# Artificial Intelligence, Machine Learning, and Data Science

Samatrix Consulting Pvt Ltd

Samatrix.io

# Course Introduction

# Course Objective

- Learn the concepts of
  - Data Science
  - Data Science Processes
  - Machine Learning
  - Artificial Intelligence

# Data Science – Learning Objective

- Introduction to Data Science:
  - Defining Data Science and Big Data,
  - Benefits and Uses of Data Science and Big Data,
  - Facets of Data, Structured Data, Unstructured Data, Natural Language, Machine generated Data, Graph based or Network Data, Audio, Image, Video, Streaming data,
  - Data Science Process,
  - Big data ecosystem and data science, Distributed file systems, Distributed programming framework, data integration framework, machine learning framework, No SQL Databases, scheduling tools, benchmarking tools, system deployments

# Data Science Process – Learning Objective

- Data Science Processes:
  - Six steps of data science processes, define research goals, data retrieval, cleansing data, correct errors as early as possible, integrating – combine data from different sources, transforming data,
  - Exploratory data analysis,
  - Data modeling, model and variable selection, model execution,
  - Model diagnostic and model comparison, presentation and automation.

# Machine Learning – Learning Objective

- Introduction to Machine Learning:
  - What is Machine Learning, Learning from Data, History of Machine Learning, Big Data for Machine Learning, Leveraging Machine Learning,
  - Descriptive vs Predictive Analytics,
  - Machine Learning and Statistics
  - Artificial Intelligence and Machine Learning,
  - Types of Machine Learning – Supervised, Unsupervised, Semi-supervised, Reinforcement Learning,
  - Types of Machine Learning Algorithms,
    - Classification vs Regression Problem,
    - Bayesian,
    - Clustering
    - Decision Tree
    - Dimensionality Reduction,
    - Neural Network and Deep Learning,
    - Training machine learning systems

# AI – Learning Objective

- Introduction to AI:
  - What is AI,
  - Turing test
  - cognitive modelling approach
  - law of thoughts
  - the relational agent approach
  - the underlying assumptions about intelligence
  - techniques required to solve AI problems
  - level of details required to model human intelligence
  - history of AI

# Learning Pedagogy

- The course has been split into Sections, Subsections and Units
- Quiz after every subsection that every participant should practice
- Graded and timed test after every section
- Midterm during midway of the course
- End-term after the course
- Assignments and Projects during the course. Please submit on time

Samatrix.io

# Introduction to Data Science

Samatrix.io

# Introduction to Data Science

- One of the biggest challenges that every company, irrespective of the size, across all the industries faces is managing and analyzing the data.
- The ability to manage and analyze has provided the organizations a competitive edge over their competitors.
- Every business struggle with finding a pragmatic approach to capture the relevant information about their products, services, customers, and suppliers.
- In the era of globalization and global supply chains, the markets have become very complicated.
- In the pursuit of gaining a competitive edge with customers, the companies continue to innovate and develop more new products and find innovative ways to reach their customer in the global market

# Introduction to Big Data

- To gain insights about the market and customer preference, every company collects the data from a variety of sources such as documents, customer service records, pictures, videos, sensors, social media, and clickstream data generated from website interaction.
- The availability of newer and more powerful mobile devices has created the potential for developing new data sources.
- To manage the intersection of all the different types of data, traditional data management techniques such as Relational Database Management System (RDMS) are not sufficient.
- Even though the RDBMS has been regarded as one-size-fits-all-solution, the requirement of handling data that varies in type and timeliness, have created the necessity of managing data differently.
- Big Data is the latest trend that has emerged to handle these complexities.

# Data Science and Big Data

- Big data helps organizations gather, store, manage, manipulate the large and complex data at the right speed, at the right time to gain the right insights.
- Data Science provides methods to analyze the data and extract meaningful insights out of the data.
- The relationship between big data and data science is the same as the relationship between crude oil and oil refinery.
- Even though the data science and big data have evolved from statistics and data management techniques, they are considered to be an independent field of study.

# Characteristics of Big Data

- We can define big data as the data source that has at least the three shared characteristics, often referred to as three Vs
  - Volume: An extremely high volume of data
  - Velocity: Extremely high velocity of data
  - Variety: An extremely high variety of data
- In addition to the three characteristics, the fourth V, veracity, extremely high accuracy of data, make the big data different from the traditional data management systems.
- Every aspect of big data that include data capture, storage, search, curation, share, transfer, and visualization, require specialized techniques.

# Tools and Techniques

- Data science finds its roots in statistics.
- It helps organizations deal with the massive data that is produced today.
- Data science requires a knowledge in computer science and statistics.
- A typical job description of a data scientist includes the ability to work with big data and experience in machine learning, computing, algorithm building.
- These job descriptions require the ability to use Hadoop, map reduce, pig, hive, Scala, Spark, R, Python, and Java among others.
- This job description is different from that of a statistician.
- During this course, we will slowly introduce these tools.
- However, our main focus would be on Python. Python has emerged the best language for data science because of the availability of data science specific libraries.
- Every popular NoSQL database offers a python specific API.
- So, the popularity of python has been steadily increasing in the world of data science

# Benefits and Uses of Data Science and Big Data

Samatrix.io

# Data Science in Industries

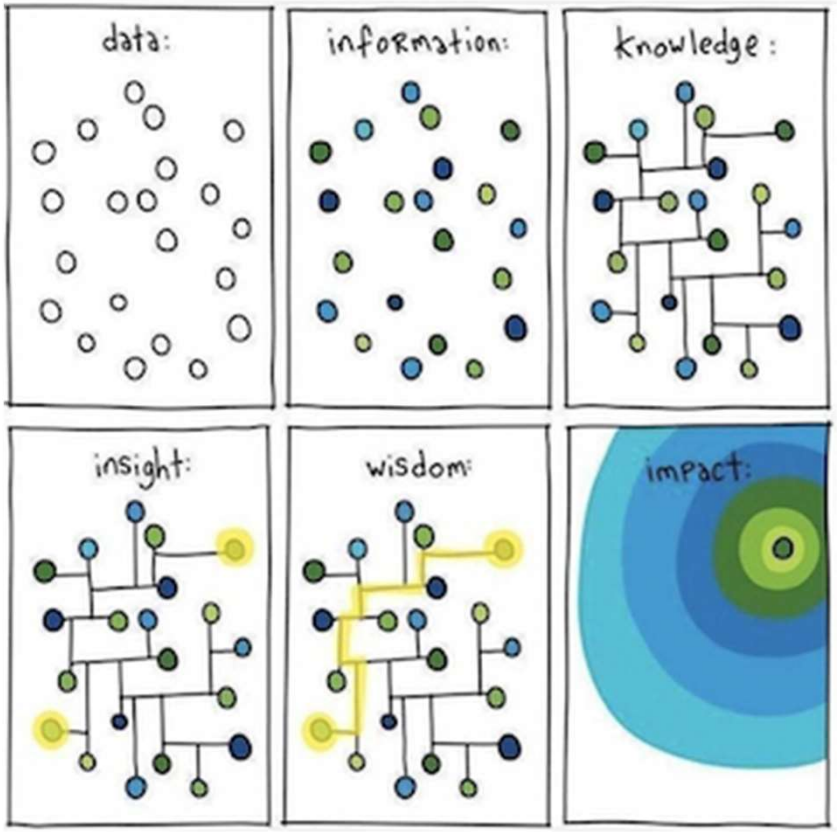
- Today, data science and big data have their footprints in every commercial and non-commercial organization.
- Commercial companies in every industry segment use data science and big data to gain insights about their customers, staff, processes, and products.
- Many companies such as Amazon use the data science and big data techniques in marketing such as targeted marketing, digital advertising, and recommendation systems for cross selling and up selling.
- Organizations use data science for general customer relationship management to analyze the ever-changing customer preferences and behavior to manage attrition and maximize the expected customer value.

