Priyanka Neogi

What is Data Industry and how this Data Industry actually works?

- Inside data industry there are 3 things that fall into place.

  i. Big data

  ii. Data science

  iii. Data analytics

1. BIG DATA:

- It is also referred to as data engineering. Data engineering and big data are entirely synonymous.
- Big data pertains to a compilation of extensive and intricate datasets that are excessively voluminous, complex, and swift-moving to be efficiently controlled and handled using conventional data processing techniques.
- These datasets frequently originate from diverse origins, including social media, sensors, devices, online transactions, and additional sources.

TOOLS USED :

1. HADOOP ECOSYSTEM :

- Hadoop stands as an open-source framework designed for the distributed storage and processing of vast datasets across clusters of computers. Hadoop presents itself in multiple versions—Hadoop 1X, Hadoop 2X, and the current market offering, Hadoop 3X. The entire system operates through distributed computation.

- In 2004, Doug Cutting commenced the development of Hadoop within the Nutch project. The central storage system employed by Hadoop to manage and store substantial data volumes across multiple nodes in a cluster is termed HDFS, or Hadoop Distributed File System.

- Doug Cutting's contributions played a pivotal role in establishing HDFS as a dependable and adaptable file system, adeptly fulfilling the storage needs of Hadoop's distributed computing model.

- Hadoop is embraced by various distributors, including Cloudera and Hortonworks (now integrated into Cloudera).

- When dealing with vast data that surpasses the capabilities of a single computer, Hadoop divides it into smaller segments, distributing each segment to multiple computers for independent processing. Once computations are completed, Hadoop

combines the results, thereby furnishing a comprehensive solution. This approach significantly expedites tasks compared to the slower process of relying on a single computer.

- Hadoop functions as a means to simplify substantial challenges by orchestrating a collaborative effort among numerous computers, functioning as a cohesive unit.

2. HIVE :
   - Serving as an additional data warehousing solution, HIVE operates atop the Hadoop framework. It simplifies the intricate workings of Hadoop's infrastructure, enabling individuals accustomed to SQL to interrogate and dissect data within a distributed computational setting.
   - In situations where handling a substantial volume of data directly becomes challenging, HIVE emerges as the solution.
   - HIVE facilitates the sorting, structuring, arrangement, and comprehension of extensive datasets. It systematically organizes the entirety of your data, allowing for inquiry through queries to extract meaningful insights.

3. HBase :
   - Functioning as a distributed storage solution catering to structured data, this platform operates as a NoSQL-based database. Similar to how Bigtable capitalizes on the distributed data storage capabilities of the Google File System, Apache HBase delivers Bigtable-like functionalities by leveraging Hadoop and HDFS.
   - HBase represents a distributed, scalable, and NoSQL database specifically designed to manage extensive volumes of sporadic and partially structured data, granting immediate read and write access. Its foundation lies atop the Hadoop Distributed File System (HDFS), seamlessly integrating into the larger Hadoop ecosystem.
   - Employing a strategy of breaking down substantial datasets into smaller fragments, HBase allocates each fragment onto a separate high-powered machine.
   - When the need arises to locate specific information within the dataset, HBase is summoned, automatically directing its search to the relevant machine.
   - Operating at remarkable speeds, HBase swiftly retrieves the desired information and presents it. Its efficiency ensures that users can effortlessly uncover their required information without encountering confusion.

4. Spark :
   - Presently, SPARK offers compatibility with PYTHON in addition to its support for JAVA, SCALA, and R programming languages.
   - Serving as a processing engine/framework, it boasts a multitude of components such as Spark Streaming, Spark SQL, and MLlib (Machine Learning Library).

- SPARK constitutes an open-source framework for data processing and analytics, meticulously crafted to manage substantial data volumes while executing computations swiftly and effectively. Within its unified platform, diverse data processing tasks like batch processing, real-time stream processing, machine learning, and graph processing find their place.
- Operating as a potent tool, SPARK empowers users to efficiently manipulate vast datasets, ensuring rapid processing and the execution of intricate undertakings like pattern analysis and predictive tasks. It functions as a high-velocity data engine, enabling users to accomplish a myriad of tasks without enduring lengthy waits.
- SPARK's prominence arises from its capacity to seamlessly handle data from various origins, engage in real-time processing, and promptly deliver insights. Comparable to an exceptionally intelligent aide for data analysis and processing, SPARK rises to meet significant challenges in the realm of information. It facilitates diverse tasks, encompassing data analysis on colossal scales, pattern identification, predictive modeling, and beyond.
- In lieu of relying on a solitary computer, SPARK harnesses the power of numerous interconnected machines to collaboratively tackle substantial tasks.

5. KAFKA :

- Functioning as a messaging system tailored for streaming datasets, this technology stands as a high-throughput, fault-tolerant, and distributed platform.
- Its core purpose revolves around facilitating the seamless transmission and reception of data streams among diverse software systems or components.
- Kafka serves as the conduit for transmitting messages and data across various software systems. It guarantees the secure conveyance of information between locations, even when confronted with substantial data loads and rapid activity.
- Confluent Cloud represents a cloud-based configuration, while an on-premises(local) setup is also feasible.
- This platform emerges as an open-source stream processing solution, meticulously architected to manage real-time, high-volume data streams. It offers a distributed and fault-tolerant structure that optimizes the dissemination, subscription, storage, and processing of streams of records.

6. OOZIE :

- Oozie operates as a system for scheduling workflows, meticulously engineered to oversee and harmonize intricate data processing assignments within the Hadoop ecosystem and other distributed frameworks.
- Its architecture relies on XML. Oozie proves invaluable for scenarios where the execution of tasks is required once or multiple times a day, week, month, etc., within the Hadoop ecosystem.

