



# ***Data Overview***

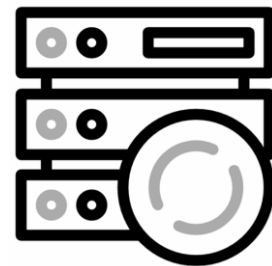


# AGENDA

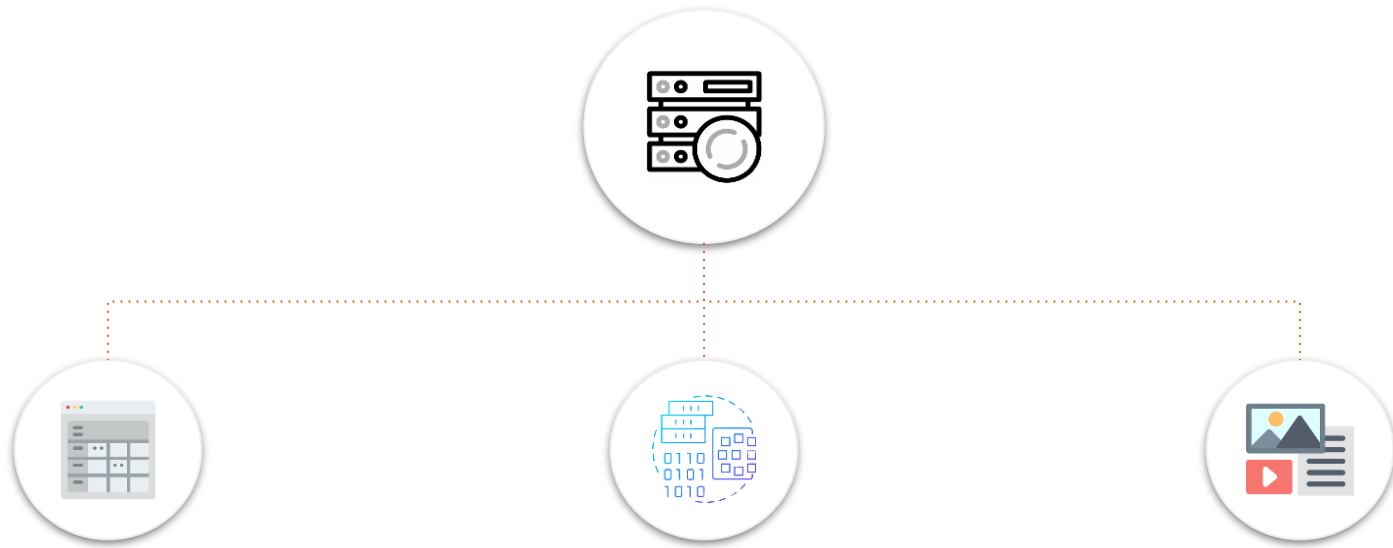
- ● **What is Data**
  - Data Analytics
  - Data Tooling (e.g. Hadoop)
  - Data Management
  - Data Visualization
  - Data Ethics and Privacy

# What is Data?

- The *New Oil* of 21st century - a hidden gold mine
- The building block of the modern digital world
- The raw material for innovation
- That what helps us make sense of the world around us and make better decisions
- It helps create better products and services with insights



# Types of Data



## Structured

*Well-organized, with a fixed schema such as customer data in relational database*

## Semi-Structured

*Partially organized, with some tags or metadata such as XML documents*

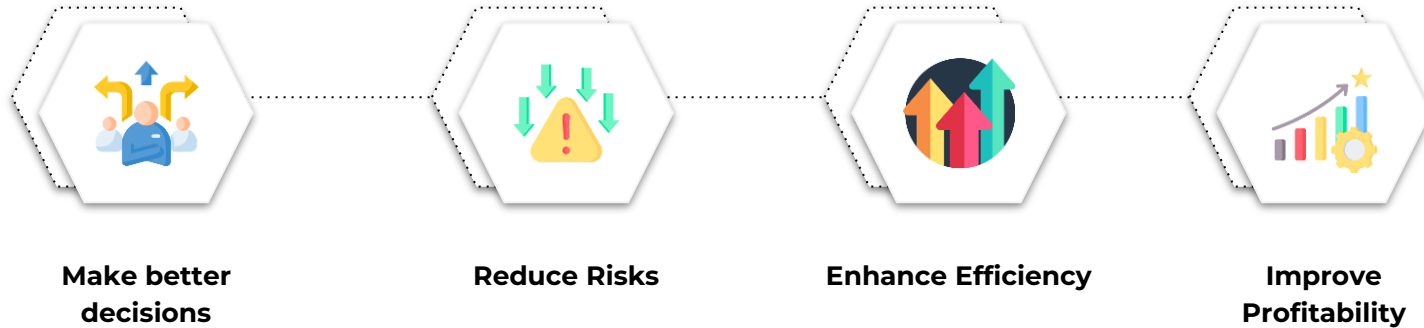
## Un-Structured

*Not organized, in its native format such as Text documents, images, videos*

# Types of Data

Data type	Organization	Storage	Analysis	Examples
<b>Structured</b>	<i>Predefined format</i>	<i>Relational databases, spreadsheets, data warehouses</i>	<i>Easy</i>	<i>Customer records, product inventory data, financial transaction data, sensor data</i>
<b>Semi-structured</b>	<i>Some organizational properties</i>	<i>XML, JSON, markup languages</i>	<i>More difficult</i>	<i>HTML documents, email messages, logs, social media posts</i>
<b>Unstructured</b>	<i>No organizational properties</i>	<i>Text documents, images, audio files, video files</i>	<i>Most difficult</i>	<i>Books, articles, poems, songs, movies</i>

# Role of Data in Decision Making

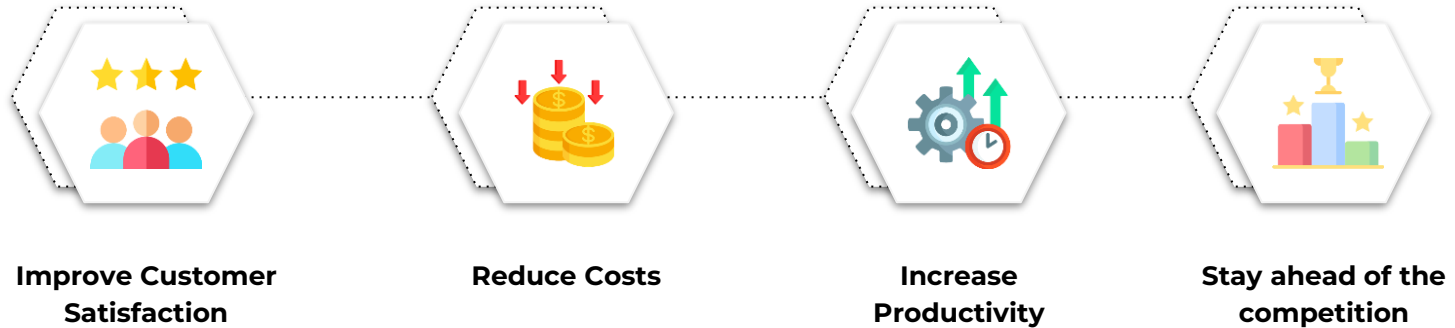


As per **Signal AI**, an estimated loss of revenue to the tune of

**\$ 4.62 Trillion**

occurred in **2020** alone from **not applying data to business decision making process**

# Role of Data in Problem Solving

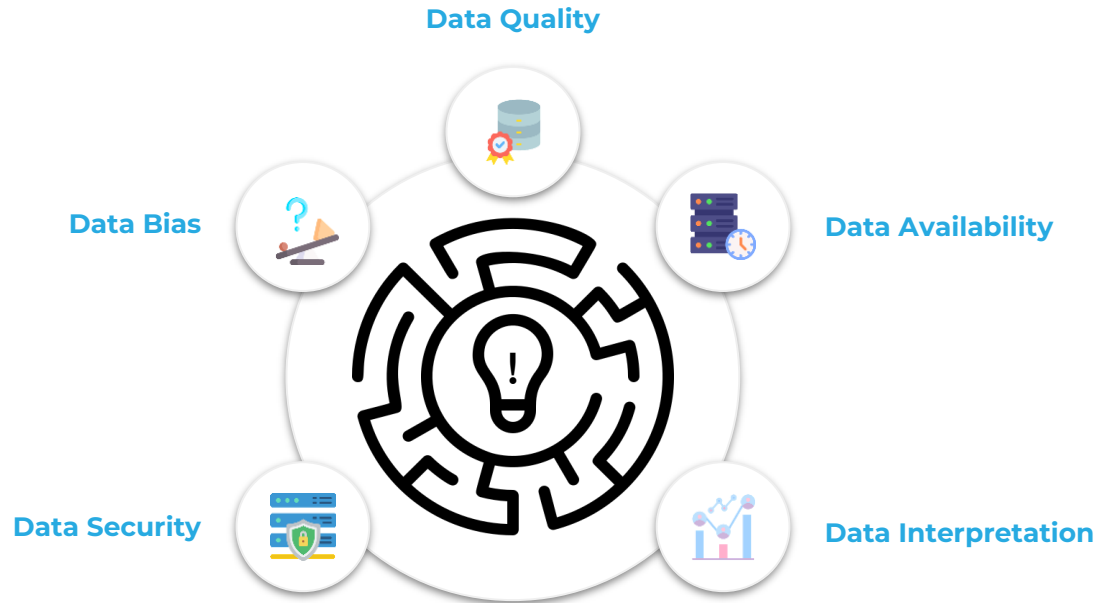


According to a study by **MIT Sloan Management Review**,

**95%**

**executives believe that data and analytics will be essential** to their company's future success

# Challenges in Data-driven Decision Making & Problem Solving





# Data Sources and their Characteristics

*Data that is generated or collected within an organization*



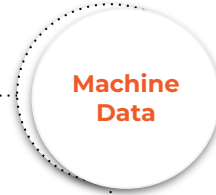
- Customer Records
- Financial Data
- Product Data
- Operational Data

