BIG DATA System

Learning Objectives and Learning Outcomes

Learning Objectives	Learning Outcomes
Introduction to digital data and its types 1. Structured data: Sources of structured data, ease with structured data, etc. 2. Semi-Structured data: Sources of semi-structured data, characteristics of semi- structured data. 3. Unstructured data: Sources of unstructured data, issues with terminology, dealing with unstructured data.	a) To differentiate between structured, semi-structured and unstructured data. b) To understand the need to integrate structured, semi-structured and unstructured data.

Agenda

Types of Digital Data

- **Structured**
  - Sources of structured data
  - Ease with structured data
- **Semi-Structured**
  - Sources of semi-structured data
- **Unstructured**
  - Sources of unstructured data
  - Issues with terminology
  - Dealing with unstructured data

Classification of Digital Data

Digital data is classified into the following categories:

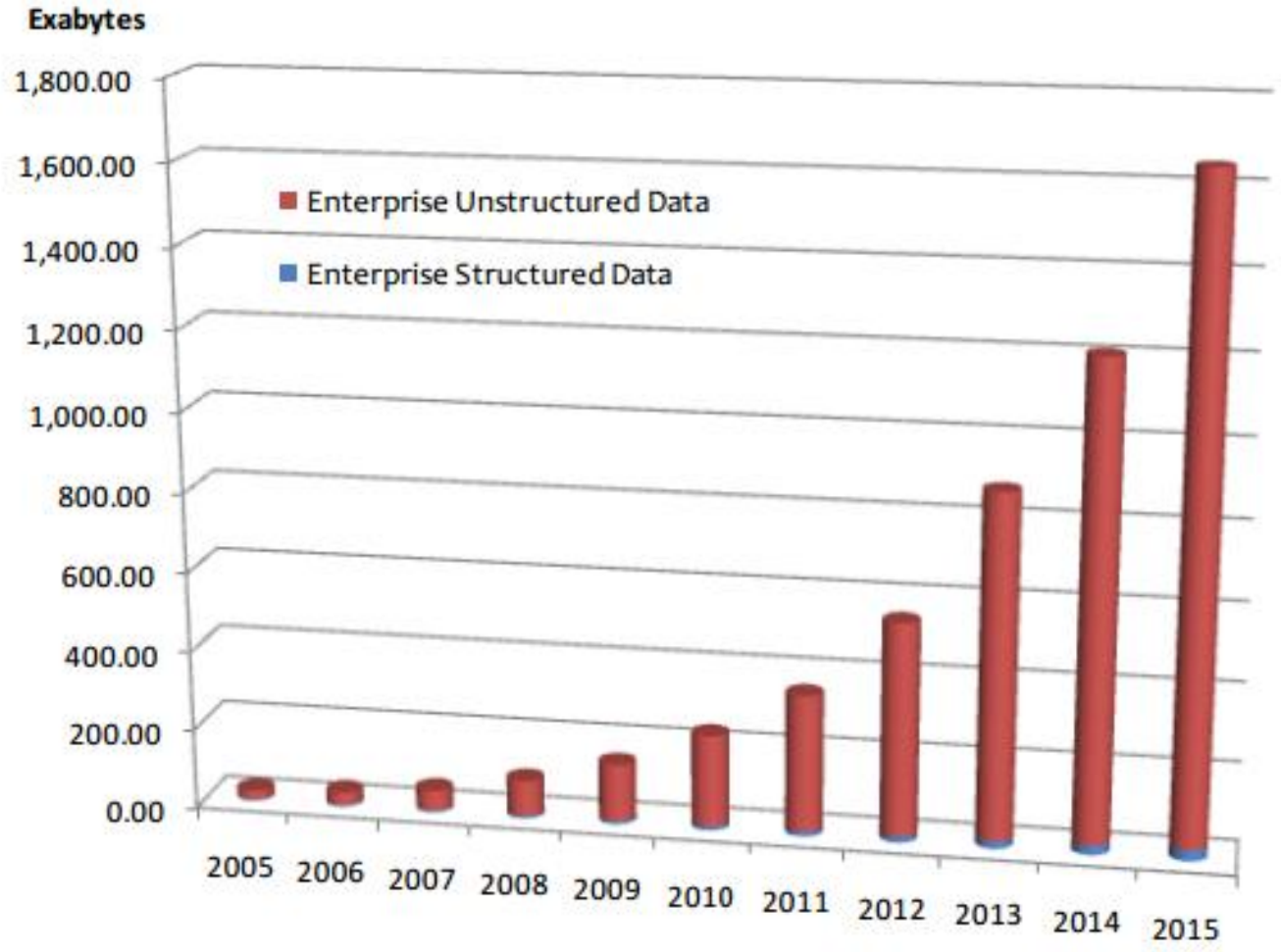
- Structured data- This is the data which is in an organized form(e.g, rows and columns) and can be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. Data stored in databases is an example of structured data.