- Oozie simplifies the creation, scheduling, and administration of workflows encompassing numerous steps. These steps entail executing a spectrum of actions, including launching MapReduce jobs, handling Pig scripts, executing Hive queries, and more. The platform ensures these actions are executed in a synchronized and dependable manner.
- In the domain of extensive data, Oozie plays a similar role for intricate data processing tasks. It aids in delineating a sequence of operations that warrant execution, such as conducting data analysis tasks or transformations.
- Oozie expertly manages the orchestration of task initiation at optimal times and guarantees their completion prior to transitioning to subsequent stages. Consequently, it facilitates the automation and governance of data workflows, circumventing the need for manual intervention.
- In contemporary contexts, the utilization of Airflow has become prevalent.

7. SQOOP:

- This tool functions as a conduit, enabling the seamless exchange of data both from SQL databases to HADOOP and vice versa.
- Known as Sqoop (short for SQL-to-Hadoop), it stands as an open-source data integration solution with the primary purpose of transferring data between relational databases and Apache Hadoop.
- Its efficiency lies in its capability to adeptly retrieve data from structured sources, such as relational databases (examples include MySQL, PostgreSQL, Oracle, and more), and transport it into the Hadoop ecosystem, particularly the Hadoop Distributed File System (HDFS) and Hive.
- Sqoop comes equipped with command-line utilities and connectors, streamlining the data transfer process. It effectively bridges the gap between conventional relational database systems and the expansive realm of Hadoop, thus simplifying the bi-directional movement of data.
- Notably, it boasts the intelligence to identify newly added or modified data since the last transfer, enabling incremental transfers that eliminate redundancy.
- Within the Hadoop environment, your data resides in well-organized repositories such as HDFS and Hive, making retrieval and utilization straightforward. Should the need arise to transport data back to a database without compromising its integrity, Sqoop adeptly fulfills this role, ensuring a smooth and efficient transfer.

8. CLOUD:

- Regardless of the industry you find yourself in, you'll inevitably find a need to store your data on one of the prominent cloud platforms like AWS, Azure, or GCP.
- Among these platforms, AWS stands out with an extensive market presence within the realm of cloud solutions. It boasts a comprehensive array of over 270 services.

Each of the aforementioned cloud providers offers a distinct suite of services catering to areas such as Data Science, Data Analytics, and Operations.

- This term pertains to a network comprising remote servers, highly capable computers interconnected and hosted on the internet. These servers are purpose-built to furnish a diverse range of computing services, encompassing storage, computational power, database management, networking, and software applications, accessible to individuals and organizations through internet connectivity.
- It denotes a versatile and expandable computing environment that is accessible via the internet. This virtualized infrastructure enables both users and businesses to tap into computing resources without the need for extensive localized infrastructure and the associated maintenance overhead.

9. OPS :

- In the realm of operations (Ops), both Docker and Kubernetes play a significant role. Within the context of both Docker and Kubernetes, "Ops" encompasses the methodologies, procedures, and toolsets employed for the management, deployment, scaling, and monitoring of applications encapsulated in containers.
- This domain revolves around operational practices that streamline pipeline construction, automate the creation of test cases, and ensure seamless deployment of projects with almost zero downtime. Additionally, it encompasses the dynamic scaling of entire machine configurations to accommodate increased loads as necessary.
- The term "Ops" finds common usage in the technology domain, denoting the collective activities, procedures, and strategies associated with governing and upkeeping various facets of software systems, infrastructure, and services.
- These methodologies are implemented to assure the dependability, security, and scalability of software applications and infrastructure at every stage of their existence.
- "Ops" pertains to the realm of computers, websites, and applications. It maintains the smooth functionality of all these components, ensuring websites remain accessible when desired, apps sustain uninterrupted operation during usage, and overall security and stability are upheld.

10. ZOOKEEPER :

- ZooKeeper stands as an open-source distributed coordination service, frequently employed in expansive distributed systems. Its primary role revolves around the management and orchestration of diverse processes and components within a system, guaranteeing their seamless collaboration.

- Within the Hadoop ecosystem, which predominantly features sessionless systems, ZooKeeper assumes a pivotal role by establishing a connection point to gather feedback and facilitate interactions with other systems.
- Functioning as a trustworthy repository of information, ZooKeeper plays a crucial part in facilitating harmonious interaction among distributed systems. It functions as the "glue" that binds together various elements of a intricate system, ensuring their effective cooperation even in challenging circumstances.
- ZooKeeper undertakes the role of an organizer and coordinator for computer programs. When numerous programs, such as applications and websites, necessitate synchronized operation, ZooKeeper steps in to facilitate their coordination, similar to how an organizer coordinates tasks.
- Its pivotal function is to facilitate seamless collaboration, adeptly manage alterations, and maintain operational continuity even when faced with obstacles along the way.

11. ETL :

- By amalgamating all the aforementioned elements, you should possess the capability to construct a project, specifically an ETL Project. ETL, which stands for "Extract, Transform, and Load," represents a project type characterized by this process.
- ETL constitutes a crucial procedure in data management and analytics, facilitating the movement and alteration of data from one location to another. This frequently involves transferring data from source systems to data warehouses or other designated target systems.
- At its core, ETL is a mechanism that gathers data from diverse origins, cleanses it, enhances its utility, and deposits it in a repository where it can be effectively employed for learning or decision-making purposes.

Here's a comprehensive technical dissection of the ETL process:

Extract:

- **Data Extraction:** During this stage, data is sourced from a range of systems, including databases, spreadsheets, APIs, and logs.
- **Data Profiling:** An analysis is conducted to grasp the structure, quality, and relationships inherent within the data.

Transform:

- **Data Transformation:** The data undergoes a transformation to align with the structure and prerequisites of the intended target system. This phase encompasses activities such as cleansing, filtering, aggregation, and enrichment.
- **Data Mapping:** Mapping rules are established to harmonize the attributes of source data with those of the target data.