# Data Science in Industries

- The financial institutions use data science to manage credit risk through credit scoring, predict the stock market, detect fraud, and manage the workforce.
- The human resource department in the organization uses people analytics to screen candidates, monitor the mood of the employees, reduce employee attrition, and improve employee engagement and productivity.
- Many retailers such as Walmart and Amazon use data science tools and techniques throughout their business that includes marketing and global supply chain management.

# Data Science For Business

- The primary objective of the course is to help you view business opportunities from data science perspective.
- How the data can be used to gather information and knowledge.
- You should be able to transform data into actionable insights.
- Based on the insights, by using wisdom take decision and actions to create an impact.




Samatrix.io

# Case Study – Hurricane Frances

- Consider an example from a New York Times story from 2004:

  *"Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Ben- tonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons … predictive technology. A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen,' as she put it. (Hays, 2004)"*

# Case Study – Hurricane Frances

- Why in the case of natural calamity, the data-driven prediction might be useful.
- People, who are in the path of a hurricane, might be interested to buy bottled water.
- But this is an obvious point, why we need data science tools and techniques to discover this fact.
- Wal-Mart executives might be interested in predicting how the hurricane will impact the sales so that the company can make the necessary arrangements.
- There could be some coincidence whereas the sales of a particular newly released movie CD went up during the week but the impact on sales was nationwide not just the areas impacted by the hurricane.
- Ms. Dillman is referring to more useful information than some general patterns.

# Case Study – Hurricane Frances

- Using data science, the data analyst team could discover the patterns that were not obvious.
- By analyzing the huge volume of Wal-Mart data from prior similar situations, the analyst could identify the surge in unusual local demand for few products and rush stocks ahead of hurricane's landfill.
- In the actual scenario, the same thing happened. The New York Times (Hays, 2004) reported that: *"… the experts mined the data and found that the stores would indeed need certain products—and not just the usual flashlights.*
- *'We didn't know in the past that strawberry PopTarts increase in sales, like seven times their normal sales rate, ahead of a hurricane,' Ms. Dillman said in a recent interview. 'And the pre-hurricane top-selling item was beer.'"*

# Facets of Data

# Variety of Data Types

- Variety is one of the basic principles of big data.
- It is also one of the four characteristics of big data.
- The data scientist should be able to manage a variety of data types.
- The information from various sources of data from bank transactions to tweets to images to videos should be integrated for analysis and data management.

# Variety of Data Types

- Based on the business problem, you may come across different facets of data.
- You would require different data management tools and techniques to analyze and extract results for each flavor of data.
- Certain situations such as monitoring traffic data require real-time data management and analysis techniques whereas other situations such as data analysis to determine unsuspected patterns require massive historic data collection.
- In certain situations, we need to integrate data from a variety of sources for our analysis.

# Main Categories of Data

- The main categories of data are
  - Structured
  - Unstructured
  - Natural language
  - Machine-generated
  - Graph-based
  - Audio, video, and images
  - Streaming

# Why Understanding Data Type is Important

- As soon as a new project is assigned, it is always tempting to jump into the exploration of statistical and machine learning models to get results faster before applying data science.
- However, without understanding the data, you would waste your time and energy in implementing the solutions that are not suitable for the given data type.
- Whenever you are assigned, you should spend time in analyzing the data according to the different categories of data.

# Structured Data

- Structured data has defined length and format.
- Number, dates, strings (such as name and address, etc.), are some of the examples of structured data.
- Structured data is usually stored in a database and can be queried using a structured query language (SQL).
- Traditionally the companies have been collecting data from sources such as customer relationship management (CRM) data, operational enterprise resource planning (ERP) data, and financial data.
- Structured data is about 20% of the data that we currently have in the overall system.

# Structured Data - Example

| | Indicator ID | Dimension List | Timeframe | Numeric Value | Missing Value Flag | Confidence Inte |
|---|---|---|---|---|---|---|
| 2 | 214390830 | Total (Age-adjusted) | 2008 | 74.6% | | 73.8% |
| 3 | 214390833 | Aged 18-44 years | 2008 | 59.4% | | 58.0% |
| 4 | 214390831 | Aged 18-24 years | 2008 | 37.4% | | 34.6% |
| 5 | 214390832 | Aged 25-44 years | 2008 | 66.9% | | 65.5% |
| 6 | 214390836 | Aged 45-64 years | 2008 | 88.6% | | 87.7% |
| 7 | 214390834 | Aged 45-54 years | 2008 | 86.3% | | 85.1% |
| 8 | 214390835 | Aged 55-64 years | 2008 | 91.5% | | 90.4% |
| 9 | 214390840 | Aged 65 years and over | 2008 | 94.6% | | 93.8% |
| 10 | 214390837 | Aged 65-74 years | 2008 | 93.6% | | 92.4% |
| 11 | 214390838 | Aged 75-84 years | 2008 | 95.6% | | 94.4% |
| 12 | 214390839 | Aged 85 years and over | 2008 | 96.0% | | 94.0% |
| 13 | 214390841 | Male (Age-adjusted) | 2008 | 72.2% | | 71.1% |
| 14 | 214390842 | Female (Age-adjusted) | 2008 | 76.8% | | 75.9% |
| 15 | 214390843 | White only (Age-adjusted) | 2008 | 73.8% | | 72.9% |
| 16 | 214390844 | Black or African American only (Age-adjusted) | 2008 | 77.0% | | 75.0% |
| 17 | 214390845 | American Indian or Alaska Native only (Age-adjusted) | 2008 | 66.5% | | 57.1% |
| 18 | 214390846 | Asian only (Age-adjusted) | 2008 | 80.5% | | 77.7% |
| 19 | 214390847 | Native Hawaiian or Other Pacific Islander only (Age-adjusted) | 2008 | DSU | | |
| 20 | 214390848 | 2 or more races (Age-adjusted) | 2008 | 75.6% | | 69.6% |

Samatrix.io

# Structured Data - Pros

- The structured data is highly organized. It can be easily understood by the machine language. The data stored in relational database can be easily and quickly searched and manipulated.
- The business users can use structured data relatively easily. They need not understand various data types and relationships among them. It is easy to develop self-service tools for the business user.
- The relational database has been around for a long time now. Several advanced tools have been developed and tested to manage structured data. It offers Data managers a variety of advanced tools and techniques.

