**6.1.1**

Human-Generated Data comprises all files and data that we create every day. Again, this data can be structured, semi-structured, or unstructured. For instance, human-generated structured data can be any piece of input data entered into a device. It can be personal details, survey responses, and so on. This can include emails, social media usage, messages we send and receive, word processing documents, presentation documents, spreadsheets, audio files, video files, and so on.

Sensors:

Machine-Generated data is data collected by sensors. Sensors are the managers of the physical computational devices in our ecosystem and are responsible for the act of sensing and acquiring data from the environment. The type of data can be categorized based on data origin, it would be structured if it is from a device, semi-structured if it is from a log file, and unstructured when it is a video or image file.

AI or Machine Learning:

The competitive advantage of advanced analytics can be multiplied by building a strong AI capability. There is far more data being generated today than humans can analyze in any meaningful way. Techniques like machine learning, predictive analytics, and data visualization can help us find meaning by digging deeper into large data sets and improving the speed and accuracy of decision making. How important is leveraging AI and advanced analytics in driving value creation and future growth? In your organization’s transformation journey, it can mean the difference between success and failure. Some practical applications of analytics and AI include:

Tracking and forecasting relevant exponential technology trends. Doing this proactively will help you determine how and when to act, to make better decisions, and stay ahead of competitors.

Using predictive analytics to minimize decisions based on intuition or outdated models. This is what analytics sage Avinash Kaushik calls the HiPPO effect: relying on the Highest Paid Person’s Opinion instead of relevant data.

The ability to benchmark and track the progress and speed of individual innovation projects through development phases and predict future outcomes and revenues.

ioT and Big Data:

When organizations are grabbing hold of the data for analysis purposes, IoT is acting as a major source for that data, and this is the point where the role of big data in IoT comes into the picture. Big data analytics is emerging as a key to analyzing IoT generated data from “connected devices” which helps to take the initiative to improve decision making.

The role of big data in IoT is to process a large amount of data on a real-time basis and store them using different storage technologies.

Transactional Data:

Transactional Data is information that is captured from transactions. It records the time of the transaction, the place where it occurred, the price points of the items bought, the payment method employed, discounts if any, and other quantities and qualities associated with the transaction. Transactional data is usually captured at the point of sale.

In other words, transactional data is data generated by various applications while running or supporting everyday business processes of buying and selling. A large and intricate web of point-of-sale servers, security software, ATM, and payment gateway data exists, originating from every device used to complete a financial transaction.

Market Analysis:

Definition of your industry – Pick what your business is going to do and base everything around that

Size – Decide how large the business is going to be, who your competitors are and how you will compete

Rate of growth – How do you plan to expand, estimating the growth of the company and how long until you reach certain goals

Potential for growth – How big can your company become and survive for as long as possible

Trends in your industry – The way you will keep up with each trend and create new ones yourself

Sustainability of your industry - Making sure your business survives and takes

calculated risks.

**6.1.2**

System performance analysis, also sometimes referred to as profiling, performance analysis as it relates to the process of evaluating how a particular software program is functioning. This process normally begins with how the program loads and what happens when each step is using the program is executed. The object of performance analysis is to ensure the software program is working at greatest efficiency and to identify and correct any issues that may negatively impact efficiency.

System performance analysis includes:

* Identify common system configuration problems -
* Analyse the performance of the underlying platform and find performance bottlenecks -
* User account access and monitoring -
* Admin/User role changes -

Data required by the system analyst:

* CPU and memory utilization -
* Memory and socket interconnect bandwidth -
* Cycles per instruction – how many cycles it takes for each instruction
* Cache misses rates
* Type of instructions executed
* Storage device access metrics

Market analysis:

Market analysis is a **quantitative** and **qualitative** assessment of a market for a business.

It could look at the size of the market for both:

•Volume

•Value

Customer segmentation analysis, buying patterns, competition, and the economic environment in terms of barriers to entry and regulation could be included.

User monitoring:

All platforms that you use daily have a user monitoring system built into it.

Windows, Mac, Android, IOS, Facebook, SnapChat, YouTube, Chrome, Google, Twitter, Discord, TikTok, Spotify, Netflix, Amazon,

This also includes – Your School, This College and your future employer.

User monitoring is a crucial part of data analytics

Ebay etc.

**User monitoring**

|  |  |
| --- | --- |
| The Good: | The Bad: |
| Increased security | Invading privacy of people |
| An employer can check and monitor the employees’ work efficiency | Reduce employee motivation |
| Can prevent people from accessing malware | Efficiency may slow down if they are being monitored to make sure the work is perfect. |
| It can track a location | Old devices may slow down due to the software that is used to monitor it |
| You can perfect what is being used more often and adapt it for the users | It can spread distress |

Target market analysis – a systematic and comprehensive assessment that allows you to identify important characteristics of your target market and group them into categories based on those characters.

4Ps:

Product – quality design, packaging brand

Price – retail price, payment plans, discounts, credit terms

Promotion – advertising, emails, personal selling, public relations

Place - retail location, delivery, downloads, distribution

1. Strategic decisions:

Strategic decisions are major choices of actions and influence whole or a major part of business enterprise. They contribute directly to the achievement of common goals of the enterprise. They have long-term implications on the business en­terprise.

They may involve major departures from practices and procedures being followed earlier. Strategic decisions are unstructured and thus, a manager must apply his business judge­ment, evaluation and intuition into the definition of the problem. These decisions are based on partial knowledge of the environmen­tal factors which are uncertain and dynamic. Such decisions are taken at a higher level of management.

2. Tactical decisions:

These decisions relate to the implementation of strategic decisions. They are directed towards developing divi­sional plans, structuring workflows, establishing distribution chan­nels, acquisition of resources such as men, materials, and money. These decisions are taken at the middle level of management.

3. Operational decisions:

These decisions relate to day-to-day op­erations of the enterprise. They have a short-term horizon as they are taken repetitively. These decisions are based on facts regarding the events and do not require much business judgement. Operational decisions are taken at lower levels of man­agement. As the information is needed for helping the manager to take rational, well-informed decisions, information systems need to fo­cus on the process of managerial decision making.

**6.2.1**

Data Formats:

Data types are the classification or categorizing of data items.

Data types represent a kind of value which determines what operations can be performed on that data. Numeric, non-numeric and Boolean data are the most used data types. However, each programming language has its own classification reflecting its programming philosophy.

Date – The time and date

Integer – Numbers, and switched strings, but without symbols. Floats, complex numbers

Real – Used as an approximation or real number, used for decimal points, used for correct calculations.

Character – Parameter for CHAR, UNICODE, and special symbols. Used to search through a file and replace any special character.

String – Words, letters and symbols,

Boolean – True or False

JSON – file type where you put values in brackets and easy to encrypt

Fixed width text file – When you format excel into a text file, it will automatically determine the widths and insert spaces to pad to that width as necessary.

CSV – A normal text file that’s just rows, and the columns are separated by commas

ASCII – mostly used for measurement files, and it is stored using only characters, numbers, punctuation, tabs, and carriage return characters.

XML – It carries data and it’s not predefined. It simplifies data sharing, transport, platform changing, and availability.

