data scientist is an analytics professional who is responsible for collecting, analyzing and interpreting data to help drive decision-making in an organization.

A data scientist finds a way to present the data in a useful form when compared to the data available in the data set. They work with both structured and unstructured data.

The basic responsibilities of a data scientist include the following activities:

- gathering and preparing relevant data to use in analytics applications;
- using various types of analytics tools to detect patterns, trends and relationships in data sets;
- developing statistical and predictive models to run against the data sets; and
- creating data visualizations, dashboards and reports to communicate their findings.

3. Differences Between Data Science and Business Intelligence

Factors	Business Intelligence	Data Science
Concept	Deals with data analysis on the business platform	Consists of several data operations in various domains
Scope	BI analyzes past data	Past data is analyzed for future predictions.
Data	Handling static and structured data	Both structured & unstructured data that is also dynamic
Data Storage	Data stored mostly in data-warehouses	Data utilized is distributed in real time cluster.
Procedure	BI helps companies to solve questions.	Questions are both curated and solved by data scientists.
Tools	MS Excel, SAS BI, Sisense, Micro strategy.	Python R, Hadoop/Spark, SAS, TensorFlow.

What Is Business Intelligence (BI)?

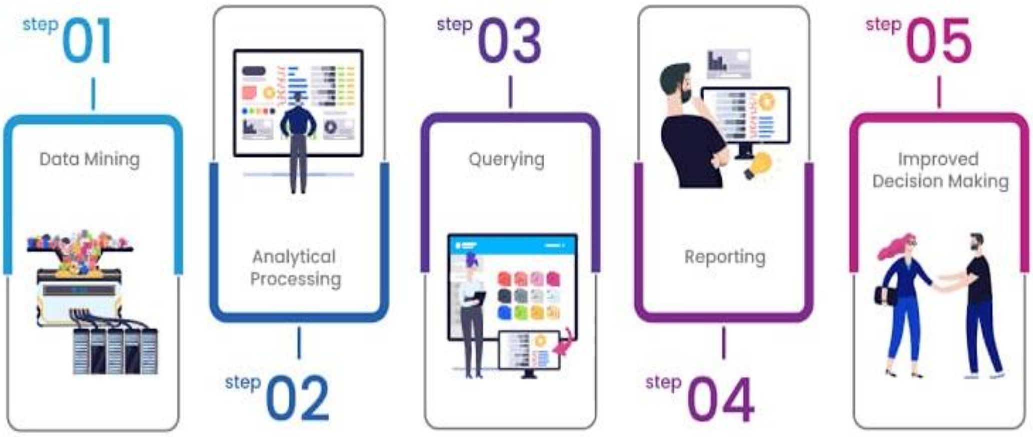
Business intelligence (BI) refers to the procedural and technical infrastructure that collects, stores, and analyzes the data produced by a company's activities.

BI is a broad term that include data mining, process analysis, performance standards, and descriptive analytics.

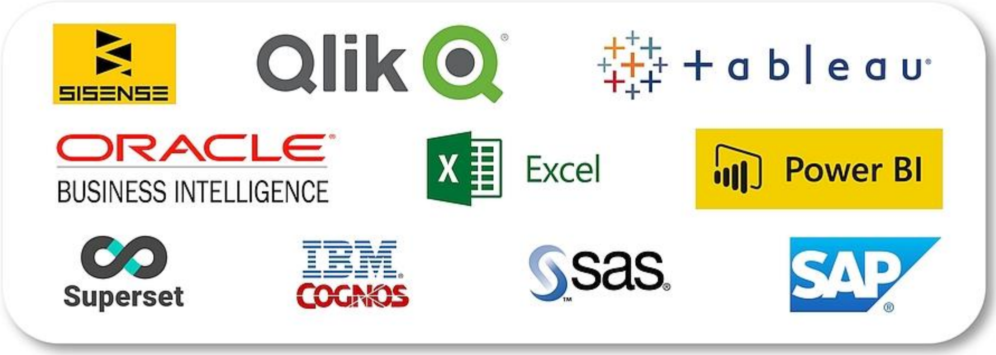
A business intelligence strategy is generating a Blueprint for deciding how you will use data in your company. You need a strategy because merely choosing the right technology, and implementing a software platform is not enough to realize a return on investment.

