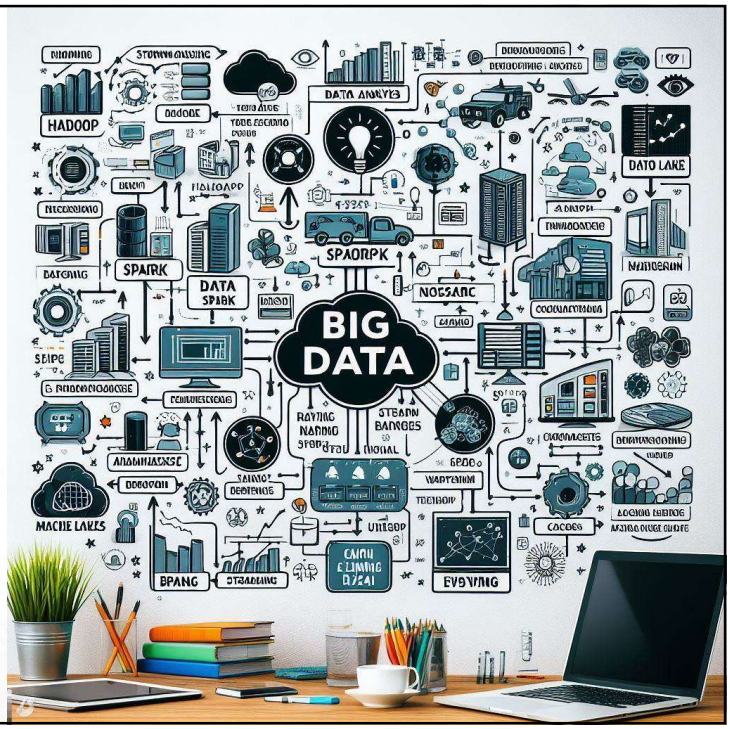
Understanding Big Data



Chapter 01

Instructor: Houssein Dhayne
houssein.dhayne@net.usj.edu.lb

Table of Contents

Introduction to Big Data

- **Definition:** Big Data means dealing with really large amounts of information from many different places.
- **Why It's Needed:** We use Big Data when our usual ways of handling information are not enough. It helps us when we need to put together different types of data or work with a lot of messy, unorganized information.

Evolution of Big Data Science

- **Historical Context:** Big Data isn't new; it has evolved from historical challenges in managing and analyzing large datasets.
- **Roots:** It grew from old problems of handling lots of information, like counting people in a census or figuring out insurance calculations.
- **Development:** Over the years, we've used more advanced technology to get even better at working with big sets of data.

Shift in Analytic Approaches

- **Statistical Foundation:** Traditional analytics relied on statistics for approximating population measures through sampling.
- **Computational Shift:** Big Data introduces computational approaches, enabling the processing of entire datasets.
- **Sampling Unnecessary:** Advances in computational science eliminate the need for sampling even in massive datasets.

Interdisciplinary Nature

- **Skills Needed:** Working with Big Data involves using math, stats, computer skills, and knowing about the subject you're studying.
- **Different Views:** People might see Big Data differently based on their expertise.
- **Always Changing:** Because technology keeps improving, what we consider a Big Data problem can change.

Big Data Today

- **Tech Evolution:** Big Data has changed as our computers and software get better.
- **From Big to Normal:** What was once a huge amount of data (like one gigabyte) is now something we handle every day.
- **Where Data Comes From:** Big Data usually comes from apps, sensors, and other sources all collecting information.

Applications and Benefits

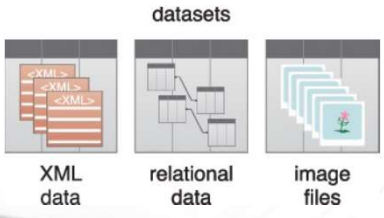
- **Insights and Benefits:** It's useful in many areas, like: *operational optimization, actionable intelligence, new market identification, accurate predictions, fault and fraud detection, detailed records, improved decision-making, and scientific discoveries.*
- **Considerations:** Despite the benefits, adopting Big Data analytics requires careful consideration of associated issues, which will be discussed in Part II.