- Semi-structured data- This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program, for example, emails, XML, markup languages like HTML etc.,

- 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 for example, memos, chat rooms, powerpoint presentations, images, videos, letters etc,.

Approximate Distribution of Digital Data

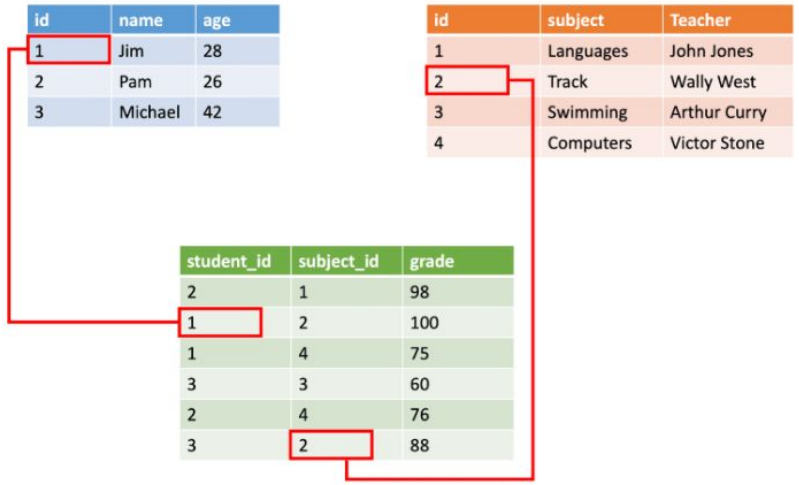
Approximate percentage distribution of digital data



Structured Data

Structured Data

- This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program.
- In structured data, all row in a table has the same set of columns.
- Data stored in databases is an example of structured data.