Excel – Data saved in columns and rows

File based Structure:

The systems that are used to organize and maintain data files are known as file-based data systems. These file systems are used to handle single or multiple files and are not very efficient.

Functionalities:

The functionalities of a File-based Data Management System are as follows:

* A file-based system helps in basic data management for any user.
* The data stored in the file-based system should remain consistent. Any transactions done in the file-based system should not alter the consistency property.
* The file-based system should not allow any illegal or potentially hazardous operations to occur on the data.
* The file-based system should allow concurrent access by different processes, and this should be carefully coordinated.
* The file-based system should make sure that the data is uniformly structured and stored so it is easier to access it.

Advantages:

The file based system is not complicated and is simple to use, because of this, this system is quite inexpensive.

Because the file-based system is simple and cheap, it is normally suitable for home users and owners of small businesses.

Since the file-based system is used by smaller organizations or individual users, it stores comparatively lesser amount of data. Hence, the data can be accessed faster and more easily

Disadvantage:

The file-based system is limited to a smaller size and cannot store large amounts of data.

This system is relatively uncomplicated, but this means it cannot support complicated queries, data recovery etc.

There may be redundant data in the file-based system as it does not have a complex mechanism to get out of it.

The data is not very secure in a file-based system and may be corrupted or destroyed.

The data files in the file-based system may be stored across multiple locations. Consequently, it is difficult to share the data easily with multiple users

**6.3 - Data systems**

Data wrangling - Core Functions - Data entry and maintenance - Visualization

Data Wrangling:

Data wrangling – also called data cleaning, data remediation, or data munging – refers to a variety of processes designed to transform raw data into more readily used formats. The exact methods differ from project to project depending on the data you’re leveraging and the goal you’re trying to achieve.

Some examples of data wrangling include:

* Merging multiple data sources into a single dataset for analysis.
* Identifying gaps in data (for example, empty cells in a spreadsheet) and either filling or deleting them.
* Deleting data that’s either unnecessary or irrelevant to the project you’re working on.
* Identifying extreme outliers in data and either explaining the discrepancy or removing them so that analysis can take place.

Structuring or Structure

Raw data is typically unusable in its raw state because it’s either incomplete or mis formatted for its intended application. Data structuring is the process of taking raw data and transforming it to be more readily used. The form your data takes will depend on the analytical model you use to interpret it.

Cleaning:

Data cleaning is the process of removing inherent errors in data that might distort your analysis or render it less valuable. Cleaning can come in different forms, including deleting empty cells or rows, removing outliers, and standardizing inputs. The goal of data cleaning is to ensure there are no errors (or as few as possible) that could influence your final analysis.

Enriching:

Once you understand your existing data and have transformed it into a more usable state, you must determine whether you have all of the data necessary for the project at hand. If not, you may choose to enrich or augment your data by incorporating values from other datasets. For this reason, it’s important to understand what other data is available for use.

If you decide that enrichment is necessary, you need to repeat the steps above for any new data.

Validating:

Data validation refers to the process of verifying that your data is both consistent and of a high enough quality. During validation, you may discover issues you need to resolve or conclude that your data is ready to be analyzed. Validation is typically achieved through various automated processes and requires programming.

Output:

Once your data has been validated, you can publish it. This involves making it available to others within your organization for analysis. The format you use to share the information – such as written reports or electronic files – will depend on your data and the organization’s goals.

**The importance of data wrangling**

Any analysis a business performs will ultimately be constrained by the data that informs them. If data is incomplete, unreliable, or faulty, then analyses will be too – diminishing the value of any insights gleaned.

Data wrangling seeks to remove that risk by ensuring data is in a reliable state before it’s analyzed and leveraged. This makes it a critical part of the analytical process.

It’s important to note that data wrangling can be time-consuming and taxing on resources, particularly when done manually. This is why many organizations institute policies and best practices that help employees streamline the data cleanup process – for example, requiring that data include certain information or be in a specific format before it’s uploaded to a database.

For this reason, it’s vital to understand the steps of the data wrangling process and the negative outcomes associated with incorrect or faulty data.

Core Functions:

Input – Add information to the database such as JSON or SQL tables

Search – Look up data from the tables by either querying through SQL or using loc/iloc in Pandas

Save – Save structured and unstructured data from the data set

Integrate – Integrate the SQL table or JSON file into another system or integrate them to be used with other queries from the main system.

Organize(index) - Similar way that we can search and save those searches using the index method. This way it becomes easier for us to find the set file when we need to.

Output – Finalize the file and output as a chosen extension such as Xml, Xls, Txt, Sql, DB, or JSON file.

Feedback Loop – A feedback loop will create a cycle of interaction between a user and a service provider, to gather, analyze and act on direct customer insights to improve the delivery of that service. It doesn’t need a big investment or high levels of technical skills. Just an open approach to learning, testing, and feeling out what’s right for the team.

Data entry and maintenance:

Data entry is the process of inputting data or information into the computer using devices such as a keyboard, scanner, disk, and voice.

Data entry is a type of clerical work that involves using various processes like typing and voice recording for entering data into computers. Data entry clerks work in industries such as healthcare, finance, retail and transportation.

A data entry job entails working as a handler of different types of electronic data and operating devices that professionals use to enter and edit data, such as a keyboard. There are a number of occupations in this industry, including typist, coder, transcriber or work processor.

Jobs in this industry have several different payment methods. If you start working in this industry, you could be paid by project, keystrokes per minute, keystrokes per word, keystrokes per hour or receive an hourly wage. Typically, the payment rate in data entry is based on your typing speed – fast typists are likely to earn more money in this industry.

Single-level directory:

The single-level directory is the simplest directory structure. In it, all files are contained in the same directory which makes it easy to support and understand. A single level directory has a significant limitation, however, when the number of files increases or when the system has more than one user. Since all the files are in the same directory, they must have a unique name. If two users call their dataset test, then the unique name rule violated.

Advantage:

* Since it is a single directory, its implementation is very easy.
* If the files are smaller in size, searching will become faster.
* The operations like file creation, searching, deletion, updating are very easy in such a directory structure.

Disadvantage:

* There may be a chance of a name collision because two files can have the same name.
* Searching will become time consuming if the directory is large.
* This cannot group the same type of files together.

Two-level directory:

As we have seen, a single directory often leads to confusion of file names among different users. The solution to this problem is to create a separate directory for each user. In the two-level directory structure, each user has their own user files directory (UFD). The UFDs have similar structures, but each list only the files of a single user. System’s master file directory (MFD) searches whenever a new user id=s logged in. The MFD is indexed by username or account number, and each entry points to the UFD for that user.

Advantages:

* We can give full path like /Username/directory name/.
* Different users can have the same name directory as well as the file name
* Searching of files becomes easier due to pathnames and user grouping

Disadvantages:

* A user is not allowed to share files with other users
* Still, it’s not very scalable, two files of the same type cannot be grouped together in the same user

Tree-structure-directory:

Once we have seen a two-level directory as a tree of height 2. The natural generalization is to extend the directory structure to a tree of arbitrary height.

This generalization allows the user to create their own subdirectories and to organize their files accordingly.

A tree structure is the most common directory structure. The tree has a root directory, and every file in the system has a unique path.

