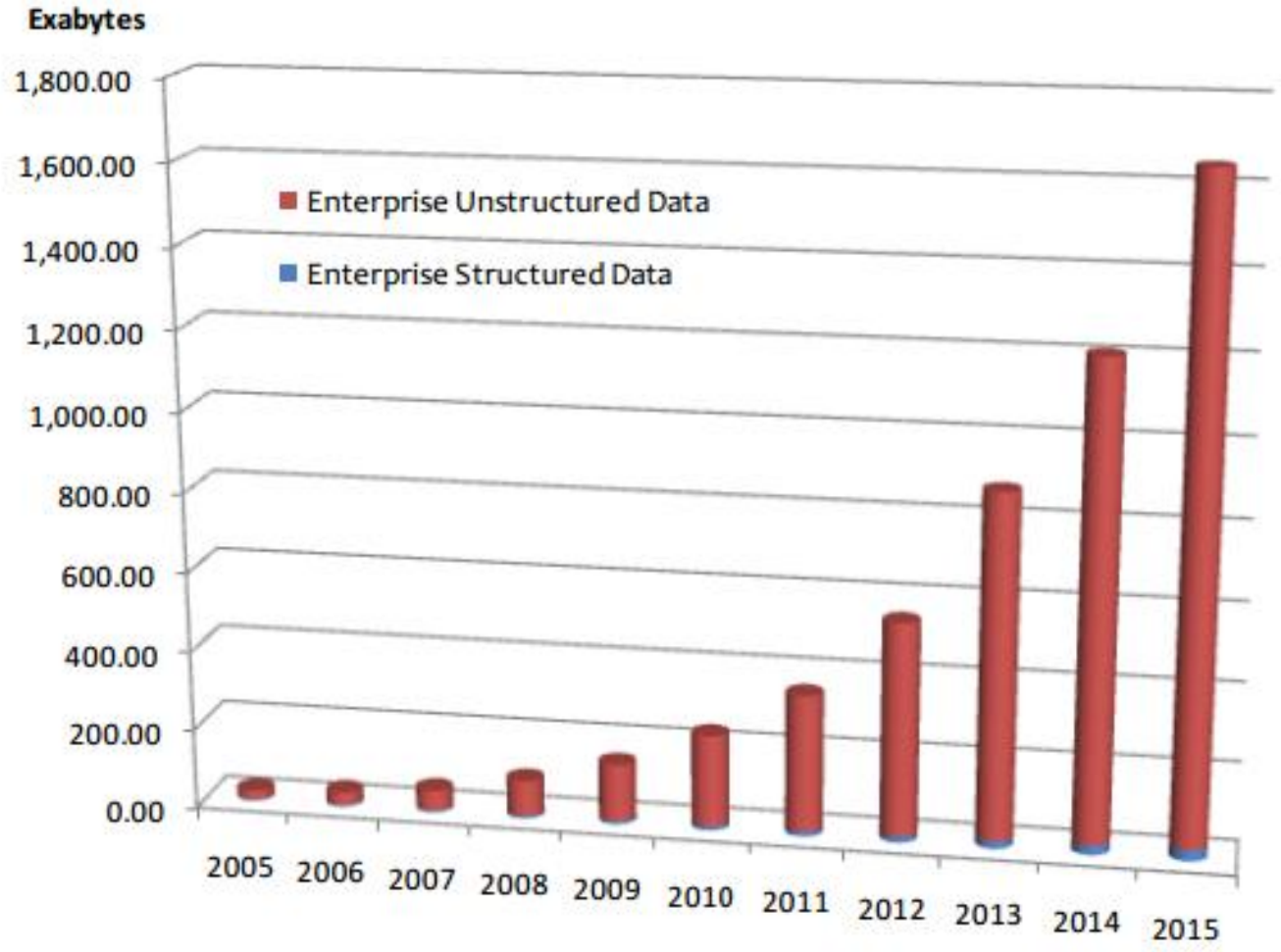
UNIT 1

- Data continues to be a **precise and irreplaceable asset** for enterprises
- Data is **present internal and outside four walls** of the enterprise
- Data present in **Homogeneous sources as well as in heterogeneous sources**

Approximate Distribution of Digital Data

Approximate percentage distribution of digital data



- Need of the hour is to **understand, manage, process** and **take the data for analysis to draw valuable insights**

Data -> Information
Information ->Insights

Classification of digital data

1. Unstructured data –
   - Data which **do not conform to a data model** or a form which **can not be used** by computer program.
   - 80-90% data is in this form
   - PPT, Images , Videos, body of an email etc

2. Semi structured data- **does not conform to a data model** but **has some structure** . **Not** in a form which can be used by a **computer program**.
   - Email, XML, HTML etc

3. Structured data

- Data in organised form
- Understandable by a computer program
- Eg- Data stored in databases