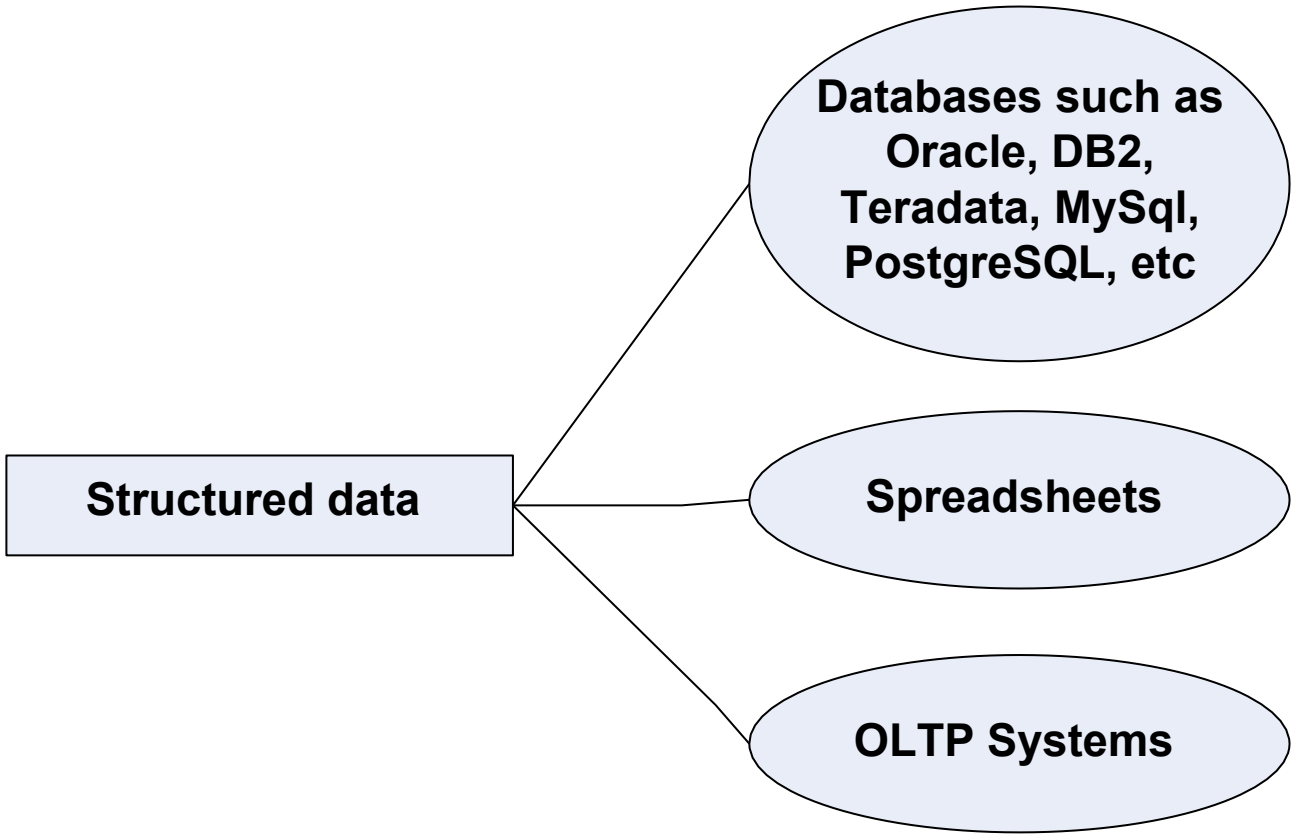


id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

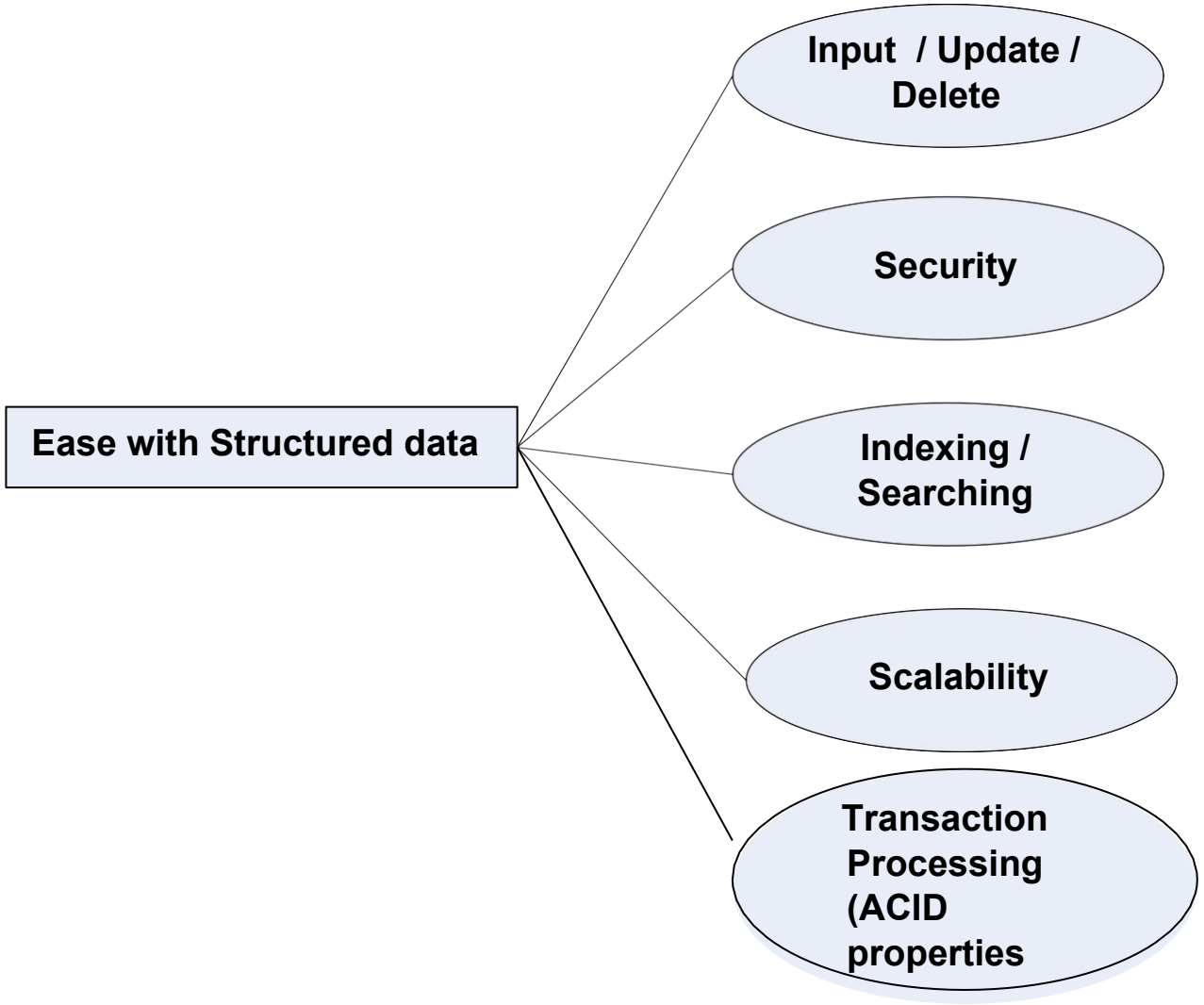
id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88