Concepts and Terminology

Datasets

Collections of related data with shared attributes.

- tweets stored in a flat file
- a collection of image files in a directory
- an extract of rows from a database table stored in a CSV formatted file
- historical weather observations that are stored as XML files.



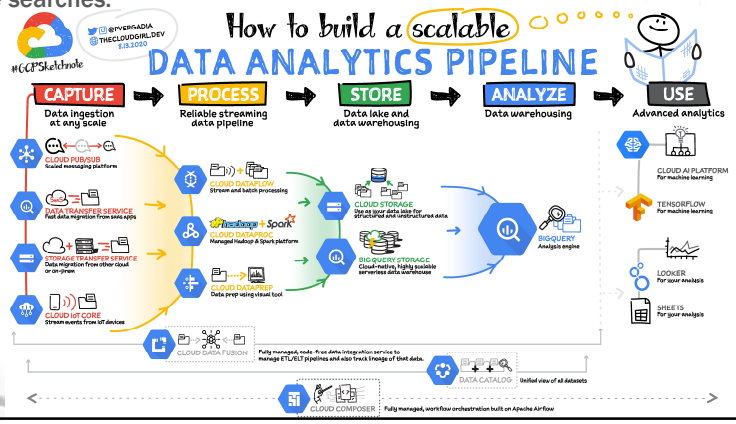
Data Analysis and Data Analytics Symbols

- **Data Analysis**
- Process: Examining data to find facts, patterns, and insights.
  - The overall goal of data analysis is to support better decision-making.

- **Data Analytics**
- Discipline: Manages the complete data lifecycle, including collection, analysis, and governance.

Big Data Analytics Lifecycle

- Lifecycle Stages
  - Identifying, procuring, preparing, and analyzing large amounts of raw, unstructured data.
  - Goal: Extract meaningful information for identifying patterns, enriching enterprise data, and large-scale searches.