*Data that is collected from outside of an organization*



- Market Data
- Social Media Data
- Weather Data
- Government Data

*Data that is generated by machines*



- Sensor Data
- Machine Log Data
- Message Queues
- APIs

*Data that is collected from the humans*

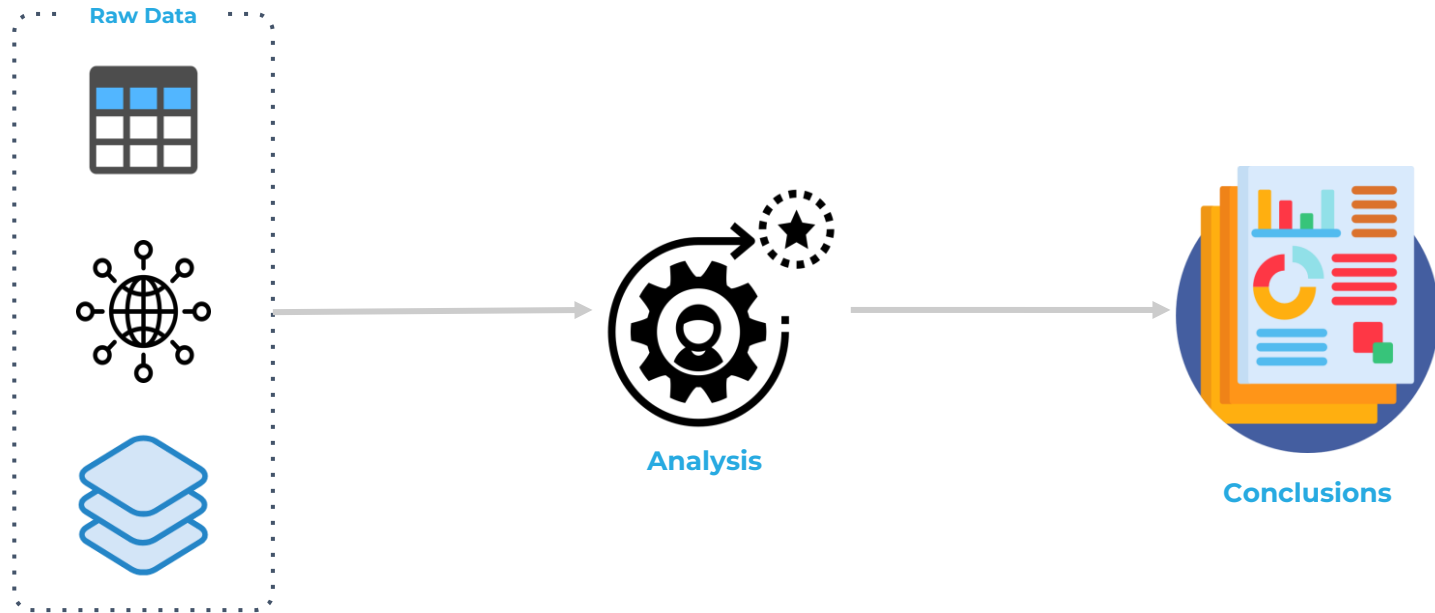


- Survey Data
- Interview Data
- Focus Group Data

# AGENDA

- What is Data
- **Data Analytics**
- Data Tooling (e.g. Hadoop)
- Data Management
- Data Visualization
- Data Ethics and Privacy

# Data Analytics



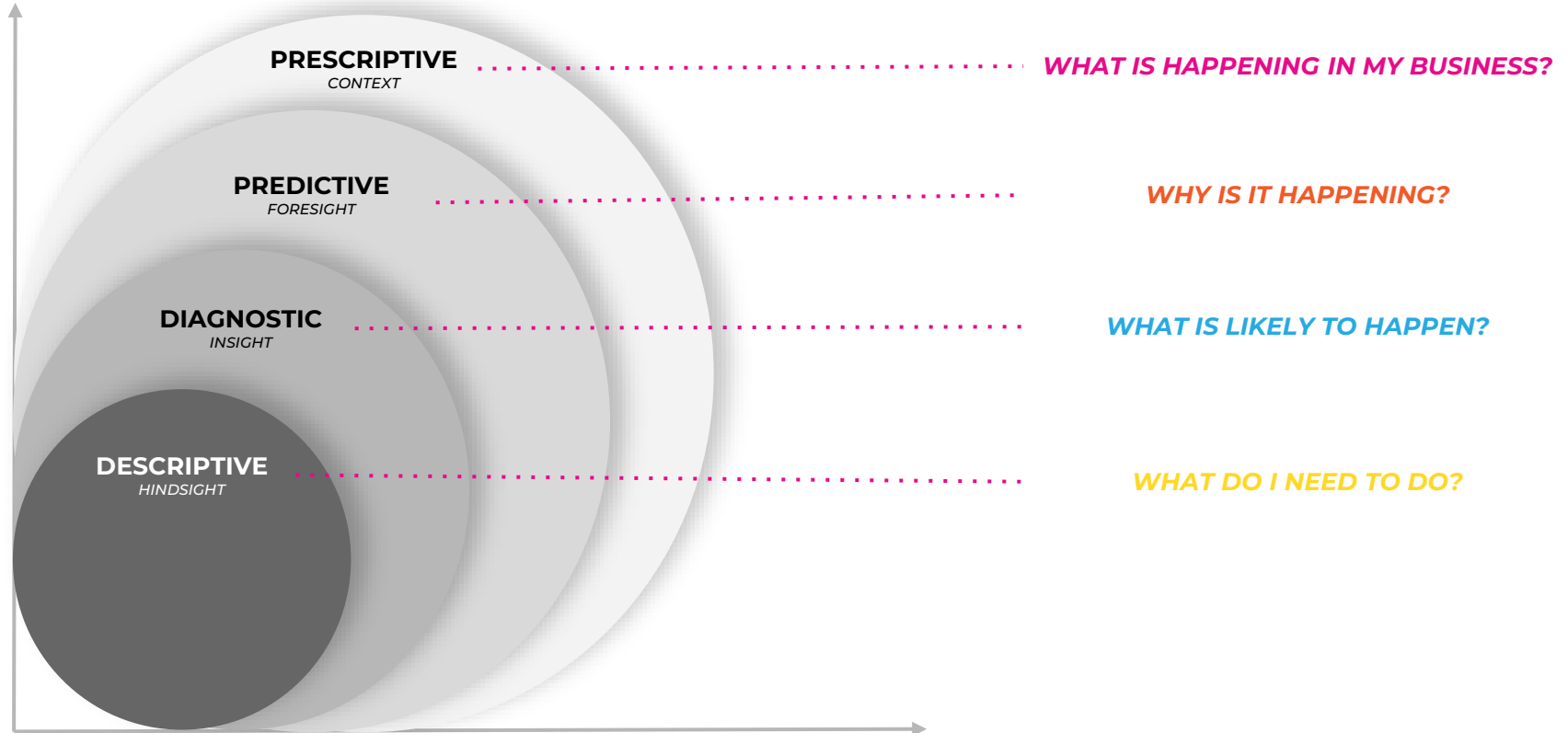
# Data Analytics - What & Why?

- ❑ Data analytics is the science of analyzing raw data to make conclusions about that information
- ❑ It helps a business:
  - ❑ optimize its performance
  - ❑ perform more efficiently
  - ❑ maximize profits, or
  - ❑ make more strategically-guided decisions.
- ❑ The techniques and processes of data analytics have been transformed into fine-tuned algorithms and automated processes that work over raw data for human consumption

# Approaches to Data Analytics

- ❑ Various approaches to data analytics include:
  - ❑ looking at what happened - *descriptive analytics*
  - ❑ why something happened - *diagnostic analytics*
  - ❑ what is going to happen - *predictive analytics*, or
  - ❑ what should be done next - *prescriptive analytics*.
- ❑ For the best possible outcomes, it relies on a variety of software tools including:
  - ❑ Spreadsheets
  - ❑ Data Visualization tools
  - ❑ Reporting tools
  - ❑ Data Mining programs
  - ❑ Open-source Languages