Structured data

-> When is data structured? –Data conforms to a pre defined schema/structure

-> Think **structured data** and **think data model**

|

-> Context of **RDBMS- Conforms to the relational model – rows/columns**

-> **Cardinality ratio** – Number of rows/columns

-> **Degree**- Number of columns

Steps

1) Design a relation/table

-> **Fields to stor**e -> **Data type**

2) **Constraints**- we would like our **data to conform**

– Unique

- Not Null

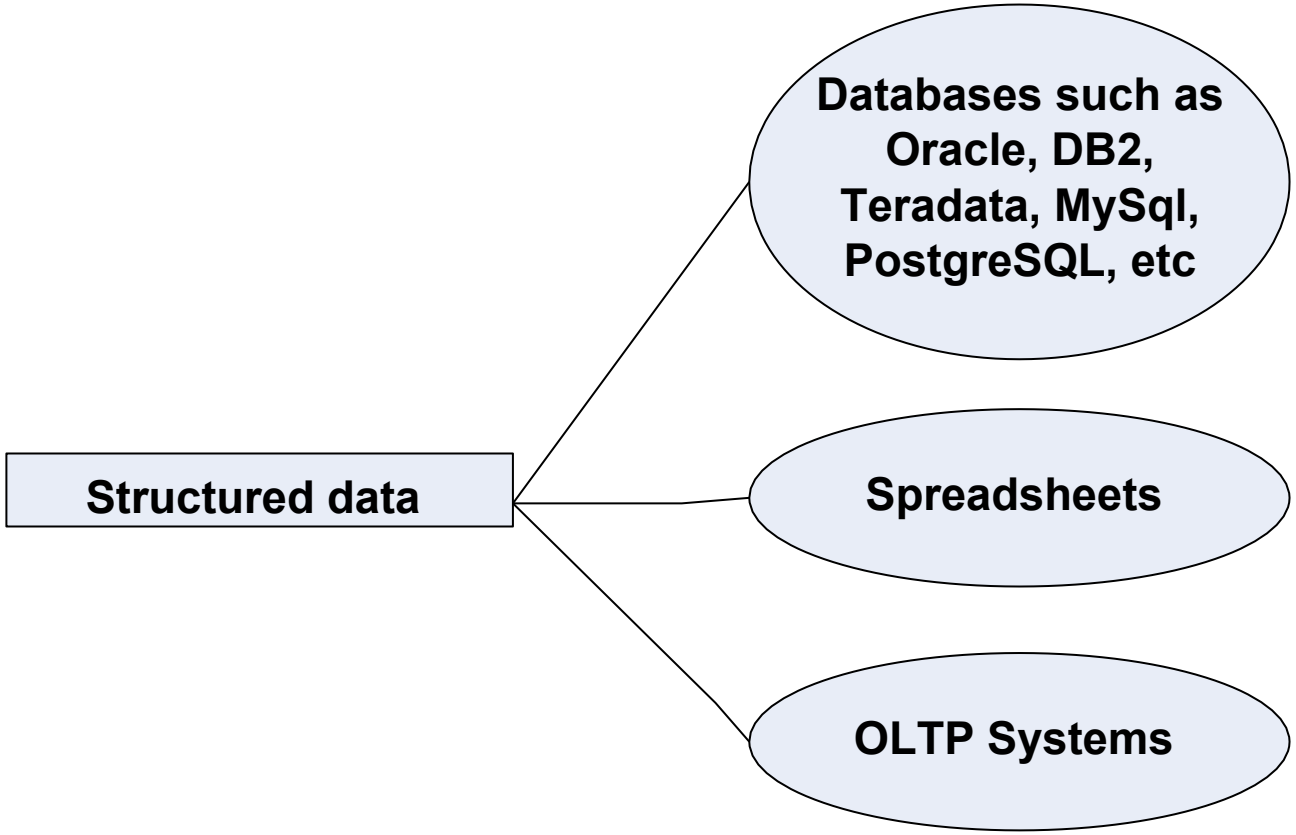
– business constraints

– permissible values a column should access

3) Figure

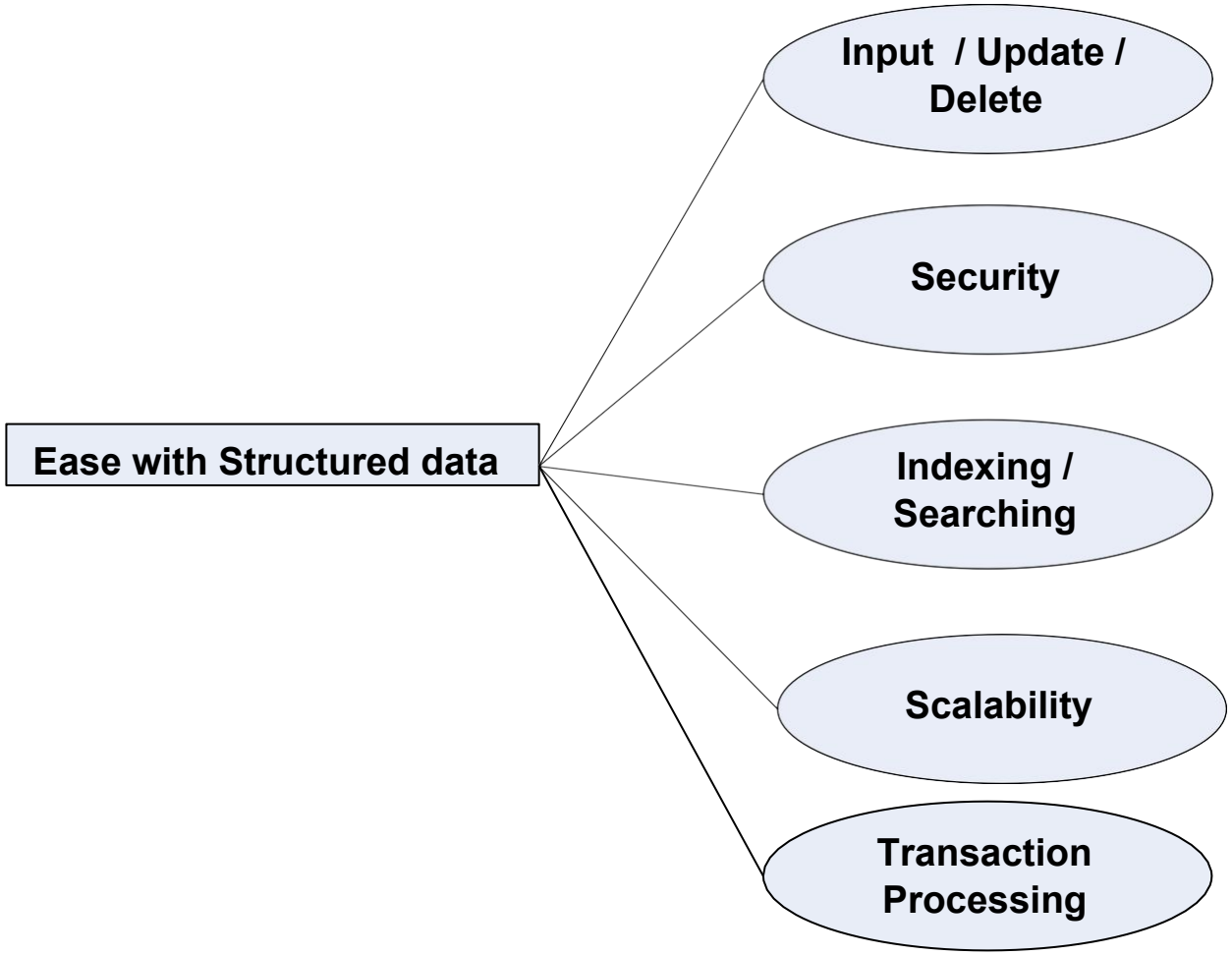
4)**Referential Integrity constraints**

Sources- When data is structured- leverage on available RDBMS



Database store operational data generated and collected by day to day business activities

Ease with Structured Data



ACID

Atomicity- T**ransaction happens in its eternity or none of it all**

Consistency- If **same information** is stored at **two or more** places **they are in complete agreement**

Isolation- **Resource allocation to transaction happen** such that the **transaction gets the impression** that it is the **only transaction happens in isolation**

Durability- **Changes made to the database during transaction is permanent**

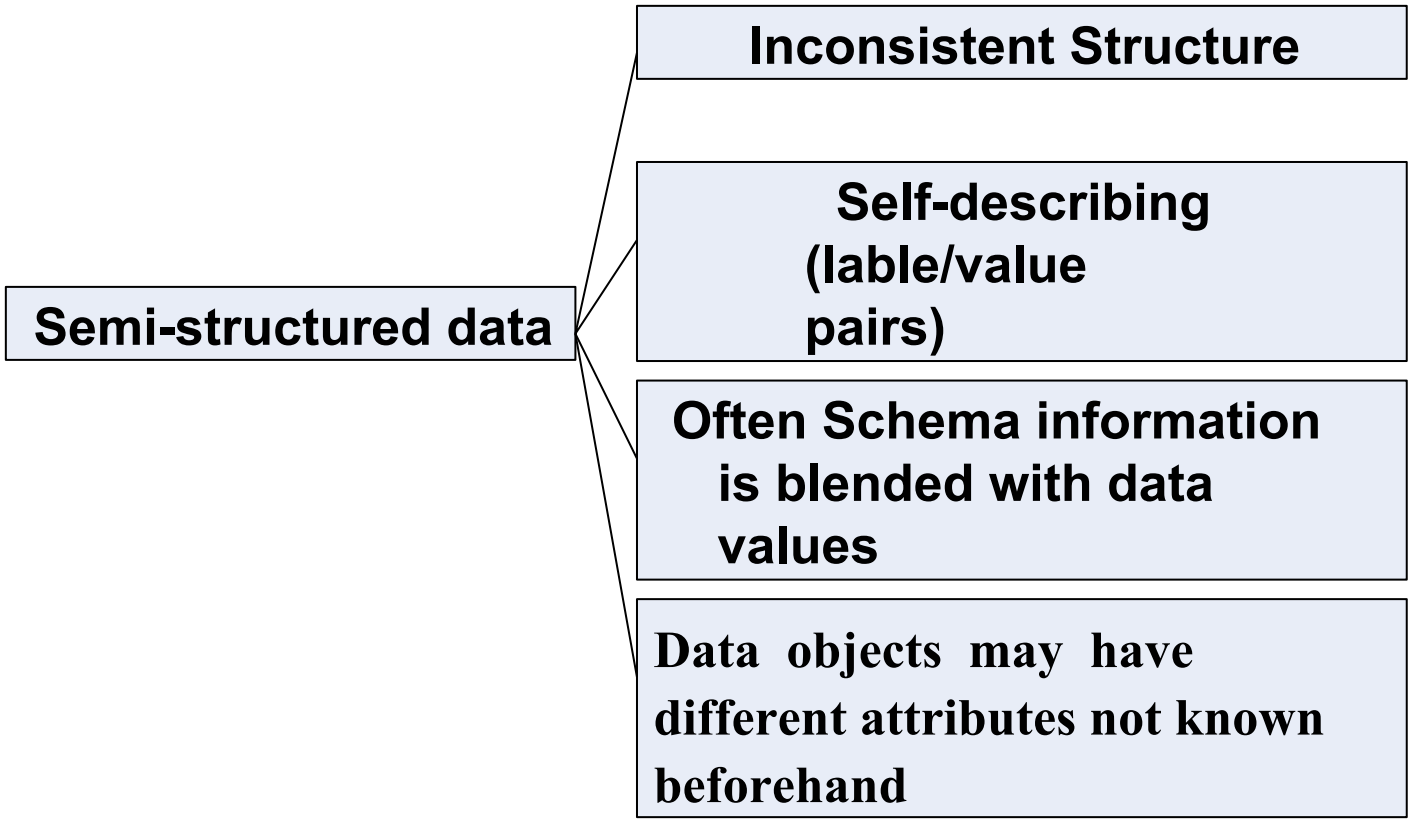
Semi structured data

- **Self describing structure**
  Example, emails, XML, markup languages like HTML, etc

