

# Lecture 1

## Fundamentals of Big Data and Data Analytics

Dr. Lydia Wahid

# Agenda

-  What is Big Data?
-  Big Data characteristics
-  Data types processed by Big Data solutions
-  Big Data Sources
-  Basic Concepts
-  What is Data Analytics?
-  Categories of Data Analytics
-  Adoption of Analytics in Business
-  Data Analytics Lifecycle
-  Applications



# What is Big Data?

# What is Big Data?

- Big Data is a field dedicated to the **processing**, **analysis**, and **storage** of large collections of data.
- By large collections of data, we mean, collections of datasets whose volume, velocity or variety is so large that it is difficult (or even impossible) to process, analyze, and store using the traditional techniques.

# What is Big Data?

- The amount of data worldwide has been growing ever since the invention of the World Wide Web.
- We have **search engines** that need to look through billions of websites to return a particular information.
- Then came the **social networks** that have billions of users that create all types of transactions and content.

# What is Big Data?

- In addition, **businesses** and **governmental institutions** record every transaction of each customer, vendor, and supplier and thus have been accumulating data.
- We have also the data generated by **sensors** embedded in devices such as smartphones, energy smart meters, automobiles.
- Also, we have the data generated daily from **satellite imagery** and **communication networks**.

# What is Big Data?

- The result is an explosive growth in the amount of data.
- This phenomenal growth of data generation means that the amount of data in a single repository can be numbered in terabytes (1,024 gigabytes) or petabytes (1,024 terabytes).
- The term *big data* also refers to such massive amounts of data.



# Big Data characteristics



# Big Data characteristics

➤ The five Big Data characteristics (named 5 V's) can be used to help differentiate data categorized as “Big” from other forms of data:

1. Volume
2. Velocity
3. Variety
4. Veracity
5. Value

# Big Data characteristics

## 1. Volume:

- The volume of data refers to the size of data managed by the system.
- In Big Data environments, high data volumes impose different data storage and processing demands, as well as additional data preparation, organization and management processes.

## 2. Velocity:

- The velocity of data refers to speed at which data is created, accumulated, ingested, and processed.
- In Big Data environments, data can arrive at fast speeds, and enormous datasets can accumulate within very short periods of time.

# Big Data characteristics

## 3. **Variety:**

- Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions.
- Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage.

## 4. **Veracity:**

- Veracity refers to the quality or fidelity of data.
- Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise.

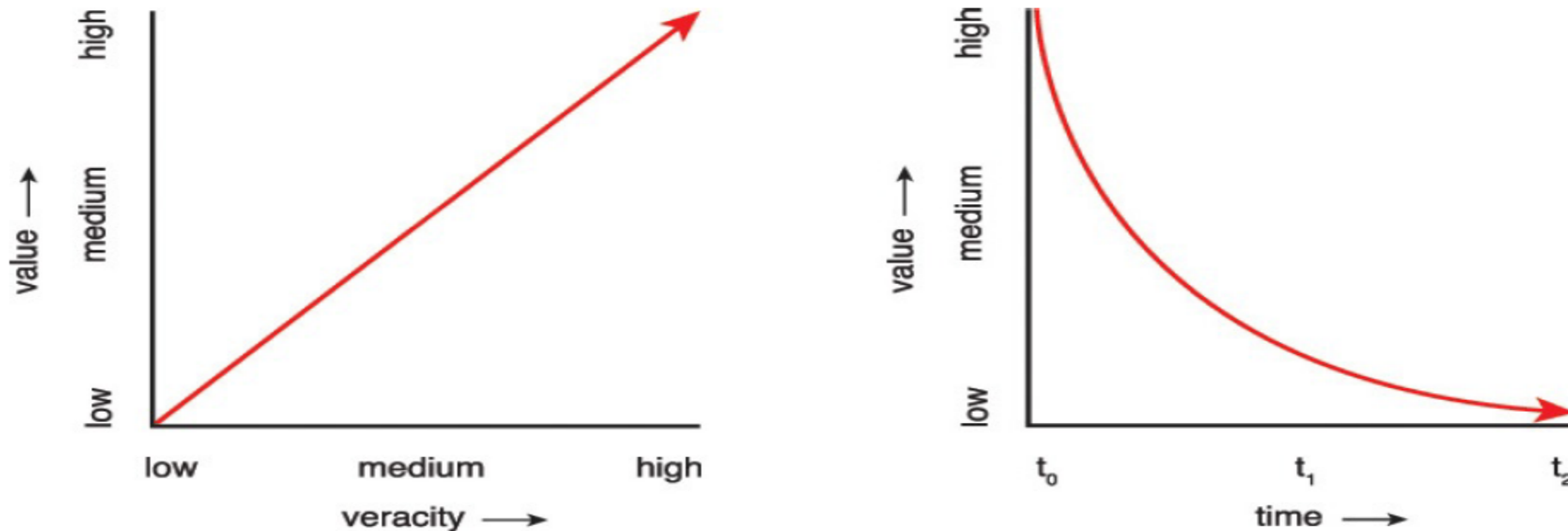
# Big Data characteristics

## 5. Value:

- Value is defined as the usefulness of data for an enterprise.
- The value characteristic is related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business.
- Value is also dependent on how long data processing takes; the longer it takes for data to be turned into meaningful information, the less value it has for a business.

# Big Data characteristics

- The following figure provides two illustrations of how value is impacted by the veracity of data and the time of generated results:



**Figure 1.15** Data that has high veracity and can be analyzed quickly has more value to a business.



# Data types processed by Big Data solutions

# Data types processed by Big Data solutions

- These data types refer to the internal organization of data and are sometimes called data formats. The primary types of data are:
  1. Structured data
  2. Unstructured data
  3. Semi-structured data

# Data types processed by Big Data solutions

## 1. Structured data:

- Structured data is data that adheres to a pre-defined **data model** or **schema** and is often stored in tabular form.
- It is used to capture relationships between different entities and is therefore most often stored in a relational database.

	branch	employee_type	number
1	Wichita Falls	Back Office	19
2	Wichita Falls	Credit Specialist	20
3	Wichita Falls	Financial Sevices Sales	16
4	Wichita Falls	Business Development Manager	1
5	Wichita Falls	Head of Sales Group	2
6	San Antonio	DSA	1
7	San Antonio	Back Office	56
8	San Antonio	Deputy Regional Director	2
9	San Antonio	Credit Specialist	96
10	San Antonio	Financial Sevices Sales	20