# Types of Data Analytics



# Types of Data Analytics

## DESCRIPTIVE

*WHAT IS HAPPENING IN MY BUSINESS?*

- Comprehensive, accurate and live data
- Efficient visualization

## PREDICTIVE

*WHAT IS LIKELY TO HAPPEN?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

## DIAGNOSTIC

*WHY IS IT HAPPENING?*

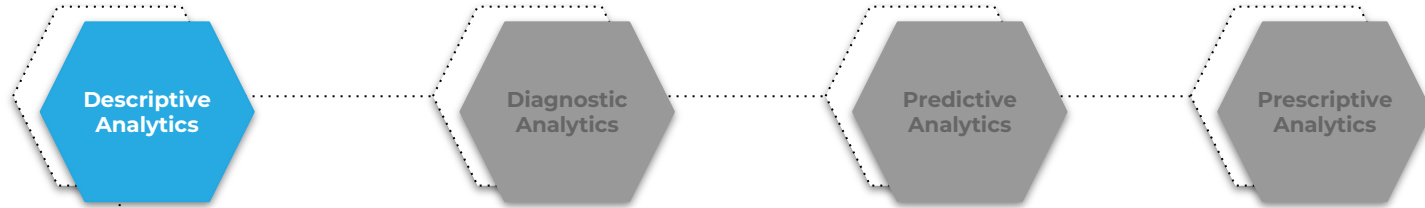
- Ability to drill down to the root-cause
- Ability to isolate *noise*

## PRESCRIPTIVE

*WHAT DO I NEED TO DO?*

- Recommended actions and strategies based on champion / challenger testing outcomes
- Advanced analytical techniques for specific recommendations

# Exploring Data Analytics Techniques and Algorithms



**Data Visualization:** It involves using charts, graphs, and other visual representations to display data



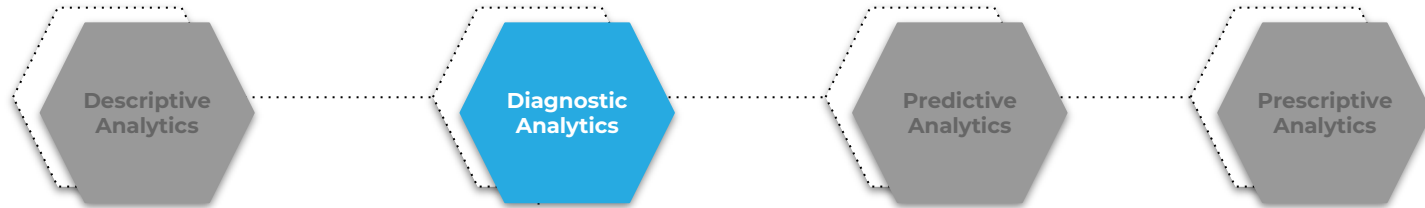
**Statistical Analysis:** It involves using statistical methods to analyze data and identify patterns



**Time Series Analysis:** It involves using statistical methods to identify patterns over time



# Exploring Data Analytics Techniques and Algorithms



**Correlation Analysis:** It involves identifying the relationships between different variables

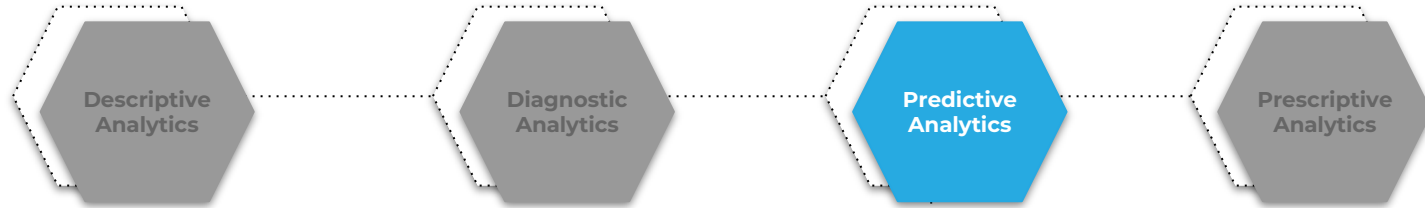


**Regression Analysis:** It involves using statistical methods to predict the value of one variable based on the value of another variable



**Anomaly Detection:** It involves identifying key data points that are unusual or unexpected

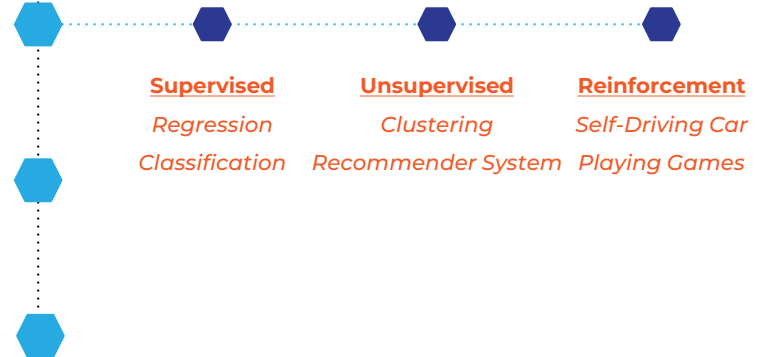
# Exploring Data Analytics Techniques and Algorithms



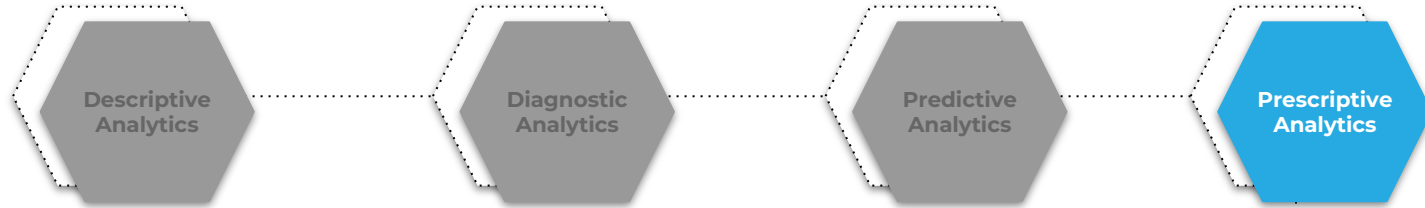
**Machine Learning:** It involves using algorithms to learn from data and make predictions

**Deep Learning:** It is a type of machine learning that uses artificial neural networks to make predictions

**Natural Language Processing:** It is a type of machine learning identifies parts of speech, entities, sentiment, and other aspects of text



# Exploring Data Analytics Techniques and Algorithms



**Optimization Modeling:** It involves using mathematical models to find the best solution to a problem

**Simulation:** It involves creating a virtual model of a system to test different scenarios

**Decision Support System:** It involves providing decision-makers with information and tools to help them make better decisions

# Data Analytics in Retail Industry

**Data analytics can help retailers to understand their customers better, identify trends, and make better marketing decisions**

**EXAMPLE**

Walmart uses data analytics to track customer behavior in its stores and online. This data helps Walmart to identify which products are popular, which products are not selling well, and which products customers are likely to buy together. They use this information to optimize their inventory, pricing, and marketing campaigns.



# Data Analytics in Healthcare Industry

**Data analytics can help healthcare providers to improve patient care, reduce costs, and identify fraud**

**EXAMPLE**

The Mayo Clinic uses data analytics to track patient outcomes and identify areas where care can be improved. The clinic also uses data analytics to identify patients who are at risk for developing certain diseases. This information helps the clinic to provide preventive care to these patients and improve their overall health and well-being.



# Data Analytics in Energy Industry

**Data analytics can help energy companies to optimize their operations, reduce costs, and improve customer service**

**EXAMPLE**

Duke Energy uses data analytics to monitor the performance of its power plants. This data helps them to identify problems early on and take corrective action before they cause major disruptions. Duke Energy also uses data analytics to optimize its power grid and improve the reliability of its service.



# Data Analytics in Government Sector

**Data analytics can help government agencies to improve service delivery, efficiency, effectiveness, and transparency**

**EXAMPLE**

The US Department of Defense uses data analytics to track the progress of its projects and identify areas where it can improve its efficiency. The department also uses data analytics to identify potential risks and threats.



# Data Analytics in Manufacturing Industry

**Data analytics can help manufacturers to improve quality and productivity by helping to remove bottlenecks and check leakages**



General Electric uses data analytics to monitor the performance of its machines. This data helps GE to identify problems early on and take corrective action before they cause major disruptions. GE also uses data analytics to optimize its production processes and improve the quality of its products.