# Structured Data - Cons

- The world is not made up of structured data. The natural data is unstructured. The structured data has a predefined structure and can be used only for the intended purpose. Due to this predefined structure the flexibility and use cases are limited.
- The structured data is stored in data warehouses and relational databases. The schema structure of both is very rigid. Any change in data needs would require updates in all the structured data. This results in massive expenditure in terms of time and resources.

# Unstructured Data

- Unstructured data does not follow any specified format, rather it is stored in its native format.
- Unstructured data is not processed until it is used.
- That is why unstructured data is also known as schema-on-read.
- A database schema is **the skeleton structure that represents the logical view of the entire database**.
- Unstructured data is the most prevalent data.
- It is everywhere.
- Approximately 80% of the data is unstructured.

# Unstructured Data

- Unstructured data has no pre-defined model so it is difficult to deconstruct and cannot be organized in relational databases.
- Instead, we use non-relational, or NoSQL databases to manage unstructured data.
- We can also use a data lake to manage data in a raw and unstructured format.

# Examples of Unstructured Data

- Satellite images, weather data, remote sensing images, Google Earth, etc.
- Scientific data include seismic imagery, atmospheric data, etc.
- Photographs and video including security, surveillance, and traffic video
- Radar or sonar data including vehicular, meteorological, and oceanographic seismic profiles.
- Social media data generated from social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flicker
- Mobile data including text messages and location information
- E-mail, survey results, documents, and logs

# Unstructured Data - Pros

- Unstructured data can provide a much deeper understanding of the customer behavior and intent whereas the structured data provide the bird eye-view of the customer.
  - For example, data science techniques help understand customer buying habits, timings, spending patterns, sentiments towards certain products and services.
- Unstructured data collected from sensors provide real-time data and helps in predictive analytics.
  - For example, sensors installed on a machine can detect the anomalies in the manufacturing process and adjust itself to avoid a costly breakdown.
- Since the unstructured data is stored in native format and is not defined until needed, the purpose of the data is adaptable. Due to which data can be used for a wider range of use cases.
- Since the data is not defined at the time of accumulation, the data can be collected quickly and easily
- Organizations prefer to store the unstructured data in a cloud data lake, which is a cost-effective and scalable solution.

# Unstructured Data - Pros

- Its supports the data which lacks a proper format or sequence.
- The data is not constrained by a fixed schema.
- Very Flexible due to absence of schema.
- Data is portable.
- It is very scalable.
- These type of data have a variety of business intelligence and analytics applications.

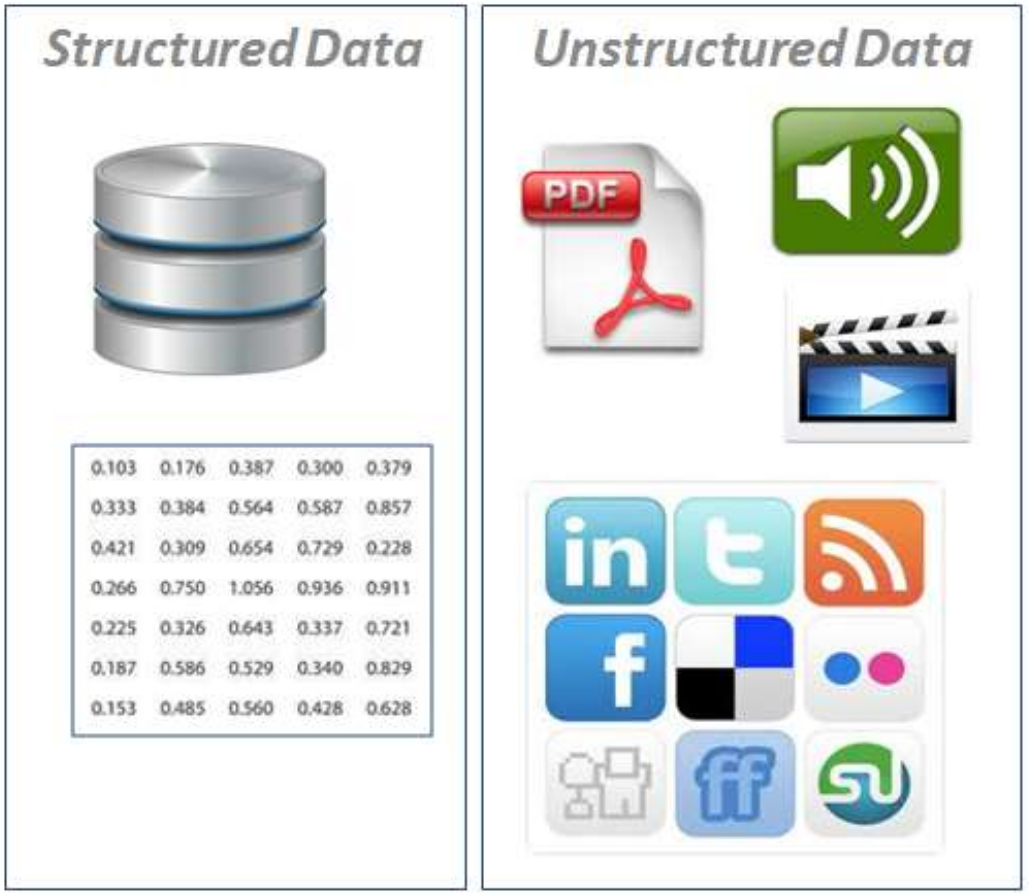
# Unstructured Data - Cons

- One of the biggest drawbacks of unstructured stats is the requirement of data science expertise.
- A normal user cannot prepare and analyze the data because the data is unorganized and unstructured
- In order to manipulate the unstructured data, the data manager needs the specialized tools that are still in their infancy.
- Indexing the data is difficult.