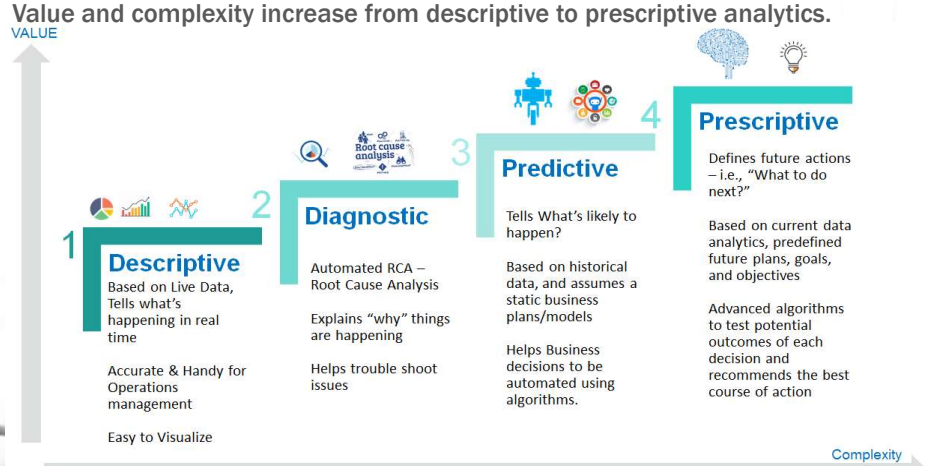


Examples

- **Business Sector:**
  - A retail company aims to enhance its customer experience by analyzing **unstructured data from social media, customer reviews, and sales data**. Through data preparation and analysis, they extract meaningful information **about customer preferences, sentiment, and purchasing patterns.**
  - **Outcome:** The company gains insights **into popular products, customer satisfaction drivers, and trends.** This information aids in optimizing inventory, tailoring marketing strategies, and improving overall operational efficiency,
- **Public Sector:**
  - A city government aims to enhance public safety by **analyzing unstructured data from various sources, including crime reports, traffic data, and social media**. They identify and procure relevant data, prepare and analyze it **to identify patterns and trends related to crime and traffic incidents**.
  - **Outcome:** The city government **gains insights into high-crime areas, traffic congestion patterns, and areas with a high risk of accidents**. This information strengthens their focus on service delivery, allowing them to allocate resources more effectively, improve emergency response times, and enhance overall public safety.

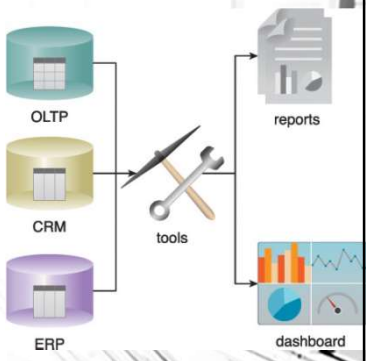
Types of Analytics

- Analytic Categories
  - Descriptive, Diagnostic, Predictive, and Prescriptive Analytics.
- Value and complexity increase from descriptive to prescriptive analytics.

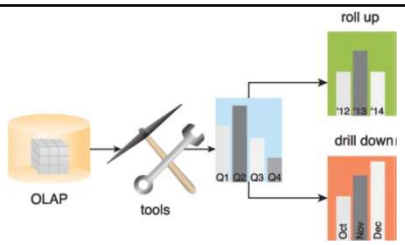


Descriptive Analytics

- Characteristics
  - Answers questions about past events, contextualizing data to generate information.
- Sample Questions
  - Sales Volume: What was the sales volume over the past 12 months?
  - Monthly Commission: What is the monthly commission earned by each sales agent?
  - Population Distribution: What is the historical trend in population distribution for a city over the last 20 years?
- Value and Skillset
  - Provides the least worth and requires a relatively basic skillset.
    - Data Collection , Basic Statistical Knowledge, Data Visualization, Database Management.

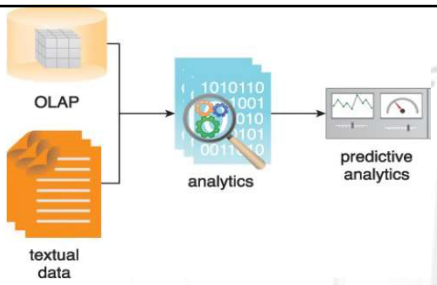


Diagnostic Analytics



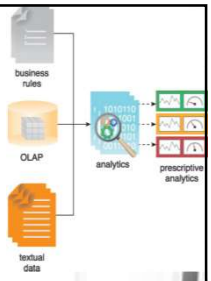
- Purpose and Questions
  - Determines the cause of past events, focusing on the reason behind the phenomenon.
- Sample Questions
  - Sales Comparison: Why were Q2 sales less than Q1 sales?
  - Support Calls: Why more calls from the Eastern region than the Western region?
  - Patient Re-Admission: Why an increase in patient re-admission rates?
- Value and Skillset
  - Provides more value than descriptive analytics but requires a more advanced skillset.
    - Advanced Statistical Analysis, Data Mining, Database Querying, Critical Thinking, Domain Knowledge, Programming Skills, Data Visualization

Predictive Analytics



- Purpose and Questions
  - Attempts to determine the outcome of a future event.
- Sample Questions
  - Loan Default: Chances of a customer defaulting on a loan after missing a payment?
  - Patient Survival: Patient survival rate if Drug B is administered instead of Drug A?
  - Product Purchases: If a customer buys Products A and B, chances of buying Product C?
- Data and Techniques
  - Involves large datasets, internal and external data, and various analysis techniques.
  - It provides greater value and requires a more advanced skillset than both descriptive and diagnostic analytics.
    - Machine Learning, Statistical Modeling, Predictive Modeling, Algorithm Development, Pattern Recognition, Data Mining, Domain Knowledge.