Features

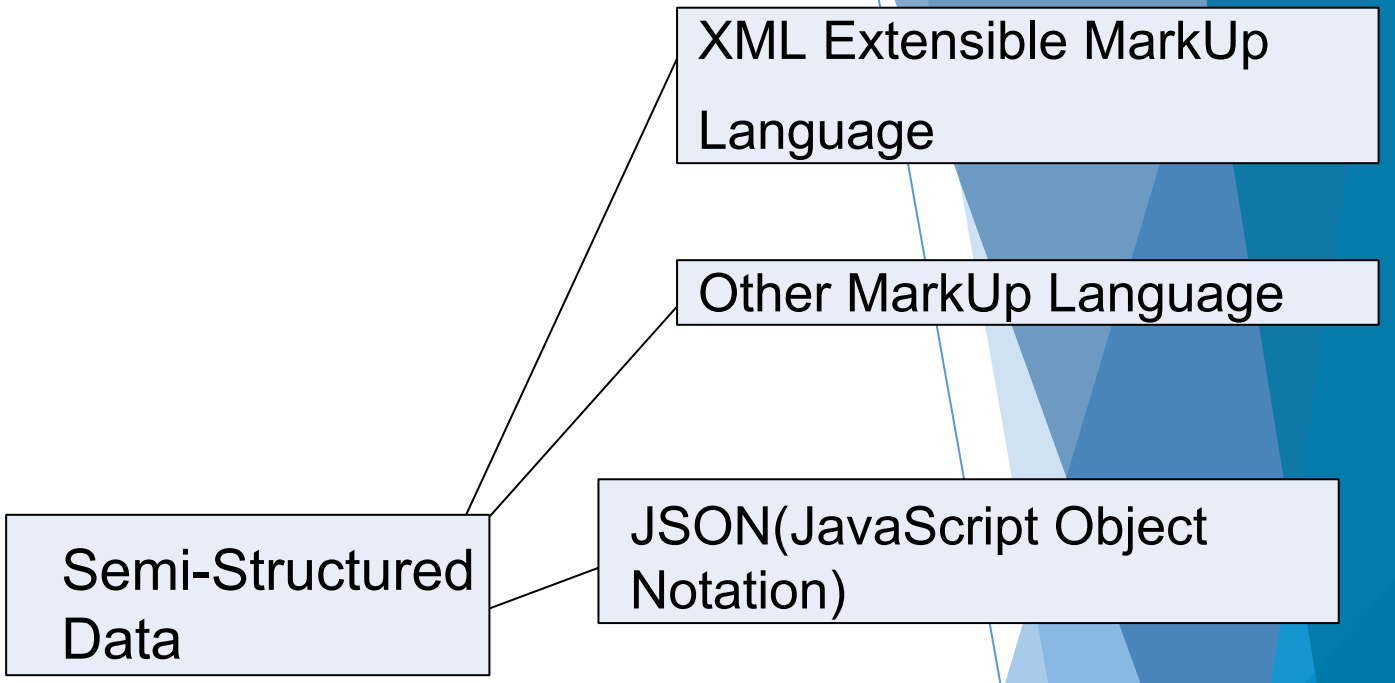
- **Does not conform to data model**
- Use **tags to segregate semantic elements**
- Tags are **also used to enforce hierarchies of records and fields within data**
- **No separation between data and schema**
- **Entities belong to the same class** and also grouped together **need not necessarily have same attributes- not necessary same order**

Characteristics of Semi-structured Data



Sources

- XML- **popularised by web services**
- JSON- t**ransmit data between server and web application**.
- Mongodb store data natively



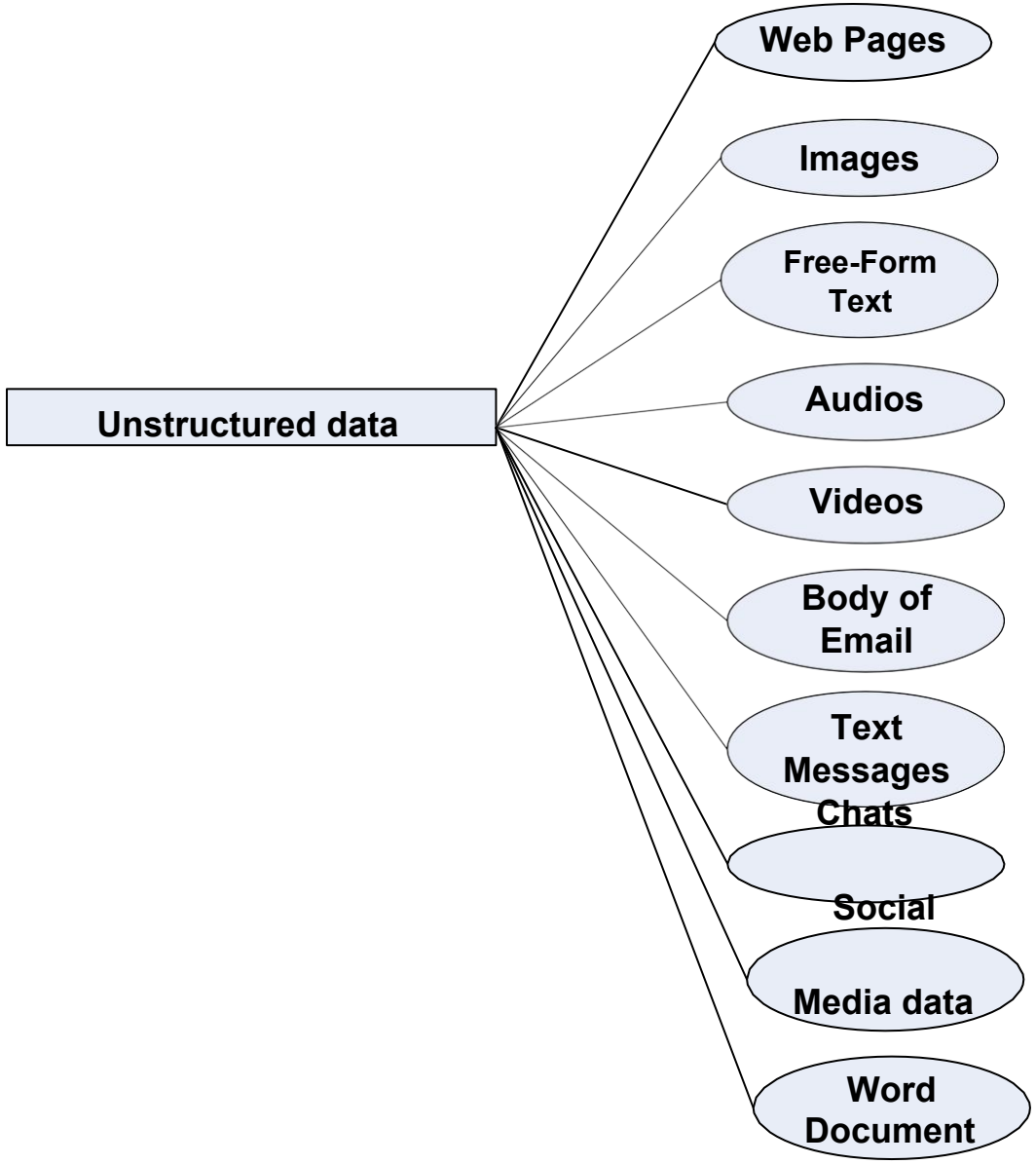
Unstructured Data

- This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.

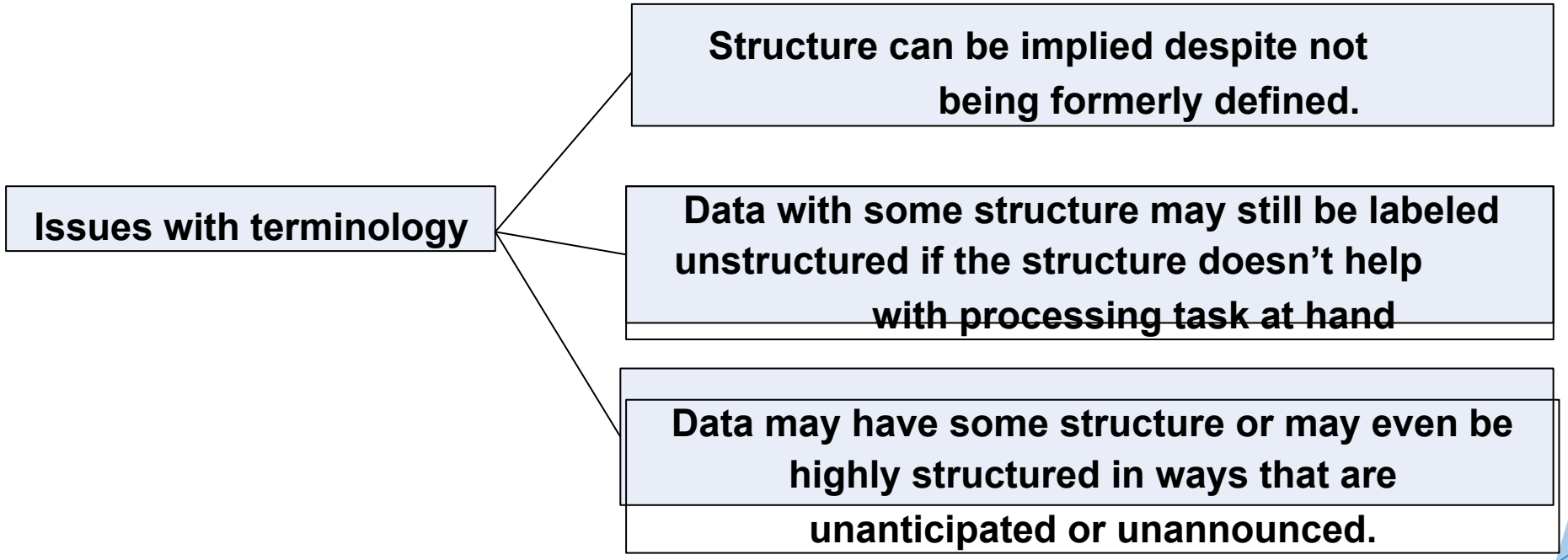
- About 80–90% data of an organization is in this format.