# Structured vs Unstructured Data

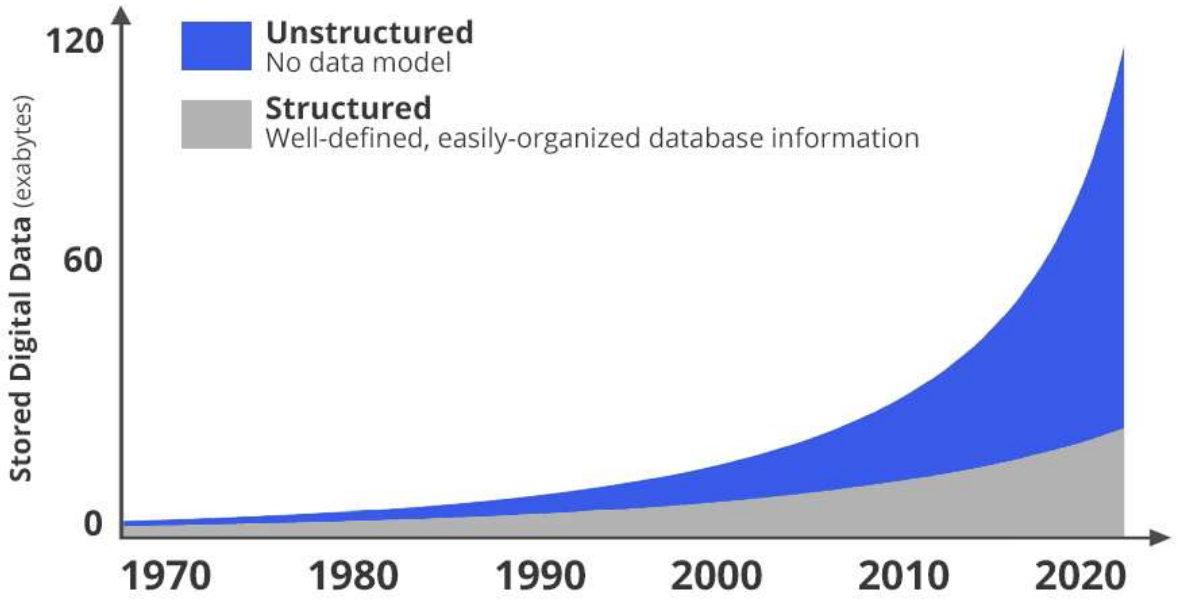
| | Structured Data | Unstructured Data |
|---|---|---|
| Who | Self Service Access | Requires Data Science Expertise |
| What | Only Selected Data Types | Variety of Data Types |
| When | Schema-on-write | Schema-on-read |
| Where | Commonly stored in Data Warehouse | Commonly stored in data lakes |
| How | Predefined Format | Native Format |

# Structured vs Unstructured Data




Samatrix.io

# Structured vs Unstructured Data




Samatrix.io

# Natural Language

- According to the theories from neuropsychology, philosophy of language, and linguistic, a natural language is any language that has evolved in humans due to use and repetition without conscious planning.
- Speech and singing are different forms of natural language.

# Human Language

- Human language is highly complex and diverse.
- The human can express himself in infinite ways and there are hundreds of languages across the globe.
- Every language has its own grammar and rules.
- Moreover, every language has its own regional accent.
- Many people mix words from different languages and use abbreviated words while speaking and writing.

# Human Machine Interaction

- Researchers from computer science and computational linguistics have been working for many decades to fill the gap between human communication and computer understanding.
- Recently due to increased interest in human-machine interaction, big data, computing power, and enhanced algorithms, significant advancements have taken place.
- Humans speak, write, and understand languages such as English, Hindi, French, and German.
- The native language of computers is machine code or machine language that constitutes millions of zeros and ones.

# Natural Language Processing

- Natural language processing uses natural language rules to convert the unstructured data into a computer recognizable form.
- Natural language processing helps the computers understand, interpret, and manipulate the human's natural language and communicate with humans in their own language.
- Teaching machines to understand human communication is a challenging task.
- Today if you say, "Alexa play party songs". Alexa will play a song for you and in its reply tell you the name of the song.

# Natural Language Processing

- This interaction has been made possible by natural language processing.
- In this interaction, as soon as the device hears your voice, it is activated.
- It understands your unspoken intent, executes the action, and provides you feedback in well-formed human language within five seconds.
- Natural language processing helps machines analyze staggered text and speech data that is generated every day on social media platforms or medical records and add useful numeric structure for downstream applications such as speech recognition or text analytics.

# Machine Generated Data

- A big component of data is created by machine without any human intervention.
- For example, every time a Boeing 787 flies, it generates half a terabyte of data.
  - Every part of the plane generates data and constantly update the inflight crew and ground staff about its status. This machine-generated data comes from various sensors installed on several parts of the plane.
- Machine-generated data could be both structured as well as unstructured

# Machine Generated Data

- Internet of Things (IoT):
  - These devices collect data, connect to other devices or networks, and execute services on their own.
  - These smart devices are used everywhere: home, cars, cities, remote areas, the sky, the ocean. They all are connected and generate data.
- Sensor data:
  - Radiofrequency ID (RFID) tags have become a popular technology.
  - RFID uses a small chip to track the objects from distance.
  - The RFID can track the containers as they move from one location to another in the supply chain.
  - Companies have been using RFID technology for supply chain management and inventory control.
  - Our smartphones contain sensors such as GPS that can capture location data to understand consumer behavior.

# Machine Generated Data

- Weblog data:
  - The servers, applications, and networks create a log of their activities.
  - That creates a massive amount of useful data.
  - This data can be used for proactive server or application maintenance activities or predict security breach
- Point of sale data:
  - On the billing counter, as soon as the cashier scans the bar code and the data associated with the product is generated.
  - With many people shopping so many products, a huge amount of data is generated.
- Financial Data:
  - Every financial transaction has a digital footprint now.
  - Whether the transaction is related to the banking or stock market, data is generated.
  - Some of the data is machine-generated whereas some data is human-generated.

# Machine Generated Unstructured Data

- Satellite images, weather data, remote sensing images, Google Earth, etc.
- Scientific data include seismic imagery, atmospheric data, etc.
- Photographs and video including security, surveillance, and traffic video
- Radar or sonar data including vehicular, meteorological, and oceanographic seismic profiles.

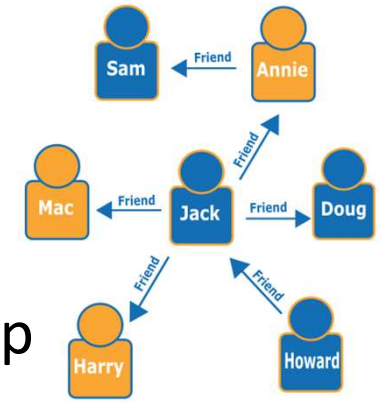
# Human Generated Data

- While interacting with computers, we generate data that is known as human-generated data. This data type includes email, messages, documents, presentations, spreadsheets, audio and video, and images. We create and share this day every day.
- Human-generated data is one of the fastest-growing data. It contains highly valuable and relevant information that is critical for data analysis.
- The human-generated data may not be very big on its own. But considering millions of other users generating such data, the size is astronomical. This data type also has a real-time component that can be used to understand the patterns and predict outcomes.

# Human Generated Data - Examples

- Input Data:
  - We input a lot of information into a computer.
  - For example, while filling an online registration or application form, we input a name, age, and address.
  - This data is used to understand basic consumer behavior.
- Click-stream data:
  - Every time you click a link on a website, click-stream data is generated.
  - This data is used to understand the online behavior of an online visitor and buying patterns.
- Gaming-related data:
  - While playing games, every move you make, generates gaming-related data.
  - This data helps understand how the end-user moves throughout a game.
  - This is used to optimize the existing games and develop new games

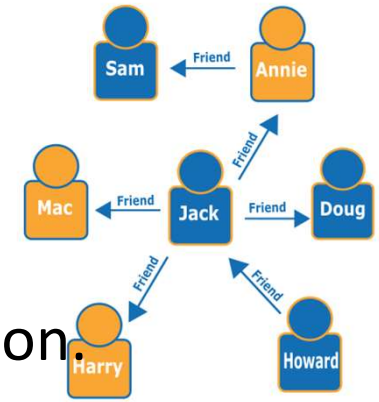
# Graph Based or Network Data




- In graph theory, the graph represents a pair-wise relationship between objects.
- Graph databases are used to store and navigate relationships.
- The value of the graph depends upon the relationships. To store and navigate the graphical data, the graph databases use nodes, edges, and properties.
- Data entries are stored in nodes and the relationship between the entities is stored in edges.
- The graph database is queried using specialized query languages such as SPARQL.