# Data types processed by Big Data solutions








## 2. Unstructured data:

- Data that **does not conform to a data model** or data schema is known as unstructured data.
- One of the most common types of unstructured data is text collected in a wide range of forms, including Word documents, email messages, PowerPoint presentations, survey responses, transcripts of call center interactions, and posts from blogs and social media sites.
- Other types of unstructured data include images, audio and video files

# Data types processed by Big Data solutions

## 2. Unstructured data:

### Unstructured data types

 Text files and documents	 Server, website and application logs	 Sensor data	 Images
 Video files	 Audio files	 Emails	 Social media data

Example of unstructured data:  
video about Antarctica expedition



# Data types processed by Big Data solutions

## 3. Semi-structured data:

- Semi-structured data has a defined level of structure and consistency, but is not relational in nature.
- XML and JSON files are common forms of semi-structured data.
- Due to the textual nature of this data and its conformance to some level of structure, it is more easily processed than unstructured data.

```
        rdf:resource="#_55D3DE366B2AD032" />
    </cim:Substation>
- <cim:EnergyConsumer rdf:ID="_2963867E4A4B1669">
    <cim:EnergyConsumer.pfixed>19.78</cim:EnergyConsumer.pfixed>
    <cim:EnergyConsumer.qfixed>06.10</cim:EnergyConsumer.qfixed>
    <cim:EnergyConsumer.LoadArea
        rdf:resource="#_9C1602456B178B75" />
    <cim:Equipment.MemberOf_EquipmentContainer
        rdf:resource="#_A9D1427B3784CD78" />
    <cim:Naming.name>4711</cim:Naming.name>
    ..
```



# Big Data Sources

# Big Data Sources

- In any big data application, before the data is processed and analyzed, it must be captured from the raw data sources into the big data systems and frameworks. Some of the examples of raw big data sources include:
- **Logs:** Logs generated by web applications and servers which can be used for performance monitoring
  - **Transactional Data:** Transactional data generated by applications such as eCommerce, Banking and Financial
  - **Social Media:** Data generated by social media platforms
  - **Sensor Data:** Sensor data generated by Internet of Things (IoT) systems
  - **Mobile Data:** which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones

# Big Data Sources

- In any big data application, before the data is processed and analyzed, it must be captured from the raw data sources into the big data systems and frameworks. Some of the examples of raw big data sources include (cont.):
- **Clickstream Data:** Clickstream data generated by web applications which can be used to analyze browsing patterns of the users
  - **Surveillance Data:** Sensor, image and video data generated by surveillance systems
  - **Medical data:** Medical data generated by medical and healthcare applications
  - **Network Data:** Network data generated by network devices such as routers and firewalls



# Basic Concepts

# Basic Concepts

- The use of the terms “**Data Science**”, “**Data Analytics**”, and “**Data Mining**” are becoming increasingly common along with “**Big Data.**”
- **Data Science:**
  - Data Science is a field focused around all what is related to data.
  - The term “Science” implies knowledge gained by systematic study.



# Basic Concepts

## ➤ Data Analytics:

Take Actions

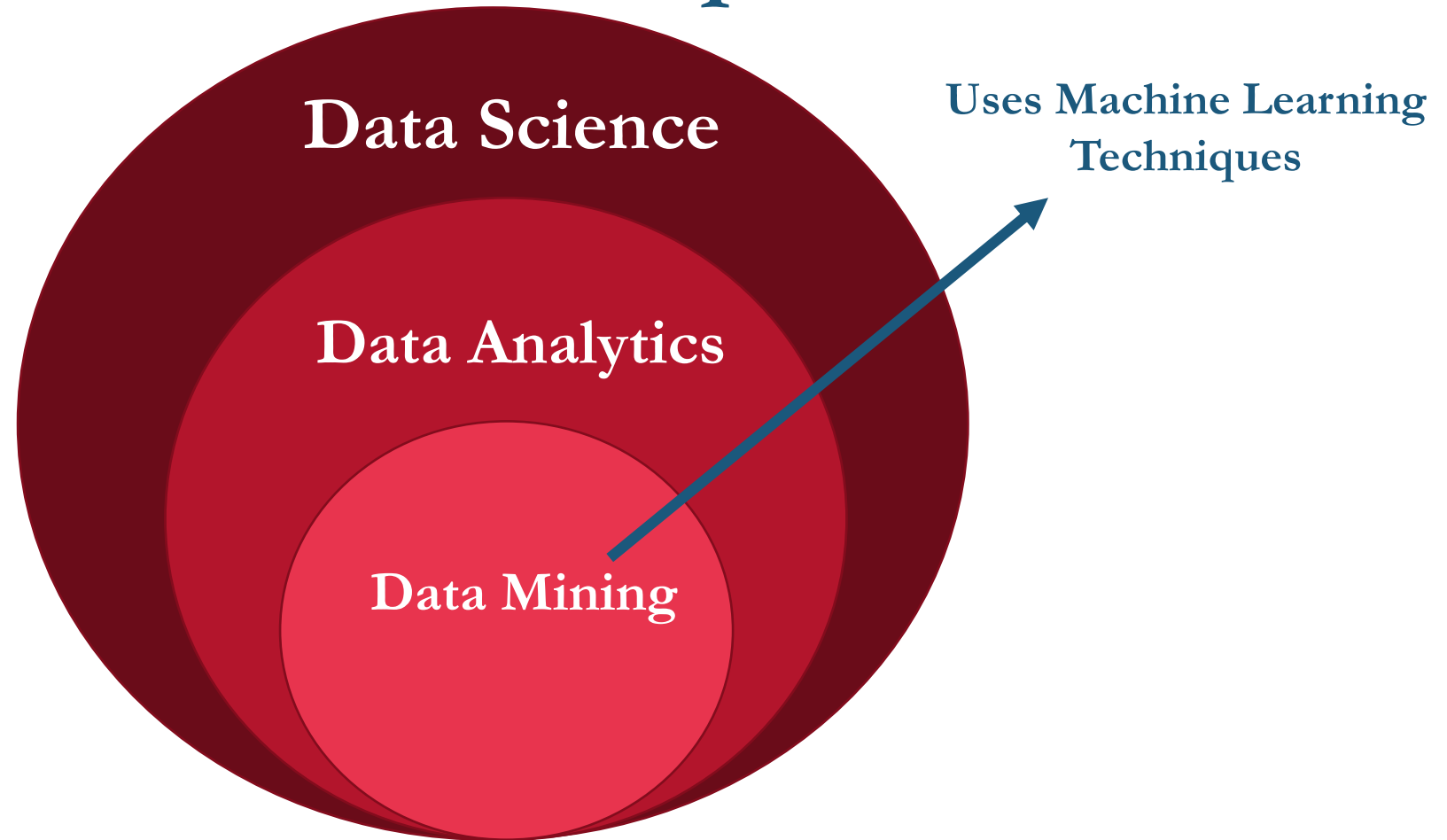
- Data analytics is defined as the application of computer systems to analyze large data sets for the **support of decisions** and **taking the right actions**.
- Data analytics helps analysts **draw conclusions** from the data.