- Example: memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

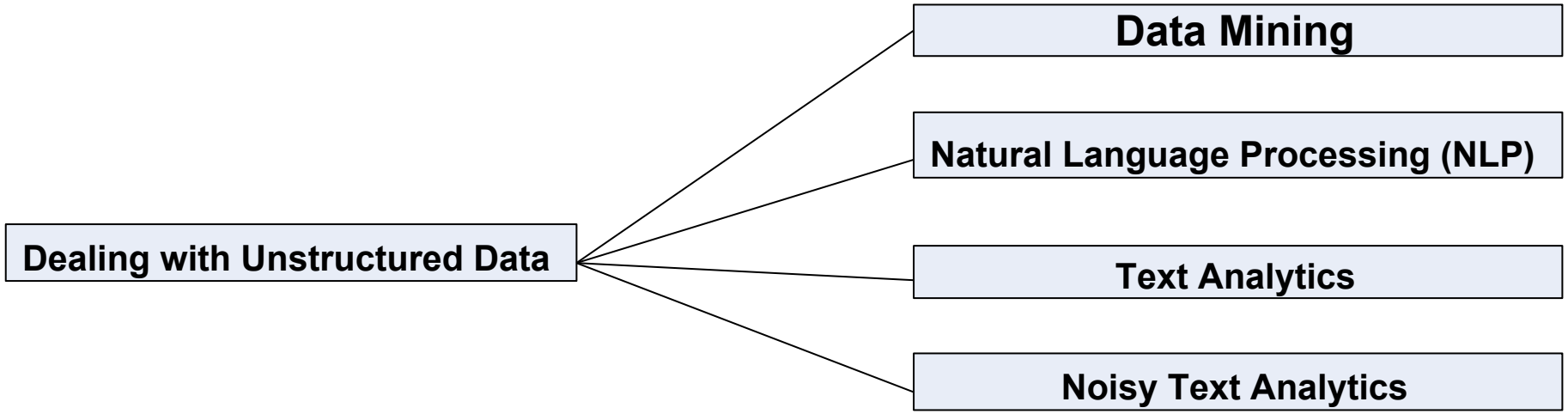
Sources of Unstructured Data



Issues with terminology - Unstructured Data



Dealing with Unstructured Data



Unstructured data

Data Mining

First we **deal with large data sets.**

Second **use methods at the intersection of AI, machine learning and statistics and database** to **unearth consistent patterns** in large data set and s**ystematic relation between variables**

Association rule mining- Market basket analysis- **what goes with what- bread , cheese**

Regression analysis- P**redict relationship between two variables – dependant and independent variable – value to be predicted – dependant**

Collaborative filtering – Predicting user preference based on preferences of group of users

Place Me in the Basket

Structured	Unstructured	Semi-Structured

Following words are to be placed in the relevant basket:

Email
MS Access
Images
Database
Chat conversations
Relations/Tables
Facebook
Videos
MS Excel
XML

Answer:

Structured	Unstructured	Semi-Structured
MS Access	Email	XML
Database	Images	
Relations/Tables	Chat conversations	
MS Excel	Facebook	
	Videos	

B. Match the Following

1.

Column A	Column B
NLP	Content analytics
Text analytics	Text messages
UIMA	Chats
Noisy unstructured data	Text mining
Data mining	Comprehend human or natural language input
Noisy unstructured data	Uses methods at the intersection of statistics, AI, machine learning & DB
IBM	UIMA

Answer:

Column A	Column B
NLP	Comprehend human or natural language input
Text analytics	Text mining
UIMA	Content analytics
Noisy unstructured data	Text messages
Data mining	Uses methods at the intersection of statistics, AI, machine learning & DBs
Noisy unstructured data	Chats
IBM	UIMA

What is Big Data?

Big data is **defined as collections of datasets whose volume, velocity or variety is so large that it is difficult to store, manage, process and analyze the data using traditional databases and data processing tools**.

In the recent years, there has been an e**xponential growth in the both structured and unstructured data generated by information technology, industrial, healthcare, Internet of Things, and other systems**

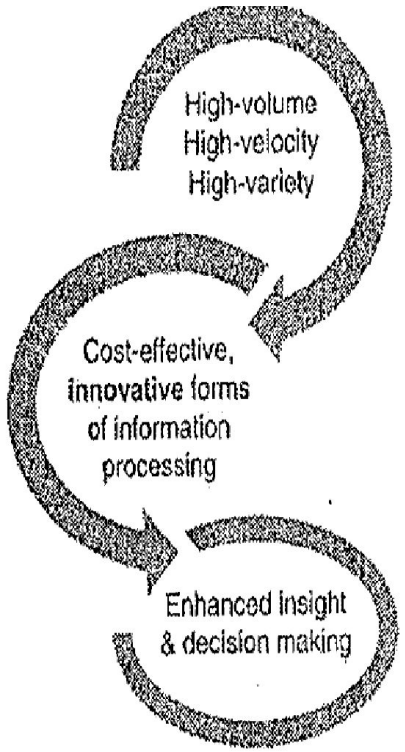
Definition of Big Data

- Big data **refers to**
  - **datasets whose size is typically beyond the storage capacity** of and also
  - **complex for traditional database software tools**

- Big data is **anything beyond the human & technical infrastructure** needed **to support storage, processing and analysis.**

 Big data is

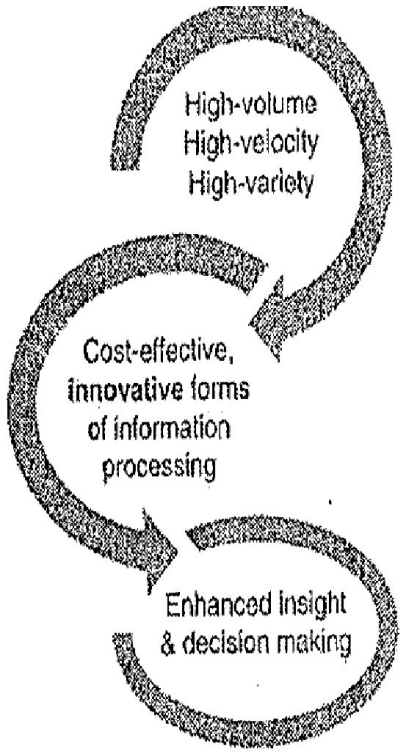
- high **volume, high-velocity and high-variety information** assets that
- demand **cost effective, innovative forms of information processing**
- for **enhanced insight and decision making**.



Part I of the definition:

talks about **voluminous data** that may have **great variety** will **require a good speed/pace for storage, preparation, processing and analysis**.

Part II of the definition:
talks about **embracing new techniques and technologies** to **capture** (ingest), **store, process, persist, integrate and visualize** the high-volume, high-velocity, and high-variety data.



Part III of the definition:

talks about **deriving deeper, richer and meaningful insights and then using these insights to make faster and better decisions** to gain **business value** and thus a **competitive edge.**

Data —> Information —> Actionable intelligence —> Better decisions —>Enhanced business value

Below are some key pieces of data from the report:

- Facebook users share nearly 4.16 million pieces of content
- Twitter users send nearly 300,000 tweets
- Instagram users like nearly 1.73 million photos
- YouTube users upload 300 hours of new video content
- Apple users download nearly 51,000 apps
- Skype users make nearly 110,000 new calls
- Amazon receives 4300 new visitors
- Uber passengers take 694 rides
- Netflix subscribers stream nearly 77,000 hours of video

Big Data analytics **deals with collection, storage, processing and analysis of this massive scale data**.

Specialized **tools and frameworks are required for big data analysis when:**

(1) the **volume of data involved is so large that it is difficult to store, process and analyze data on a single machine**

(2) the **velocity of data is very high and the data needs to be analyzed in real-time**,

(3) there is **variety of data involved, which can be structured, unstructured or semi-structured**, and is collected from multiple data sources,

(4**) various types of analytics need to be performed to extract value** from the data such as descriptive, diagnostic, predictive and prescriptive analytics

Big data analytics is **enabled by several technologies such as cloud computing, distributed and parallel processing frameworks, non-relational databases, in-memory computing**, for instance.