Advantages:

* Very general, since full pathname can be given.
* Very scalable, the probability of collision is less.
* Searching becomes very easy, we can use both absolute paths as well as relatives.

Disadvantages:

* Every file does not fit into the hierarchical model, files may be saved into multiple directories.
* We cannot share files.

If the linked page is located in the same directory: Your link should only list the file name (I.e. mystuff.html)

If the linked page is located in a HIGHER directory: Your link would list “../” before the file name to indicate “move to the parent directory”

If the linked page is located in a sub-directory: Your link would list the sub directory name before the file name (I.e., mysubdirectory/mystuff.html)

**Online Data Entry:**

* Ensuring the accuracy and relevancy of all data entered databases.
* Highlighting any discrepancies in data and reports to management.
* Entering sales data, market survey information and personal information into computer databases.
* Handling confidential information in line with the firm’s data security protocols.
* Performing data review and clean-up processes.

**Data Security:**

* VPN (virtual private network)
* SSL (Secure socket layer)
* TLS (Transport layer security)
* SFTP (Secure file transport protocol)
* SSH (Secure Shell)

**Risk of Data Entry Errors**

Risk of data entry errors – There are many risks of entering the wrong data or figures. It can cause a backlog of issues when wrong data is processed. Therefore, most data collection and entry firms use data validation process before the data becomes information. This process adds sub checks on the data so it's being validated by the system first and then the team manager or department manager before being pushed to the new file. There are also rules where data validation needs to be on separate files so the original file does not get overwritten at any time.

**Time to create the entry screen and enter data**

All the data entry clerk jobs involve working with bespoke apps which only show them partial information they need to enter from one set to another.

This process can be time consuming also because you need to create an app which with the minimum UI which the data entry clerks can use without seeing all the sensitive data.

This is also important because it allows for a faster workflow when you get the right UI right for the dataset they will be working with.

Using Python’s TK Inter, or other UI frameworks, we can create such entry screens for people to understand and use.

In the UI screen you can also add validation which can check if the same information from the rows is inputted inside of the other row. This way you can add 2 factor validation for the data entry process.

**Data Visualization:**

|  |  |  |
| --- | --- | --- |
| Methods | Usage | Examples |
| Graphs/charts | See the difference between different aspects on the chart. |  |
| Data tables | See data more detailed and can be used to calculate a total for each row or column. |  |
| Reports | Information by an organization that represents how they are doing financially. |  |
| Infographics | Infographics are graphic visual representations of information, data, or knowledge intended to present information quickly and clearly. It can help people spot patterns. |  |

**6.3.2**

Business Information Systems

Business information systems are sets of inter-related procedures using IT infrastructure in a business enterprise to generate and disseminate desired information. Such systems are designed to support decision making by the people associated with the enterprise in the process of attainment of its objectives.

In today’s business world, there are varieties of information systems such as transactions processing systems (TPS), office automation systems (OAS), management information systems (MIS), decision support system (DSS), and executive information systems (EIS), Expert System (ES) etc.

Business Information System:

Business Information systems are sets of inter-related procedures using IT infrastructure in a business enterprise to generate and disseminate desired information. Such systems are designed to support decision-making by the people associated with the enterprise in the process of attainment of its objectives.

In today’s business world, there are varieties of information systems such as transactional systems such as transaction processing systems (TPS), office automation systems (OAS), management information systems (MIS), decision support system (DSS), and executive information systems (EIS), Expert System (ES) etc.

What are business intelligence tools?

Business Intelligence (BI) tools are types of application software that collect and process large amounts of unstructured data from internal and external systems, including books, journals, documents, health records, images, files, emails, videos and other business sources. While not as flexible as business analytics tools, BI tools provide a way of amassing data to find information primarily through queries. These tools also help prepare data for analysis so that you can create reports, dashboards, and data visualizations. The results give both employees and managers the power to accelerate and improve decision-making, increase operational efficiency, pinpoint new revenue potentials, identify market trends, report genuine KPIs and identify new business opportunities.

Typically used for more straightforward querying and reporting of business data, business intelligence tools can combine a broad set of data analysis applications, including ad-hoc analysis and querying, enterprise reporting, online analytical processing (OLAP), mobile BI, real-time BI, operational BI, cloud and software as a service BI, open-source BI, collaboration BI and location intelligence. It can also include data visualization software for designing charts, as well as tools for building BI dashboards and performance scorecards that display business metrics and KPIs to bring company data to life in easy-to-understand visuals.

KPI – Key Performance Indicator

Financial Planning and analysis tools (FP&A)

FP&A is the budgeting, financial forecasting, and financial analysis and decision-making that support the financial health and strategy of every company. This team is responsible for a variety of activities, including periodic financial close and consolidation, strategic and annual planning, monthly forecasting, cash flow forecasting, financial forecasting, financial modeling, and what-if scenario planning and analysis.

It’s the job of the FP&A team to understand the financial performance of the company – past, present and future. They apply context and narrative to what is happening in the business and why it’s happening. Their comprehensive financial lens allows them to spot new opportunities and anticipate unexpected threats to position the company for success. This puts them at the strategic center of the organization, with business leaders all over the company looking to them to provide critical insights and recommendations on how to navigate the business forward.

Customer Relationship Manager (CRM)

Simply put, CRM is a technology that helps manage a company’s interactions with customers and prospects, with an aim to improve business relationships. CRM tools help in contact management, sales management, enhancing productivity, improving profitability and more.

With the right solution, you can focus on relationships with customers, partners, service users, suppliers and colleagues throughout their lifecycle. At the same time, it helps you identify and nurture new leads, convert those into customers, and provide support and services.

What challenges do businesses face without one?

* Difficulty following the pace of lead flow
* Unorganized customer data
* Lack of planning and unclear growth path
* Time-consuming processes
* Poor customer service
* Inability to manage long-term accounts

Benefits:

Automated Tasks: A CRM tool helps automate mundane and repetitive everyday tasks like scheduling emails and creating reports, freeing up employees’ time so that they can focus on engaging with leads, resolving issues and closing deals.

Improved Communication: With access to the same customer data, sales and support reps can offer the same high level of service to everyone. This ensures customers don’t have to start fresh and explain their unique issues and preferences all over again in case they are not communicating with the same agent each time.

Enhanced Lead Management: You can identify new leads quickly and easily and categorize them based on multiple factors such as age, location or interest. Sales reps can prioritize leads based on closing likelihood, identify leads that need more nurturing and work towards improving their quality.

Increased Referrals: When you understand your customers better, upselling and cross-selling opportunities become clear and the chances of winning new business increase. Better services lead to happy customers who not only spend more but also refer your company to others.

Improved Analytical Reporting: CRM tools eliminate the chances of miscalculated data and miscommunications. With all information stored in one place, it becomes easier to analyze data and generate automatic reports. These help you make informed and effective decisions to increase credibility and long-term profitability.

Examples of CRMs:

**HubSpot** – best for high email send limits – free, but to upgrade for Sales, Marketing, and Service, it costs £16.80 per user, per month.

Pros – When it comes to users, contacts, and data storage, it’s effectively limitless

Cons – Extremely limited customer support (at least, until you become a paying customer)

**Freshworks** – Used to engage with customers – free, but to upgrades the prices differ from £11 to £25.

Pros – Unlimited users, contacts, and storage. The interface is a genuine pleasure to use and navigate.