Prescriptive Analytics

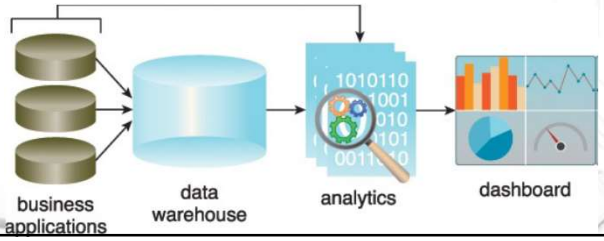


- Building on Predictive Analytics
  - Prescribes actions to be taken based on predictive analytics results.
- Sample Questions
  - Drug Comparison: What actions should be taken to maximize the effectiveness and optimize outcomes among the three drugs?
  - Stock Trading: Best time to trade a particular stock?

- Value and Skillset
  - Provides the most value but requires the most advanced skillset and specialized tools.
    - Decision Science, Optimization Techniques, Simulation Modeling, Risk Management, Business Strategy, Advanced Quantitative Analysis, Domain-Specific Expertise, Data Interpretation.
- Example
  - With prescriptive analytics, the e-commerce company can develop strategies to maximize sales based on the predictions. For example, if the predictive model indicates a surge in demand for a specific product during a certain season, **prescriptive analytics might recommend increasing inventory levels for that product, optimizing pricing strategies** to gain advantage from the predicted trend.

Business Intelligence (BI)

- Insight into Enterprise Performance
  - Analyzes data from business processes and information systems.
  - Used by management for directing business decisions and enhancing performance.
- Components
  - BI applies analytics to large amounts of data, often **consolidated into an enterprise data warehouse.**
  - Output is surfaced to **a dashboard** for result analysis and refinement of queries.

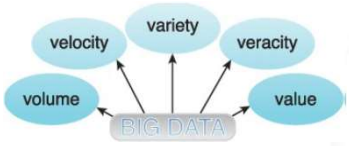


Key Performance Indicators (KPI)

- Metric for Business Success
  - Measure success within the business context.
  - Linked to strategic goals for direction.
- Example
  - In a hospital, a Key Performance Indicator (KPI) is like a goal that measures the success of treating emergency patients promptly.
    - If the hospital achieves the target (e.g., 90% treated within 30 minutes), it's doing well.
    - Falling below the target (e.g., 85%) signals potential issues, prompting the need for improvements.

Big Data Characteristics

Big Data Characteristics

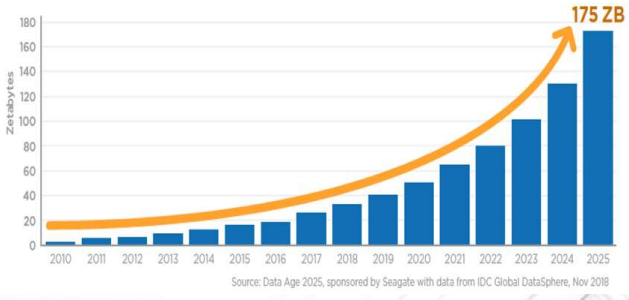


- Differentiating Big Data
  - Five characteristics (Volume, Velocity, Variety, Veracity, Value) impacting solution design.
- The Five Vs
  - Volume: Substantial and ever-growing data.
  - Velocity: Data arrives at fast speeds, requires elastic and available processing solutions.
  - Variety: Multiple formats and types of data, challenging for integration and processing.
  - Veracity: Quality or fidelity of data, distinguishing signal from noise.
  - Value: Usefulness of data for an enterprise, impacted by fidelity and processing time.