Samatrix.io

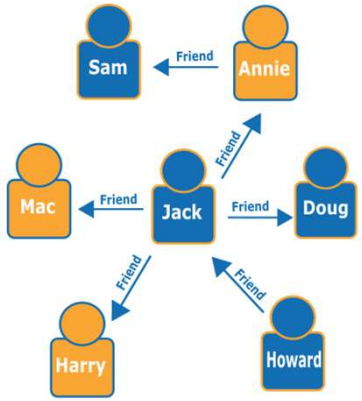
# Graph Based or Network Data




- An edge consists of a start node, end node, type, and direction.
- The edge is also used to describe the parent-child relationships, ownerships, and actions.
- A node can have an unlimited number and kind of relationships.
- Using the graph database, a graph can either traverse along with specific edge type or an entire graph.
- The relationships between the nodes are persisted in the database.
- They are not calculated at the query time.
- So, the traversing in the graph database is very fast.

# Graph Based or Network Data – Use Cases

- The use cases such as social networking, recommendation engines, and fraud detection use the graph databases extensively.
- In these use cases, there is a need to build a relationship between the data and the ability to query these relationships quickly
- Refer to the following picture that shows an example of a social network graph.
- If you have people (nodes) and relationships among them (edge), you can easily find the “friends of friends” of a particular person.



# Graph Based or Network Data – Use Cases

- Fraud Detection:
  - Graph databases are used for fraud detection.
  - Graph relationships can be used for real-time processing of financial and purchase transactions.
  - For example, if a person is using the same email address and credit card as recorded in a known fraud case, you can detect easily using graph databases.
  - You can easily detect the cases where multiple people are associated with one personal email address or cases where multiple people, who are residing at a different physical address, are sharing some IP addresses.
- Recommendation Engines:
  - Using a graph database, you can easily store the relationship between several categories such as customer interests, friends, and purchase history.
  - Using these relationships, you can make product recommendations based on the purchase history of other users with similar interests.

# Audio, Images and Videos

- The advancement in multimedia technology, high-speed internet, and smart devices are changing the digital data landscape.
- A huge amount of digital data in the form of text, audio-video, and images is generated every second.
- With technology, the trend is on the increase.
- It is important for any organization to equip itself to address the needs of data analytics for multimedia content.
- Audio, images, video, and other multimedia data analytics have a challenging task for data scientists.
- Tasks such as recognizing an object in an image or video look trivial for humans but are challenging for the machines.

# Streaming Data

- The term “streaming” means a continuous data stream that is never-ending and that does not have a beginning or an end.
- The streaming data can be utilized without downloading.
- Streaming data has become an important part of our day to day life.
- We can gather real-time data from online gaming, social media, eCommerce, GPS, and IoT devices.
- In order to gain a competitive advantage, every company wants to take a lead in extracting accurate customer insights from streaming data.
- Real-time data processing data science technologies such as Kafka and Kinesis are used to collect and analyze the data in real-time.

# Data Science Processes

Samatrix.io

# Six Steps

- In order to be successful in a data science project at a low cost, it is important to follow a structured approach.
- A typical data science structured approach involves six steps that every project should undertake before taking up any data science project.


Samatrix.io

# 1. Setting the research goal

- Before starting any data science project, the first step is to identify the problem statement.
- Along with the problem statement, you should understand why there is a problem and how the solution will help meet the final business objectives.
- This step will help you understand the data that you would require, where would you find it, and the required architecture of the system.

# 1. Setting the research goal

- For example, if you got a data science assignment from the sales department of a company. Before getting into a solution space you need to understand the problem. You can do so by asking the right questions to understand the sales process and customer.
  1. Who the customers are?
  2. What are the company's unique selling points?
  3. How to predict customer buying behavior?
  4. Which segments are working well and which ones are not and why?
  5. What are the opportunity losses and why?
- Such a question will help you define the problem and understand the business and the data flow across the organization.

# 2. Retrieve data

- For a data science project, you would be handling a huge amount of data.
- You need to collect lots of data from lots of sources.
- Available data may not meet all your requirements and may not match your problem statement.
- The data may be available within your organization as well as other organizations also.
- The required data for your project should be available at the given sources.
- You also need to ensure that you have the required access to retrieve data from the sources.
- The data quality is also important.
- Before retrieving the data, you need to ensure how trustworthy the data is and what is the data quality.

# 3. Data preparation

- Raw data that is available at multiple sources may have some inconsistencies that need to be taken care of before using it for your data science project.  Examples of issues that a data scientist face are
  - Some important values might be missing
  - Some values might be corrupted. It may have invalid entries
  - Due to time zone differences, the user data may have inconsistencies in the date-time data
  - Due to inconsistencies in capturing data, some data may have date range error.
- The data from the different sources and having different formats and structures need to be integrated.
- Several data science techniques have certain requirements of data format.
- They require data in a separate format than the data that is available.
- So, some data conversion is necessary.

# 3. Data preparation

- The following steps are often required.
  - Data cleansing: remove false value and inconsistencies from the data source
  - Data integration: enrich data by combining information from multiple sources
  - Data transformation: change the data to a suitable format that can be used in your model

# 4. Data exploration

- In this phase, you build a deeper understanding of the data.
- You try to understand each variable and relationship among the variables in the data set.
- You access the data distribution and presence of outliers (if any).
- Descriptive statistics, data visualization, and sample modeling are methods used for data exploration.
- We often call this step as Exploratory Data Analysis (EDA).

# 5. Data modeling

- The models, domain knowledge, and insights from the data exploration phase are used to answer the research questions.
- In this phase, you select a model from statistics, machine learning, etc.
- It is an iterative process where you select variables from the model, execute the model and perform model diagnostics.

# 6. Presentation and automation

- In this final phase, the outcome of the data science process is implemented in an information system or business process to meet the initial objective.
- In this step, you may also present your findings to your client or business.
- You may present data in various forms such as presentations and research reports.
- If the outcome of the process is required in any other project, you may also automate the execution of the process.