## ➤ Data Mining:

Extract Knowledge

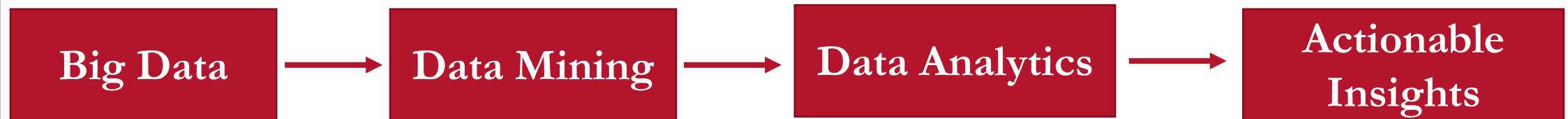
- The goal of data mining is to **extract knowledge** from data.
- In this context, knowledge is defined as **interesting patterns** that are generally valid, novel, useful, and understandable to humans.
- Data Analytics uses Data mining techniques to help achieve its goals.

# Basic Concepts



# Basic Concepts

➤ Therefore, the sequence of progression looks as follows:



# Basic Concepts

## ➤ Example:

- A data engineer decides to look into a supermarket's raw sales data (**Big Data**).
- After reviewing the data, the engineer discovers a high correlation of people buying burgers and fries on Sunday afternoon. (**Data Mining**)
- Data analytics can look into the correlation of people buying burgers and fries on Sunday afternoon and offer valuable insights to create targeted advertising campaigns. (**Data Analytics**)



# What is Data Analytics?

# What is Data Analytics?

- **Raw data** does not have a meaning until it is processed into useful **information**.
- This information obtained is then organized and structured to infer **knowledge** about the system and/or its users, its environment, and its operations and progress towards its objectives, thus making the systems smarter and more efficient.
- Data Analytics is this process of creating information and knowledge from raw data to find actionable insights.

# What is Data Analytics?

- **Data Analytics** is a broad term that includes the **processes, technologies, frameworks** and **algorithms** to extract meaningful insights from data.
- **Data Analytics** encompasses the **management of the complete data lifecycle**, which includes collecting, cleansing, organizing, storing, analyzing and governing data.
- In **Big Data** environments, data analytics has developed methods that allow analytics to occur through the use of highly scalable distributed technologies and frameworks that are capable of analyzing large volumes of data from different sources.

# What is Data Analytics?

- The choice of the technologies, algorithms, and frameworks for data analytics is driven by the analytics goals of the application.
- For example, the goals of the data analytics task may be:
  - (1) to **predict something** (for example sales for the coming month, whether it will rain on a particular day),
  - (2) to **find patterns** in the data (for example, finding the top 10 coldest days in the year or finding which pages are visited the most on a particular website),
  - (3) **finding relationships** in the data (for example, finding similar news articles, finding similar patients in an electronic health record system, finding related products on an eCommerce website, or finding correlation between news items and stock prices).
  - ..etc.



# What is Data Analytics?

- **Data Analytics** enable data-driven *decision-making* with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone.
- Different kinds of organizations use data analytics tools and techniques in different ways. Take, for example, these three sectors:
  - In **business-oriented environments**, data analytics results can lower operational costs and facilitate strategic decision-making.
  - In **the scientific domain**, data analytics can help identify the cause of a phenomenon to improve the accuracy of predictions.
  - In **service-based environments** as public sector organizations, analytics can help focusing on delivering high-quality services by driving down costs.



# Categories of Data Analytics

# Categories of Data Analytics

- There are four general categories of analytics that are distinguished by the results they produce:
  - Descriptive Analytics
  - Diagnostic Analytics
  - Predictive Analytics
  - Prescriptive Analytics

# Categories of Data Analytics

## ➤ Descriptive Analytics: *“What has happened?”*

- Descriptive analytics are carried out to **answer questions about events that have already occurred.**
- These help in describing patterns in the data and present the data in a summarized form.
- Sample questions can include:
  - What was the sales volume over the past 12 months?
  - What is the number of support calls received as categorized by severity and geographic location?
  - What is the monthly commission earned by each sales agent?

# Categories of Data Analytics

## ➤ Diagnostic Analytics: *“Why did it happen?”*

- Diagnostic analytics aim to **determine the cause of a phenomenon** that occurred in the past using questions that focus on the reason behind the event.
- The goal of this type of analytics is to **determine what information is related to the phenomenon** in order to enable answering questions that seek to determine why something has occurred.
- Sample questions can include:
  - Why were Q2 sales less than Q1 sales?
  - Why have there been more support calls originating from the Eastern region than from the Western region?
  - Why was there an increase in patient re-admission rates over the past three months?

# Categories of Data Analytics

## ➤ Predictive Analytics: *“What is likely to happen?”*

- Predictive analytics includes **predicting the occurrence of an event** or the likely outcome of an event or **forecasting the future values** using prediction models.
- Sample questions can include:
  - What are the chances that a customer will default on a loan if they have missed a monthly payment?
  - What will be the patient survival rate if medicine B is administered instead of medicine A?
  - If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?