Volume

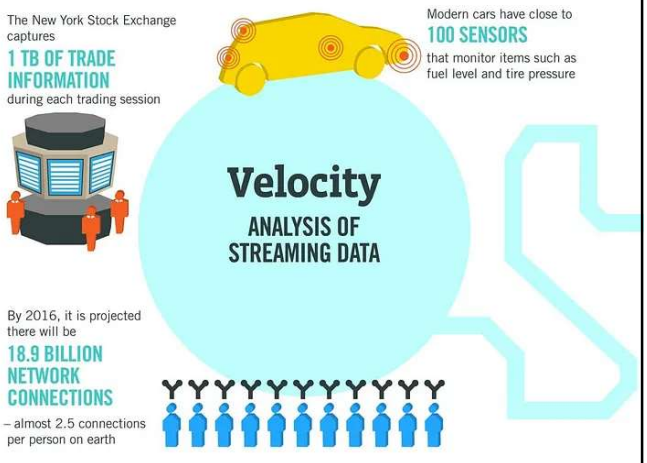
- Challenge: Processing and storing substantial and ever-growing data.
- Examples: Online transactions, scientific experiments, sensor data, social media.
- Visualization: Organizations and users world-wide create over 2.5 EBs of data a day. As a point of comparison, the Library of Congress currently holds more than 300 TBs.

kilobytes (KB),
megabytes (MB),
gigabytes (GB),
terabytes (TB),
petabytes (PB),
exabytes (EB),
zettabytes (ZB),
yottabytes (YB).



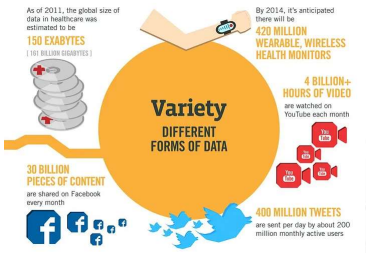
Velocity

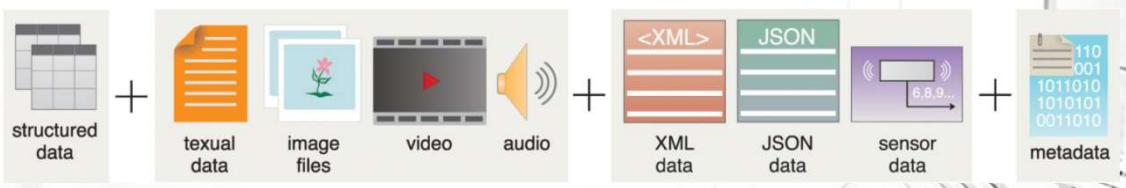
- Definition: Data arriving at high speeds, requiring elastic and available processing solutions.
- Examples: Daily generation of tweets, video uploads, emails, and sensor data.
- Enterprise Impact: Influences data processing solution design.



Variety

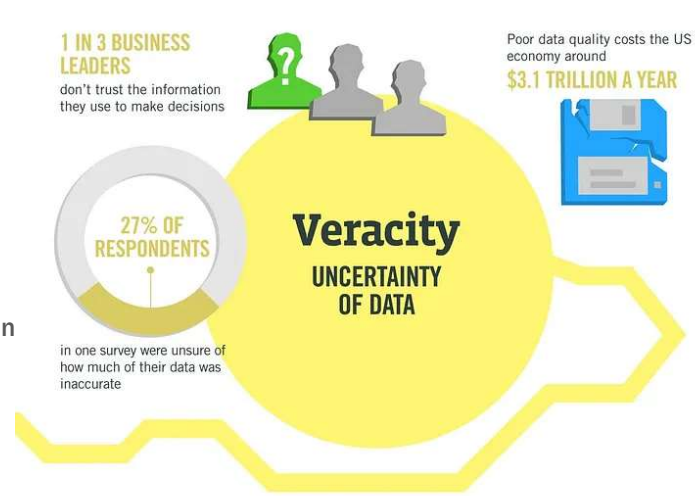
- Definition: Multiple formats and types of data requiring support.
- Challenges: Integration, transformation, processing, and storage.
- Examples: Structured, textual, image, video, audio, XML, JSON, sensor data.