- **Data Quality:** Rigorous checks are conducted to ensure the precision, uniformity, and entirety of the data.

Load:

- **Data Loading:** The transformed data is loaded into the designated target system, which could encompass a data warehouse or a database.
- **Data Indexing:** Indexes and structures are generated to optimize data access and querying, thereby enhancing efficiency.
- Through these meticulously executed stages, the ETL process ensures that data is not only successfully migrated but also refined to meet the demands of the desired application, contributing to informed decision-making and meaningful insights.

12. PROGRAMMING LANGUAGE : (base)

- Proficiency in modular coding and code optimization is imperative.
- Venturing into the domains of Data Industry, Big Data, Data Analytics, and Data Science necessitates a strong grasp of programming languages.

1. Mastery of JAVA - Particularly JAVA 10 (which incorporates SCALA's features).
2. Adeptness in SCALA - (Notably, it is constructed atop the JAVA programming language.)
3. Proficiency in PYTHON - (As of 2021, PYTHON has emerged as a prominent language in Data Engineering. This shift is attributed to PYTHON's extensive array of frameworks and libraries.)

- In these domains, expertise in these programming languages is paramount, facilitating the execution of tasks ranging from modular coding to efficient data handling and analysis.

13. DATA STRUCTURE ALGORITHM (DSA) : (base)

- The ability to construct logic for diverse algorithms using data structures is essential.
- In situations demanding code optimization, the creation of enhanced code versions, or the reduction of time complexity, a solid understanding of Data Structure Algorithms (DSA) proves invaluable.
- Emphasizing the importance of minimizing both space and time complexity in addressing the given problem statement.

14. STRUCTURED QUERY LANGUAGE (SQL) : (base)

- Proficiency in Structured Query Language (SQL) holds paramount importance across various roles in the data industry, spanning from data science and data analysis to data engineering. SQL forms the backbone of databases.
- SQL serves as a query language enabling the issuance of diverse queries to a range of databases such as DB2, Oracle, SQL Server, MySQL, and PostgreSQL. This is

primarily because, ultimately, data is often stored in a database system, whether it's a SQL or NoSQL database.

- Despite the variety of databases, the syntactical variations among them are typically minor, ensuring a degree of consistency in working with different database systems.

2. DATA SCIENCE:

- The definitions of artificial intelligence and data science are largely comparable.
- The foundational elements of data science are Python Programming, Data Stucture Algorithm (DSA) and Structured Query Language(SQL)
- The core elements of data science encompass mathematics, machine learning, deep learning, reinforcement learning, operations (Ops), cloud computing, computer vision, and natural language processing.

TOPICS TO KNOW :

1. PROGRAMMING LANGUAGE:

- Possessing this knowledge is of utmost importance as it serves as the foundation of Data Science.
- Undoubtedly, Python stands as one of the most widely adopted programming languages in the realm of data science.
- Python provides an extensive assortment of libraries and frameworks catering to tasks like data manipulation (NumPy, pandas), data visualization (Matplotlib, Seaborn), machine learning (scikit-learn, TensorFlow, PyTorch), among others.
- Its user-friendly nature and the robust support offered by its vast community render it an excellent option for both newcomers and seasoned professionals in the field.

2. DSA :

- Once again, this forms the fundamental underpinning of Data Science – the ability to proficiently structure data.
- For data scientists, a robust comprehension of data structures and algorithms holds significant worth.
- In the realm of complex algorithms, a firm grasp of these concepts translates to enhanced efficiency and efficacy in tasks such as data manipulation, analysis, and modeling. Furthermore, it empowers informed decision-making when selecting the most suitable techniques for diverse data science endeavors.

3. SQL :

- Within the realm of data science, SQL assumes a critical role as it offers the tools necessary to engage with and manipulate data housed within relational databases – a prevalent method for storing structured data.
- Although it might not supplant more intricate programming languages in intricate analysis and modeling tasks, its user-friendly nature and adaptability render it an essential skill for any data scientist.
- A proficiency level extending to an intermediate level is necessary.

4. MATHEMATICS :

- A comprehensive understanding of mathematics is essential for comprehending algorithms. Machine learning algorithms not only encompass theoretical understanding but also involve mathematical implementations.
- Grasping algorithms in a way that encompasses their mathematical foundations simplifies the process of implementation.
- Proficiency in mathematics equips data scientists to construct precise models, devise efficient algorithms, and render informed judgments grounded in data analysis. This proficiency also fosters an appreciation of the limitations and underlying assumptions of various methods, a crucial aspect for sidestepping common pitfalls during data interpretation and modeling.
- A robust grasp of mathematical concepts pertinent to their specific domain significantly amplifies data scientists' capabilities within their field.
- This knowledge furnishes the requisite theoretical framework and tools for comprehending, modeling, analyzing, and interpreting data.
- The pertinent areas within the mathematical spectrum include Probability, Statistics, Calculus, and Linear Algebra.
- A multitude of machine learning concepts are rooted in Probability. For instance, Decision Trees and Random Forests are built upon probabilistic principles.
- Linear and Logistic Regression hinge on Linear Algebra.
- All these machine learning algorithms are firmly grounded in mathematics, and a firm grasp of mathematical concepts significantly simplifies the process of implementing them.

5. MACHINE LEARNING:

The classification is as follows:

1. Supervised Learning
2. Unsupervised Learning
3. Semi-Supervised Learning

- A comprehensive understanding encompassing theory, mathematics, and practical application is pivotal in this context.
- This domain revolves around the deployment of algorithms and statistical models to empower computer systems in enhancing their proficiency within a designated task through data-driven learning.
- Within the realm of data science, machine learning methodologies are harnessed to scrutinize and decipher extensive datasets, unearthing valuable insights and discernible patterns that guide decision-making, predictive analyses, and automation.