Business intelligence (BI) and its application uses a software to convert reams of information into bite-sized insights to inform decision-making. The software receives data from a company's enterprise resource planning ERP system and other data sets via a sync tool or API. The BI tool then analyses the data sets and presents findings in reports and dashboards.

How Business Intelligence works:



BI **tools and software's** are in a wide variety of forms.



some common types of BI solutions are:

- ***Spreadsheets:*** Spreadsheets like Microsoft Excel and Google Docs are some of the most widely used BI tools.
- ***Reporting software:*** Reporting software is used to report, organize, filter, and display data.
- ***Data visualization software:*** Data visualization software translates datasets into easy-to-read, visually appealing graphical representations to quickly gain insights.
- ***Data mining tools:*** Data mining tools "mine" large amounts of data for patterns using things like artificial intelligence, machine learning, and statistics.
- ***Online analytical processing (OLAP):*** OLAP tools allow users to analyze datasets from a wide variety of angles based on different business perspectives.

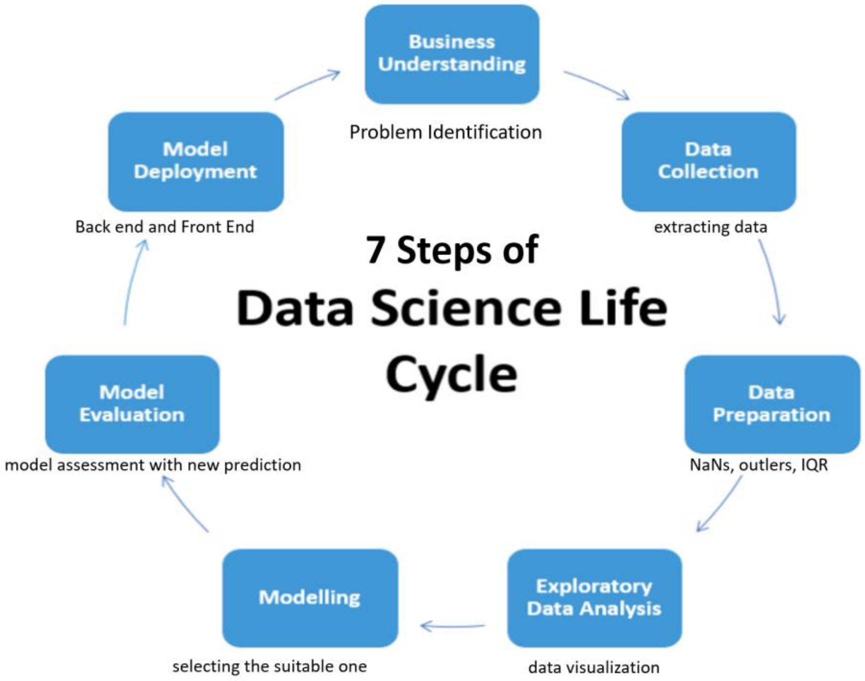
Now days the BI is cloud cost management which is on top of the mind for enterprises, and companies will continue to invest in the cloud but with more caution. And they have help. Active metadata platforms like Atlan are helping reduce costs by optimizing data processing and removing unused data to make space.

Factors	Business Intelligence	Data Science
Concept	Deals with data analysis on the business platform	Consists of several data operations in various domains
Scope	BI analyzes past data	Past data is analyzed for future predictions.
Data	Handling static and structured data	Both structured & unstructured data that is also dynamic
Data Storage	Data stored mostly in data-warehouses	Data utilized is distributed in real time cluster.
Procedure	BI helps companies to solve questions.	Questions are both curated and solved by data scientists.
Tools	MS Excel, SAS BI, Power BI, Tableau, Sisense, Micro strategy.	Python R, Hadoop/Spark, SAS, TensorFlow.

4. Lifecycle of Data science:

When data professionals start to work on new project they involve data analysis, in phases of lifecycle. They are

1) Problem Identification and Business understanding
2) Data collection and understanding
3) Data cleaning and preparation - Preprocessing
4) Exploratory Data Analysis
5) Model Building
6) Model evaluation and interpretation
7) Deployment and communication Findings



we look in more detail at each of the steps.