# Big Data Ecosystem and Data Science

Samatrix.io

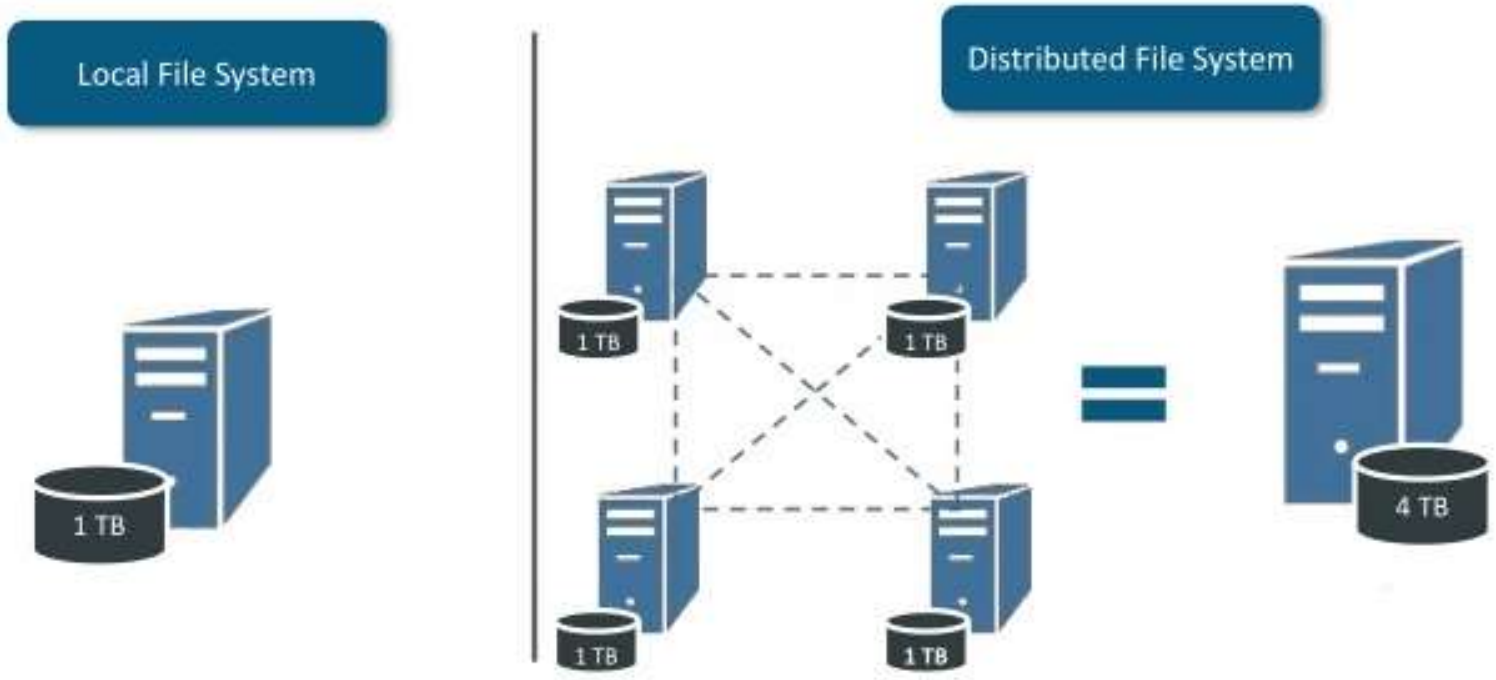
# Introduction

- A data scientist can use many big data tools and frameworks.
- But this is an evolving field.
- New technologies appear rapidly.
- For a project, the data scientist uses many technologies but not all the technologies.
- Based on functionalities and goals, we can group the bigdata ecosystem into the technologies.

# Distributed File System

- A distributed file system is similar to a normal file system.
- In the case of a distributed file system, many individual computers or servers are networked together across geographies as if they are a single file system.
- The distributed computing environment shares resources ranging from memory to network and storage.
- On the distributed file system, you can perform almost all the actions such as storing, reading, deleting, and securing the files, that you can do on the single file system.

# Distributed File System




Samatrix.io

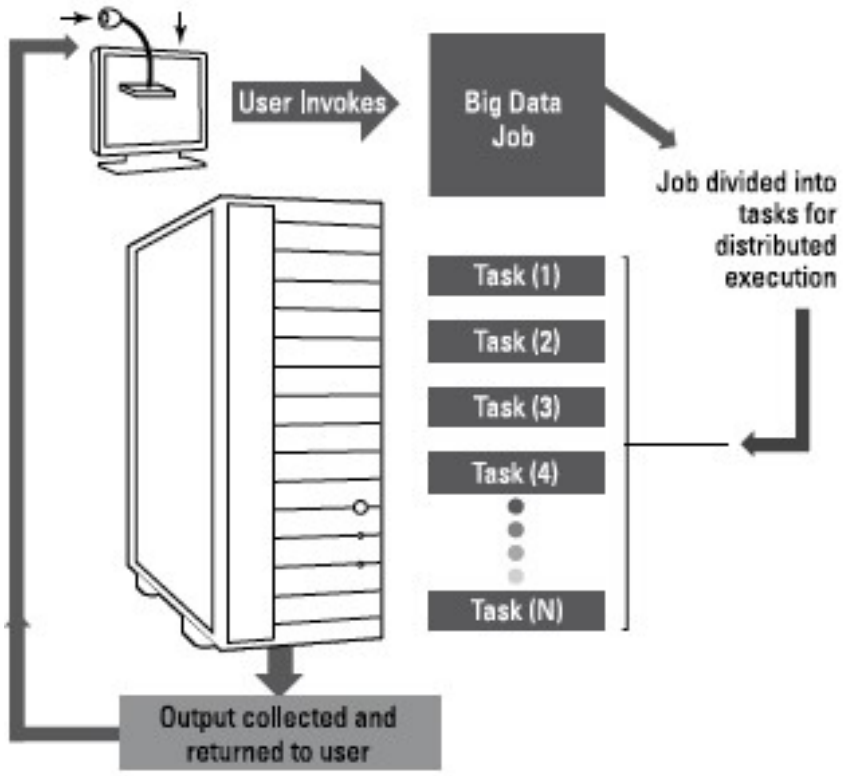
# Distributed File System

- The advantages of the distributed file systems are as follows:
  - A file larger than the computer disk can be stored
  - The files are replicated automatically across networked servers for redundancy or parallel operations. Such complexities are hidden from the users
  - The distributed file systems can be scaled easily without adding the memory or storage restrictions

# Distributed File System

- Earlier, to achieve the scale, we used to move everything to another server with more memory, storage, and better CPU.
- This process is known as vertical scaling.
- However, with a distributed file system, you can add another small server.
- The principle of horizontal scaling has made the scaling potential virtually limitless.
- Hadoop File System (HDFS) is the best-known distributed file system.
- HDFS is an open-source implementation of the Google File System.
- HDFC is the most popular distributed file system among users.
- Other commonly used distributed file systems are: Red Hat Cluster File System, Ceph File System, and Tachyon File Systems

# Distributed Programming Framework




Samatrix.io

# Distributed Programming Framework

- Once the data is stored in the distributed file system, the data scientist would use it for analyzing and solving business problems.
- To enable parallel processing of data, code is transferred to nodes. The tasks are performed on the node where the data is stored.
- In the case of distributed hard disk, we do not move the data to the program, rather we move the program to the data.
- That is why while working with a normal general-purpose programming language such as C, Python, or Java, the data scientist has to deal with the complexities of distributed programming such as restarting jobs that have failed and tracking the results from different sub-processes.
- The distributed programming frameworks such as Apache MapReduce, Apache Pig, Apache Spark, Apache Twill, Apache Hama are few open-source frameworks that help work with the distributed data and deal with the challenges the distributed file system carries.