Cons – Free plan is hampered its lack of daily bulk email sends.

**Capsule** – offers intelligent pipeline views and straightforward contact management features. - free for up to 2 people, however you get all sales, marketing and service for £12 per month, per user.

Pros – Integrates seamlessly with Google apps, as well as key software products such as Xero and MailChimp

Cons – No email marketing functionality included

**Bitrix24** – Business platform to structure your business – Free for up to 12 users pre Covid-19, now it isn’t.

Pros – Easy to set up and get to grips with as well as offering a striking visualization of your pipeline

Cons – Each email you send will have a Bitrix24 watermark at the bottom.

**Insightly** – Helping better track and manage leads and close deals faster – Free for up to 2 users.

Pros – 14-day free trail available for the Professional version to test it out.

Cons – Limited storage capacity. It supports only two users.

The Tesco Clubcard is Tesco’s way of rewarding its most loyal customers. For every pound you spend online or in-store, you’ll get a point. These points, when accumulated, can be traded in for Tesco Clubcard vouchers.

It allowed Tesco’s to collect more data and be able to market based on the data collected from the customers.

I believe this is an appropriate solution to use because it allows them to make more money and advertise more appropriately.

Apple has something called an Apple ID. If you’ve ever used iTunes, you’ll have an Apple ID – and if you’ve ever used an Apple device, you’ll have to register it using an Apple ID, too. These unique IDs synchronize across devices, remember music and film selections, and provide personalized recommendations based on what’s you've been watching or listening to.

This allowed Apple to collect more data on their customers and recommend specific offers for specific groups of people.

I believe this is an appropriate solution for Apple to use for their products, as it can make life easier for their customers by having every one of their products synced through the use of an individual ID.

McDonald’s has an app that allows their customers to see offers and items, as well as order food online. It takes geographical data and sales data from orders.

This allowed McDonalds to collect data from customers to see where the most popular places are for McDonalds. It allows them to notify the people using the app for offers and limited time products.

This is an appropriate solution for McDonalds to use for their products because it allows them to advertise more and freely to existing customers and make more money.

**Data Model:**

Conceptual data model

Logical data model

Physical data model

Hierarchical data model

Relational model

What is a data model?

Data modeling is the act of exploring data-oriented structures. Like other modeling artifacts data models can be used for a variety of purposes, from high-level conceptual models to physical data models.

From the point of view of an object-oriented developer data modeling is conceptually like class modeling. With data modeling you identify entity types whereas with class modeling you identify classes. Data attributes are assigned to entity types just as you would assign attributes and operations to classes. There are associations between entities, like the associations between classes – relationships, inheritance, composition, and aggregation are all applicable concepts in data modeling.

Traditional data modeling is different from class modeling because it focuses solely on data – class models allow you to explore both the behavior and data aspects of your domain, with a data model you can only explore data issues. Because of this focus data modelers tend to be much better at getting the data “right” than object modelers. However, some people will model database methods (stored procedures, stored functions, and triggers) when they are physical data modeling.

**Conceptual Data Model**

The conceptual data model is a structured business view of the data required to support business processes, record business events, and track related performance measures. This model focuses on identifying the data used in the business but not its processing flow or physical characteristics. This model’s perspective is independent of any underlying business applications. For example, it allows businesspeople to view sales data, expense data, customers, and products – business subjects that are in the integrated model and outside of the business model.

The conceptual data model represents the overall structure of data required to support the business requirements independent of any software or data storage structure. The characteristics of the conceptual data model include:

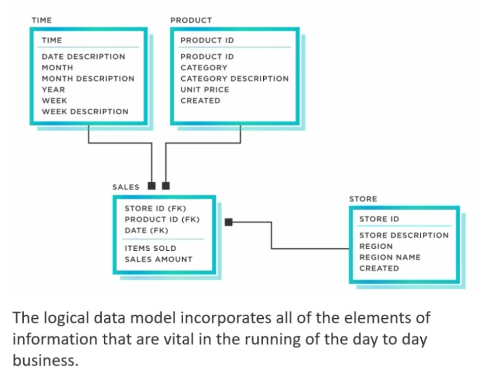
* An overall view of the structure of the data in a business context.
* Features that are independent of any database or physical storage structure.
* Objects that may not ever be implemented in physical databases. There are some concepts and processes that will not find their way into models, but they are needed for the business to understand and explain what is needed in the enterprise.
* Data needed to perform business processes or enterprise operations.

The conceptual data model is a tool for business and IT to define:

* Data requirements scope
* Business terms and measures across different business units and those that are agreed upon for enterprise-wide usage.
* Names, data types, and characteristics of entities and their attribute.

**Logical Data Model**

A logical data model establishes the structure of data elements and the relationships among them. It is independent of the physical database that details how the data will be implemented. The logical data model takes the elements of conceptual data modeling a step further by adding more information to them.



Each box in the picture is an entity.

Components of a Logical Data Model:

Entities – Each entity represents a set of things, people, or concepts relevant to a business.

Relationships – Every relationship represents an association between two of the above entities.

Attributes – Each attribute is a descriptive piece, characteristic or any other information that is useful to further describe an entity.

These are the most important characteristics of a logical data model:

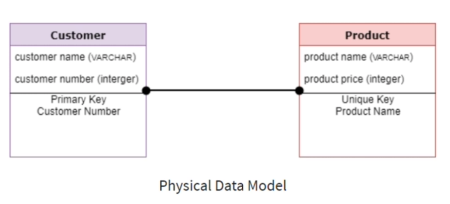
* A logical data model can describe the data needs for each individual project. Yet, it is designed to seamlessly integrate with other logical data models should the project demand it to do so.
* A logical data model can be developed and designed independently from the database management system. The type of database management system does not affect it that much.
* Data attributes contain data types with exact length and precision.
* In logical data modeling, no primary or secondary key is defined. At this level of modeling, it is required to verify and tweak connector details that were set prior to defining relationships.
* A logical data model is like a graphical representation of the information requirements of a business area. It is not a database or database management system itself.
* A logical data model is independent of any physical data storage device, such as a file system.
* A logical data model must be designed to be independent of technology, so as not to be affected by the rapid changes in technology.

Advantages:

* As data remains stable over time, a logical data model is also a stable one and highly conductive to data re-use and physical data sharing, which ultimately leads to reduced storage of redundant data.
* Components of a logical data model can be recycled, re-used, and adapted as more teams weigh in with their (often changing) needs.
* Costs associated with building and maintaining a logical data model are offset in the long run by the advantages it confers, not least by identifying and integrating all business needs and rules at the outset.
* User requests for making changes can be minimized by being proactive.
* Logical data models can be used for impact analysis, as each and every business process plus rule is connected within it.
* As objects in the logical data model bear textual definitions in business language, it makes it easier to maintain and access system documentation.

Physical Data Model

A physical data model describes a database-specific implementation of the data model. It offers database abstraction and helps generate the schema. This is because of the richness of metadata offered by a Physical data model. The physical data model also helps in visualizing database structure by replicating database column keys, constraints, indexes, triggers, and other relational database management system (RDBM) features.