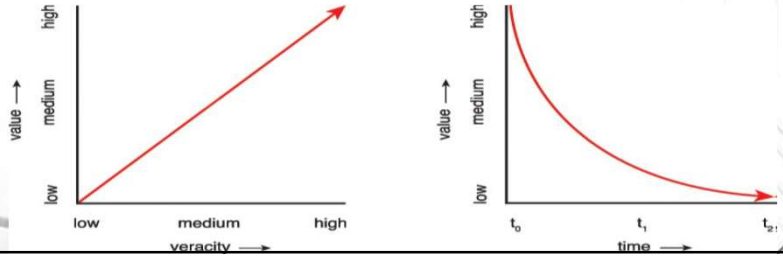
Veracity

- Definition: Quality or fidelity of data.
- Assessment: Crucial for processing, leading to data cleaning activities.
- Signal-to-Noise Ratio: High veracity indicates valuable data
- the veracity of data depends on the source and the method of acquisition



Value

- Value Definition: The usefulness of data for a business.
- Veracity Connection: Higher data fidelity increases business value.
- Time Dependency: Data processing time impacts value.
- Inversely Related: Value decreases with longer data-to-information processing times.
- Impact on Decision-Making: outdate results hinder decision quality and speed.



Value

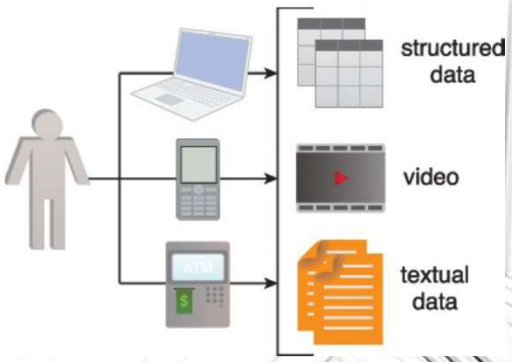
Recommendation System:

- Consider a scenario where a user is exploring a selection of clothing items. If the recommendation system can quickly process the user's preferences and browsing behavior in real-time, it can suggest complementary accessories or matching items, increasing the likelihood of a successful sale.
- However, if there's a delay in processing this information, and recommendations are **not delivered in real-time, the user might lose interest or move on to another platform**, diminishing the value of the recommendation system for the business.

Different Types of Data

Different Types of Data

- Classification: Structured, unstructured, and semi-structured.
- Impact: Different data types influence processing and storage in Big Data solutions.

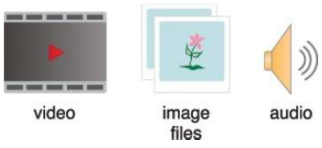


Structured Data

- Definition: Conforms to a data model, often tabular.
- Storage: Typically in relational databases.
- Examples: Banking transactions, invoices, customer records.