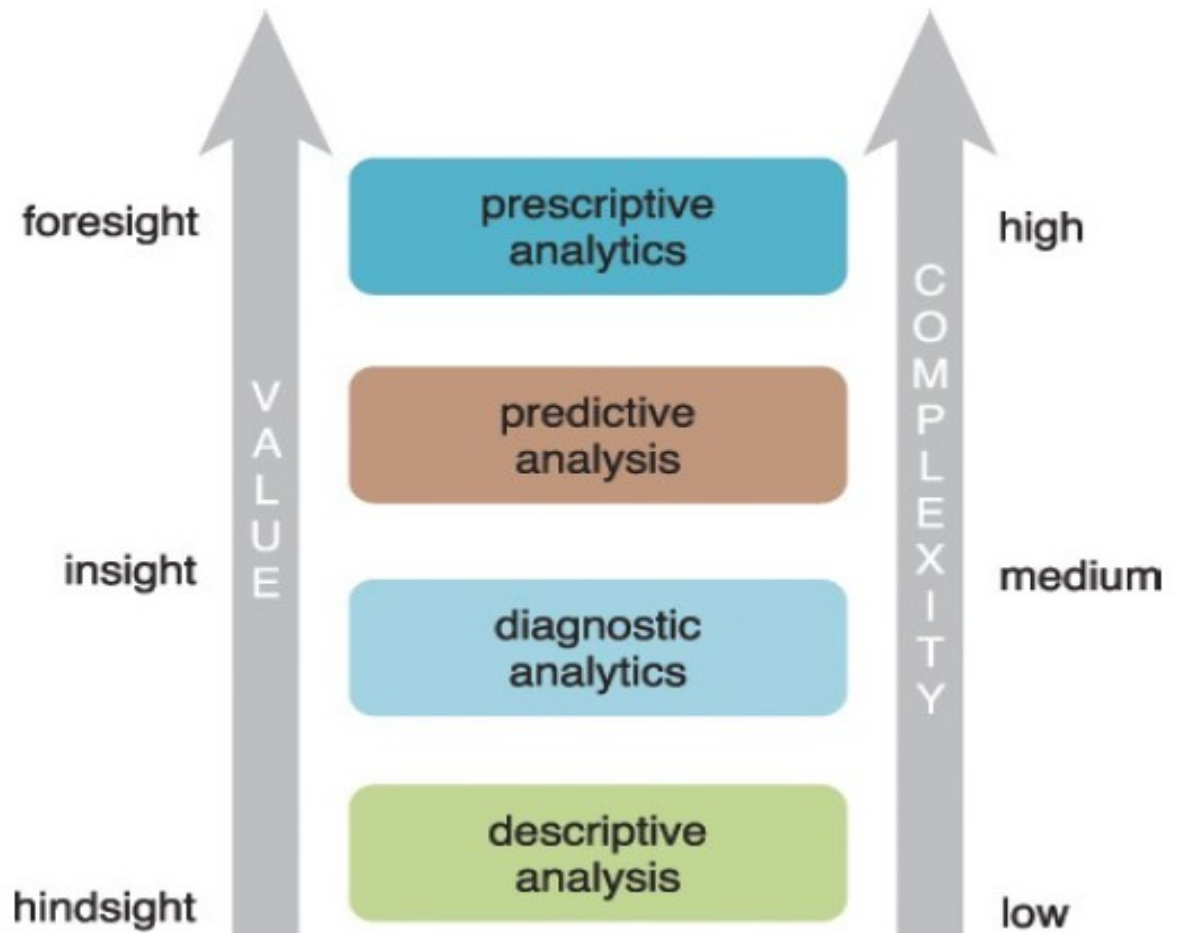
# Categories of Data Analytics

## ➤ Prescriptive Analytics: *“What can we do to make it happen??”*

- Prescriptive analytics build upon the results of predictive analytics by **prescribing actions that should be taken.**
- Prescriptive analytics uses multiple prediction models to **predict various outcomes and the best course of action for each outcome.**
- Sample questions can include:
  - Among three medicines, which one to produce?
  - When is the best time to trade a particular stock?

# Categories of Data Analytics

- The shown figure demonstrates the increase in **Value** and **Complexity** from descriptive to prescriptive analytics.
- **Hindsight:** understanding of a situation after it has happened.
- **Insight:** gain an accurate and deep understanding of something.
- **Foresight:** the ability to predict what will happen or what will be needed in the future.







# Adoption of Analytics in Business

# Terminologies

## ➤ Business Intelligence (BI):

- BI enables an organization to **gain insight into the performance** of an enterprise **by analyzing data** generated by its business processes and information systems.
- The results of the analysis can be used by management to steer the business in an effort to **correct detected issues** or otherwise **enhance organizational performance**.

# Terminologies

## ➤ Key Performance Indicators (KPIs):

- A KPI is a metric that can be used to **measure success** within a particular business context.
- KPIs therefore act as **quantifiable reference** points for measuring a specific aspect of a business' overall performance.

# Terminologies

## ➤ Key Performance Indicators (KPIs):

- KPIs are often displayed via a **KPI dashboard** as shown in the figure.
- The dashboard consolidates the display of multiple KPIs and compares the **actual measurements** with **threshold values** that define the acceptable value range of the KPI.



KPI dashboard

# Terminologies

## ➤ **Key Performance Indicators (KPIs):**

- Each department within an enterprise will use different KPI types to measure success based on specific business goals and targets.
- Examples of KPIs:
  - Monthly Sales Growth,
  - Average Profit Margin,
  - Lifetime value of a customer,
  - Customer retention

# Terminologies

## ➤ Performance Indicators (PIs):

- PIs are different in that they simply track the status of a specific business process while KPIs track whether you hit business objectives/targets, and metrics track processes.

# Terminologies

## ➤ **Critical Success factors (CSFs):**

- Examples of Critical Success factors
  - Training and education.
  - Quality data and reporting.
  - Management commitment, customer satisfaction.
  - Staff Orientation.
  - Continuous improvement.