1. Problem Identification and Business understanding: Many developments in the world first started with the question of "why".

a good data science life cycle starts with "why". It is essential to understand the business objective clearly because that will be your final goal of the analysis.

In this first phase of data analytics, the stakeholders regularly perform the following tasks — examine the business trends, make case studies of similar data analytics, and study the domain of the business industry. The entire team makes an assessment of the in-house resources, the in-house infrastructure, total time involved, and technology requirements. Once all these assessments and evaluations are completed, the stakeholders start formulating the initial hypothesis for resolving all business challenges in terms of the current market scenario.

2. ***Data collection and understating:*** After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used for this commercial enterprise problem, and different information. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.

3. ***Data preparation and preprocessing:***. Generally, data scientists or business/data analysts make that phase. This includes

- Selecting the relevant data,
- Integrating the data by merging the data sets,
- Cleaning them,
- Treating the missing values by either removing them or imputing them,
- Treating erroneous data by removing them,
- Checking outliers using box plots and handle them with statistical approach

4. Exploratory Data Analysis: This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and heat maps. Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

5.Modeling Building: Data modelling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

6. Model Evaluation: Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved. **Common metrics used to evaluate models:**
Classification metrics: Precision-Recall, ROC-AUC, Accuracy, Log-Loss
Regression metrics: MSE, MAE, R Square, Adjusted R Square
Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric. We can construct

more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

7. Model Deployment: The model after a rigorous assessment is at the end deployed in the preferred structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste. For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.

5. Pros and Cons of Data Science

Data science is an upcoming field, having multiple opportunities that has pros and cons.

1. Pros of Data Science

The various benefits of data science are as follows:

a) *it's in demand*

b) *abundance of positions*

c) *A Highly Paid Career*

d) *Data Science is Versatile*

e) *Data Science Makes Data Better*

f) *Data Scientists are Highly Prestigious*

g) *No More Boring Tasks*

h) *Data Science Makes Products Smarter*

i) *Data Science can Save Lives.*

j) *Data Science Can Make You A Better Person*

The detailed explanation for various benefits of data science are as follows:

a) *it's in demand*

Data science is greatly in demand. prospective job seekers have numerous opportunities. it is the fastest growing job on linkedin and is predicted to create 11.5 million jobs by 2026. this makes data science a highly employable job sector.

b) *2. abundance of positions*

there are very few people who have the required skill-set to become a complete data scientist. this makes data science less saturated as compared with other it sectors.

Therefore, Data Science is a vastly abundant field and has a lot of opportunities. The field of Data Science is high in demand but low in supply of Data Scientists.

c) *A Highly Paid Career*

Data Science is one of the most highly paid jobs. According to Glassdoor, Data Scientists make an average of $116,100 per year. This makes Data Science a highly lucrative career option.

d) *Data Science is Versatile*

There are numerous applications of Data Science. It is widely used in health-care, banking, consultancy services, and e-commerce industries. Data Science is a very versatile field. Therefore, you will have the opportunity to work in various fields.

e) *Data Science Makes Data Better*

Companies require skilled Data Scientists to process and analyze their data. They not only analyze the data but also improve its quality. Therefore, Data Science deals with enriching data and making it better for their company.

f) *Data Scientists are Highly Prestigious*

Data Scientists allow companies to make smarter business decisions. Companies rely on Data Scientists and use their expertise to provide better results to their clients. This gives Data Scientists an important position in the company.

g) *No More Boring Tasks*

Data Science has helped various industries to automate redundant tasks. Companies are using historical data to train machines in order to perform repetitive tasks. This has simplified the arduous jobs undertaken by humans before.

h) *Data Science Makes Products Smarter*

Data Science involves the usage of Machine Learning which has enabled industries to create better products tailored specifically for customer experiences.