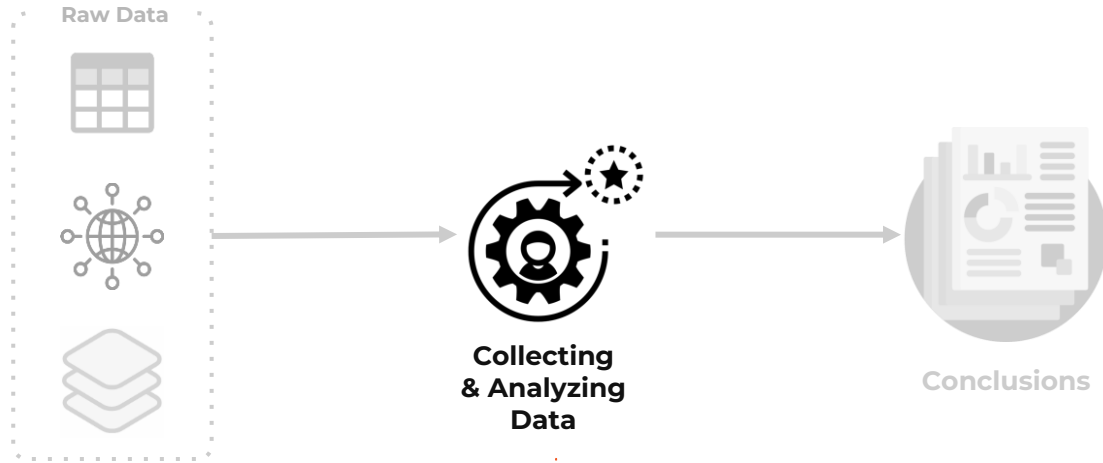




# AGENDA

- What is Data
- Data Analytics
- **Data Tooling (e.g. Hadoop)**
- Data Management
- Data Visualization
- Data Ethics and Privacy

# Overview of Data Tooling



*A collection of software and hardware tools used to  
collect, store, process, and analyze data*

# Managing & Analyzing Large Datasets using Data Tooling



It can be used to manage and analyze data of all sizes, but it is especially important for managing and analyzing **large datasets\***

1

**Collect** data from a variety of sources

2

**Store** data at a centralized location

3

**Process** data quickly and efficiently

4

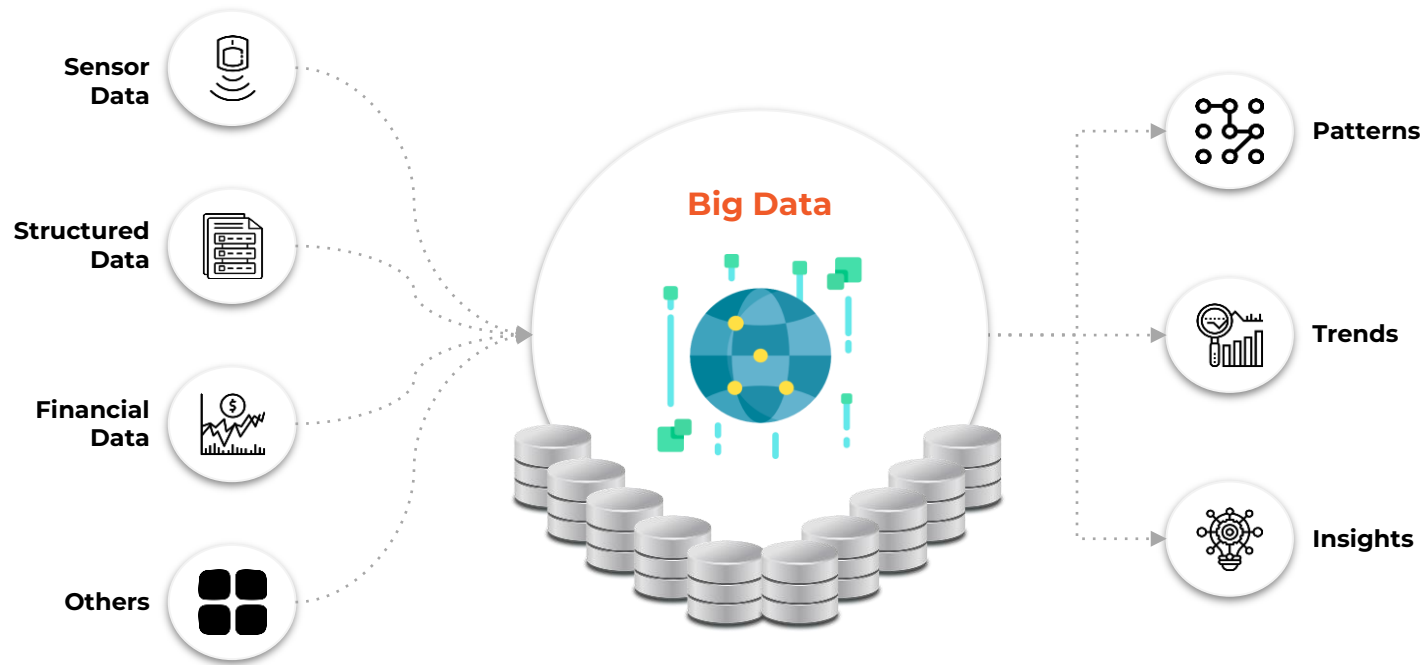
**Analyze** data to identify patterns and trends

5

**Visualize** data to make it easier to understand

*\* datasets that are too large to be processed and analyzed using traditional methods*

# What is Big Data?



# What is Big Data?

## Volume

Can you find the information you are looking for?

## Velocity

Information gain momentum and opportunities evolve in real-time

## Value

Can you find it when you need it the most?

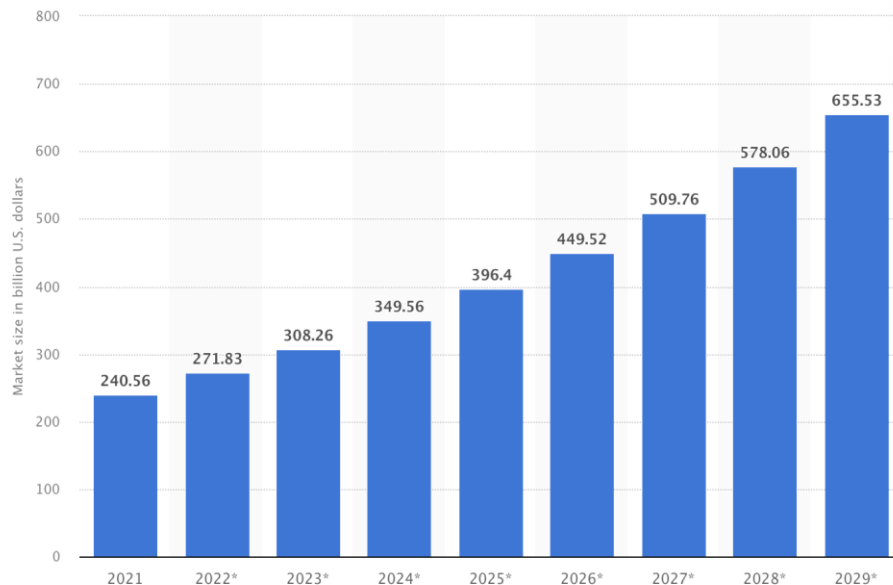
## Variety

Is your data is balanced?

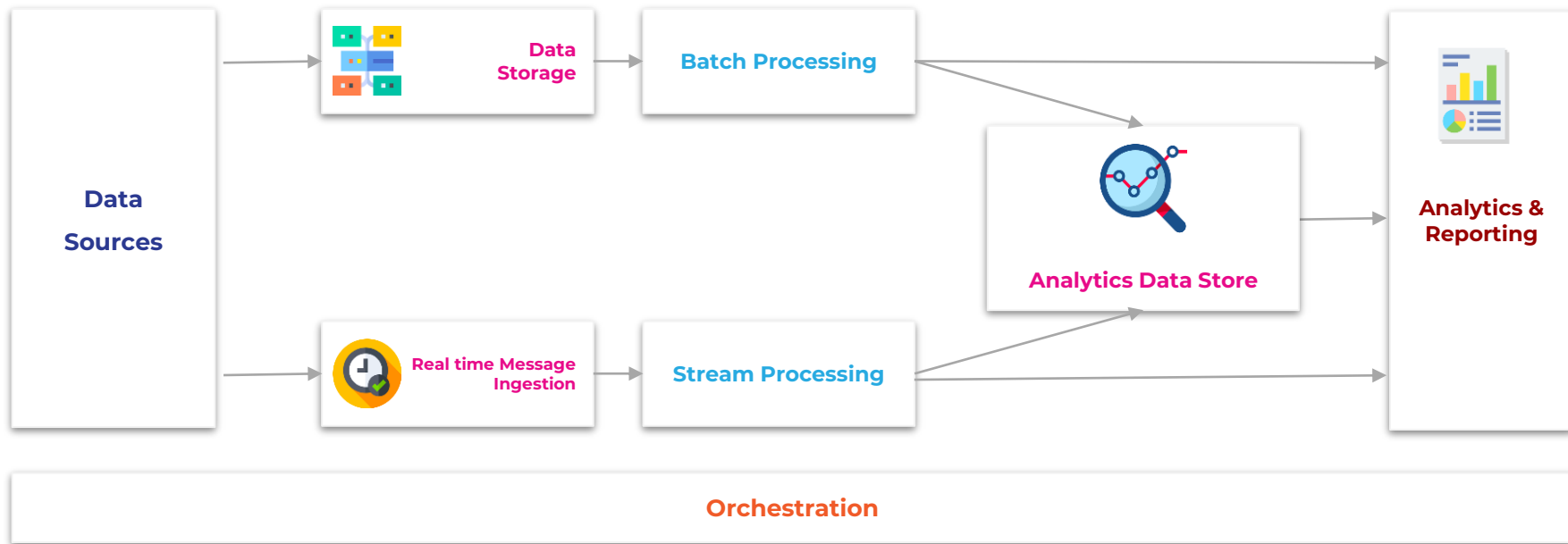
## Veracity

Are you dealing with information or disinformation?