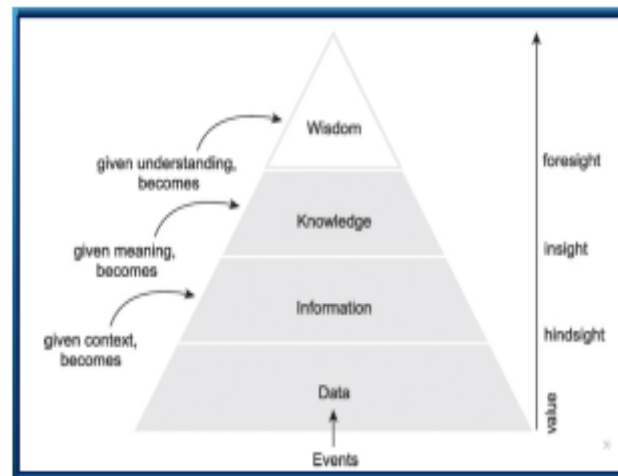
# Adoption of Analytics in Business

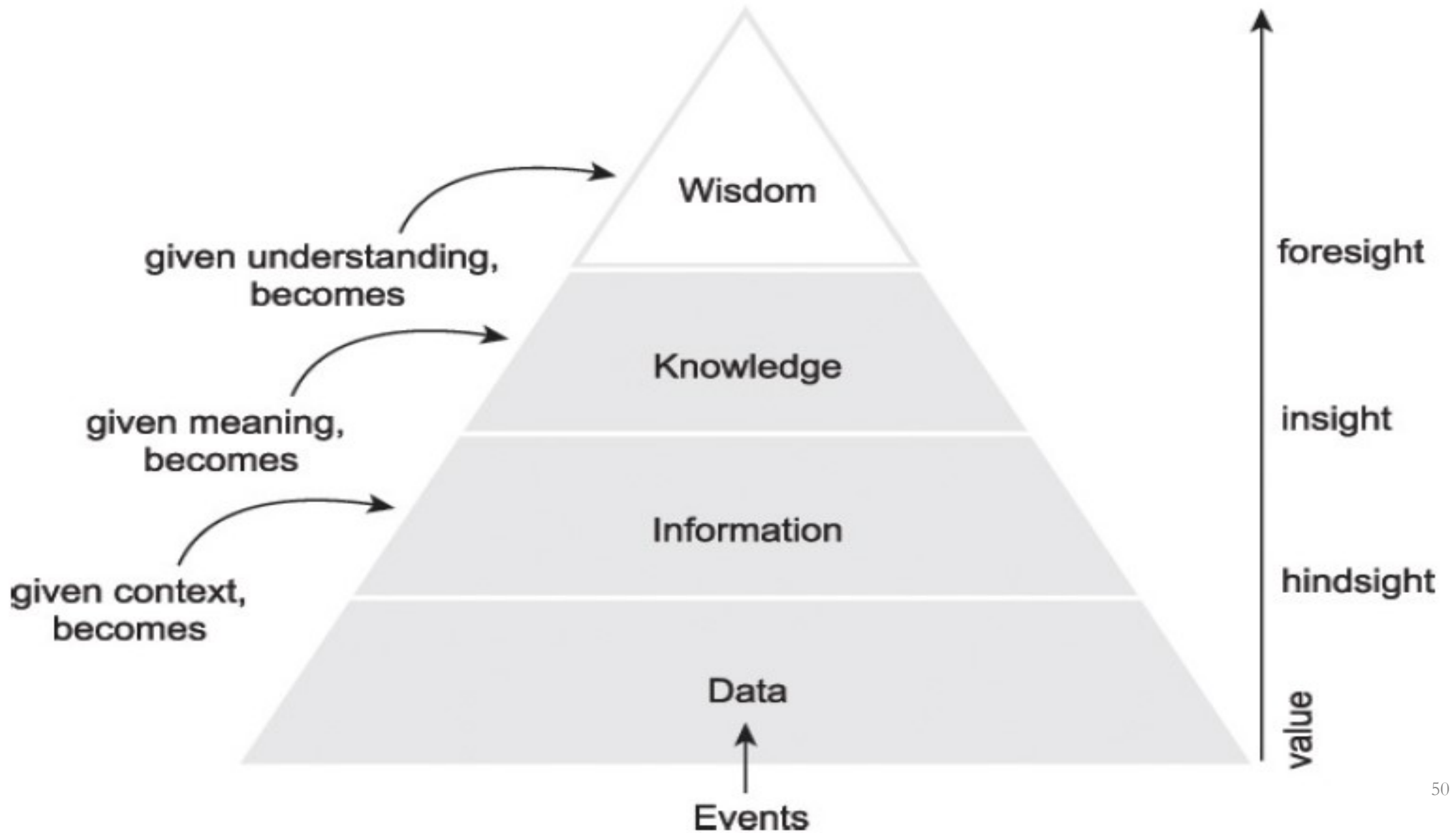
- Companies need to expand their BI activities **beyond retrospective** reflection on internal information extracted from their corporate information systems.
- Recognizing that external data brings additional context to their internal data allows a corporation to **move up the analytic value chain** from hindsight to insight with greater ease.
- With appropriate tooling, a company can **develop analytic results that provide foresight**.
- **This is the power of Big Data**—enriching corporate perspective beyond introspection.



# Adoption of Analytics in Business

- The transition from hindsight to foresight can be understood through the lens of the DIKW (Data Information Knowledge Wisdom) pyramid depicted in the figure:





# Adoption of Analytics in Business

## ➤ Examples of DIKW pyramid:

- **Example1:**
  - **Data:** 1815 feet, CN Tower, Toronto
  - **Information:** The CN Tower in Toronto is 1815 feet tall
  - **Knowledge:** Elevator can be used to ascend the building since CN tower is around 147 floors from the info of its height
  - **Wisdom:** If the elevator is not working, then do not use the stairs and go another day

# Adoption of Analytics in Business

## ➤ Examples of DIKW pyramid:

- Example2:

- **Data:**

- Employees – Ben; Anna; Mark; Kathy; Rose; Jack; Jane ...
- Departments – Accounting, Sales, Human Resources,....

- **Information:**

- Ben, Anna, Mark works in the Accounting Dept.
- Kathy, Rose, Jack works in the Sales Dept.
- Jane works in the Human Resources Dept.

- **Knowledge:**

- The company should hire a number of employees in all Depts.

- **Wisdom:**

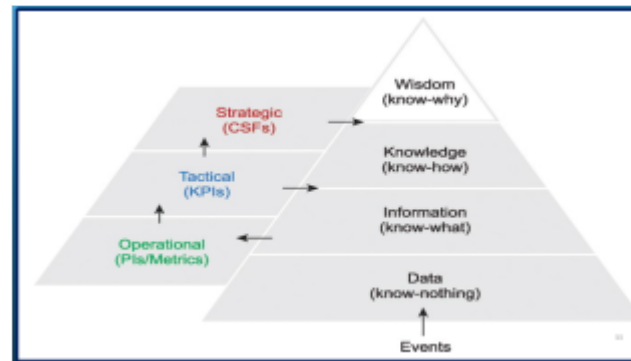
- HR Dept. has only one Employee to handle the recruitment process, so we need to focus now on hiring for the HR Dept.