For example, Recommendation Systems used by e-commerce websites provide personalized insights to users based on their historical purchases. This has enabled computers to understand human-behavior and take data-driven decisions.

i) *Data Science can Save Lives*

Healthcare sector has been greatly improved because of Data Science. With the advent of machine learning, it has been made easier to detect early-stage tumors. Also, many other health-care industries are using Data Science to help their clients.

j) *Data Science Can Make You A Better Person*

Data Science will not only give you a great career but will also help you in personal growth. You will be able to have a problem-solving attitude. Since many Data Science roles bridge IT and Management, you will be able to enjoy the best of both worlds.

b. Disadvantages of Data Science

The various disadvantages to this field are

a. *Data Science is Blurry Term.*
b. *Mastering Data Science is near to impossible*
c. *Large Amount of Domain Knowledge Required*
d. *Arbitrary Data May Yield Unexpected Results*
e. *Problem of Data Privacy*

The detail explanation for various disadvantages to this field are

a. *Data Science is Blurry Term* - it is very hard to write down the exact meaning of a Data Scientist.
b. *Mastering Data Science is near to impossible* - It's a mixture of many fields, like Statistics, Computer Science and Mathematics. Mastering is difficult.
c. *Large Amount of Domain Knowledge Required* - A person with a considerable background in Statistics and Computer Science will find it difficult to solve/predict a health issue.
A health-care industry working on an analysis of genomic sequences will require a suitable employee with some knowledge of genetics and molecular biology.
d. *Arbitrary Data May Yield Unexpected Results* :A Data Scientist analyzes the data and makes careful predictions in order to facilitate the decision-making process. Many times, the data provided is arbitrary and does not yield expected results. This can also fail due to weak management and poor utilization of resources.

e. *Problem of Data Privacy*

For many industries, data is their fuel. Data Scientists help companies make data-driven decisions. However, the data utilized in the process may breach the privacy of customers.The personal data of clients are visible to the parent company and may at times cause data leaks due to lapse in security. The ethical issues regarding preservation of data-privacy and its usage have been a concern for many industries.

6. Skills Required for Data Science:

Technical skills

Data science is a rapidly growing field, and as such, the skills required for a data scientist are constantly evolving. However, certain technical skills are considered essential for a data scientist to possess. These skills are often listed prominently in job descriptions and are highly sought after by employers.

These skills include programming languages such as Python and R, statistics and probability, machine learning, data visualization, and data modeling. Many of these skills can be developed through formal education and business training programs, and organizations are placing an increasing emphasis on them as they continue to expand their analytics and data teams.

1. Prepare data for effective analysis

One important data scientist skill is preparing data for effective analysis. This includes sourcing, gathering, arranging, processing, and modeling data, as well as being able to analyze large volumes of structured or unstructured data.

The goal of data preparation is to present data in the best forms for decision-making and problem-solving. This skill is crucial for any data scientist as it enables them to take raw data and make it usable for analysis and insights discovery. Data preparation is an essential step in the data science workflow, and data scientists should be familiar with various data preparation tools and best practices.

2. Data visualization

Data visualization is a powerful tool for data scientists to effectively communicate their findings and insights to both technical and non-technical audiences.

Having a strong understanding of the benefits and challenges of using data visualization, as well as basic knowledge of market solutions, allows data scientists to create clear and informative visualizations that effectively communicate their insights.

This skill includes an understanding of best practices and techniques for creating data visualizations, and the ability to share results through self-service dashboards or applications.

Self-service analytics platforms allow data scientists to surface the results of their data science processes and explore the data in a way that is easily understandable to non-technical stakeholders, which is crucial for driving data-driven decisions and actions.

3. Programming

Data scientists need to have a solid foundation in Programming languages such as Python, R, and SQL. These languages are used for data cleaning, manipulation, and analysis, and for building and deploying machine learning models.

Python is widely used in the data science community, with libraries such as Pandas and NumPy for data manipulation, and Scikit-learn for machine learning. R is also popular among statisticians and data analysts, with libraries for data manipulation and machine learning.

SQL is a must-have for data scientists as it is a database language and allows them to extract data from databases and manipulate it easily.

4. Ability to apply math and statistics appropriately

Exploratory data analysis is a crucial step in the data science process, as it allows data scientists to identify important patterns and relationships in the data, and to gain insights that inform decisions and drive business growth.