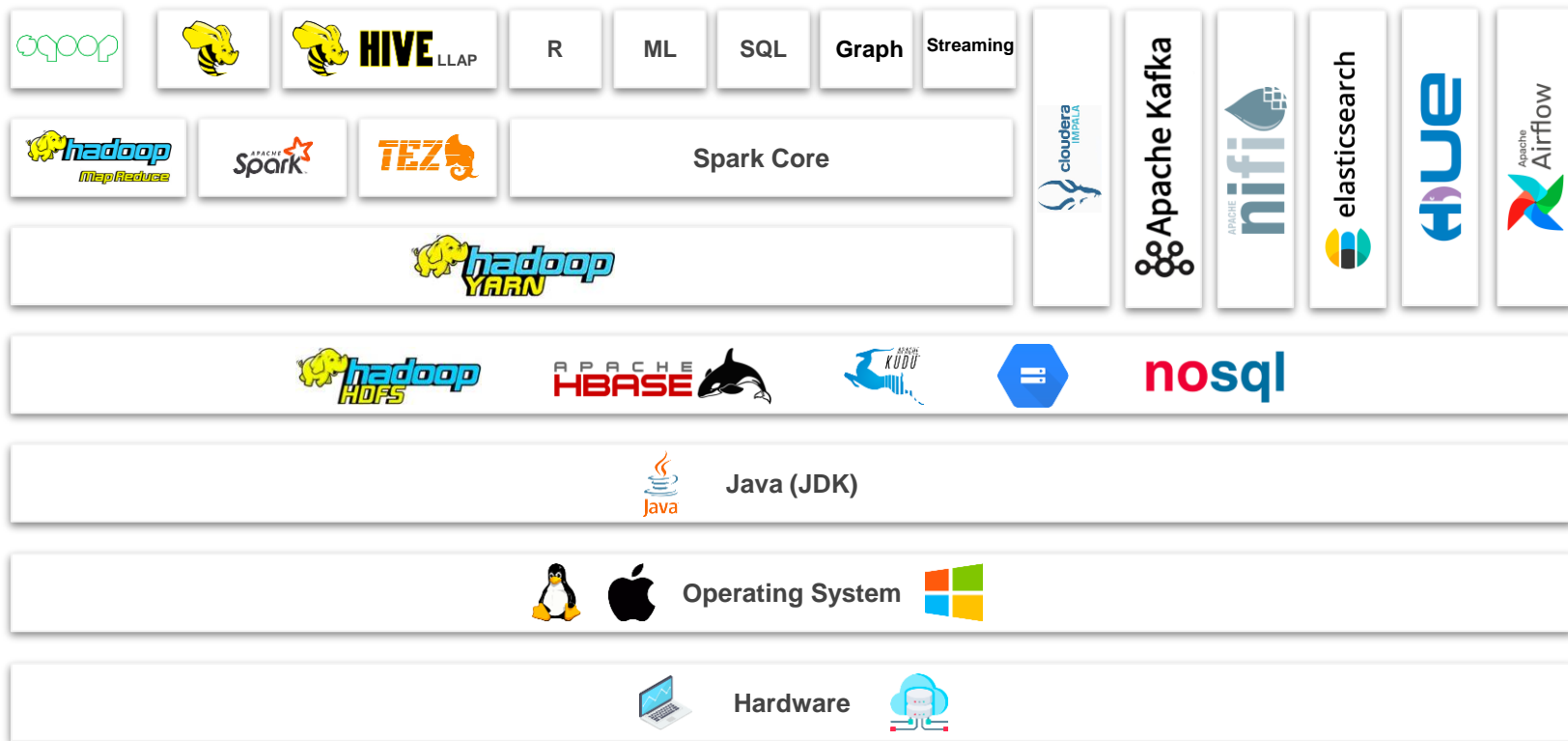
Global Big data market size (revenue) from 2021 to 2029 (st)




# Big Data Architecture



# Understanding the Big Data Ecosystem



# AGENDA

- What is Data
- Data Analytics
- Data Tooling (e.g. Hadoop)
-  • **Data Management**
- **Data Visualization**
- **Data Ethics and Privacy**

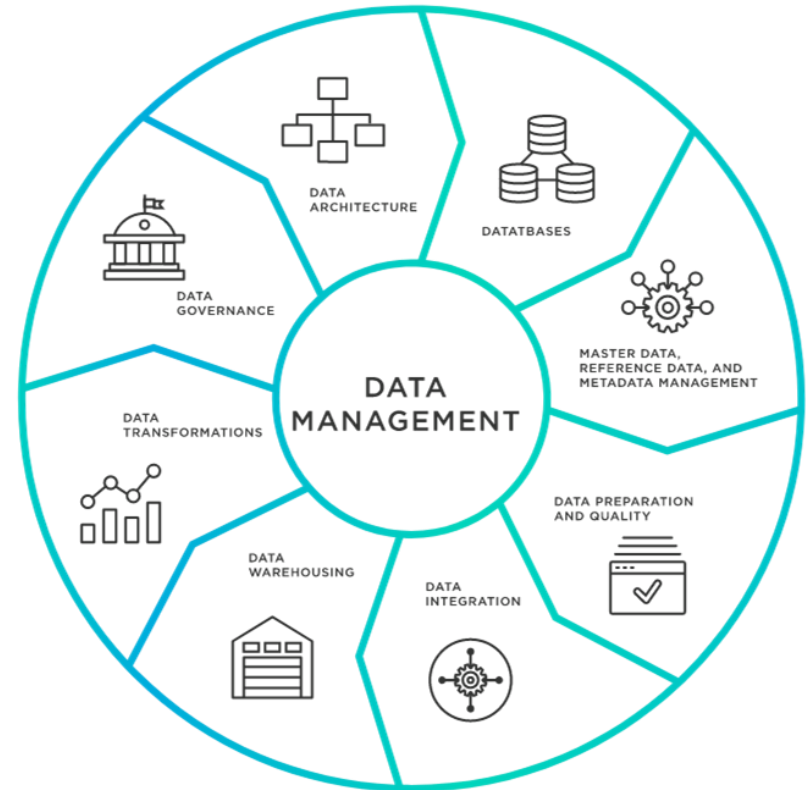


# What is Data Management?

Processes and procedures used for storing, organizing, and protecting their data.

It is important for a number of reasons, including:

- **Data is valuable:** Data is a valuable asset for organizations. It can be used to make decisions, improve efficiency, and identify new opportunities.
- **Data is complex:** Data can be complex and difficult to manage. It can be stored in a variety of formats, and it can be difficult to keep track of changes.
- **Data is vulnerable:** Data is vulnerable to security threats. It can be stolen, corrupted, or lost.



# Significance of effective Data Management



## PROTECT DATA

Can be done through a variety of measures, such as encryption, access control, and disaster recovery planning



## BETTER DECISION

Can be done through a variety of tools and techniques, such as data analysis and data visualization



## IMPROVED EFFICIENCY

Can be done by automating tasks and reducing the time it takes to access and use data