The dots on both sides mean that there are many different sales.

If it was just a line, it would mean one sale.

Characteristics of a physical data model:

* The physical data model describes data need for a single project or application though it may be integrated with other physical data models based on project scope.
* Data Model contains relationships between tables that address cardinality and nullability of the relationships.
* Developed for a specific version of a database management system (DBMS), location, data storage or technology to be used in the project.
* Columns should have exact datatypes, lengths assigned and default values.
* Primary and foreign keys, views, indexes, access profiles, and authorizations, etc. Are defined.

**6.4**

**Data Management:**

Data is only useful if it’s high-quality

Big data is at best inconsequential. In the worst-case scenario, it can lead companies to make costly mistakes.

IBM estimates that bad data costs the U.S. economy $3.1 trillion per year. Those costs come from the time employees must spend correcting bad data and errors that cause mistakes with customers.

**Primary data** is raw data that hasn’t been altered.

**Secondary data** is data that has been altered and ready for use.

Highest role of a network is not admin, it’s the supervisor.

Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively. The goal of data management is to help people, organizations, and connected things optimize the use of data within the bounds of policy and regulation so that they can make decisions and take actions that maximize the benefit to the organization. A robust data management strategy is becoming more important than ever as organizations increasingly rely on intangible assets to create value.

The work of data management has a wide scope, covering factors such as how to:

•Create, access, and update data across a diverse data tier

•Store data across multiple clouds and on premises

•Provide high availability and disaster recovery

•Use data in a growing variety of apps, analytics, and algorithms

•Ensure data privacy and security

•Archive and destroy data in accordance with retention schedules and compliance requirements.

•A formal data management strategy addresses the activity of users and administrators, the capabilities of data management technologies, the demands of regulatory requirements, and the needs of the organization to obtain value from its data.

The six Vs of big data

Big data is a collection of data from a various source, often characterized by what’s become known as the 3Vs: Volume, Variety, Velocity. Over time, more Vs have been added.

Volume – The amount of data from myriad sources.

Variety – The types of data: structured, semi-structured, unstructured.

Velocity – The speed at which big data is generated.

Veracity – The degree to which big data can be trusted.

Value – The business value of the data collection.

Variability – The ways in which the big data can be used and formatted.

**Value**

•This V describes what value you can get from which data and how big data gets better results from stored data.  
   
For example, I enriched the database by postal code area for a Dutch retailer. Based on the specific customer information, the retailer decided which location for a new store would have the best connection with the target group. Enrichment allows you to make predictions. My customer also chose the layout of the store and the offer to suit the specific wishes of (potential) shoppers.  
   
Also, a good way to value your big data is to work with people. They give a name and face to different customer groups and are a very powerful way of making organizations more customer-oriented. Personas were devised because there was a need to profile the many website visitors, thus increasing the user-friendliness of these sites.  
   
You can create personas based on available customer behavior data. For the Van Gogh Museum, for example, personas have been created to bring the different visitor types to life.

**Volume**

•Volume is an obvious feature of big data and is mainly about the relationship between size and processing capacity. This aspect changes rapidly as data collection continues to increase. Just like the IT capacity for storage and processing.  
   
Walmart, a company with an incredible amount of data, is building the largest private cloud in the world to handle large amounts of data per hour. With the Data Café program, they model, manipulate and visualize this information to gain insight into their shoppers. A practical example: during Halloween, sales analysts could see that, although a special new cookie was very popular in most stores, there were two stores where it was not selling at all. This was quickly picked up and it turned out that the cookies were accidentally not placed on the shelves. It was resolved immediately.  
   
When you talk about big data people often only think of volume, but there are also the five other Vs that can help you make data valuable: These Vs are also important in enriching smaller databases.  
   
In addition, with data volume it can also be "high-dimensional": you can ask big questions about small data.

**Velocity**

•Speed of data generation and frequency of delivery.

•The data flow is massive and continuous which is valuable to researchers as well as business for decision making for strategic competitive advantages and ROI.

•For processing of data with high velocity tools for data processing known as Streaming analytics were introduced.

The velocity adds value to the organization.

**Variety**

The V of variety describes the wide variety of data that is being stored and still needs to be processed and analyzed. New types of data from social networks and mobile devices, among others, complement existing types of structured information. For example: audio and video files, photos, GPS data, medical files, instrument measurements, graphics, web documents, bonus cards and internet search behavior. Unstructured data such as voice and social media make processing and categorizing data extra complicated. How do you ensure you are only taking the data that helps target your audience?

Predict donorship  
 An example: a charity has a database of households. These include features such as car ownership, value under the Valuation of Immovable Property Act (WOZ) and whether people are donors or not. I linked this data to the Mentality segmentation tool. I then searched in that database for the features that your donor company can predict. So, I calculated which households had a high chance of becoming a donor and the charity undertook targeted fundraising actions.

Predicting political views.

**Veracity**

Veracity shows the quality and origin of data, allows it to be considered questionable, conflicting or impure, and provides information about matters you are not sure how to deal with. In short: the truth and authenticity of the data, and what can you do with it? In a sense, it is a hygiene factor. By showing the veracity of your data, you show that you have taken a critical look at it.

Everything belonging to a company's core process is reliable, the rest is contaminated. You must take this pollution into account. You must be convinced that the data you have selected will also work properly and will be sufficient. It is a lot of monotonous but necessary work.

**Variability**

Finally, variability: to what extent, and how fast, is the structure of your data changing? And how often does the meaning or shape of your data change?

For example, take the newspaper subscription benefit: an internet subscription costs £50, a paper subscription £100 subscription, and a paper and internet subscription £100. One option is illogical. If you offer these options to people, most people choose a paper and internet subscription, which seems more advantageous. But if you take away the illogical choice: an internet subscription for £50 or a paper and internet subscription for £100, then many people will choose the internet subscription.  
   
In this way, the composition of a questionnaire or, for example, unsubscribe buttons changes how things appear to people and thus the outcome. In purely technical terms this means: if you change variables, your model will also change.

**Data assurance**

Data assurance checks for and corrects errors that might occur as data is communicated between a host and a storage array. DA capabilities are presented at the pool and volume group level in System Manager. The data assurance feature increases data integrity across the entire storage system.

How does data assurance increase confidence in data?

Organizations may not be willing to share their data if they are concerned about data quality, how it will be used, its security or what will happen to it after use. Where an organization wants to limit data use and reuse to a defined purpose, data assurance practices (for example, clear documentation and licensing) can help allay fears of the data being misused, misunderstood or mishandled.

Organizations collecting and reusing data often must spend a lot of time and effort verifying the quality and completeness of data before it can be used or shared onwards. The same organization may be concerned about verifying sources of the data and establishing rights and consent for that data to be used in new ways or shared. Data assurance practices and tools can be used to assess and demonstrate trust in data sources, in various ways, for example checking against standards or verifying provenance.

Many organizations collect, use and share different sets of data at the same time and in different ways. At each step in its lifecycle, data must be tested and verified according to what quality is expected and how it will be used. Data assurance tools and practices can help to provide confidence in that data being fit for purpose, and trustworthy, throughout its lifecycle.