Some examples of big data are listed as follows:

- Data generated by social networks including text, images, audio and video data
- Click-stream data generated by web applications such as e-Commerce to analyze user behavior
- Machine sensor data collected from sensors embedded in industrial and energy systems
for monitoring their health and detecting failures
- Healthcare data collected in electronic health record (EHR) systems
- Logs generated by web applications
- Stock markets data

Characteristics of Big Data

Volume

Big data is a **form of data** whose **volume is so large that it would not fit on a single machine** therefore specialized tools and frameworks are required to store process and analyze such data.

The volumes of **data generated by modern IT, industrial, healthcare, Internet of Things, and other systems is growing exponentially driven by the lowering costs of data storage.**

Though **there is no fixed threshold for the volume of data to be considered as big data, however, typically, the term big data is used for massive scale data that is difficult to store, manage and process using traditional databases and data processing architectures**.

What is Big Data?

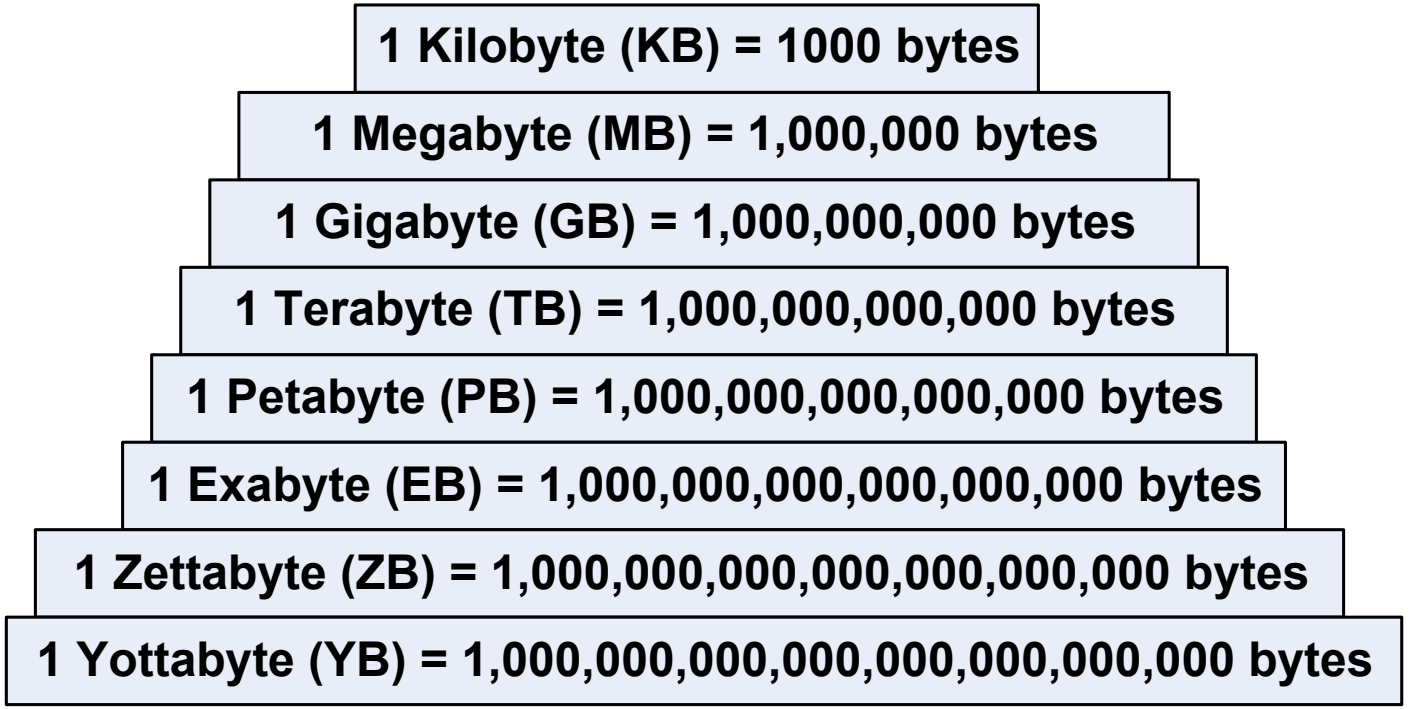
Volume

Bits-Bytes-Kilobytes-Megabytes-Gigabytes-Terabytes-Petabytes-Exabytes-Zettabytes-Yottabytes

Table 2.2 Growth of data

Bits	0 or 1
Bytes	8 bits
Kilobytes	1024 bytes
Megabytes	1024^2 bytes
Gigabytes	1024^3 bytes
Terabytes	1024^4 bytes
Petabytes	1024^5 bytes
Exabytes	1024^6 bytes
Zettabytes	1024^7 bytes
Yottabytes	1024^8 bytes

A Mountain of Data



Where does this data get generated

-Multitude of sources

- XLS, DOC, Youtube videos, Chat conversations, customer feedback CCTV coverage

1. **Typical Internal Data Sources- Data present within an organization firewall**
   - **Data Storage-** File systems, SQL, NoSQL..
   - **Archives-** Achieves of scanned documents , paper archives, customer correspondence records, patient's health records, Student admission records and so on.

2. **External Data Sources- Data residing an organization firewall**
   - **Public Web**: Wikepedia, weather, regulatory, census

3. **Both internal and external data sources**

- *Sensor data:* Car sensors, smart electric meters, office buildings, air conditioning units, refrigerators, and so on.
- *Machine log data:* Event logs, application logs, Business process logs, audit logs, clickstream data, etc.
- *Social media:* Twitter, blogs, Facebook, LinkedIn, YouTube, Instagram, etc.
- *Business apps:* ERP, CRM, HR, Google Docs, and so on.
- *Media:* Audio, Video, Image, Podcast, etc.
- *Docs:* Comma separated value (CSV), Word Documents, PDF, XLS, PPT, and so on.

Velocity

Velocity of data refers to **how fast the data is generated**.

Data generated by **certain sources can arrive at very high velocities, for example, social media data or sensor data.**

Velocity is another **important characteristic of big data and the primary reason for the exponential growth of data.**

High velocity of **data results in the volume of data accumulated** to become very large, in short span of time.

Some applications can have **strict deadlines for data analysis** (such as trading or online fraud detection) and the **data needs to be analyzed in real-time**.

Velocity

Batch → Periodic → Near real time → Real-time processing

Variety

Variety refers to the **forms of the data**.

Big data comes in **different forms such as structured, unstructured or semi-structured, including text data, image, audio, video and sensor data.**

Big data systems need to be **flexible** enough to **handle such variety of data**

- **Structured data**: example: traditional transaction processing systems and RDBMS, etc.

- **Semi-structured data**: example: Hyper Text Markup Language (HTML), eXtensible Markup Language (XML).

- **Unstructured data**: example: unstructured text documents, audio, video, email, photos, PDFs, social media, etc.

Veracity

Veracity refers to **how accurate is the data.**

To **extract value** from the data, the **data needs to be cleaned to remove noise**.

Data-driven applications can **reap the benefits** of big data only **when the data is meaningful and accurate.**

Therefore, cleansing of data is important so that incorrect and faulty data can be filtered out.

Value

Value of data **refers to the usefulness of data for the intended purpose**.

The **end goal of any big data analytics system is to extract value from the data.**

The **value of the data** is also related to the **veracity or accuracy of the data.**

For some applications value also depends on how fast we are able to process the data.

Analytics is a broad term that **encompasses the processes, technologies, frameworks and algorithms to extract meaningful insights from data.**

Analytics is this **process of extracting and creating information from raw data by filtering, processing, categorizing, condensing and contextualizing the data.**

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.**

The choice of the technologies, algorithms, and frameworks for analytics is driven by the analytics

The goals of the analytics task may be:

(1) **to predict something** (for example **whether a transaction is a fraud or not**, whether **it will rain on a particular day, or whether a tumor is benign or malignant**),

(2) **to find patterns in the data** (for example, finding the **top 10 coldest days** in the year, finding **which pages are visited the most on a particular website**, or finding the most searched celebrity in a particular year),

(3) **finding relationships in the data** (for example, finding similar news articles, finding similar patients in an electronic health record system

The National Research Council [1] has done a characterization of **computational tasks for massive data analysis** (called the seven “giants").

These computational tasks include:
(1) Basis Statistics, (2) Generalized N-Body Problems, (3) Linear Algebraic Computations, (4) Graph-Theoretic Computations, (5) Optimization,
(6) Integration and (7) Alignment Problems.