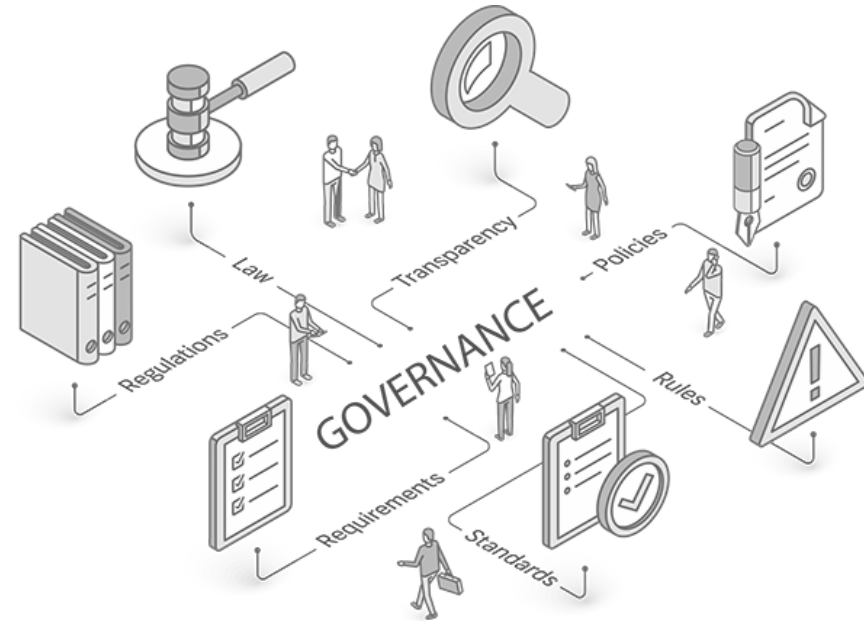


## NEW OPPORTUNITIES

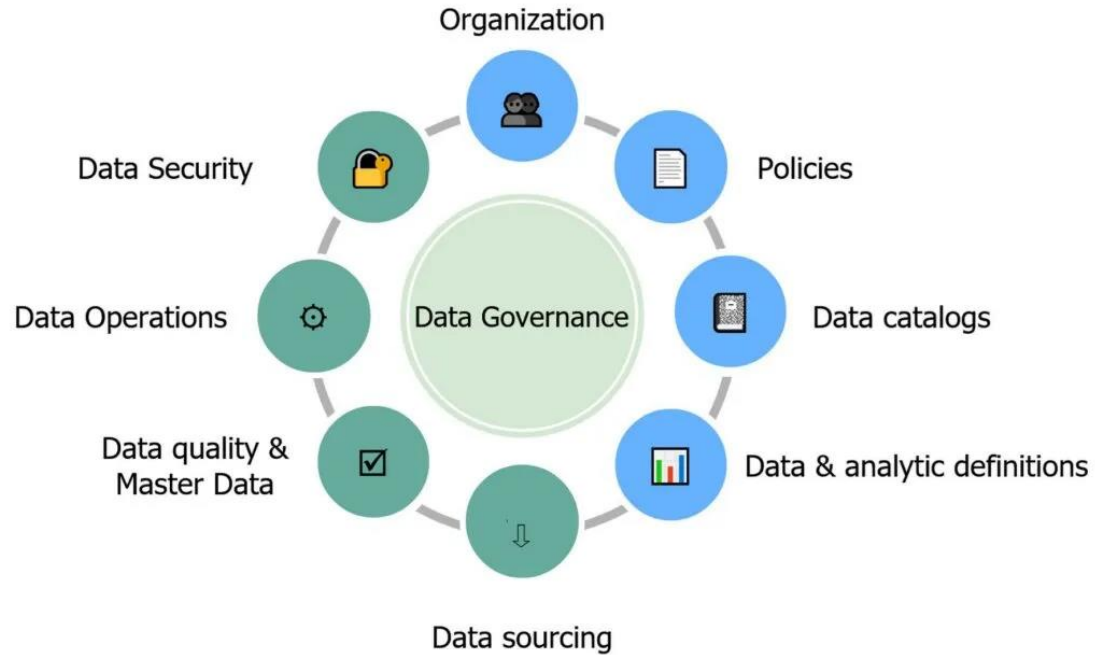
Can be done by analyzing data to identify trends and patterns

# Overview of Data Governance

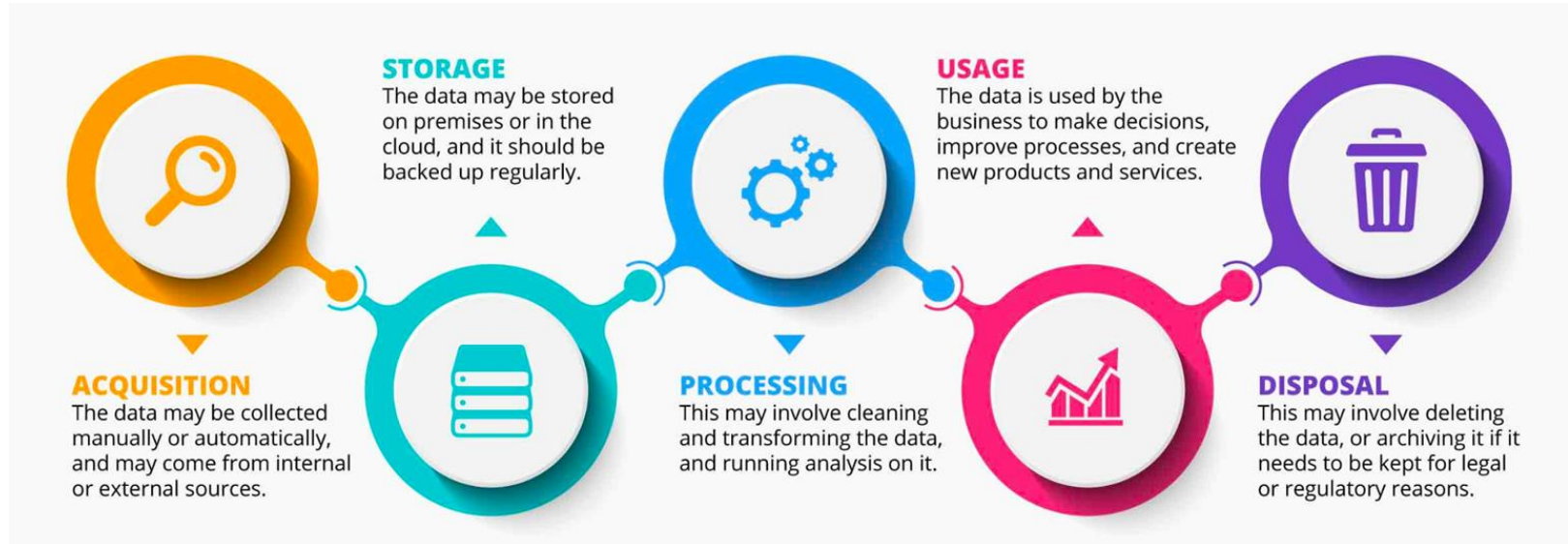
- Data governance is a collection of processes, roles, policies, standards, and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals.
- It establishes the processes and responsibilities that ensure the quality and security of data used across the organization.
- Data governance defines who can take what action, upon what data, in what situations, and using what



# Overview of Data Governance

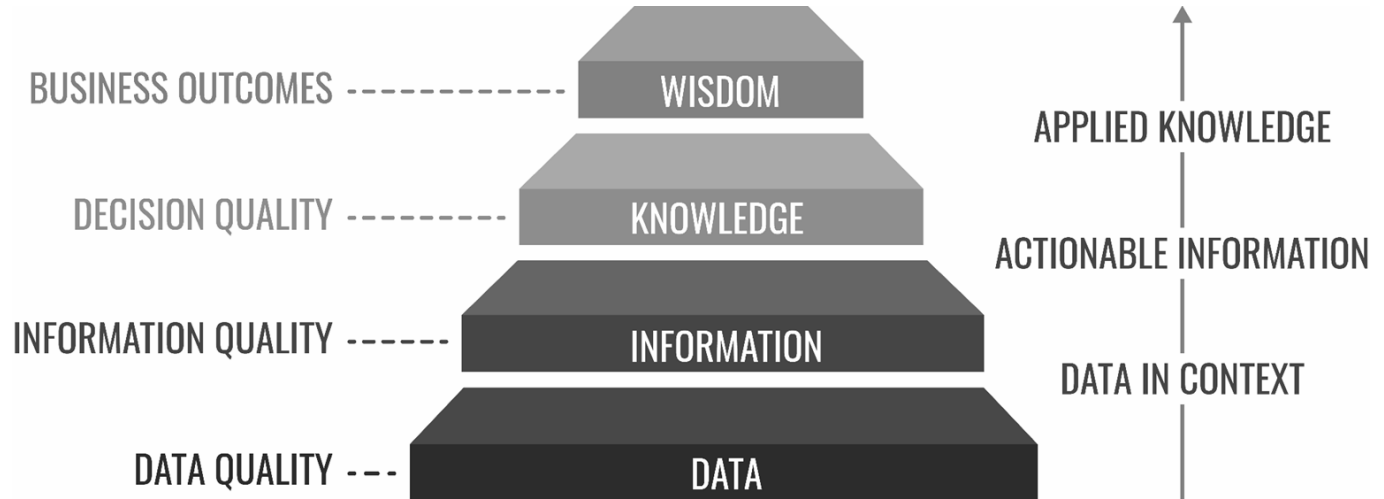


# Data Lifecycle Management



# What is Data Quality?

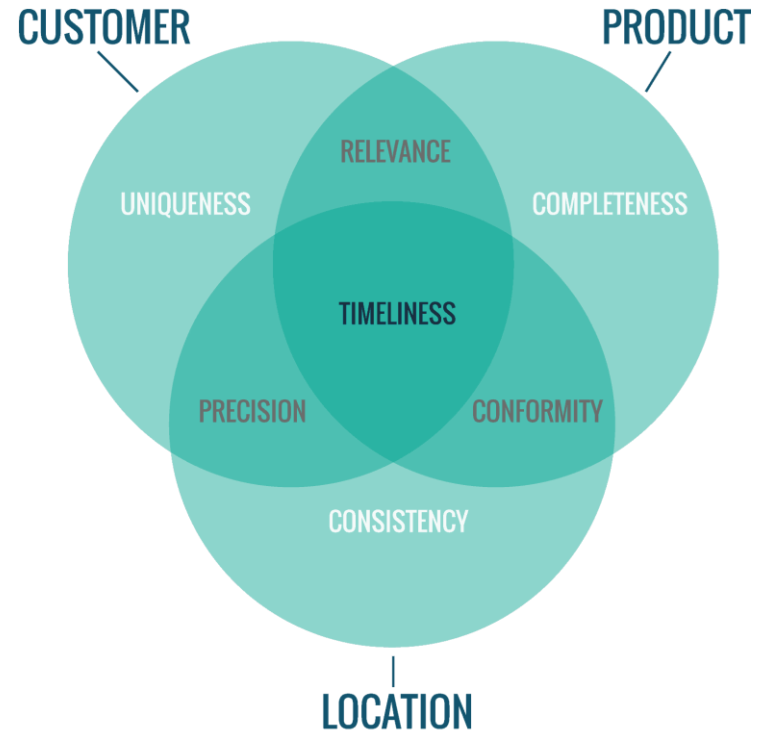
- Data is of high quality, if it is fit for the intended purpose or use
- Data is of high quality, if it correctly represents the real-world construct being modeled