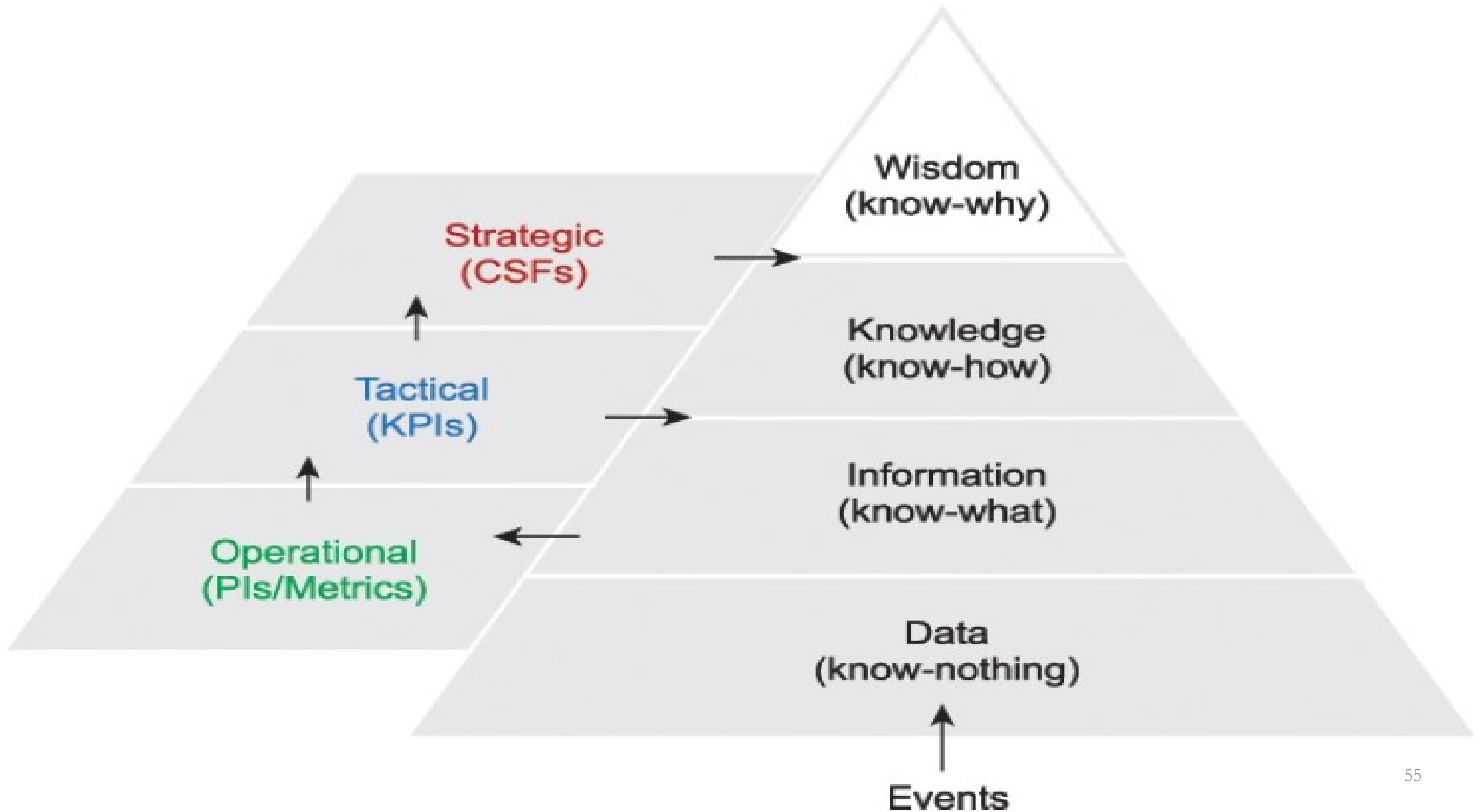
# Adoption of Analytics in Business

- A business operates as a layered system—the top layer is the **strategic layer** occupied by C-level executives and advisory groups.
- The middle layer is the **tactical or managerial layer** that seeks to steer the organization in alignment with the strategy.
- The bottom layer is the **operations layer** where a business executes its core processes and delivers value to its customers.
- These three layers often exhibit a degree of independence from one another, but each layer's goals and objectives are influenced by another.

# Adoption of Analytics in Business

- The following DIKW pyramid illustrates alignment with Strategic, Tactical and Operational corporate levels:





# Adoption of Analytics in Business

- For instance, **at the operational level**, metrics are generated that simply report on *what* is happening in the business. *Data → Information*.
- **At the managerial level**, this information can be examined through the lens of corporate performance to answer questions regarding *how* the business is performing. *Information → Knowledge*.
- This information may be further enriched to answer questions regarding *why* the business is performing at the level it is. The **strategic layer** can then provide further insight to help answer questions of which strategy needs to change or be adopted in order to correct or enhance the performance. *Knowledge → Wisdom*.



# Adoption of Analytics in Business

- Big Data has ties to business architecture at each of the organizational layers.
- Big Data enhances value as it helps convert data into information and provide meaning to generate knowledge from information.



# Data Analytics Lifecycle

# Data Analytics Lifecycle

➤ The Data analytics lifecycle can be divided into the following nine stages:

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analytics
8. Data Visualization
9. Utilization of Analytics Results

# Data Analytics Lifecycle

## 1. Business Case Evaluation:

- Each Data analytics lifecycle must begin with a **well-defined business case** that presents a clear understanding of the *justification*, *motivation* and *goals* of carrying out the analysis.
- The Business Case Evaluation stage requires that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks.

# Data Analytics Lifecycle

## 2. Data Identification:

- Identifying the datasets required for the analysis project and their sources.
- Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be **internal** and/or **external** to the enterprise.

# Data Analytics Lifecycle

## 3. Data Acquisition and Filtering:

- During the Data Acquisition and Filtering stage, the data is **gathered** from all of the data sources that were identified during the previous stage.
- The acquired data is then subjected to automated **filtering** for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.