6. DEEP LEARNING :

- In data science, deep learning techniques are harnessed to unravel complex patterns and insights from large datasets.
- This technology is instrumental in tasks like feature extraction, dimensionality reduction, and predictive modeling.
- By leveraging deep learning in data science, professionals can address intricate challenges and derive valuable insights from diverse and unstructured data sources.

7. REINFORCEMENT LEARNING :

- Reinforcement learning holds a significant role within the scope of data science.
- It's a machine learning paradigm that centers around training agents to make sequential decisions in an environment to maximize cumulative rewards.
- This approach is particularly valuable for scenarios where an agent learns through trial and error, interacting with its environment and adjusting its actions based on feedback.

8. COMPUTER VISION :

- Computer vision (CV) plays a crucial role within the domain of data science.
- It's a field that focuses on enabling computers to interpret and understand visual information from the world, similar to how humans perceive and interpret images and videos.

9. NATURAL LANGUAGE PROCESSING :

- Natural Language Processing (NLP) is a critical component of data science that focuses on enabling computers to understand, interpret, and interact with human language in a way that is both meaningful and valuable.

10. OPS:

- Ops in data science involves tasks such as data pipeline management, version control for code and data, automation of repetitive tasks, monitoring of models and data processes, deployment of machine learning models into production, and scaling infrastructure to handle growing data volumes.

11. CLOUD :

- Cloud services such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) offer a wide range of tools and services that support data storage, processing, analysis, and machine learning, making them integral to modern data science workflows.

3. DATA ANALYTICS:

- To put it differently, it's also referred to as business analytics.
- Data Analytics, Data Science, and big data operate collaboratively. Many correlations exist among data analytics, data science, and big data. These three elements coexist within the data industry. Data scientists collaborate with both the data analytics and big data teams.

1. SQL

- SQL holds significant importance within the realm of data analytics.
- SQL plays a vital role in data analytics by allowing analysts to interact with structured data stored in relational databases, thus facilitating the extraction of insights.

2. PYTHON :

- Python's versatility and extensive ecosystem of libraries make it an indispensable tool for data analysts.
- Its capabilities span data manipulation, visualization, analysis, machine learning, and more, contributing significantly to the field of data analytics.

3. DASHBOARD :

- Within the realm of dashboards, the following options are available:

1. Tableau
2. Power BI
3. Qlik

- Power BI - It stands out with the largest market share and serves as a versatile tool for creating dashboards. Power BI is a versatile tool that empowers organizations to gain insights from their data, make data-driven decisions, and communicate findings effectively through interactive visuals and reports.

- Tableau - Tableau is a widely used data visualization and business intelligence tool that helps users create interactive and shareable visualizations, reports, and dashboards. It allows users to transform raw data into meaningful insights, making it a valuable tool for data analysis and decision-making.

- Qlick - Qlik (formerly known as QlikView and Qlik Sense) is a data visualization and business intelligence platform that enables users to create interactive and intuitive visualizations, dashboards, and reports. Qlik focuses on providing self-service analytics and data discovery capabilities.

4. EXCEL :

- While Excel has its advantages for smaller-scale data tasks, it may not be suitable for handling very large datasets, complex analyses, or advanced visualization needs.
- While Excel has its advantages for smaller-scale data tasks, it may not be suitable for handling very large datasets, complex analyses, or advanced visualization needs. Nonetheless, it remains a widely accessible tool for various aspects of data analytics due to its familiarity and ease of use.

5. STATISTICS :

- Statistics provides the tools and methodologies to transform raw data into meaningful insights, validate assumptions, and support data-driven decision-making in various industries and domains.
- It plays a crucial role in data analytics by providing the foundation for understanding and drawing meaningful insights from data.

6. MACHINE LEARNING :

- In the context of data analytics, machine learning plays a significant role in automating and enhancing the process of extracting insights and making predictions from large and complex datasets.

THE BIG DATA INDUSTRY AND ITS FUNCTIONING :

- Data consistently arrives in diverse forms, exhibiting various variations and differing velocities. Consequently, efforts are directed towards categorizing and quantifying this data.

- Numerous data sources are accessible in the market to house data. For instance, data can be stored in formats such as Excel sheets, document files, notebooks, and PDFs. For instance, if dealing with a movie, it would be placed in an MP4 system, while audio content would be placed in an MP3 system. Ultimately, all forms, whether it's a JPG, MP4, or PDF, are considered as data.

- Within the industry, data is eventually stored using various file formats:

1. jSON
2. XML
3. ORC
4. RC

- Consequently, data can exist in any of these aforementioned formats, which pertain to different file formats.

DIFFERENT FILE FORMATS :

1. CSV :

- CSV stands for "Comma-Separated Values." It is a simple and widely used file format for storing tabular data, such as spreadsheets or databases.
- In a CSV file, each line represents a row of data, and within each line, fields are separated by commas (or other delimiters) to define the columns.

Here's a basic example of what a CSV file might look like:

```
Name, Age, City, Occupation
John, 25, New York, Engineer
Jane, 30, Los Angeles, Doctor
Mark, 22, Chicago, Student
```

2. PDF :

- PDF stands for "Portable Document Format." It is a widely used file format for presenting and exchanging documents, regardless of the software, hardware, or operating systems used by the viewer.
- PDF files are designed to accurately display and preserve the formatting, fonts, images, and other elements of a document, regardless of the device or platform on which they are viewed.
- It's worth noting that while PDFs are generally intended for preserving layout and formatting, they are not as easily editable as some other document formats.
- To edit the content of a PDF, specialized software is often required. However, simple modifications like adding comments, annotations, and form field entries can be done using various PDF readers.