# Data Integration Framework


Samatrix.io

# Data Integration Framework

- Data integration is an important step for any data science project.
- The data integration framework combines data from various sources and data in various formats before presenting meaningful and valuable information to the users.
- In the traditional data warehouse environments, the Extract, Transform, and Load (ETL) technologies have been used.
- However, the elements of the big data platform manage the data differently from the traditional relational database.
- Scalability and high performance are the basic requirements for managing structured and unstructured data.
- Every component of the big data ecosystem ranging from Hadoop to NoSQL database has its own approach for extracting, transforming, and loading the data.
- Apache Airflow, Apache Kafka, Apache Flume, Chukwa, Scribe, Talend Open Studio are some of the examples of open source data integration frameworks.

# Machine Learning Framework

- After collecting the data in the required formats, the data scientists focus on extracting the insights.
- In this phase, they use machine learning, statistics, and applied mathematics.
- To deal with high volume and complex data, we need specialized machine learning frameworks and libraries.
- A machine learning framework is an interface, library, or tool that helps the data scientists build machine learning models without getting into the underlying mathematical and statistical algorithms.

# Machine Learning Framework

The most popular machine learning frameworks are as follows:

- **Tensorflow**
  - is a python library that is provided by Google.
  - It is an open-source python library that is used for numerical computation using data flow graphs.
- **Keras**
  - is python based open-source neural-network library that can run on the top of TensorFlow, Theano, R, Microsoft Cognitive Toolkit, or PlaidML.
  - Keras is capable of conducting fast experimentation with deep neural networks.
  - It is a user-friendly, modular, and extensible.

# Machine Learning Framework

- **Scikit-learn**
  - is one of the most popular machine-learning frameworks.
  - Scikit-learn can easily implement supervised and unsupervised learning algorithms.
  - The library has modules for classification, regressions, and clustering algorithms.
  - Scikit-learn can interoperate easily with Python numerical and scientific libraries: NumPy and SciPy.
- **Apache Spark MLlib**
  - interoperates well with NumPy in Python and R libraries.
  - You can also use HDFC, HBase, or any other Hadoop data source that makes MLlib easy to plug into the Hadoop workflows.
  - MLlib leverages iterative computation and contains high-quality algorithms.
  - Due to which MLlib can provide better results as compared to one-pass approximations used on MapReduce.

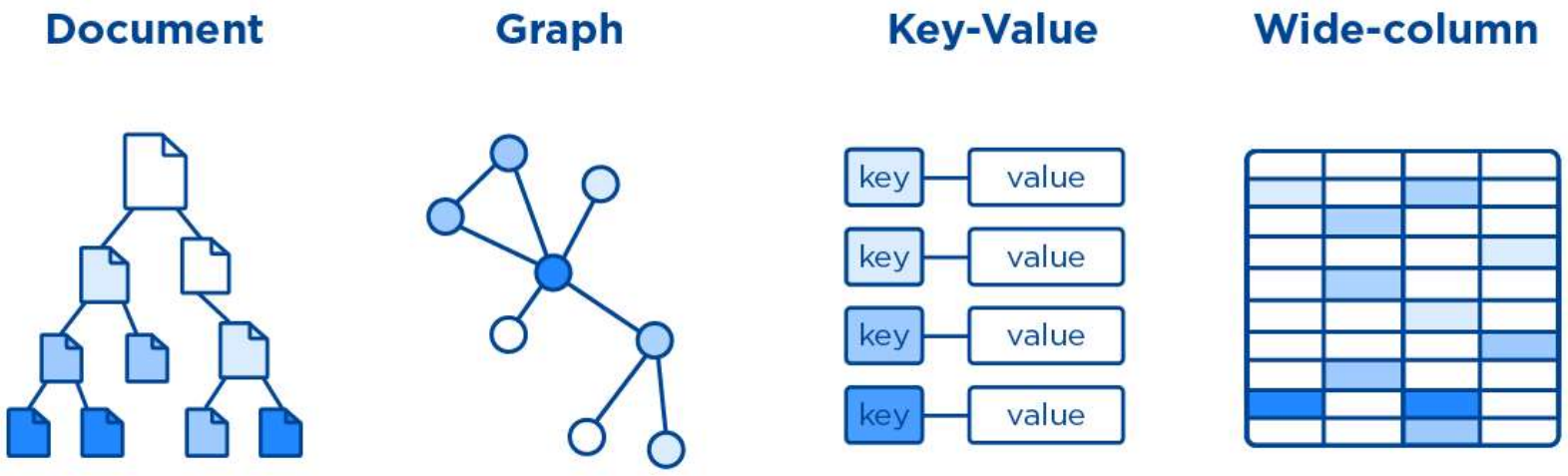
# Machine Learning Framework

- In addition to the frameworks given above, there are other popular machine learning frameworks such as Azure ML Studio, Google Cloud ML Engine, Torch, and Amazon Machine Learning frameworks.

# NoSQL Database

- To store a huge amount of data, the data scientist needs specialized software that can manage and query this data.
- Traditionally, relational databases such as Oracle SQL, MySQL, and Sybase IQ to manipulate and retrieve the data.
- With the emergence of new data types especially streaming, graphs, and unstructured datasets, the traditional databases cannot scale well.
- However, NoSQL databases can allow the endless growth of the data.
- The term “NoSQL” stands for “non SQL” or “not only SQL”.
- NoSQL databases can store data in a format that is different from relational tables.
- NoSQL databases can store the relational data in a more effective way than in a relational table because in the case of a NoSQL database there is no need of splitting the data between the tables.