**Data validation**

•Data validation is an essential part of any data handling task whether you’re in the field collecting information, analyzing data, or preparing to present data to stakeholders. If the data isn’t accurate from the start, your results won’t be accurate either. That’s why it’s necessary to verify and validate data before it is used.

•While data validation is a critical step in any data workflow, it’s often skipped over. It may seem as if data validation is a step that slows down your pace of work, however, it is essential because it will help you create the best results possible. These days data validation can be a much quicker process than you might’ve thought. With data integration platforms that can incorporate and automate validation processes, validation can be treated as an essential ingredient to your workflow rather than an additional step.

**Why Validate?**

•Validating the accuracy, clarity, and details of data is necessary to mitigate any project defects. Without validating data, you run the risk of basing decisions on data with imperfections that are not accurately representative of the situation at hand.

•While verifying data inputs and values is important, it is also necessary to validate the data model itself. If the data model is not structured or built correctly, you will run into issues when trying to use data files in various applications and software.

•Both the structure and content of data files will dictate what exactly you can do with data. Using validation rules to cleanse data before use helps to mitigate “garbage in = garbage out” scenarios. Ensuring the integrity of data helps to ensure the legitimacy of your conclusions.

**Data Verification**

Data verification is the process of checking data for accuracy after a data migration. There are different types of verification.

Full verification, where all the data is checked.

Sampling verification, where a small sample of the data is checked

Data verification can be both expensive and time-consuming to carry out.

When data is migrated from a data warehouse for use in a big data processing system, the data needs to be checked to ensure that it is accurate. Everything from spelling errors to inaccurate numbers to data loss could jeopardize a big data project.

One method of verifying the data is comparing data in one system to the migrated data in the other one-to-one, but this can be time-consuming and the costs of running two systems can be expensive.

It is also possible to check just a subset of the data, but a sample cannot possibly represent all the data. Administrators must weigh the tradeoff between keeping the time and expense of data verification down while ensuring accuracy. Automating the process is one solution.

**Data Reliability**

Data-driven decisions” has become quite the buzzword, but don’t let that distract from the fact that making business decisions based on accurate data is vitally important. Over the last decade, businesses have gained unprecedented access to quantitative and qualitative data about their operations, customers, and prospects. With all this information at our fingertips, it’s become best practice for every team, from marketing to customer success, to make data-driven decisions. Why trust your gut or talk about hypotheticals if you can use hard facts and figures to guide your next move?

In some ways, access to this information has helped level the playing field among small, medium, and more mature businesses. The data is out there if you’re able to tap into it and know how to use it. But this also presents a problem.

You may be collecting and storing a sizable amount of information within your data stack, but are you certain that these data sets — individually and when combined for analysis — are always complete, accurate, and up to date? If your answer is not a confident “yes,” then your business can’t reliably make data-driven decisions.

Data reliability is the foundation for confident decision-making and a successful data team.

**Data Redundancy**

Data redundancy occurs when the same piece of data is stored in two or more separate places and is a common occurrence in many businesses. As more companies are moving away from soiled data to using a central repository to store information, they are finding that their database is filled with inconsistent duplicates of the same entry. Although it can be challenging to reconcile — or even benefit from — duplicate data entries, understanding how to reduce and track data redundancy efficiently can help mitigate long-term inconsistency issues for your business.

How does data redundancy occur?

Sometimes data redundancy happens by accident while other times it is intentional. Accidental data redundancy can be the result of a complex process or inefficient coding while intentional data redundancy can be used to protect data and ensure consistency — simply by leveraging the multiple occurrences of data for disaster recovery and quality checks.

Types of big data:

* Structured
* Unstructured
* Semi-structured

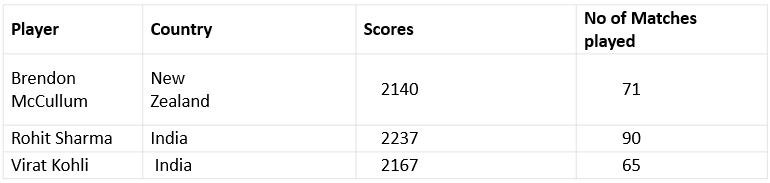
**Structured Data**

*My definition of structured data is that it is ready to use data that’s already stored in databases in an ordered manner.*

Structured Data is used to refer to the data, which is already stored in databases, in an ordered manner. It accounts for about 20% of the total existing data and is used the most in programming and computer-related activities.

There are two sources of structured data – machines and humans. All the data received from sensors, weblogs, and financial systems are classified under machine-generated data. These include medical devices, GPS data, data of usage statistics captured by servers and applications and the huge amount of data that usually move through trading platforms, to name a few.

Human-generated structured data mainly includes all the data a human input into a computer, such as his name and other personal details. When a person clicks a link on the internet, or even makes a move in a game, data is created- this can be used by companies to figure out their customer behavior and make the appropriate decisions and modifications.



**Unstructured Data**

*My definition of unstructured data is that it’s primary data/raw data that is hard to understand, has no patterns and has no clear format.*

While structured data resides in the traditional row-column databases, unstructured data is the opposite- they have no clear format in storage. The rest of the data created, about 80% of the total account for unstructured big data. Most of the data a person encounters belongs to this category- and until recently, there was not much to do to it except storing it or analyzing it manually.

Unstructured data is also classified based on its source, into machine-generated or human-generated. Machine-generated data accounts for all the satellite images, the scientific data from various experiments and radar data captured by various facets of technology.

Human-generated unstructured data is found in abundance across the internet since it includes social media data, mobile data, and website content. This means that the pictures we upload to Facebook or Instagram handle, the videos we watch on YouTube and even the text messages we send all contribute to the gigantic heap that is unstructured data.

Examples of unstructured data include text, video, audio, mobile activity, social media activity, satellite imagery, surveillance imagery – the list goes on and on.

The unstructured data is further divided into:

Capture – It is the data based on the user’s behavior. The best example to understand it is GPS via smartphones which help the user each and every moment and provides a real-time output.

User-Generated data – It is the kind of unstructured data where the user itself will put data on the internet every movement. For example, Tweets and Re-tweets, Likes, Shares, Comments, on YouTube, Facebook, etc.

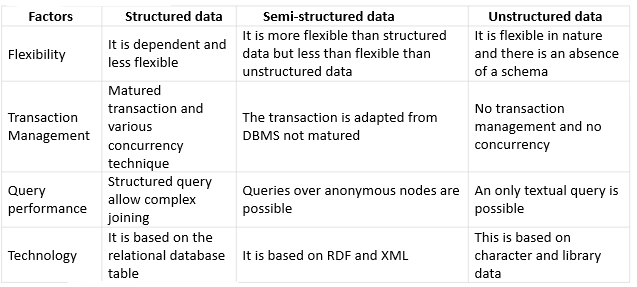
**Semi-structured data:**

*My definition of semi-structured data is that it is data that at first appears to be unstructured, but it contains some properties that make it easier to understand.*

The line between unstructured data and semi-structured data has always been unclear since most of the semi-structured data appear to be unstructured at a glance. Information that is not in the traditional database format as structured data but contains some organizational properties which make it easier to process, is included in semi-structured data. For example, NoSQL documents are considered to be semi-structured, since they contain keywords that can be used to process the document easily.