3. XML :

- XML stands for "eXtensible Markup Language." It is a widely used markup language for structuring and storing data in a format that is both human-readable and machine-readable.
- XML is not a programming language but rather a set of rules for encoding documents in a format that is easy to understand for both humans and computers.
- XML uses tags to define elements within a document, similar to how HTML is used to structure content on the web. * XML is more flexible and versatile, as it allows you to define your own tags and document structure, making it suitable for representing a wide range of data formats.

Here's a basic example of XML syntax:

```
<person>
  <name>John Doe</name>
  <age>30</age>
  <city>New York</city>
</person>
```


4. JSON :

- JSON stands for "JavaScript Object Notation." It is a lightweight data interchange format that is easy for both humans to read and write and for machines to parse and generate.
- JSON is often used to transmit data between a server and a web application, as well as to store and exchange structured data.
- JSON data is represented as key-value pairs, where each key is a string and each value can be a string, number, boolean, array, object, or null.
- JSON closely resembles the syntax used in JavaScript object literals, which is why it's commonly associated with JavaScript. However, JSON is not limited to JavaScript and can be used in various programming languages.

Here's an example of JSON syntax:

```
{
  "person": {
    "name": "John Doe",
    "age": 30,
    "city": "New York"
  }
}
```

- JSON's simplicity and readability have contributed to its widespread adoption. Many programming languages provide built-in support for parsing and generating JSON, making it a convenient choice for data interchange and storage.

5. YAML :

- YAML stands for "YAML Ain't Markup Language" (a recursive acronym) or sometimes "Yet Another Markup Language."
- It is a human-readable data serialization format that is often used for configuration files, data exchange between languages with different data structures, and other scenarios where human readability and simplicity are important.
- Unlike JSON or XML, YAML is designed to be more intuitive for humans to write and read, with a focus on minimizing special characters and using indentation for structural representation. It is often used in projects where configuration files need to be easily edited by both developers and non-developers.

Here's a basic example of YAML syntax:

```
person:
  name: John Doe
  age: 30
  city: New York
```


- YAML is often used for configuration and data serialization, it's not suitable for all scenarios. For more complex data structures or when precise control over data validation is required, other formats like JSON or XML might be better choices.

6. MP4/MP3 :

- MP4 (MPEG-4 Part 14):

MP4 is a digital multimedia container format that can store a combination of audio, video, subtitles, and images. It is one of the most popular formats for delivering video content over the internet and is widely supported by various devices and platforms.

- MP3 (MPEG-1 Audio Layer III):

MP3 is an audio file format that uses lossy compression to significantly reduce the file size of audio recordings while preserving a reasonable level of sound quality. MP3 files are widely used for storing and playing back music and other audio content.

7. JPG, PNG, GIPHY :

- JPG, PNG, and Giphy are terms related to image file formats and a popular online platform for sharing animated images (GIFs).

- JPG (JPEG - Joint Photographic Experts Group):

JPEG, often referred to as JPG, is a commonly used image file format. It uses lossy compression to reduce the file size of images while attempting to preserve visual quality. JPEG is best suited for photographs and images with smooth gradients, as it can produce smaller files without significantly compromising the perceived quality.

- PNG (Portable Network Graphics):

PNG is another popular image file format. Unlike JPEG, PNG uses lossless compression, which means it retains all the image data without any loss of quality. This format is well-suited for images that require transparency, sharp edges, and simple graphics. PNG files can be larger than equivalent JPEG files due to the lack of lossy compression.

- GIPHY (Graphics Interchange Format):

Giphy is an online platform that allows users to discover, create, and share animated images known as GIFs. GIFs are a type of image format that supports animations by displaying a sequence of frames in a loop. GIFs are often used to convey short animations, reactions, or humorous content.

8. ORC, RC, PARQUET :

- "ORC," "RC," and "Parquet" are all file formats used for storing and managing data, particularly in the context of big data processing and storage. These formats are optimized for efficient storage, retrieval, and processing of large datasets.

- ORC (Optimized Row Columnar):

ORC is a columnar storage file format developed by the Apache Hive project, which is part of the Apache Hadoop ecosystem. It is designed to improve performance and storage efficiency for queries and analytics on large datasets. ORC files store data in a column-oriented manner, allowing for better compression and faster data retrieval for specific types of queries.

- RC (Record Columnar):

RCFile (Record Columnar File) is another columnar storage file format that was also developed for the Apache Hive project. Similar to ORC, RCFile is designed to optimize performance and storage efficiency by storing data in a column-oriented layout.

RCFile was one of the earlier columnar storage formats used in Hadoop-based systems, and it has since been largely replaced by more advanced formats like ORC and Parquet.

- Parquet:

Parquet is a columnar storage file format developed as a collaboration between the Apache Parquet and Apache Arrow projects. It is designed for efficient storage and processing of large datasets in big data frameworks like Hadoop and Spark.

9. TXT :

- "TXT" stands for "text," and it refers to plain text files. A text file is a simple and common type of computer file that contains only human-readable characters, without any special formatting or binary data.
- Text files are used to store and exchange textual information, such as documents, code, configuration files, and more.

Your client has no constraints on the data formats they provide to you or an external system that generates files. Unfortunately, that external system cannot modify the file format according to your preference.

- Part 1: As an engineer, it's important for you to possess the skills to efficiently format data with various extensions, regardless of its source.
- Part 2: You find yourself dealing with data derived from diverse systems, each utilizing distinct file formats.

You receive data from the following options:

1. Cloud Platforms - AWS, Azure, GCP
2. Salesforce Platform
3. SAP System - FICO, FIFO, CRM, FICA, HANA
4. Oracle
5. Sensors - Found in vehicles, devices, satellites, these collect data.
6. Applications - Data extraction from various apps to generate required analytics.