# Data Analytics Lifecycle

## 4. Data Extraction:

- Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution.
- The need to address disparate types of data is more likely with data from external sources.
- The Data Extraction lifecycle stage, is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.

# Data Analytics Lifecycle

## 5. Data Validation and Cleansing:

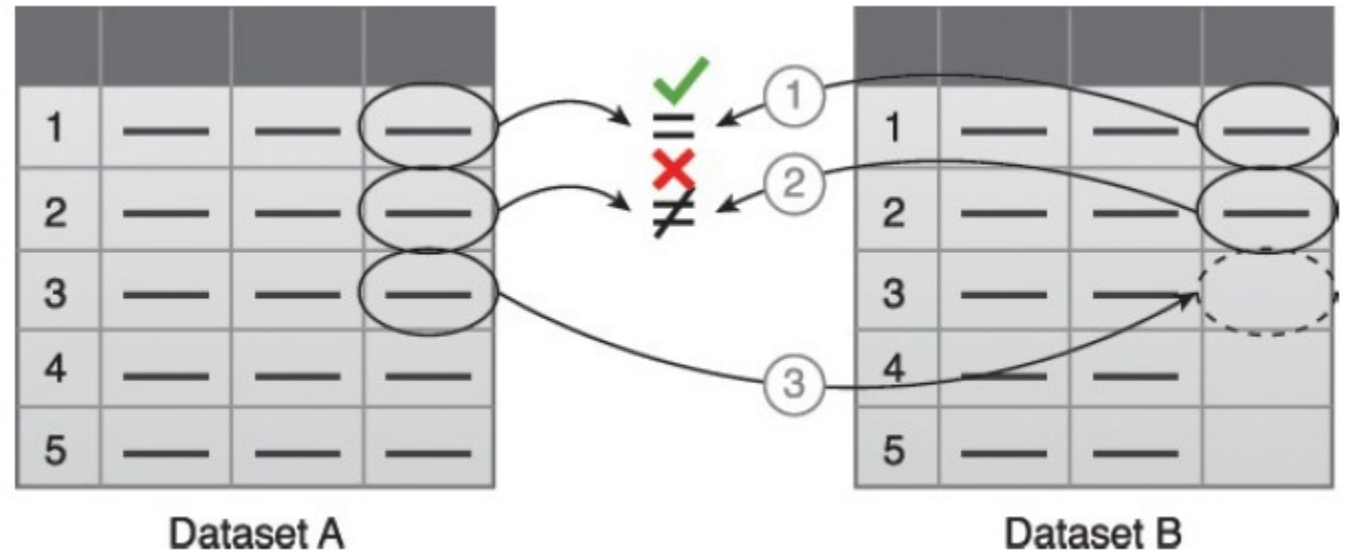
- Data input into Big Data analyses can be unstructured without any indication of validity.
- The Data Validation and Cleansing stage is dedicated to **establishing often complex validation rules** and **removing any known invalid data**.
- Big Data solutions often receive redundant data across different datasets.
- This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data.



# Data Analytics Lifecycle

## 5. Data Validation and Cleansing:

- For example, as illustrated in the figure, the first value in Dataset B is validated against its corresponding value in Dataset A.
- The second value in Dataset B is not validated against its corresponding value in Dataset A.
- If a value is missing, it is inserted from Dataset A.



# Data Analytics Lifecycle

## 5. Data Validation and Cleansing:

- Another example, the presence of invalid data is resulting in spikes.
- Although the data appears abnormal, it may be indicative of a new pattern.

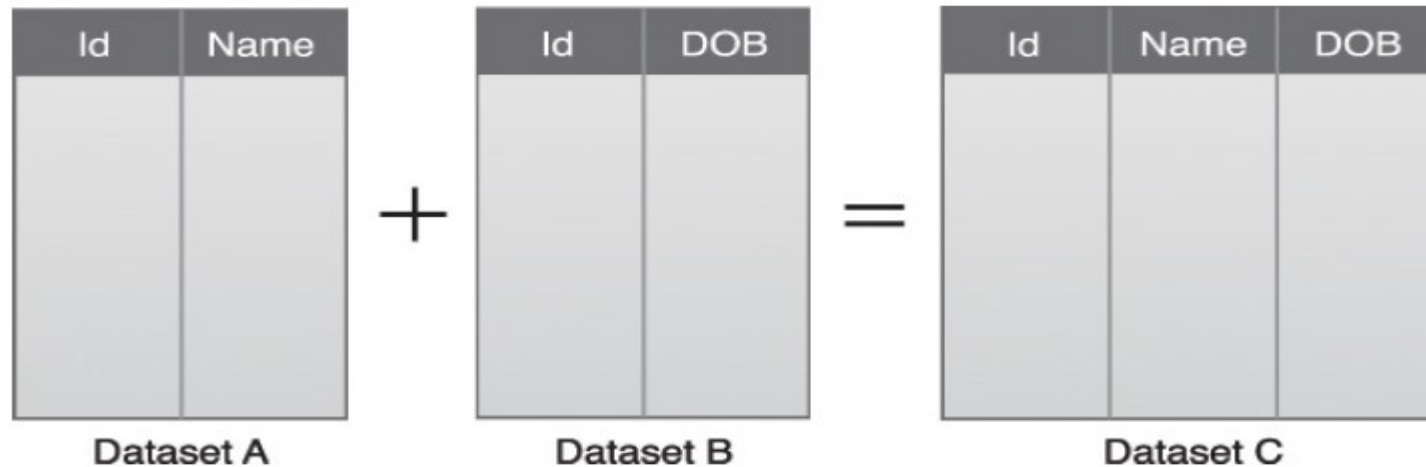


# Data Analytics Lifecycle

## 6. Data Aggregation and Representation:

- This stage is dedicated to integrating multiple datasets together to arrive at a unified view.
- Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID.