Big Data analysis has been found to have definite business value, as its analysis and processing can help a company achieve cost reductions and dramatic growth. So, it is imperative that you do not wait too long to exploit the potential of this excellent business opportunity.

**Difference between Structured, Semi-structured and Unstructured data**



Research Population:

A research population is generally a large collection of individuals or objects that is the focus of a scientific query. It is for the benefit of the population that research is done. However, due to large sizes of populations, researchers often cannot test every individual in the population because it is too expensive and time-consuming. This is the reason why researchers rely on sampling techniques.

A research population is also known as a well-defined collection of individuals or objects known to have similar characteristics. All individuals or objects within a certain population usually have a common, binding characteristic or trait.

Usually, the description of the population and the common binding characteristic of its members are the same. "Government officials" are a well-defined group of individuals which can be considered as a population and all the members of this population are indeed officials of the government.

Two types of population in research:

* Target population – refers to the ENTIRE group of individuals pr objects to which researchers are interested in generalizing the conclusions. The target population usually has varying characteristics and it is also known as the theoretical population.
* Accessible population – The accessible population is the population in research to which the researchers can apply their conclusions. This population is a subset of the target population and is also known as the study population. It is from the accessible population that researchers draw their samples.

Quantitative data:

* Countable or measurable, relating to numbers
* Tells us how many, how much, or how often
* Fixed and universal, “factual”
* Gathered by measuring and counting things
* Analyzed using statistical analysis

Qualitative data:

* Descriptive, relating to words and language
* Describes certain attributes, and helps us to understand the “why” or “how behind certain behaviors
* Dynamic and subjective, open to interpretation.
* Gathered through observations and interviews
* Analyzed by grouping the data into meaningful themes or categories

**What is Quantitative Data?**

Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it’s quantitative data. Quantitative data can tell you “How many,” “how much,” or “how often”—for example, how many people attended last week’s webinar? How much revenue did the company make in 2019? How often does a certain customer group use online banking?

To analyze and make sense of quantitative data, you’ll conduct statistical analyses.

**What is Qualitative Data?**

Unlike quantitative data, qualitative data cannot be measured or counted. It’s descriptive, expressed in terms of language rather than numerical values.

Researchers will often turn to qualitative data to answer “Why?” or “How?” questions. For example, if your quantitative data tells you that a certain website visitor abandoned their shopping cart three times in one week, you’d probably want to investigate why—and this might involve collecting some form of qualitative data from the user. Perhaps you want to know how a user feels about a particular product; again, qualitative data can provide such insights. In this case, you’re not just looking at numbers; you’re asking the user to tell you, using language, why they did something or how they feel.

Qualitative data also refers to the words or labels used to describe certain characteristics or traits—for example, describing the sky as blue or labeling a particular ice cream flavor as vanilla.

The main difference between the two types of data lies in what they tell us, how they are collected, and how they are analyzed.

**Legislation and regulatory compliance:**

* Regulatory compliance is when a business follows state, federal, and international laws and regulations relevant to its operations. The specific requirements can vary, depending largely on the industry and type of business.
* Some examples of regulatory compliance regulations include The U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA), the Sarbanes-Oxley Act of 2002, and the European Union’s General Data Protection Regulation of 2016 (GDPR).
* Regulatory compliance (adhering to government laws) differs from other aspects of corporate compliance (such as following internal policies and rules).
* While both are important to ensure integrity, safety, and ethical behavior in businesses, it helps to understand the difference.
* Regulatory compliance involves following external legal mandates set forth by state, federal, or international government. In contrast, following company policies and procedures involves following internal requirements set forth by the business. Both, however, help drive accountability in the workplace.
* To define regulatory compliance, you must have some idea about compliance plan.
* Semantically, compliance means conforming to a rule, such as a spec, policy, pattern or law. Regulatory compliance outlines the goals that organizations want to achieve in their efforts to ensure that they are aware of and take steps to comply with relevant laws, policies, and regulations.
* The number of rules and the need for operational clarity is rising. So, organizations are willingly adopting the use of consolidated and tuned sets of compliance controls. They use this approach to assure that all necessary governance requirements without the useless replication of effort and activity from resources.

**Functions of regulatory compliance:**

* Identification – Identifies possible risks that an organization may face - Addresses the jeopardies
* Prevention – Advises on how to prevent the risk management factors – implements control to protect the organization from those pitfalls
* Monitoring and detection: Observes and reports on the effectiveness of controls in the management of the organization’s risk vulnerability.
* Resolution – Determines compliance issues as they enter into the picture – Maintains the company’s input and output qualities – actively try to lessen financial wrongdoings.
* Advisory – advocate business on rules and regulation – commend what to do to prevent future loss.

The United Kingdom has noteworthy regulations. Many of these come from the European Union laws. Institutions like the Financial Conduct Authority (FCA), Environment Agency, Scottish Environment Protection Agency, Information Commissioner’s Office, Care Quality Commission, etc. also control some regulatory compliance domains.

The Data Protection Act 1998 issues necessary compliance for all large and small firms. The Freedom of Information Act 2000 observes the public sector.

The International Organization for Standardization, known as ISO and it’s ISO 19600 standard is one of the first international standards.

The roles of ISO:

* Manipulates how businesses handle regulatory compliance
* Gives a reminder of how compliance and risk should operate mutually
* Emphasizes companies to share a common framework with some subtleties to account for their variations
* Produces international standards like ISO/IEC 27002 to support organizations meet regulatory compliance with their security management and assurance best practices.
* Some local or international specialized organizations, for instance, the American Society of Mechanical Engineers (ASME) also promote standards and regulation policies. They provide a wide range of rules and instructions. Thus, they ensure compliance of the products to safety, security or design standards.

**Big Data Ethics:**

While having tons of customer information may be incredibly advantageous for digital marketers and business owners, there is a growing concern regarding the ethical use of big data for marketing and advertising purposes. Are you using big data responsibly or are you abusing the power that is in your hands? If you’re not sure of your answer, it’s time to ask yourself these questions.

* How do you decide which information will help you achieve your marketing goals, and how do you know when you have collected enough of it?
* How do you draw the line between the ethical and unethical use of big data and how do you make sure you don’t cross it?
* Have you instituted measures to protect customer information and privacy?

**Organizational Factors:**

All organizations must consider the following factors heading into Data Management for Big Data. Big Data requires Time, Skills and it's also expensive to create and evaluate.

Time:

* Data mining, organizing and analyzing is time consuming effort. It takes TEAMS of data wranglers to manage to structure unstructured data or to update them for the purpose of sales, research or analytics.

Skill:

* There are very specific skills required by the organization to make sense of the data. Most data scientists or data wranglers are qualified with a degree in the field or some form of formal training.
* Data scientists need to train in the organization policy and also, they need to adhere to the laws of the country they are mining data for.

Cost:

* This includes the cost of employees in the organization
* Cost of buying the data
* Cost of equipment's etc.

Big Data relies on the small parts working correctly inside an organization. The more efficient an organization is, it will produce reliable and verifiable data.