This characterization of computational tasks aims to **provide a taxonomy of tasks that have proved to be useful in data analysis and grouping them roughly according to mathematical structure and computational strategy**.

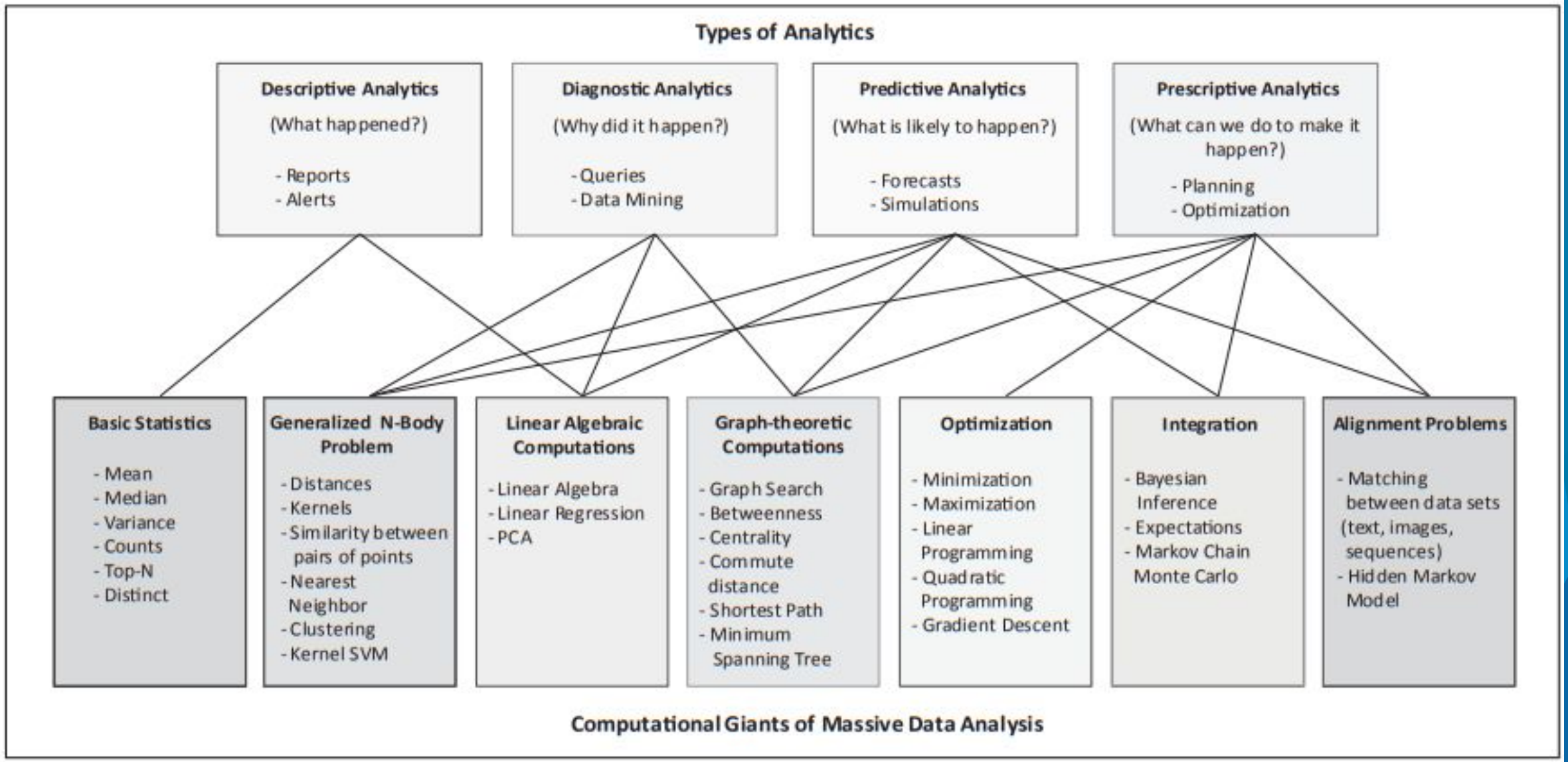


Figure 1.1: Mapping between types of analytics and computational tasks or 'giants'

Descriptive Analytics

Descriptive analytics **comprises analyzing past data to present it in a summarized form which can be easily interpreted.** Descriptive analytics aims to answer –
What has happened?

A major portion of analytics done today is **descriptive analytics through use of statistics functions such as counts, maximum, minimum, mean, topN,percentage, for instance.**

These statistics help in **describing patterns in the data and present the data in a summarized form.**

For example, computing the **total number of likes for a particular post, computing the average monthly rainfall or finding the average number**

Diagnostic Analytics

Diagnostic analytics comprises analysis of past data **to diagnose the reasons as to why certain events happened.**

Diagnostic analytics aims to answer - **Why did it happen?**

Let us consider an example of a system that **collects and analyzes sensor data from machines for monitoring their health and predicting failures.**

While descriptive analytics can be **useful for summarizing the data by computing various statistics** (such as mean, minimum, maximum, variance, or top-N), **diagnostic analytics can provide more insights into why certain a fault has occurred based on the patterns in the sensor data for previous faults**.

Predictive Analytics

Predictive analytics comprises **predicting the occurrence of an event or the likely outcome of an event or forecasting the future values using prediction models.**

Predictive analytics aims to answer - What is likely to happen?

For example, predictive analytics can be **used for predicting when a fault will occur in a machine, predicting whether a tumor is benign or malignant, predicting the occurrence of natural emergency (events such as forest fires or river floods) or forecasting the pollution levels.**

Predictive Analytics is done using predictive models which are trained by existing data. These models learn patterns and trends from the existing data and predict the occurrence of an event

Prescriptive Analytics

While predictive analytics uses prediction models to predict the likely outcome of an event, **prescriptive analytics uses multiple prediction models to predict various outcomes and the best course of action for each outcome.**

Prescriptive analytics aims to answer - **What can we do to make it happen?**

Prescriptive Analytics can **predict the possible outcomes based on the current choice of actions.**

Prescriptive analytics can be used to **prescribe the best medicine for treatment of a patient based on the outcomes of various medicines** for similar patients.

Another example of prescriptive analytics would be to suggest the **best mobile data plan for a customer based on the customer's browsing p**

Domain Specific Examples of Big Data

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, and healthcare.

Various applications of big data for each of these domains are:

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.**

For collecting data on user visits, two approaches are used.

In the first approach**, user visits are logged on the web server** which collects data such as the **date and time of visit, resource requested, user's IP address, HTTP status code, for instance.**

The second approach, called **page tagging**, uses a **JavaScript which is embedded in the web page.**

The benefit of the **page tagging approach is that it facilitates real-time data collection and analysis**.

This approach allows **third party services**, which **do not have access to the web server (serving the website) to collect and process the data**.

These **specialized analytics service providers** (such as Google Analytics) are offer **advanced analytics and summarized reports**. **user sessions, page visits, top entry and exit pages, bounce rate, most visited page**

Performance Monitoring: **Multi-tier web and cloud applications such as such as**

- **e-Commerce,**
- **Business-to-Business,**
- **Health care, Banking and Financial,**
- **Retail and Social Networking applications**

can experience rapid changes in their workloads.

Provisioning and capacity planning is a challenging task for complex multi-tier applications since each class of applications has different **deployment configurations with web servers, application servers and database servers**.

For **performance monitoring, various types of tests can be performed s**uch as

- load tests (which evaluate the performance of the system with multiple users and workload levels)
- Stress test etc

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**.

• Ad Targeting & Analytics:

Search and display advertisements are the two most widely used approaches for Internet advertising.

In search advertising, **users are displayed advertisements** ("ads"), **along with the search results**, as they **search for specific keywords on a search engine.**

Advertisers can create ads using the advertising networks provided by the search engines or social media networks.

These **ads are setup for specific keywords** which are related to the product or service being advertised.

Users searching for these keywords are shown ads along with the search results.

Display advertising, is another form of Internet advertising, in **which the ads are displayed within websites, videos and mobile applications who participate in the advertising network**

The **ad-network matches these ads against the content on the website, video or mobile application and places the ad**s.

The most commonly used **compensation method for Internet ads is Pay-per-click** (PPC), in which the **advertisers pay each time a user clicks on an advertisement**.

Advertising networks use **big data systems for matching and placing advertisements and generating advertisement statistics reports**.

- Advertises can use **big data tools for tracking the performance of advertisements, optimizing the bids for pay-per-click advertising, tracking which keywords link the most to the advertising landing page**

- 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, history of content consumed, and user ratings.**

Such applications can leverage **big data systems for recommending new content to the users based on the user preferences and interests.**

Financial

- Credit Risk Modeling: **Banking and Financial institutions use credit risk modeling to score credit applications and predict if a borrower will default or not in the future.**

Credit risk models are **created from** the **customer data that includes, credit scores obtained from credit bureaus, credit history, account balance data, account transactions data and spending patterns of the customer**.