- Across different systems, there are numerous instances where a multitude of data is extracted in varying formats.

Modes of produced data include:

1. Batch Mode
2. Real Time Mode
3. Near Real Time Mode/ Mini Batch

1 . Batch Mode :

- A system generates data periodically, with intervals that could span seconds, minutes, hours, days, or even months. Our approach involves processing this data in batches, as exemplified by the production of banking system statements.
- Batch mode refers to a data processing method where a collection of data is processed as a group or batch. In this mode, data is accumulated over a period of time and then processed all at once, which can lead to more efficient resource utilization and improved processing speeds for certain types of tasks.

2. Real Time Mode :

- Real-time mode involves working with sensors that generate continuous streaming data. For instance, government display boards continuously show pollution levels or AQI (Air Quality Index) in real-time. The process involves immediate data capture

through sensors that scan and collect data without delay, ensuring ongoing data processing.

- Real-time mode refers to the operational state in which data is processed and responded to as it is generated, without any significant delay.
- This mode is often associated with systems that deal with streaming data, such as sensors, where information is collected and acted upon instantaneously, allowing for rapid and responsive decision-making.

3. Near Real Time Mode :

- Types of data being processed within slight delay of milliseconds and microseconds.
- Near real-time mode signifies a processing approach where data is dealt with and responded to with only a slight delay, typically in a matter of seconds or minutes after its generation.
- This mode aims to provide prompt insights and actions while not achieving the immediate response of true real-time systems.

DATA IN REAL TIME :

- Numerous systems and devices, employing diverse formats, have the capability to generate a multitude of data, each with varying speeds. A vast array of data types exists. From a systemic perspective, the diversity of systems further expands, and within that context, data is obtained through various modes.
- In a real-time setting, acquiring the precise type of data necessary to construct a specific use case might not always be feasible.
- Consider the analogy of **crude oil**, wherein the extraction of petroleum, gasoline, and diesel involves separating components at distinct temperature thresholds. Just as crude oil cannot be directly used in vehicles without refining, data, often likened to oil, is akin to crude oil. It necessitates refining; utilizing raw data directly for final use cases is unfeasible.
- **DATA IS CRUDE OIL.**
- The analogy draws a parallel between data and crude oil, highlighting that, while data is valuable, it requires refinement. Similar to the process of extracting usable components from crude oil, data must be refined to align with your business use cases, enabling the creation of practical business applications.

The categories of data you will encounter include:

1. Structured data
2. Unstructured data
3. Semi-structured data

1. Structured data:

- Structured data refers to information organized according to a predetermined schema.
- For instance, if you're gathering data from an SQL system or maintaining an organized Excel spreadsheet with well-defined column names, the data can be considered structured.
- Structured data refers to information that has been organized and formatted according to a predefined schema or data model.
- This schema outlines the specific categories, relationships, and types of data that are present within the dataset.
- Examples of structured data include data stored in relational databases, well-organized spreadsheets, and other systems where data is categorized into rows and columns with clear labels and data types.
- The structured nature of this data makes it easily searchable, analyzable, and suitable for various data processing tasks.

2. Unstructured data:

- Unstructured data, on the other hand, involves sources like PDFs, web scraping from platforms such as Stack Overflow or Wikipedia.
- In these cases, the data lacks a specific schema or format; it's essentially raw text being extracted.
- An example of this is a post on LinkedIn. Here, they utilize graph algorithms to establish connections and degrees of relationships. Each degree of connection results in unstructured data due to the inclusion of emojis, images, videos, and similar content.
- Unstructured data refers to information that lacks a specific organizational schema or predefined format.
- This type of data doesn't adhere to a consistent structure, making it more challenging to classify and analyze using traditional methods.
- Examples of unstructured data include free-form text, images, audio recordings, videos, social media posts, and other content that doesn't fit neatly into rows and columns.
- Because of its lack of structure, unstructured data often requires advanced techniques like natural language processing, image recognition, and machine learning algorithms to extract meaningful insights from the raw content.

3. Semi-structured data :

- Semi-structured data falls in between these two categories. It possesses a partial schema, meaning there is a certain structure in place but it might not be consistently followed throughout the entire dataset.
- Semi-structured data is a type of information that exhibits characteristics of both structured and unstructured data. It possesses some level of organization or structure, but it doesn't conform entirely to a rigid schema like structured data does.
- While there might be certain patterns or attributes that repeat across the dataset, there's also flexibility in terms of how data elements are presented.
- Semi-structured data often includes elements like tags, labels, or attributes that provide a degree of categorization or organization. This makes it more accessible and amenable to processing compared to completely unstructured data.
- Examples of semi-structured data include XML files, JSON data, and certain types of documents where there is a mix of organized sections along with variable content.
- Because semi-structured data retains some order while allowing for variations, it's possible to use a combination of manual and automated methods to extract useful information from it.

DATA INDUSTRY WORKING STRATEGY :

1. DATA ENGINNEERS

- I am retrieving information from various systems, with distinct values arriving at varying speeds and modes, each exhibiting unique data structures. This is the point at which individuals specialized in **Data Engineering or Big Data expertise** become essential in the **first layer**. They manage a multitude of systems and diverse data forms.
- In the capacity af a **Data Scientist**, the role involves specifying the desired data type. It is focused on addressing specific business challenges, honing in on particular attributes within that context.
- When data is managed by a skilled Data Engineer, the concept of ETL (Extract, Transform, Load) comes into play.

1. **Extract:**

At this stage, data is obtained from various systems.

Engineers create code that establishes connections with all relevant systems to pull the necessary data. This extraction involves dealing with diverse data types and formats, ranging from structured to semi-structured or unstructured, and it can involve both batch and real-time data.