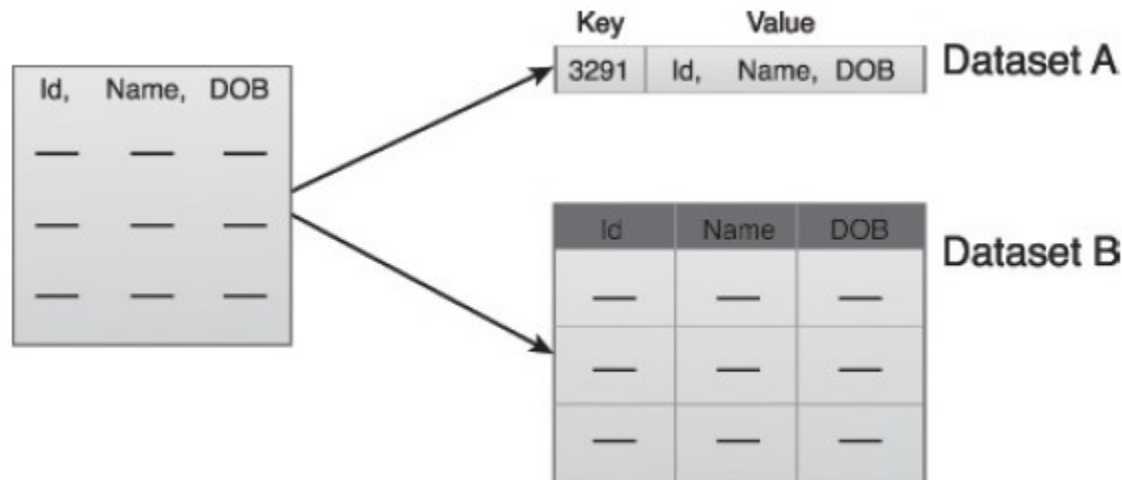
- Example:



# Data Analytics Lifecycle

## 6. Data Aggregation and Representation:

- Performing this stage can become complicated because of differences in:
  - *Data Structure* – Although the data format may be the same, the data model may be different.
  - *Semantics* – A value that is labeled differently in two different datasets may mean the same thing, for example “surname” and “last name.”
- Example:



# Data Analytics Lifecycle

## 7. Data Analytics:

- This is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature.
- Data analysis can be classified as confirmatory analysis or exploratory analysis:
  - **Confirmatory data analysis** is a *deductive* approach where the cause of the phenomenon being investigated is proposed beforehand. The proposed cause or assumption is called a hypothesis. The data is then analyzed to prove or disprove the hypothesis and provide definitive answers to specific questions.
  - **Exploratory data analysis** is an *inductive* approach that is closely associated with **data mining**. No hypothesis or predetermined assumptions are generated. Instead, the data is explored through analysis to develop an understanding of the cause of the phenomenon.

# Data Analytics Lifecycle

## 8. Data Visualization:

- The Data Visualization stage, is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users.

# Data Analytics Lifecycle

## 9. Utilization of Analytics Results:

- This stage, is dedicated to determining how and where processed analysis data can be further leveraged.
- Depending on the nature of the problems being addressed, it is possible for the analytics results to produce “**models**” that encapsulate new insights and understandings about the nature of the **patterns and relationships** that exist within the data that was analyzed.
- A model may look like a **mathematical equation** or a **set of rules**.
- Models can be used to improve business process logic and application system logic, and they can form the basis of a new system or software program.



# Applications



# Applications

- The applications of big data span a wide range of domains including (but not limited to) homes, cities, environment, energy systems, retail, logistics, industry, agriculture, Internet of Things, healthcare, education and cybersecurity.
- We will now provide an overview of various applications of big data for some domains.

# Applications

## 1. **Web:**

- **Web Analytics:**

- Web analytics deals with collection and analysis of data on the user visits on websites and cloud applications.
- Analysis of this data can give insights about the user engagement and tracking the performance of online advertisement campaigns.

- **Content Recommendation:**

- Content delivery applications that serve content (such as music and video streaming applications), collect various types of data such as user search patterns and browsing history, and user ratings.
- Such applications can leverage big data systems for recommending new content to the users based on the user preferences and interests.

# Applications

## 1. Web:

- **Performance Monitoring:**

- For performance monitoring, various types of tests can be performed such as **load tests** (which evaluate the performance of the system with multiple users and workload levels), **stress tests** (which load the application to a point where it breaks down) and **soak tests** (which subject the application to a fixed workload level for long periods of time).
- Big data systems can be used to analyze the data generated by such tests, to predict application performance under heavy workloads and identify bottlenecks in the system so that failures can be prevented.

# Applications

## 2. Financial:

- **Credit Risk Modeling:**

- Banking and Financial institutions use credit risk modeling to score credit applications and predict if a borrower will fail to pay or not in the future.
- Big data systems can help in computing credit risk scores of a large number of customers on a regular basis.

- **Fraud Detection:**

- Banking and Financial institutions can leverage big data systems for detecting frauds such as credit card frauds, money laundering and insurance claim frauds.

# Applications

3. **Internet of Things (IoT):** IoT refers to things that are connected to the Internet. The "Things" in IoT are the devices which can perform remote sensing, triggering and monitoring.
- **Intrusion Detection:**
    - These systems use security cameras and sensors (such as PIR sensors and door sensors) to detect intrusions and raise alerts.
  - **Smart Parkings:**
    - Smart parkings are powered by IoT systems that detect the number of empty parking slots and send the information over the Internet to smart parking application back-ends.
    - These applications can be accessed by the drivers from smart-phones, tablets and in-car navigation systems.

# Applications

## 3. Internet of Things (IoT):

- **Structural Health Monitoring:**

- Systems use a network of sensors to monitor the vibration levels in the structures such as bridges and buildings.
- The data collected from these sensors is analyzed to assess the health of the structures.

# Applications

## 4. Industry:

- **Machine Diagnosis & Prognosis:**

- Machine prognosis refers to predicting the performance of a machine by analyzing the data on the current operating conditions and the deviations from the normal operating conditions.
- Machine diagnosis refers to determining the cause of a machine fault.

# Applications

5. **Retail:** Retailers can use big data systems for boosting sales, increasing profitability and improving customer satisfaction.
  - **Customer Recommendations:**
    - New products can be recommended to customers based on the customer preferences, and personalized offers and discounts can be given.
    - Customers with similar preferences can be grouped and targeted campaigns can be created for customers
  - **Forecasting Demand:**
    - Big data systems can be used to analyze the customer purchase patterns and predict demand and sale volumes



# Big Data Applications

## 6. Environment:

- Weather Monitoring
- Noise Pollution Monitoring
- Forest Fire Detection
- River Floods Detection
- Water Quality Monitoring



# Thank You