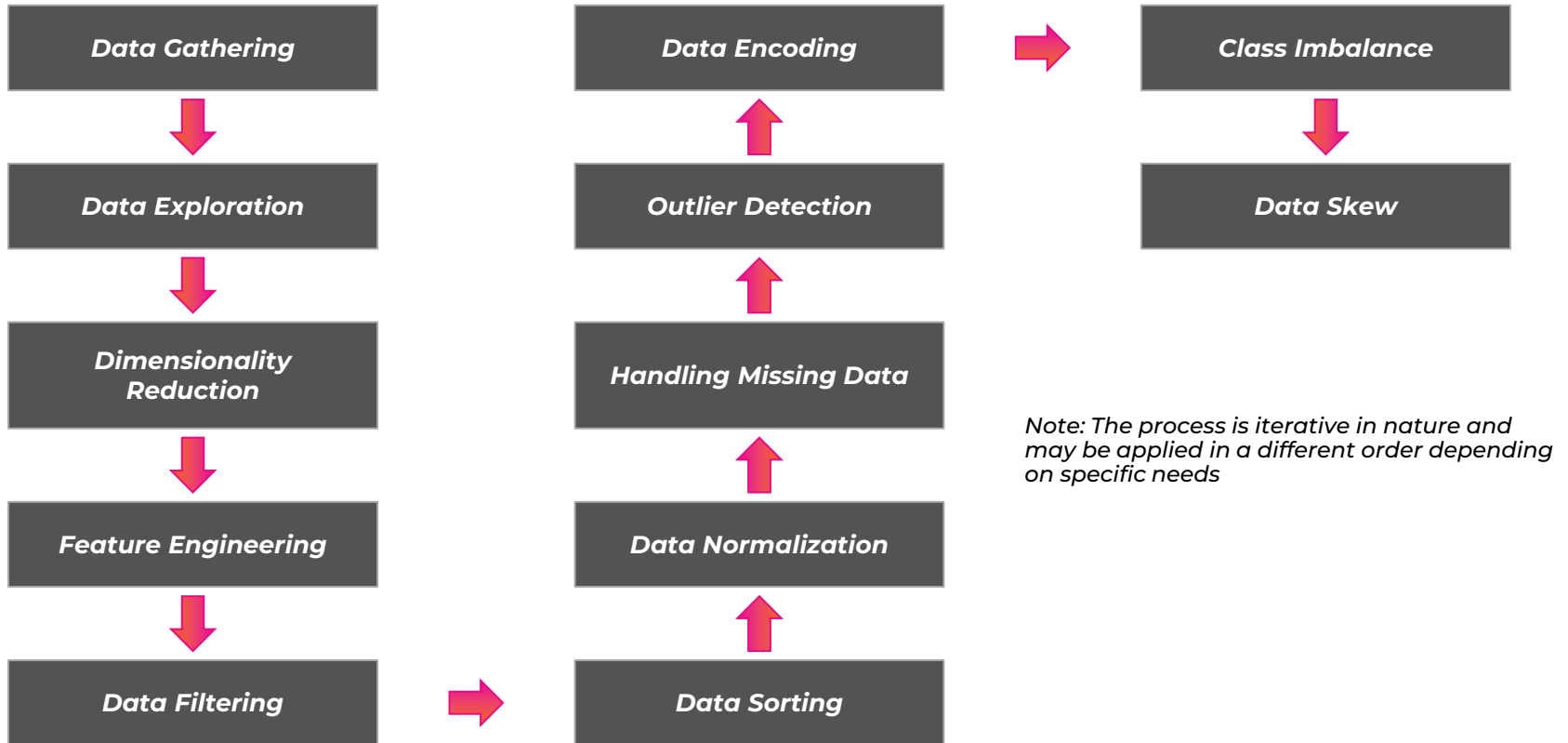
# How to ensure Data Quality?

The remedies used to prevent data quality issues and eventual data cleansing includes these disciplines:

- Data Governance
- Data Profiling
- Data Matching
- Data Quality Reporting
- Master Data Management (MDM)
- Customer Data Integration (CDI)
- Product Information Management (PIM)
- Digital Asset Management (DAM)



# How to ensure data is clean and ready for analysis?





# Introduction to Data Integration and Data Warehousing

## DATA INTEGRATION

The process of **combining data from different sources into a single, consistent view**. This can be a complex process, as the data may come from different formats, systems, and databases.



## DATA WAREHOUSING



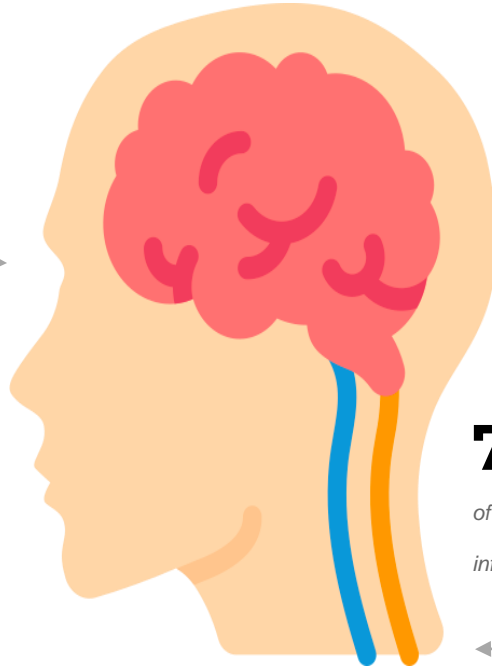
The process of **storing integrated data in a central repository**. Data warehouses are designed for analytical purposes, and they can be used to run complex queries and analysis on large amounts of data.

# AGENDA

- What is Data
- Data Analytics
- Data Tooling (e.g. Hadoop)
- Data Management
- > • **Data Visualization**
- **Data Ethics and Privacy**

# Significance of Data Visualization

**VISUALS** are processed  
**60000**  
times faster than **TEXT**



**90%**  
of information transmitted to  
the brain is **VISUAL**

**70%**  
of people respond better to **VISUAL**  
information than **TEXT**

As per Gartner

by **2025, 75%** of all newly generated data will be **unstructured** and organizations will need to find new ways to make sense of this data

As per Microsoft

**90%** of people are **visual learners**, i.e., **data visualization is a powerful way to communicate** with people

As per Gartner

**77%** of people are more likely to **read content having visuals**, meaning that data visualization can be used to **increase engagement with content**

# Various Data Visualization tools



## Factors to consider while choosing Data Visualization tool

**Ease of use:** The tool should be easy to use, even for users with no prior data visualization experience

**Flexibility:** The tool should be flexible enough to create a variety of visualizations, from simple charts and graphs to complex dashboards

**Scalability:** The tool should be scalable to handle large data sets

**Cost:** The tool should be affordable for the organization

**Integration:** The tool should be able to integrate with other tools that the organization uses, such as data warehouses and business intelligence tools