The ETL process functions as a pipeline. In the initial stage of the pipeline, code is written to extract the data.

1. **Transform:**

In the transform phase, engineers code transformations to prepare the data for its intended purpose.

Transformations are then implemented based on specific business requirements. For instance, if the received data contains a time attribute in hours and needs to be converted to minutes, this constitutes a data transformation.

1. **Load:**

This is where the transformed or extracted data finds its place. Depending on the structure of the data, there are various storage options available:

- Structured Storage: SAP systems, Oracle databases, SQL and MySQL databases, DB2 systems, PostgreSQL, and Google BigQuery.
- Unstructured Storage: Hive, HBase, MongoDB, Cassandra, Neo4j, Redis, and InfluxDB.

  The choice of storage solution is guided by the system's architecture. An architect determines where data will be sourced, what kind of transformations are necessary, and where the final data will reside. Different products have unique use cases.

  Social media platforms like LinkedIn, Facebook, and Instagram rely on graph databases due to their inherent flexibility. Textual data is best suited for Elasticsearch and Apache Solr. Certain data might find its home in document-based databases like MongoDB or Cassandra. Indexed databases, such as HBase, or data warehousing solutions like Hive, could also be utilized.

  The specific requirements of each case dictate the approach. System design precedes the execution of the extract, transform, and load operations. Once data is cleaned at the **initial level**, it is loaded into an appropriate database or file system.

2. DATA SCIENTIST

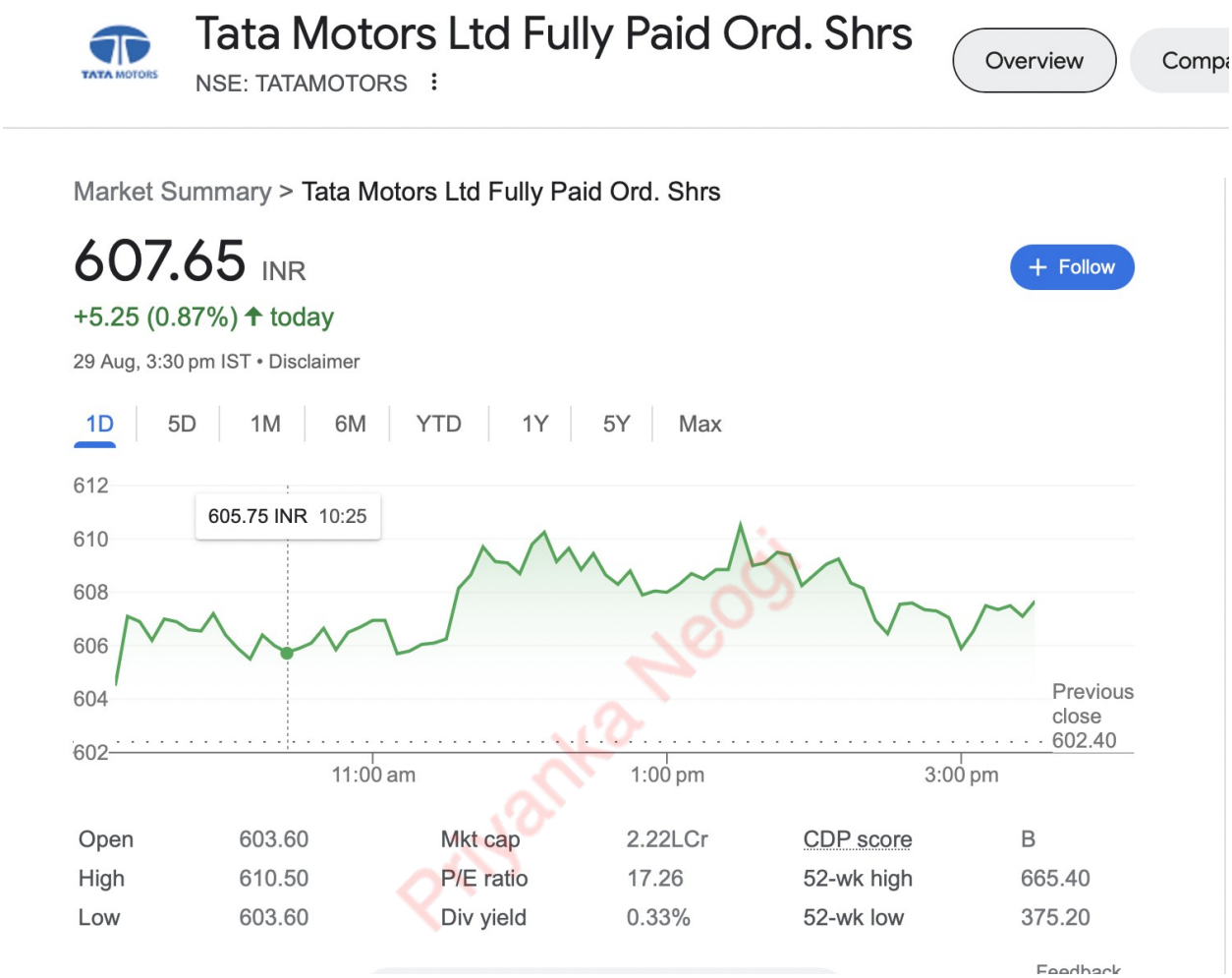
- Now enters the data scientist, who now becomes involved in assessing the business requisites for constructing both predictive and forecasting models, involving statistical scrutiny.
- A common misconception is that data scientists merely work with data presented as in Kaggle datasets in formats like Excel sheets or CSV files. They are thought to perform tasks such as data cleaning, preprocessing, exploratory data analysis (EDA), and feature engineering using libraries like pandas, numpy, and scipy.
- However, reality differs from this perception.