**Data Warehouse:**

•A data warehouse, or enterprise data warehouse (EDW), is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, artificial intelligence (AI), and machine learning. A data warehouse system enables an organization to run powerful analytics on huge volumes (petabytes and petabytes) of historical data in ways that a standard database cannot.

•Data warehousing systems have been a part of business intelligence (BI) solutions for over three decades, but they have evolved recently with the emergence of new data types and data hosting methods. Traditionally, a data warehouse was hosted on-premises—often on a mainframe computer—and its functionality was focused on extracting data from other sources, cleansing and preparing the data, and loading and maintaining the data in a relational database. More recently, a data warehouse might be hosted on a dedicated appliance or in the cloud, and most data warehouses have added analytics capabilities and data visualization and presentation tools.

**Data Lakes**

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

Why do you need a data lake?

Organizations that successfully generate business value from their data, will outperform their peers. An Aberdeen survey saw organizations who implemented a Data Lake outperforming similar companies by 9% in organic revenue growth. These leaders were able to do new types of analytics like machine learning over new sources like log files, data from click-streams, social media, and internet connected devices stored in the data lake. This helped them to identify, and act upon opportunities for business growth faster by attracting and retaining customers, boosting productivity, proactively maintaining devices, and making informed decisions.

Summarized – A storage of data that can be organized and unorganized at any scale. You can run different analytics and visualize the data processing.

**Data Mining**

Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets. Given the evolution of data warehousing technology and the growth of big data, adoption of data mining techniques has rapidly accelerated over the last couple of decades, assisting companies by transforming their raw data into useful knowledge. However, even though technology continuously evolves to handle data at a large-scale, leaders still face challenges with scalability and automation.

Data mining has improved organizational decision-making through insightful data analyses. The data mining techniques that underpin these analyses can be divided into two main purposes; they can either describe the target dataset or they can predict outcomes through the use of machine learning algorithms. These methods are used to organize and filter data, surfacing the most interesting information, from fraud detection to user behaviors, bottlenecks, and even security breaches.

When combined with data analytics and visualization tools, like Apache Spark, delving into the world of data mining has never been easier and extracting relevant insights has never been faster. Advances within artificial intelligence only continue to expedite adoption across industries.

When you look through a large amount of data and look for patterns of data that can be useful.

**Data Report**

A data report is an analytical tool used to display past, present, and future data to efficiently track and optimize the performance of a company. It combines various sources of information and is usually used both on an operational or strategic level of decision-making.

As mentioned, these reports had features of static presentation of data, manually written or calculated, but with the introduction of modern processes such as dashboard reporting, they have developed into an invaluable resource to successfully manage your sales processes, marketing data, even robust manufacturing analytics and numerous other business processes needed to stay on top of the pack.

Top 5 data analytical tools:

* KNIME – is an open source data analytics platform that supports data integration, processing, visualization, and reporting. It plugs in machine learning and data mining libraries with minimal or no programming requirements.
* Microsoft Power BI - It allows users to create and share reports, visualizations, and dashboards. Users can combine a group of dashboards and reports into a Power BI app for simple distribution. Power BI also allows users to build automated machine learning models
* Sisense - is a data analytics platform aimed at helping both technical developers and business analysts process and visualize all their business data.
* Tableau - is a data visualization and analytics platform that allows users to create reports and share them across desktop and mobile platforms, within a browser, or embedded in an application. It can run on the clouds or on-premises
* Qlik - provides a self-service data analytics and business intelligence platform that supports both cloud and on-premises deployment. The tool boasts strong support for data exploration and discovery by technical and nontechnical users alike.

**API**

An application programming interface is a computing interface to a software component or a system, that defines how other components or systems can use it.

It defines the kinds of calls or requests that can be made, how to make them, the data formats that should be used, the conversation to follow, etc.

**Meta Data**

Data that can’t be seen but is always there. Very specific details

•Every time you create, open or save a Microsoft Word document, hidden information is created and stored within the document that you may not want others outside of your organization to discover. Hidden information can also reside in other Microsoft application files, such as Excel spreadsheets or PowerPoint presentations and includes:

•

•Your Name and Initials, Company Name, Computer Name, Location of Document on Local or Network Server, Attached Template, Hidden Text, Comments, Track Changes, Non-Visible Portions of OLE Objects, File Properties and Summary Information, And more …

**Data at Rest:**

•Data at rest is one of the three states of digital data and it refers to any digital information that is stationary and contained within permanent storage devices, such as hard drives and tapes, or information reservoirs such as off-site backups, databases, archives, etc. The other states of digital data are data in motion, and data in use. Once data is transported and settles in its destination, it is called data at rest during the entire period it remains inactive. If the data needs to be utilized for whatever purpose, and is being processed, it is then classified as data in use.

Data that is stored but isn’t used.

**Data in Motion:**

Data in motion is a term used to label any digital information that is being transferred from one location to another. It is also commonly referred to as data in transit or data in flight. When the data is finally contained in one location, it becomes data at rest.

For data in motion, the information can be moved to various locations within the same computer such as copying files from one app to another, or downloading from a web browser to a local app. It can also be transferred between different computers through cloud services such as email or physically across short or long distances via portable storage devices like USB flash drives.

Data protection is a central aspect to an organization's welfare and performance. All data generated by or entrusted to an organization should be handled carefully, and this rule especially applies to data in motion security. If data in motion is ever exposed to malicious actors, the ramifications can be potentially fatal to business continuity as well as reputation, which can result in significant monetary loss.

Data in motion can be broadly classified into two categories. The first is virtual information transferred within the boundaries of a private network. This information, to an extent, is protected by firewalls and other internally established data protection measures. The second category is information being transported outside of the organization. Data in motion is the most vulnerable in the instances where it is transferred out an organization or private network as it's sometimes processed across unreliable networks like the internet or using peripheral devices that, if handled improperly, can become exposed to unauthorized viewers.

**Data in Use:**

Data in use refers to data that is not simply being passively stored in a stable destination, such as a central data warehouse, but is working its way through other parts of an IT architecture. Data in use may be in the process of being generated, amended or updated, erased, or viewed through various interface endpoints. This is a helpful term for pursuing comprehensive security for IT systems.

The idea behind protecting data in use is that data sets are vulnerable to different kinds of threats depending on where they are in an IT system. One of the most common and basic issues with data in use revolves around endpoints. Endpoints are points where data from a system gets routed to an individual device or workstation by or for an end user.

Obviously, advanced IT systems and networks need to anticipate a variety of different endpoints. Many of the issues around endpoint security or data-in-use security relate to the trend toward BYOD, where employees may be using personal devices to view corporate data. Even for mobile devices or other hardware systems that may be company-owned, companies have to look at how end users can view or capture data from secure locations.

Although professionals sometimes recommend techniques like full disk encryption or a comprehensive data leak prevention plan, others question whether total data-in-use security is ever really possible. This is partly because of the inherent setup of most operating systems, but another big issue is that even in the most secure systems, data has to be sent to endpoint displays in order to be useful. With that in mind, there is no single way to guarantee complete data safety, which is why companies tend to focus on ironclad agreements with employees and other end users rather than relying only on data-in-use protection systems.

SITES:

https://www.vertica.com/docs/9.3.x/HTML/Content/Authoring/SQLReferenceManual/DataTypes/CharacterDataTypes.html