# Creating Meaningful and Impactful Data Visualizations

## Effective Storytelling



Data



Narrative



Visuals



## Visual Perceptions for better Visualization



Size



Shape



Color Value



Texture



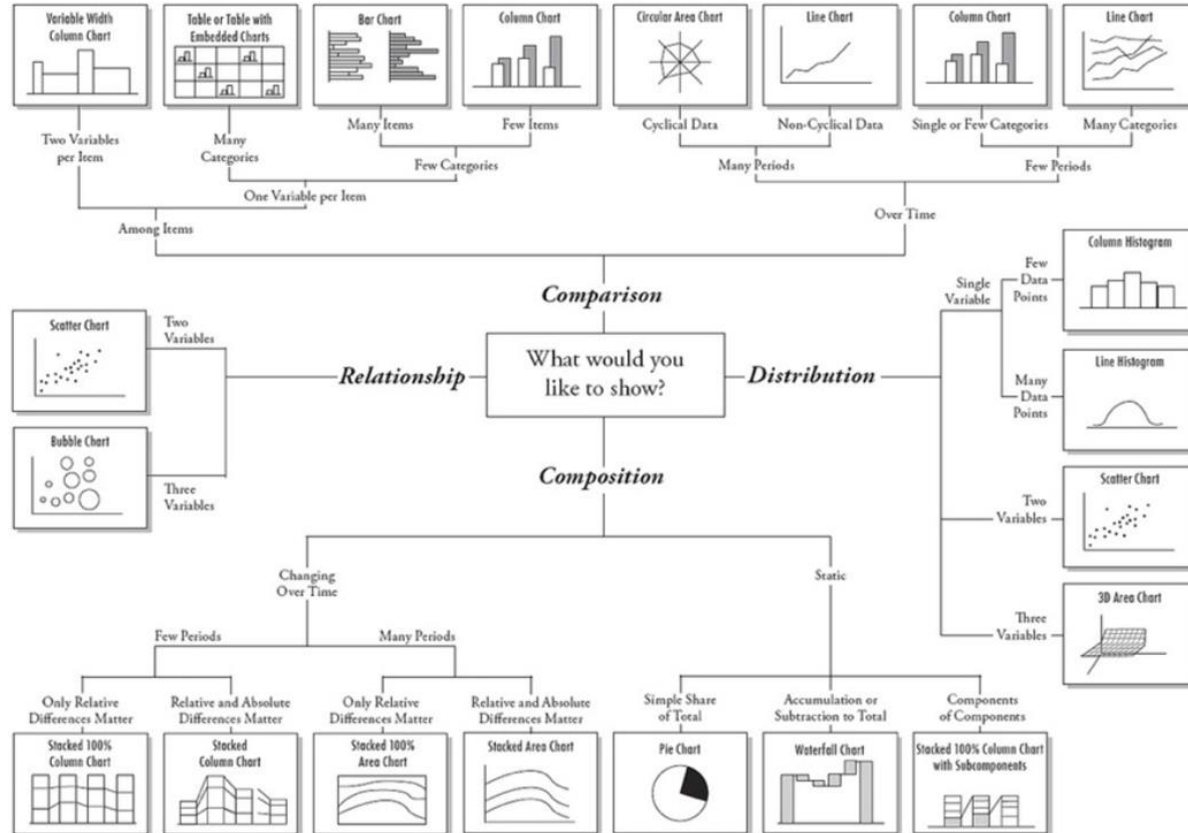
Orientation



Hue

CONTEXT is the KING

# Various Data Visualization Charts



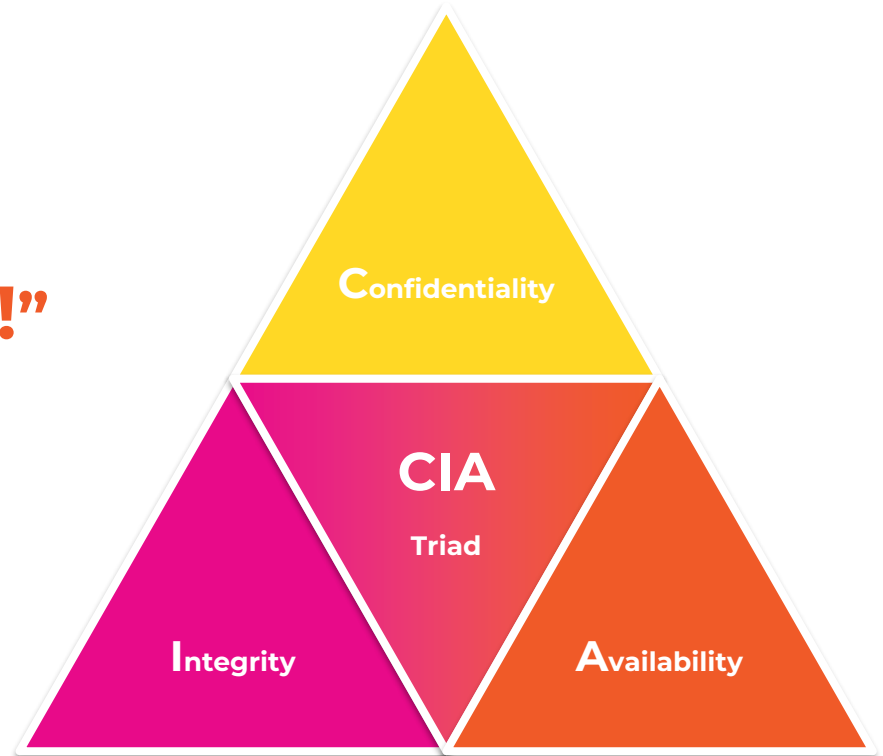
# AGENDA

- What is Data
- Data Analytics
- Data Tooling (e.g. Hadoop)
- Data Management
- Data Visualization
- **> Data Ethics and Privacy**

# Understanding the ethical considerations

Before moving ahead, always remember that

**“It’s always about CIA!”**





# Understanding the ethical considerations

## Data Security

Data is secured in the database, no matter how users connect to the protected data

## Privacy

All Personal Identifying information (PII) is masked for developers

## CI/CD Framework

Deployments should only be done through the CI/CD tooling and data can only be retrieved from a published zone

## Security Rules

Rules can be reused across tables and domains

## Role Based Framework

Each user has a personal view into the data for which they have permissions using a role based framework

## AD Syncing and SAML

Pre-defined AD Groups integrated into Snowflake Role with Okta syncing



# Overview of Data Privacy Regulations



# Overview of Data Privacy Regulations



The **GDPR** is an European Union (EU) regulation governing data protection and privacy for all individuals within the EU and the European Economic Area (EEA). It aims primarily to give control back to citizens and residents over their personal data, and to simplify the regulatory environment for international businesses by unifying regulations within the EU. The GDPR sets out a number of requirements for organizations that process personal data, including:

- Obtaining consent from individuals before collecting or processing their personal data
- Providing individuals with access to their personal data and the right to request that it be deleted
- Reporting data breaches to regulatory authorities within 72 hours
- Implementing technical and organizational measures to protect personal data

# Overview of Data Privacy Regulations



The **CCPA** is a state law of the state of California, USA that gives consumers more control over their personal data. The CCPA applies to businesses that collect or sell the personal data of its residents. The CCPA gives consumers the following rights:

- The right to know what personal data is being collected about them
- The right to delete their personal data
- The right to opt out of the sale of their personal data
- The right to non-discrimination while exercising their CCPA rights

# Overview of Data Privacy Regulations



The **LGPD** is a Brazilian law that regulates the processing of personal data by individuals and organizations. The LGPD applies to all organizations that process the personal data of Brazilian residents, regardless of where the organization is located. The LGPD sets out a number of requirements for organizations that process personal data, including:

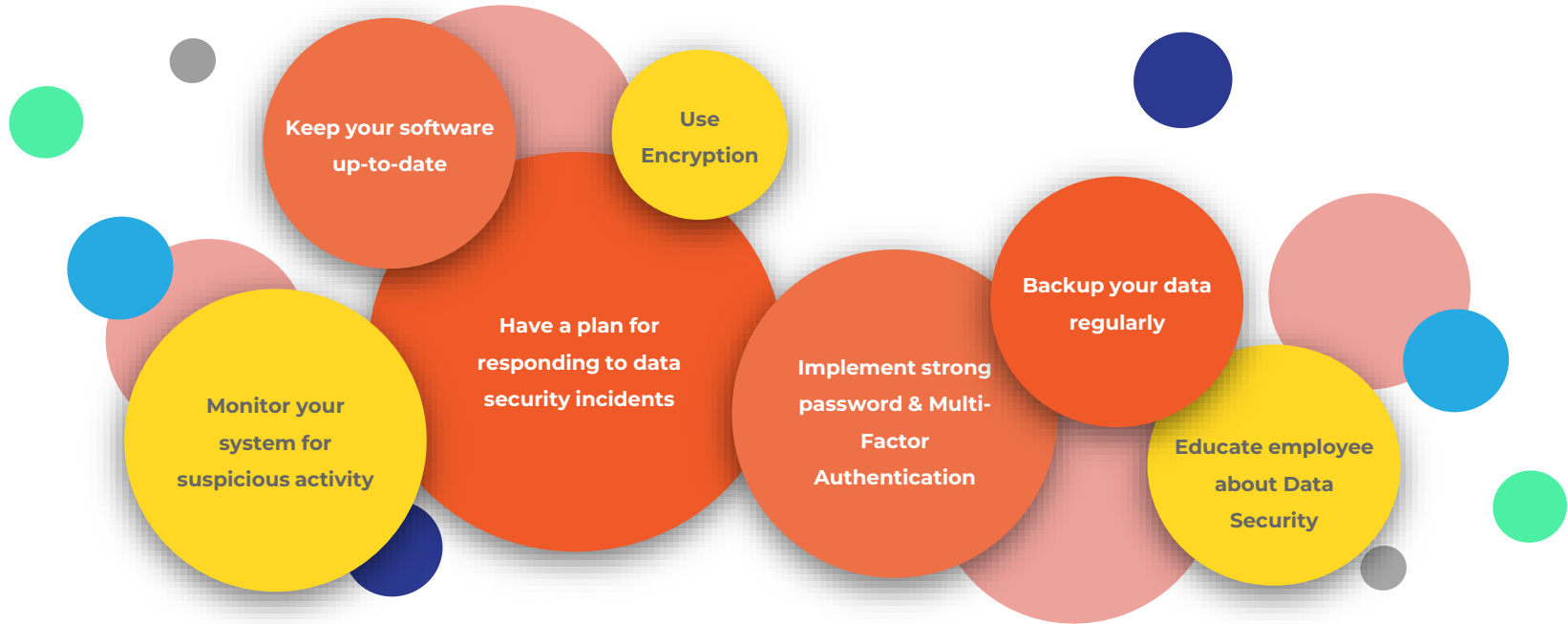
- Obtaining consent from individuals before collecting or processing their personal data
- Providing individuals with access to their personal data and the right to request that it be deleted
- Reporting data breaches to authorities in Brazilian government within 72 hours
- Implementing technical and organizational measures to protect personal data

# Importance of Data Security

Data security is the practice of protecting digital information from unauthorized access, use, disclosure, disruption, modification, or destruction. Data security is important for a number of reasons, including:



# Safeguarding Sensitive Information



# Strategies for Ethical and Responsible Data Practices

- **Obtain consent**: Before collecting or using data. It must be informed consent, meaning that the individual must be fully aware of how their data will be collected, used, and shared.
- **Be transparent**: About their data collection and usage practices. Clearly communicate what data is being collected, how it is being used, and with whom it is being shared.
- **Protect data security**: By using strong encryption and other security measures to protect data from unauthorized access, use, or disclosure.
- **Use data responsibly**: Only in ways that are consistent with the purpose for which it was collected. For example, if data is collected for marketing purposes, it should not be used for other purposes, such as making hiring decisions.
- **Be accountable**: For your data collection and usage practices. This means having a process in place to address complaints or concerns about data privacy and security.



# AGENDA

- What is Data
- Data Analytics
- Data Tooling (e.g. Hadoop)
- Data Management
- Data Visualization
- Data Ethics and Privacy

# Questions