Big data systems can help in computing credit risk scores of a large number of customers on a regular basis.

These frameworks can be used to build credit risk models by analysis of customer data

- 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.**

Real-time big data analytics frameworks can help in analyzing data from disparate sources and label transactions in real-time

- Healthcare

The **healthcare ecosystem** consists of numerous entities including **healthcare providers** (primary care physicians, specialists, or hospitals), **payers** (government, private health insurance companies, employers), **pharmaceutical, device and medical service companies, IT solutions and services firms, and patients.**

The process of **provisioning healthcare involves** massive **healthcare data** that exists in **different forms** (structured or unstructured), is **stored in disparate data** sources (such as relational databases, or file servers) and in **many different formats.**

To promote more **coordination** of care across the multiple providers involved with patients, their **clinical information is increasingly aggregated from diverse sources into Electronic Health Record (EHR) systems**.

EHRs capture and store information on patient health and provider actions including individual-level laboratory results, diagnostic, treatment, and demographic data.

Though the primary use of EHRs is to maintain all medical data for an individual patient and to provide efficient access to the stored data at the point of care, **EHRs can be the source for valuable aggregated information about overall patient populations.**

Big data systems can be used for data collection from different stakeholders (patients, doctors, payers, physicians, specialists, etc) **and disparate data sources**.

Big data **analytics systems allow massive scale clinical data analytics and facilitate development of more efficient healthcare applications, improve the accuracy of predictions and help in timely decision making**

- Internet of Things

Internet of Things (IoT) refers to **things that have unique identities and are connected to the Internet**.

The "Things" in IoT are the **devices which can perform remote sensing, actuating and monitoring**.

IoT **devices can exchange data with other connected devices and applications (directly or indirectly), or collect data from other devices and process the data.**

IoT systems **can leverage big data technologies for storage and analysis of data.**
IoT applications that can benefit from big data system

- Intrusion Detection
- Smart Parking
- Smart Roads
- Structural Health Monitoring
- Smart Irrigation

Intrusion Detection: Intrusion detection systems use **security cameras and sensors (such as PIR sensors and door sensors) to detect intrusions and raise alerts.**

Smart Parkings: Smart parkings make the **search for parking space easier and convenient for drivers**. 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**

Smart Roads: Smart roads equipped with sensors can **provide information on driving conditions, travel time estimates and alerts in case of poor driving conditions, traffic congestions and accidents**

.

Structural Health Monitoring: 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

Smart Irrigation: Smart irrigation systems can **improve crop yields while saving water.** Smart irrigation systems - use IoT devices with soil moisture sensors - determine **the amount of moisture in the soil** and **release the flow of water** -w**hen the moisture levels go below a predefined threshold.**

Environment

Environment monitoring **systems generate high velocity and high volume data**.

Accurate and **timely analysis of such data can help in understanding the current status of the environment and also predicting environmental trends**.

Weather Monitoring : Weather monitoring systems can **collect data from a number of sensor attached**. This data can then be **analyzed and visualized for monitoring weather and generating weather alerts**

Air Pollution Monitoring: Air pollution monitoring systems can monitor **emission of harmful gases (CO2, CO, NO, or NO2) by factories and automobiles using gaseous and meterological sensor**

Noise Pollution Monitoring: Due to growing urban development, **noise levels in cities have increased and even become alarmingly high in some cities**
Urban noise maps can help the **policy makers in urban planning** and making policies to **control noise levels near residential areas, schools and parks**

Forest Fire Detection: There can be different **causes of forest fires including lightening, human negligence, volcanic eruptions and sparks from rock falls**
Forest fire detection systems use a **number of monitoring nodes deployed at different locations in a forest.**

River Floods Detection: River flood monitoring system use a **number of sensor nodes that monitor the water level (using ultrasonic sensors) and flow rate (using the flow velocity sensors).**

Big data systems can be used to **collect and analyze data from a number of such sensor nodes and raise alerts when a rapid increase in water level and flow rate is detected.**

Logistics & Transportation

- Real-time Fleet Tracking: Vehicle fleet tracking systems **use GPS technology to track the locations of the vehicles in real-time**.

Big data systems can be used to **aggregate and analyze vehicle locations and routes data for detecting bottlenecks** in the supply chain such as **traffic congestions on routes, assignment and generation of alternative routes**

Shipment Monitoring:monitoring the **conditions inside containers- food spoilage**

Remote Vehicle Diagnostics: Remote vehicle diagnostic systems **can detect faults in the vehicles or warn of impending faults.**

Route Generation & Scheduling: **Modern transportation systems** are driven by **data collected from multiple sources which is processed to provide new services** to the stakeholders. such as **advanced route guidance, dynamic vehicle routing, anticipating customer demands for pickup and delivery problem**

Hyper-local Delivery: Hyper-local delivery platforms are being increasingly used by **businesses such as restaurants and grocery stores to expand their reach**. These platforms **allow customers to order products** (such as grocery and food items) using web and mobile applications and the **products are sourced from local stores**

Cab/Taxi Aggregators: On-demand **transport technology aggregators (or cab/taxi aggregators) allow customers to book cabs using web or mobile applications and the requests are routed to nearest available cabs**

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**.

Industrial machines have a **large number of components that must function correctly for the machine to perform its operations**. Sensors in machines can monitor the operating conditions

Risk Analysis of Industrial Operations: In many industries, there are **strict requirements on the environment conditions and equipment working conditions** Harmful and **toxic gases such as carbon monoxide (CO), nitrogen monoxide (NO), Nitrogen Dioxide (NO2), for instance, can cause serious health problems**. Gas monitoring systems can help in monitoring the **indoor air quality** using various gas

Production Planning and Control: Production planning and control systems **measure various parameters of production processes and control the entire production process in real-time**

Retail

Retailers can use big data systems for **boosting sales, increasing profitability and improving customer satisfaction.**

Inventory Management: **Inventory management -increasingly important** in the recent years with the growing competition. While **over-stocking of products can result in additional storage expenses** and risk -**under-stocking can lead to loss of revenue**.

Tags attached to the products allow them to be tracked in real-time so that the inventory levels can be determined accurately and products which are low on stock can be replenished

Customer Recommendations: Big data systems can be used to **analyze the customer data (such as demographic data, shopping history, or customer feedback) and predict the customer preferences**

Store Layout Optimization: Big data systems can **help in analyzing the data on customer shopping patterns and customer feedback to optimize the store layouts**

Forecasting Demand: Due to a **large number of products, seasonal variations in demands and changing trends and customer preferences, retailers find it difficult to forecast demand and sales volumes**. Big data systems can be used to **analyze the customer purchase patterns and predict demand and sale volumes**

Analytics Type	Question	Technique Used
Descriptive	What happened?	Statistics
Diagnostic	Why?	Data mining, correlation
Predictive	What will happen?	Prediction, ML
Prescriptive	What to do?	Optimization

Analytics Flow	Big Data Stack
Data collection	Data source layer
Ingestion	Kafka / Flume
Storage	HDFS / NoSQL
Processing	Spark / MapReduce
Querying	Spark SQL
Visualization	Web dashboards

Case Study: Weather Data Analysis

Using the **big data stack for analysis of weather data**

To come up with a **selection of the tools and frameworks from the big data stack** that can be used for weather data analysis, **let us first come up with the analytics flow for the application**