# NoSQL Database

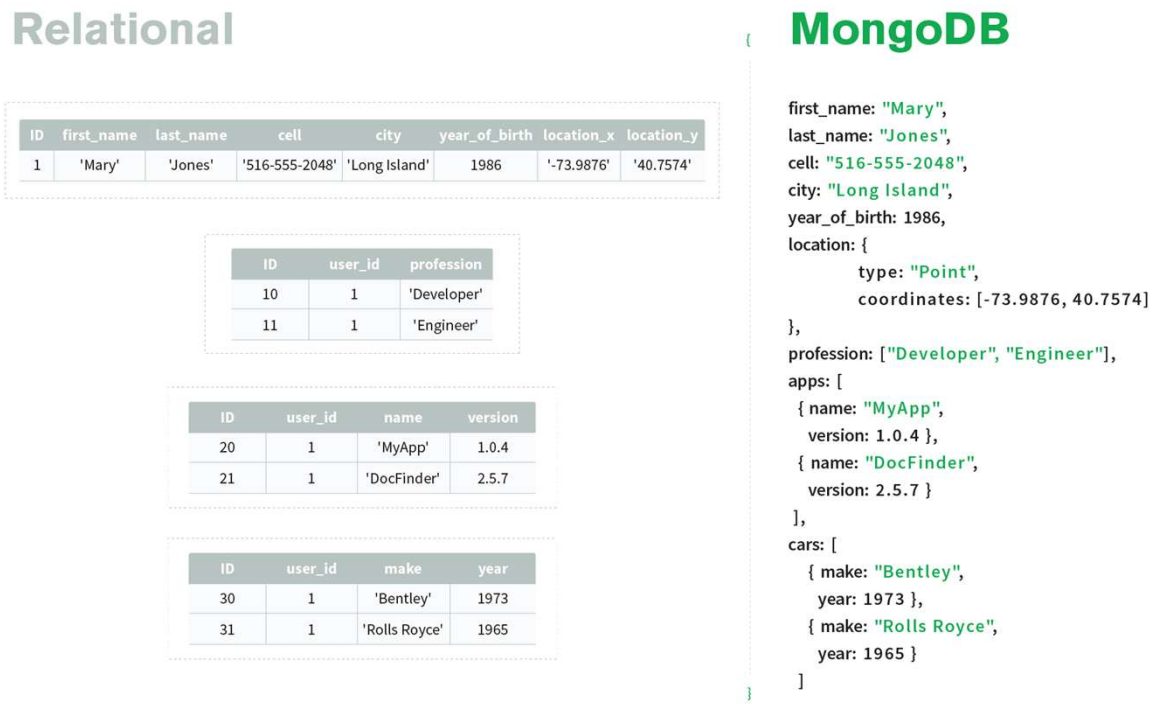



Samatrix.io

# NoSQL Database

Over time, four major types of NoSQL databases have emerged

- **Document databases**

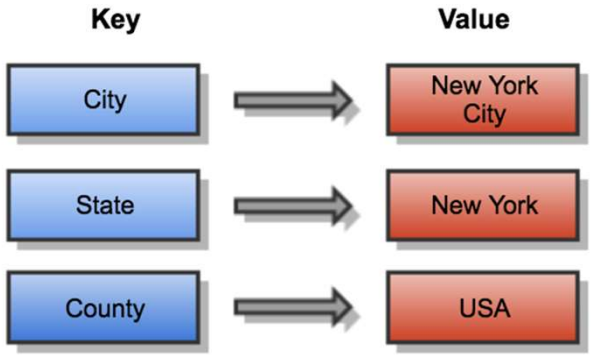



Samatrix.io

# NoSQL Database

- **Document databases**
  - can store data in document format, which is similar to JSON (JavaScript Object Notation) objects.
  - Each document has pairs of fields and values.
  - The document databases can take a variety of value types such as strings, numbers, booleans, arrays, or objects.
  - The structure aligns well with objects that the developers use in their code.
  - Due to a variety of field-value types, the document databases are termed as general-purpose databases and they can be used for a wide range of use cases.
  - **MongoDB** is the most popular NoSQL document database.

# NoSQL Database

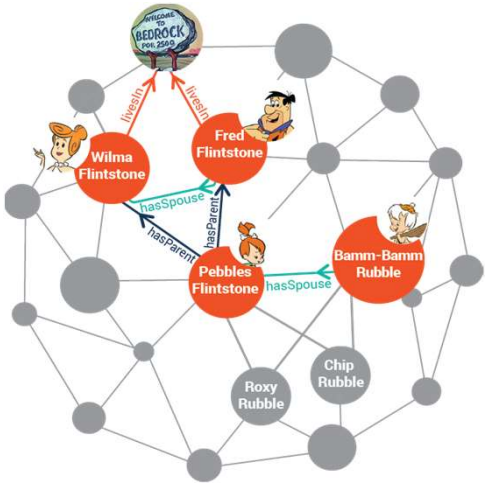



- **Key-value databases**
  - are simpler databases.
  - Each item in this database contains keys and values.
  - By referencing the key, the value can be easily retrieved.
  - Key-value databases are useful when you need to store a large amount of data without using complex queries to retrieve it.
  - Storing user preferences or caching are common use cases of key-value databases.
  - **Redis** and **DynamoDB** are examples of key-value databases.

# NoSQL Database


- **Wide-column stores**
  - use tables, rows, and dynamic columns to store the data.
  - It is more flexible than the relational databases because each row can have different columns.
  - In the wide column store, the data is stored in columns instead of rows as in a conventional relational database management system (RDBMS).
  - The names and format of the columns can vary from row to row in the same table
  - Wide-column stores can also be considered to be two-dimensional key-value databases.
  - We use wide-column stores when we need to store a large amount of data and we can predict the query pattern.
  - This database can be used to store the Internet of Things (IoT) and user profile data.
  - **Cassandra** and **Hbase** are examples of wide-column stores.

# NoSQL Database




- **Graph databases**
  - use nodes and edges to store data.
  - We store the information about people, places, and things in the nodes whereas we store the information about the relationships between the nodes in the edges.
  - We use the graph database when we need to traverse the relationship to find patterns such as social networks, fraud detection, and recommendation engines.
  - Neo4j and JanusGraph are the most popular graph databases.

# Scheduling Tools

- For any real implementation, we always have limited resources.
- On a busy server or cluster, often an application has to wait for some of its requests to be completed.
- The scheduling tools such as YARN, allocate resources to applications according to the pre-defined policies.
- Scheduling is a complex task and there is no one “best” policy.
- YARN provides a choice of schedulers and configurable policies.

# Scheduling Tools

- Three schedulers, FIFO, Capacity, and Fair schedulers, are available in YARN.
  - **FIFO scheduler** first places all the application in a queue and run them in the order of their submission (first in first out)
  - **Capacity schedulers** allow the small job to start as soon as it is submitted. It means that the larger job waits longer when compared to FIFO schedular
  - **Fair scheduler** dynamically balances resources between all the running jobs and hence there is no need to reserve a capacity

# Benchmarking Tools

- The benchmarking tools are used to optimize the big data installation.
- Before any big data installation, the performance of each tool in the installation is measured.
- The performance metrics are compared with other tools which are known as big data benchmarking.
- Using an optimized infrastructure can make a big cost difference.
- If you can optimize to reduce 10% clusters of servers, you can save the cost of 10 servers. In the majority of organizations, the benchmarking tools are not in the job scope of the data scientists.
- These activities are carried out by IT infrastructure teams.
- Yahoo Streaming benchmark, BigBench, TPC-DS are recent approaches to big data benchmarking.

# System Deployment

- Setting up the big data infrastructure is a big task.
- The system deployment tools help the deployment of new applications into the big data cluster.
- The system deployment tools automate the installation and configuration of big data components.
- In many organizations, system deployment is not the core task of data scientists.

# Thanks

Samatrix Consulting Pvt Ltd

Samatrix.io