Sources of Structured Data



Ease with Structured Data



Semi-structured Data

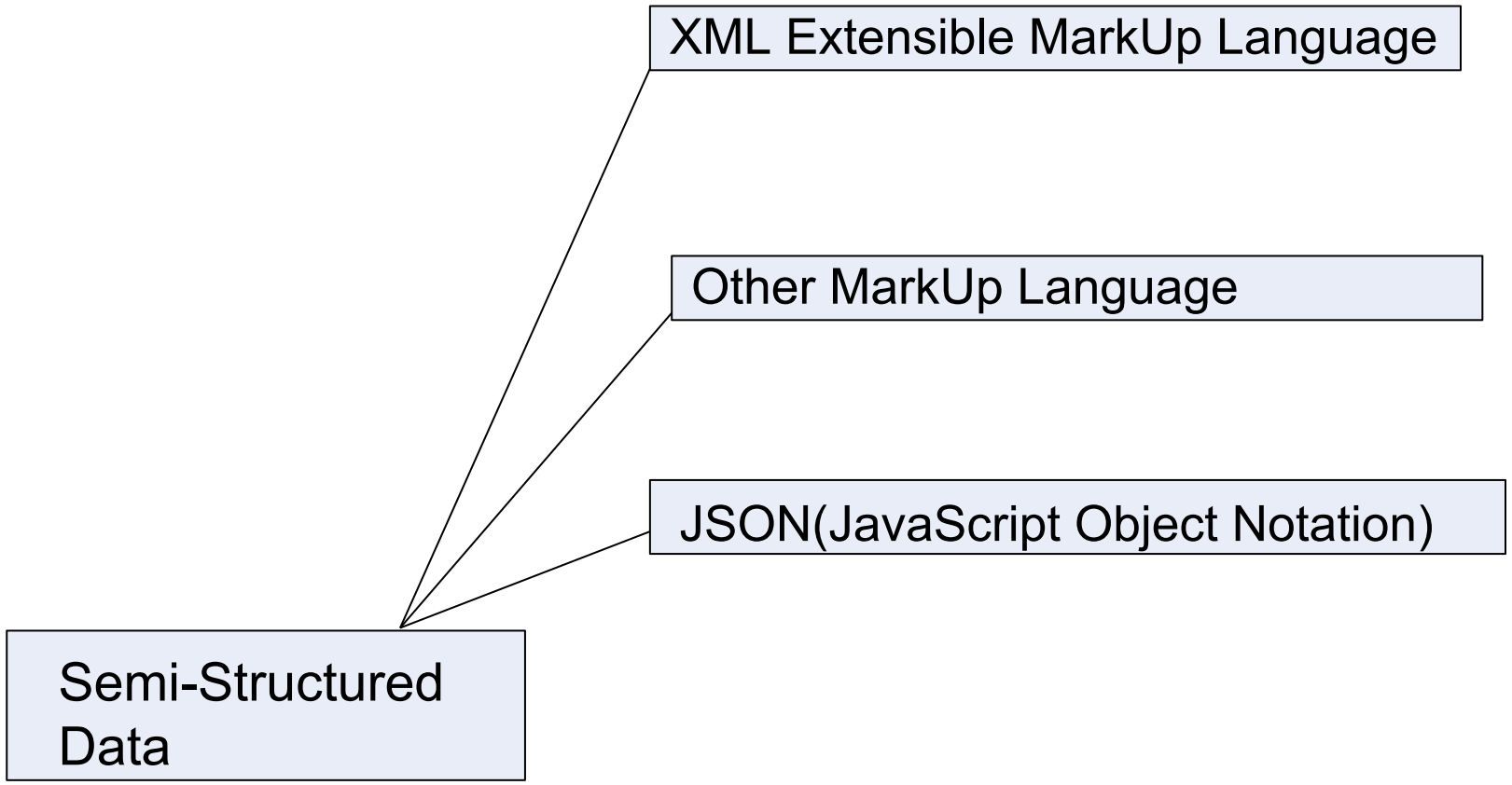
Semi-structured Data

- This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program.