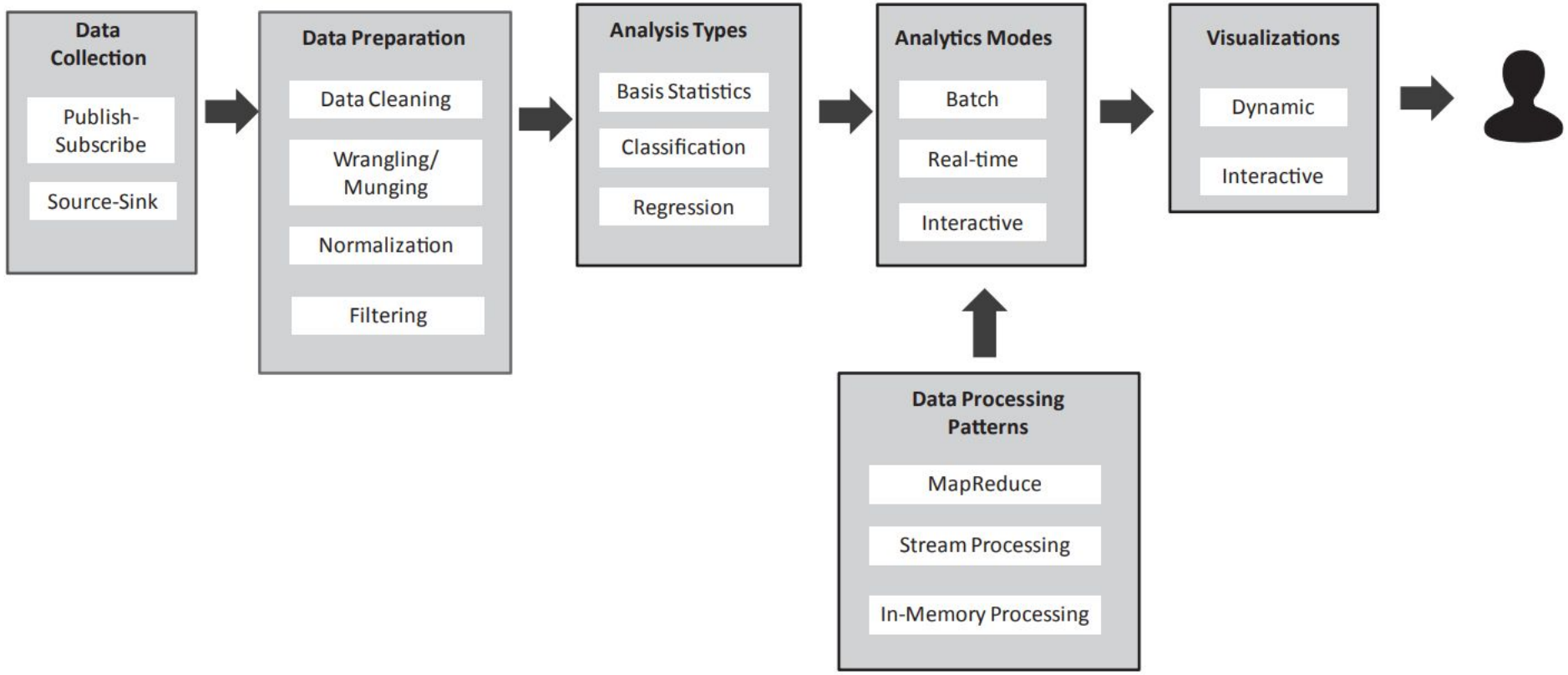


Figure 1.10: Analytics flow for weather data analysis application

Data Collection

Let us assume, we have **multiple weather monitoring stations or end-nodes equipped with temperature, humidity, wind, and pressure sensors.**

To **collect and ingest streaming sensor data generated** by the weather monitoring stations, we can use a **publish-subscribe messaging framework to ingest data for real-time analysis** within the Big Data stack and

Source-Sink connector to ingest data into a distributed filesystem for batch analysis.

Data Preparation

Since the weather **data received from different monitoring stations** can have **missing values**, use **different units** and have **different formats**, we may need to **prepare data for analysis by cleaning, wrangling, normalizing and filtering the data**

Analysis Types

The choice of the **analysis types is driven by the requirements of the application.**

Let us say, we want our **weather analysis application**

- **to aggregate data on various timescales** (minute, hourly, daily or monthly)
- to **determine the mean, maximum and minimum readings for temperature, humidity, wind and pressure**.

to **support interactive querying for exploring the data, for example, queries such as: finding the day with the lowest temperature in each** month of a year, finding the top-10 most wet days in the year, for instance.

These type of analysis come **under the basic statistics category.**

Next, we also want the application to **make predictions of certain weather events**, for example, predict the **occurrence of fog or haze**. For such an analysis, we would **require a classification model**.

Additionally, if we **want to predict values (such as the amount of rainfall), we would require a regression model**

Analysis Modes

Based on the analysis types determined the previous step, we know that the **analysis modes required for the application will be batch, real-time and interactive.**

Visualizations

The front end application for visualizing the analysis results would be **dynamic and interactive**.

Mapping Analysis Flow to Big Data Stack

Now that we have the analytics flow for the application, let us map the selections at each step of the flow to the big data stack

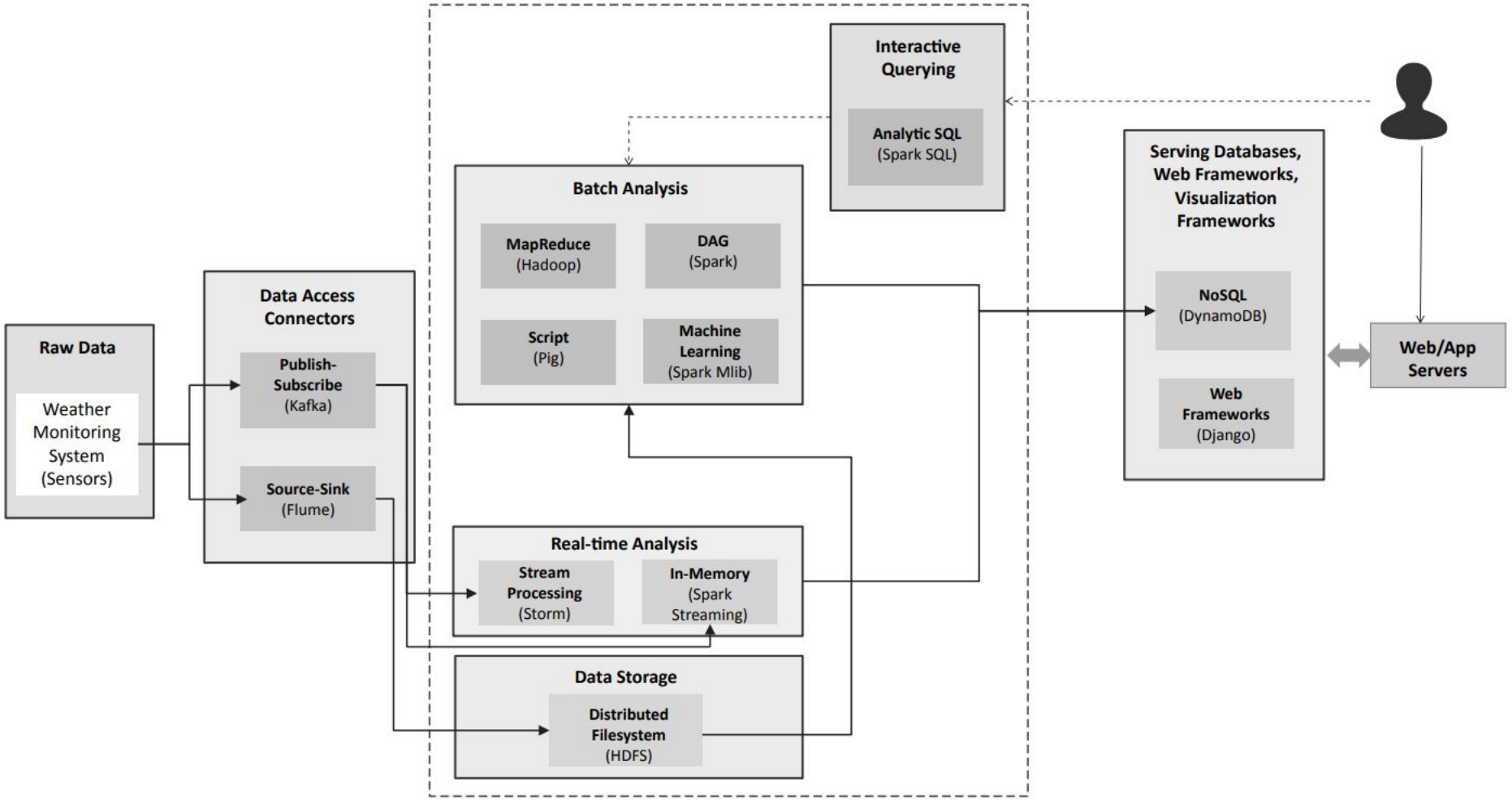


Figure 1.11: Using Big Data stack for weather data analysis

Figure shows a subset of the components of the big data stack based on the analytics flow.

To **collect and ingest streaming sensor data generated by the weather monitoring stations**, we can **use a publish-subscribe messaging framework such as Apache Kafka** (for real-time analysis within the Big Data stack).

Each weather station publishes the sensor data to Kafka

Real-time analysis frameworks such as Storm and Spark Streaming can receive data from Kafka for processing

For **batch analysis**, we can use a **source-sink connector such as Flume to move the data to HDFS.**

Once the data is in HDFS, we can use **batch processing frameworks such as Hadoop-MapReduce**

While the **batch and real-time processing frameworks are useful when the analysis requirements and goals are known upfront, interactive querying tools can be useful for exploring the data**.

We can **use interactive querying framework such as Spark SQL, which can query the data in HDFS for interactive queries**.

For presenting the **results of batch and real-time analysis, a NoSQL database such as DynamoDB can be used as a serving database**.

For developing **web applications and displaying the analysis results we can use a web framework such as Django.**