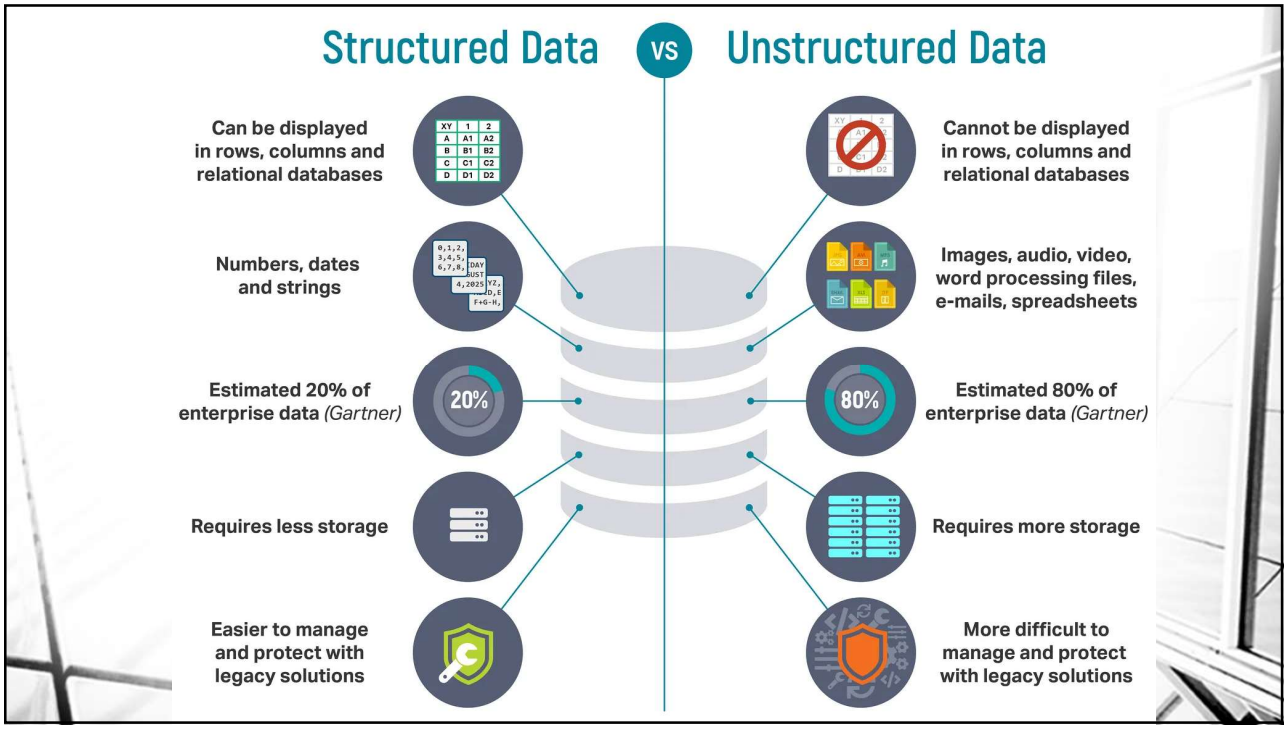
Unstructured Data

- Definition: Does not conform to a data model.
- Volume: Comprises 80% of enterprise data.
- Growth Rate: Faster than structured data.
- Examples: Textual or binary data, video, images, audio.



Semi-Structured Data

- Characteristics: Defined structure, hierarchical or graph-based.
- Format: It may not conform to a fixed schema but has tags, markers, or hierarchies.
- Examples: XML, JSON, sensor data, NoSQL databases.
- Processing: More manageable than unstructured data.



Metadata

- Definition: Provides information about a dataset's characteristics and structure.
- Importance: Essential for processing, storage, and analysis, particularly in semi-structured and unstructured data.
- Examples: XML tags, attributes in digital photographs.

Conclusion

- This chapter have laid the foundation for understanding Big Data, its interdisciplinary nature, applications, analytics types, and key characteristics.
- Emphasizing the importance of strategic decision-making, The slides guide from basic to advanced concepts, providing a comprehensive overview. that will serve as a guide for exploring the detailed and dynamic field of Big Data in following chapters.

Case Study

- ETI is an insurance company that provides health, property, marine and aviation policies.
- It has experienced **declining profits** due to factors like **fraud**, catastrophes, non-compliant regulations.
- Current systems mainly **use structured data** and are inadequate for demands.
- A committee investigated and set goals to improve risk assessment, claims processing, compliance.
- They recommend **a data-driven strategy** using enhanced analytics across functions.
- IT identified **obstacles like inability to store/process large volumes of internal/external unstructured data.**
- **Existing systems cannot handle variety/volume of data** in a timely manner.
- *Big Data is proposed to address these challenges by enabling analysis of diverse structured and unstructured data at scale.*