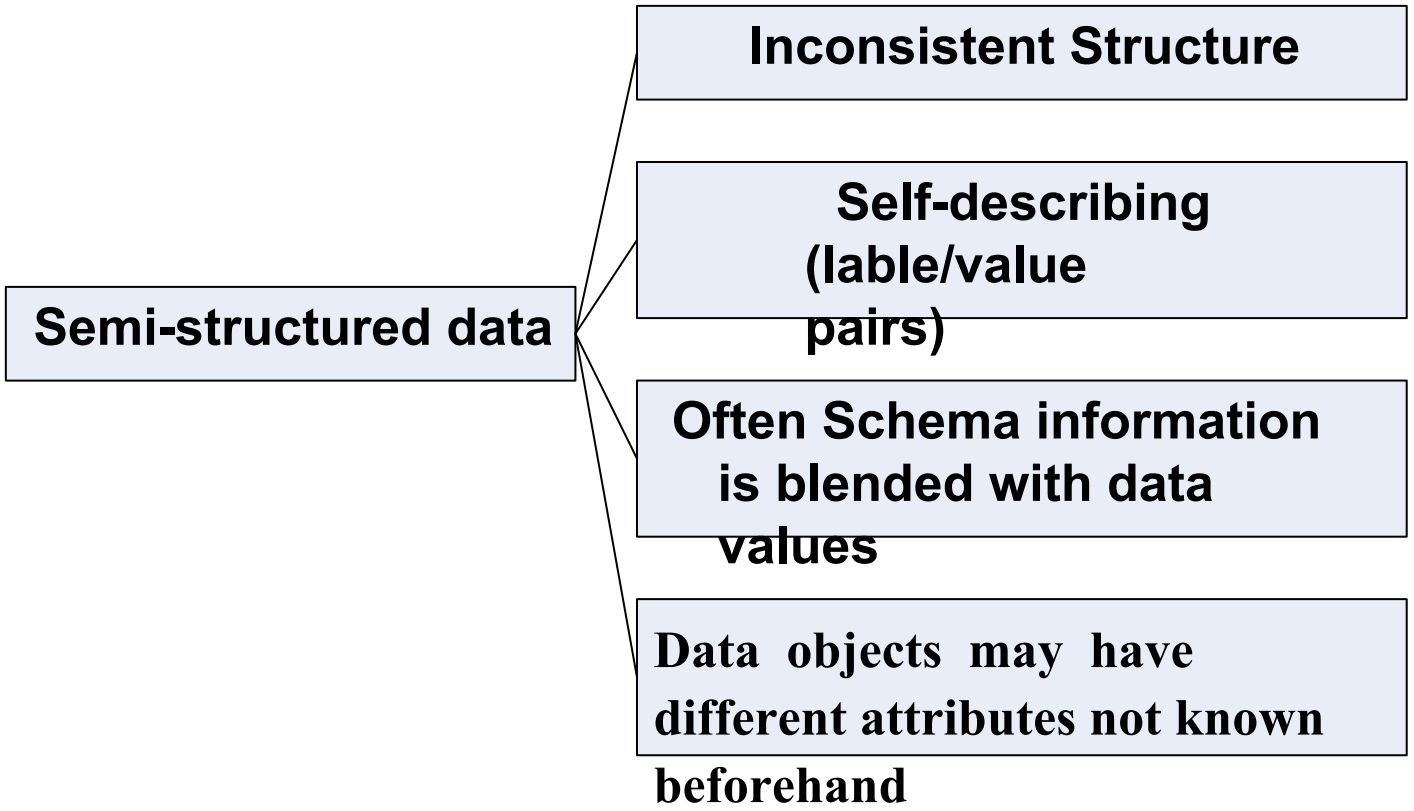
- Example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.

```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

Sources of Semi-structured Data



Characteristics of Semi-structured Data

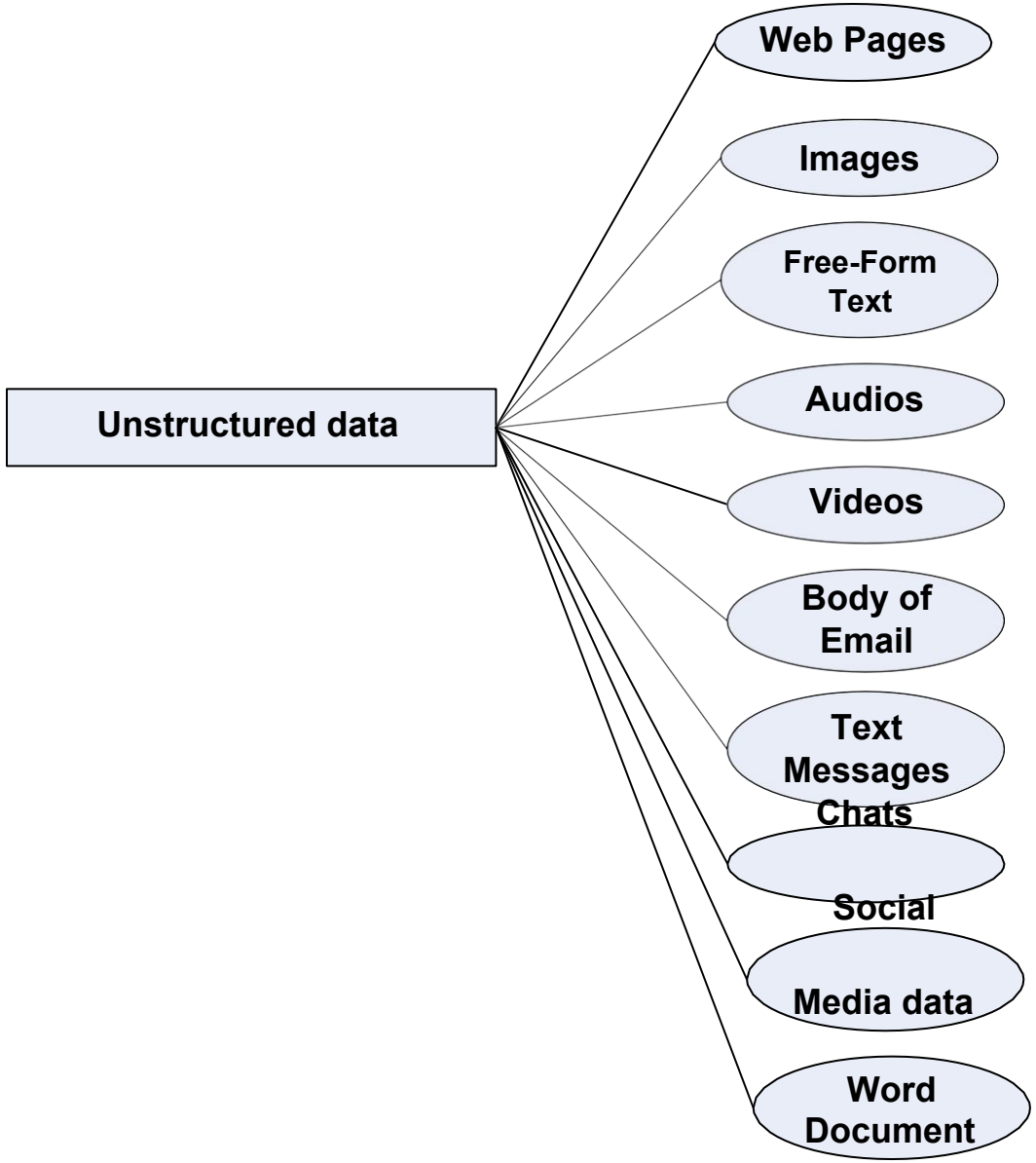


Unstructured Data

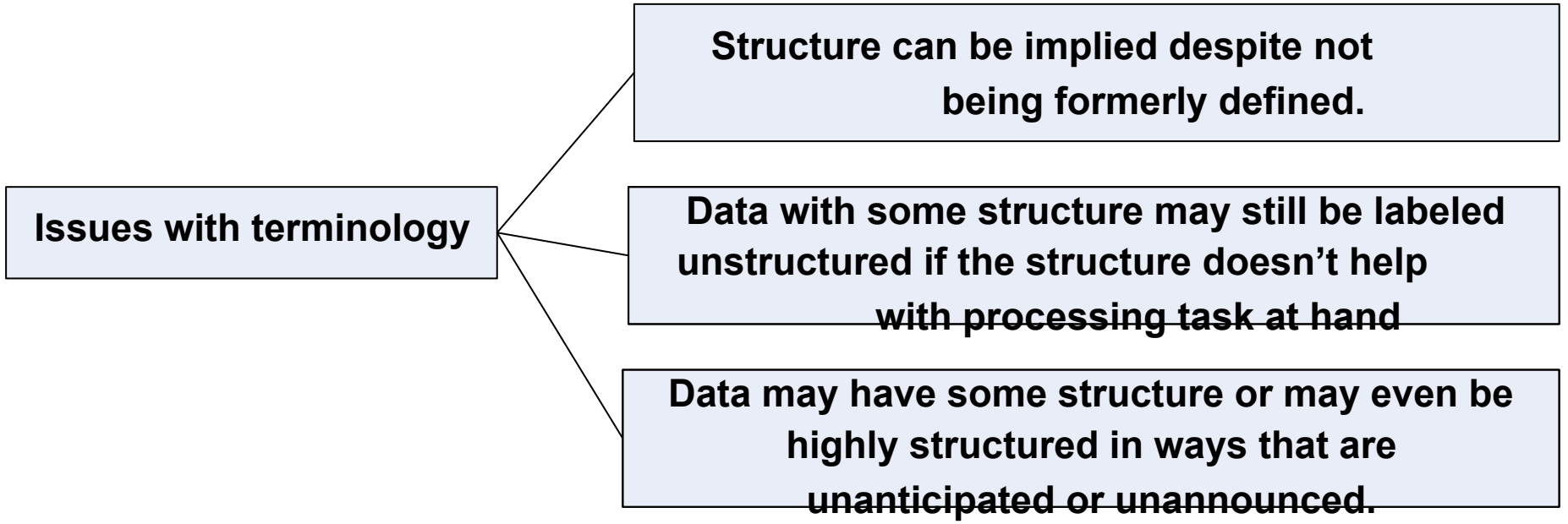
Unstructured Data

- 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.

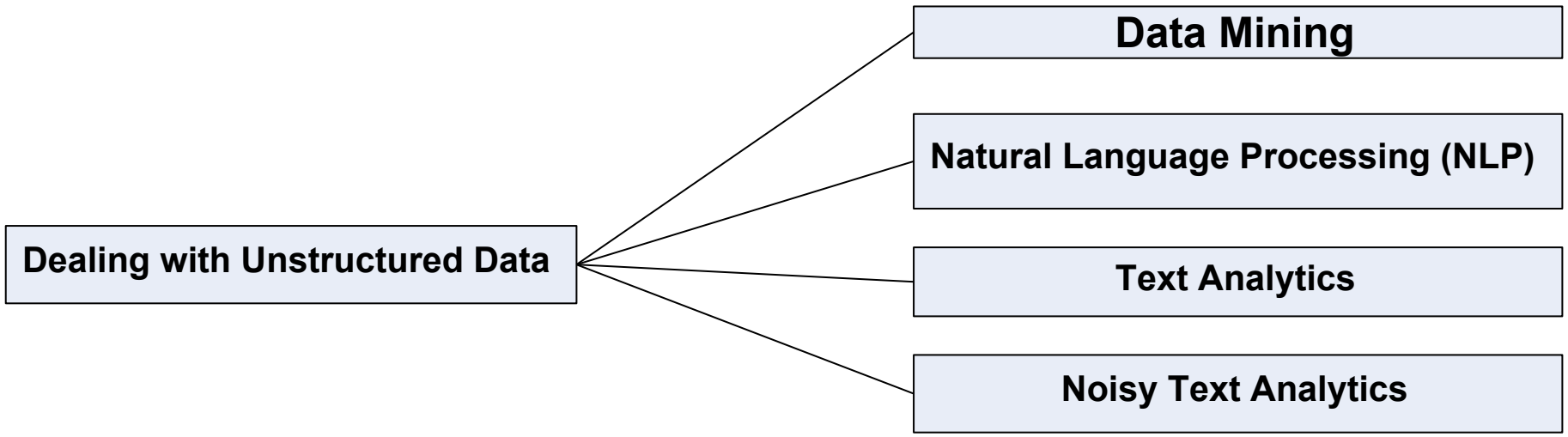
Sources of Unstructured Data



Issues with terminology - Unstructured Data



Dealing with Unstructured Data



Dealing with Unstructured Data

- **Data Mining**
  - Association Rule Mining
  - Regression Analysis
  - Collaborative Filtering

- **Text analysis and Text Mining**

- **Natural Language Processing(NLP)**

- **Noisy text Analysis**

- **Manual tagging with metadata**

- **Part-of-speech tagging**

- **Unstructured Information Management Architecture(UIMA)**

Answer a few quick questions …

Answer Me

- Which category (structured, semi-structured, or unstructured) will you place a Web Page in?

- Which category (structured, semi-structured, or unstructured) will you place Word Document in?