To perform exploratory data analysis effectively, data scientists must have a strong understanding of math and statistics. Understanding the assumptions and algorithms underlying different analytic techniques and tools is also crucial for data scientists.

Without this understanding, data scientists risk misinterpreting the results of their analysis or applying techniques incorrectly. It is important to note that this skill is not only important for students and aspiring data scientists but also for experienced data scientists.

5. Machine learning and artificial intelligence (AI)

Machine learning and artificial intelligence (AI) are rapidly advancing technologies that are becoming increasingly important in data science. However, it is important to note that these technologies will not replace the role of data scientists in most organizations.

Instead, they will enhance the value that data scientists deliver by providing new and powerful tools to work better and faster. One of the key challenges in using AI and machine learning is knowing if you have the right data. Data scientists must be able to evaluate the quality of the data, identify potential biases and errors, and determine.

Non-Technical Skills

In addition to technical skills, soft skills are also essential for data scientists to possess to succeed in the field.

These skills include

a. critical thinking,
b. effective communication,
c. proactive problem-solving,
d. intellectual curiosity.
e. Team work

These skills may not require as much technical training or formal certification, but they are foundational to the rigorous application of data science to business problems. They help data scientists to analyze data objectively, communicate insights effectively, solve problems proactively, and stay curious and driven to find answers.

Even the most technically skilled data scientist needs to have these soft skills to make an impact in any organization and stand out in a competitive job market.

Critical thinking

The ability to objectively analyze questions, hypotheses, and results, understand which resources are necessary to solve a problem, and consider different perspectives on a problem.

Effective communication

The ability to explain data-driven insights in a way that is relevant to the business and highlights the value of acting.

Proactive problem solving

The ability to identify opportunities, approach problems by identifying existing assumptions and resources, and use the most effective methods to find solutions.

Intellectual curiosity

The drive to find answers, dive deeper than surface results and initial assumptions, think creatively, and constantly ask “why” to gain a deeper understanding of the data.

Teamwork

The ability to work effectively with others, including cross-functional teams, to achieve common goals. This includes strong collaboration, communication, and negotiation skills.

7. Tools Required for Data Science:

the most popular programming languages used in the data science are

a. R,
b. Python,
c. Hadoop,
d. Structured Query Language (SQL),
e. SAS Apache Spark –
f. D3.js

1) R Programming:

R is an open-source, domain-specific language, explicitly designed for data science. Very popular in finance and academia, R is a perfect language for data manipulation, processing and visualization, as well as statistical computing and machine learning.

2) Python stands as a predominant programming language in the realm of data science. Renowned for its versatility, Python empowers professionals to analyze vast and varied datasets, whether structured, semi-structured, or unstructured.

Its user-friendly syntax and extensive library to facilitate tasks such as data analysis, cleansing, and visualization. Python's applicability extends beyond its data science capabilities to mobile, desktop, and web application development. Its robust community continually enhances its library offerings, positioning Python as a go-to tool for data science and software development endeavors. Here are the best data science tools for Python:

3) Apache Hadoop

The Apache® Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. Hadoop is perfect for high-level computations.

4) SQL Much of the world's data is stored in databases. SQL (Structured Query Language) is a domain-specific language that allows programmers to communicate with, edit and extract data from databases.

5) Statistical Analytical System SAS is a software suite that can mine, alter, manage and retrieve data from a variety of sources and perform statistical analysis on it. SAS provides a graphical point-and-click user interface for non-technical users and more through the SAS language.

3 Apache Spark

Also called "Spark," this is an all-powerful analytics engine and has the distinction of being the most used data science tool. It is known for offering lightning-fast cluster computing. Spark accesses varied data sources such as Cassandra, HDFS, HBase, and S3. It can also easily handle large datasets.

3) D3.js:

D3.js, short for Data-Driven Document, represents a premier JavaScript library tailored to craft interactive web data visualizations. Its client-centric approach facilitates dynamic data processing and visualization, integrating seamlessly with CSS for enhanced visual appeal. D3.js supports multiple data formats, from JSON to CSV, enabling users to create charts and graphs for enriched data representation.