- **In actuality, data scientists enter the picture after the Extract, Transform, Load (ETL) process.**
- They first grasp the business issue at hand and evaluate the adequacy of the available data for the intended business objective.
- Being a data scientist doesn't imply an immediate dive into building models as soon as data is obtained. Their responsibility encompasses verifying if the existing data and attributes are sufficient to proceed and if they can effectively support the final model. Until this assurance is obtained, data manipulation does not commence.
- The focus shifts towards identifying additional attributes that might be necessary. This evaluation is central to data science – determining if the available data is suitable for constructing the intended model.
- If the data falls short, the data scientist collaborates with the big data team to identify appropriate data sources. This back-and-forth interaction involves discussions and consultations, often with the involvement of an Data Architect. For instance, when extracting real-time system data, Architects might need to be consulted to devise a solution.
- ChatGPT originates from the field of Natural Language Processing (NLP). The evolution of NLP models started with Recurrent Neural Networks (RNNs), then progressed to Transformers, Attention Models, and eventually generative models. The foundation of most modern technologies, including the one powering ChatGPT, is rooted in the Attention Model, which facilitates the understanding of text and speech.
- MODELS : They are essentially mathematical equations. By inputting values for X and constants, predictions or forecasts (referred to as outcomes, often denoted as y) are obtained. In the context of models, the equation itself can be exceedingly Complex.

3. DATA ANALYSTS

- After the model is built and predictions and forecasts are made, the data analyst will then step in. The data analyst will prepare the final dashboard, which will display the predictions and forecasts in a graphical format. This will allow the buyer to make decisions based on the trends.
- The model is built. This involves collecting data, training the model, and evaluating the model.
- Predictions and forecasts are made. This involves using the model to predict future values of the data.
- The data analyst prepares the final dashboard. This involves selecting the data to be displayed, creating the graphs and charts, and adding annotations to explain the data.

- For Example - In Stock Market , The buyer makes decisions based on the trends. This involves interpreting the data and using it to make decisions about the future.



- The data analytics professional plays a crucial role in enabling users to make informed decisions by creating interactive dashboards that display trends through various types of graphs such as pie charts, scatter plots, and line plots.
- These dashboards empower businesses to efficiently supervise operations and swiftly address issues, saving valuable time.
- Additionally, within these dashboards, various ML models are integrated.

In the data industry, collaboration among data engineers, data scientists, and data analysts from different domains is essential for success.

- In any industry, the flow of data typically starts with the collection and analysis of big data. This involves examining large volumes of data to uncover patterns and trends.

- The insights gained from big data analysis are then used in the field of data science, where professionals use statistics, mathematics, programming, and problem-solving to extract valuable insights from the data.
- Finally, the data analysts utilize these insights to create interactive dashboards with various types of graphs, allowing businesses to make informed decisions and address issues efficiently.
- This sequential flow of data is considered an ideal situation in the industry

Various Application Areas in AI and Data Science:

1. AI and Data Science Applications in Social Media Platforms - Instagram, Twitter, Facebook, LinkedIn:
   - i. Within Facebook, the system employs an Image Processing or Facial Identification algorithm to automatically tag individuals when photos are uploaded, utilizing recognition from past photos.
   - ii. An implemented generative model assists in automatically completing typed text on these platforms.
   - iii. Utilizing AI techniques like 'Text to Speech' and 'Speech to Text,' these platforms allow communication through both voice and text.
   - iv. YouTube employs sentiment analysis algorithms to automatically detect and prohibit abusive comments, sparing the need for human manual review of each post.
2. AI and Data Science Applications in Gmail:
   - The Gmail service employs a generative model to automatically assign tags to emails, enhancing organization and retrieval.
3. AI and Data Science Applications in E-commerce Platforms - Amazon, Flipkart, Myntra:
   - i. E-commerce sites leverage users' search patterns to provide tailored recommendations.
   - ii. Utilizing prediction forecasting and recommendation systems, these platforms offer suggestions based on users' past purchases.
4. AI and Data Science Applications in Facial Recognition:
   - i. In instances such as DigiYatra and office attendance systems, facial recognition technology is employed. Cameras capture and recognize individuals based on their unique facial features, even if appearance has changed over time.
   - ii. Many cities' traffic management uses number plate detection for tasks like issuing speeding tickets through automated processes.
5. AI Voice Assistants - Alexa, Siri, Google Voice Assistance:
   - These virtual assistants convert spoken language to text, perform online searches, convert text back to speech, and execute tasks based on verbal instructions.
6. Google Translator:

    - Google's translation service employs AI techniques to automatically translate text from one language to another.
7. AI and Data Science in Ride-Sharing Services like OLA/UBER:
    - i. In the past, fare pricing was displayed post-ride; however, current practices involve estimating fares beforehand. This estimation considers factors like fuel prices, traffic conditions, and distance, utilizing historical data for pricing generation.
8. Zomato and Swiggy: They make a prediction regarding the cost.
9. Weather Forecasting : Weather predictions are being made for upcoming days and weeks.

Will AI replace our job?

- "AI will not replace humans, but will complement our skills and abilities."
- "AI is a tool that can be used to automate tasks and make our work more efficient."
- "AI can free up human workers to focus on more creative and strategic tasks."
- "AI can help us to solve complex problems and make better decisions."
- "AI can be a powerful ally in our quest to create a better future for all."
- As humans, we should continue to evolve our skills and knowledge to stay relevant in the workforce."
- "AI is rapidly changing the workplace, but humans still have a vital role to play. We need to be prepared to adapt to new technologies and learn new skills."
- "AI is not a threat to humanity, but it is a challenge. We need to embrace AI and use it to our advantage, not fear it."
- "The future of work is uncertain, but one thing is for sure: humans will always be needed to do jobs that require creativity, empathy, and problem-solving skills. Let's evolve ourselves to meet the demands of the future."