- State a few examples of human generated and machine-generated data.

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

Summary please…

few participants of the learning program to summarize the lecture.

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table	It is based on XML/RDF(Resource Description Framework).	It is based on character and binary data
Transaction management	Matured transaction and various concurrency techniques	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as a whole
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured data	It is more flexible and there is absence of schema
Scalability	It is very difficult to scale DB schema	It’s scaling is simpler than structured data	It is more scalable.
Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

References …

Further Readings

- http://data-magnum.com/the-big-deal-about-big-data-whats-inside-structured-unstructured-and-semi-structured-data/
- http://www.webopedia.com/TERM/S/structured_data.html
- http://en.wikipedia.org/wiki/UIMA

Thank you

Chapter 2

Introduction to Big Data

Learning Objectives and Learning Outcomes

Learning Objectives	Learning Outcomes
Introduction to big data 1. Definition of big data. 2. Challenges of big data. 3. Why big data? 4. Traditional Business Intelligence versus big data.	a) To understand the significance of big data. b) To understand the other characteristics of data that are not definitional characteristics of big data. c) To understand the challenges of big data and how to deal with the same. d) To understand what is new today.

Agenda

- Definition of Big Data
  - Volume
  - Velocity
  - Variety
- Challenges of Big Data
- Other Characteristics of Data Which are Not Definitional Traits of Big Data
- Why Big Data?
- Traditional Business Intelligence (BI) versus Big Data
  - A Typical Data Warehouse Environment
  - A Typical Hadoop Environment
  - Coexistence of Big Data and Data Warehouse

Characteristics of Data

Data has three characteristics:

1. Composition: deals with structure of data, that is, the sources of data , the granularity, the types, and the nature of the data as to whether it is static or real-time streaming.

2. Condition: The condition of data deals with the state of the data that is “can one use this data as is for analysis?” or “Does it require cleansing for further enhancement and enrichment?”

3. Context: deals with “Where has this data been generated?”, “Why was this data generated?” and so on.

2.2 EVOLUTION OF BIG DATA

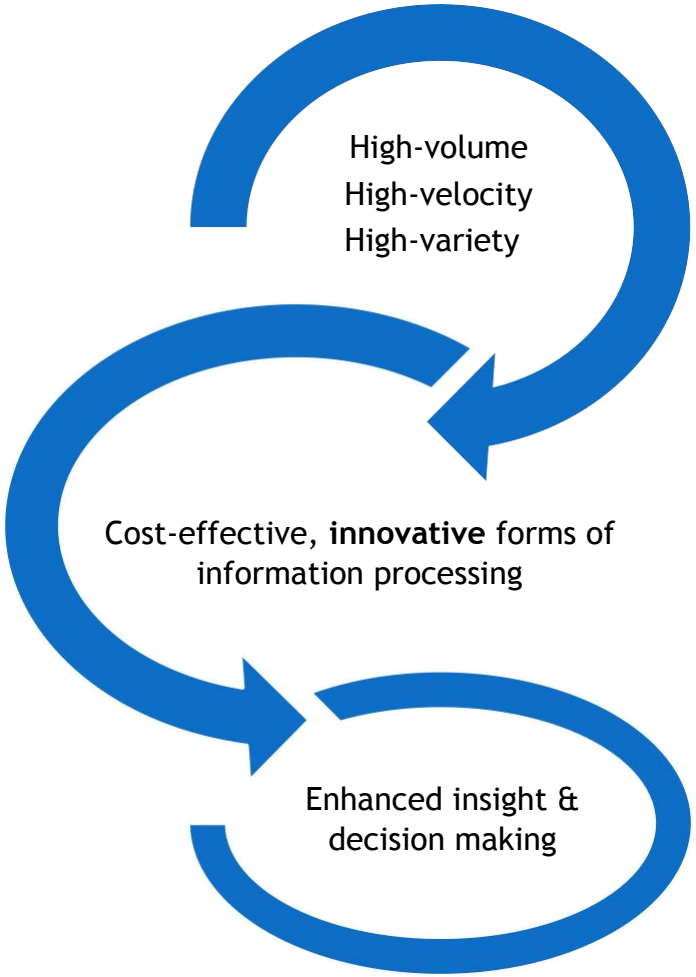
1970s and before was the era of mainframes. The data was essentially primitive and structured. Relational databases evolved in 1980s and 1990s. The era was of data intensive applications. The World Wide Web (WWW) and the Internet of Things (IoT) have led to an onslaught of structured, unstructured, and multimedia data. Refer Table 2.1.

Table 2.1 The evolution of big data

	Data Generation and Storage	Data Utilization	Data Driven
Complex and Unstructured			Structured data, unstructured data, multimedia data
Complex and Relational		Relational databases: Data-intensive applications	
Primitive and Structured	Mainframes: Basic data storage		
	1970s and before	**Relational (1980s and 1990s)**	**2000s and beyond**

Definition of Big Data

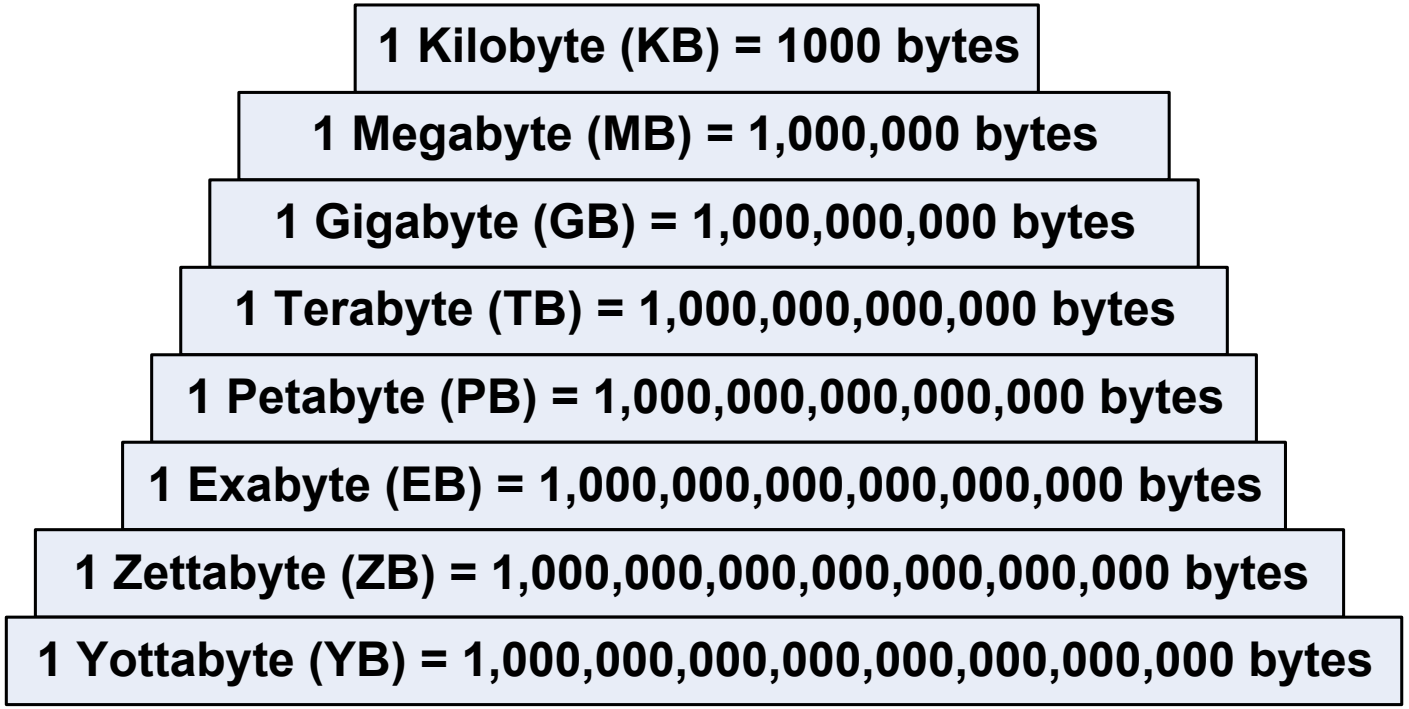
Definition of Big Data



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.

Source: Gartner IT Glossary

Volume - A Mountain of Data

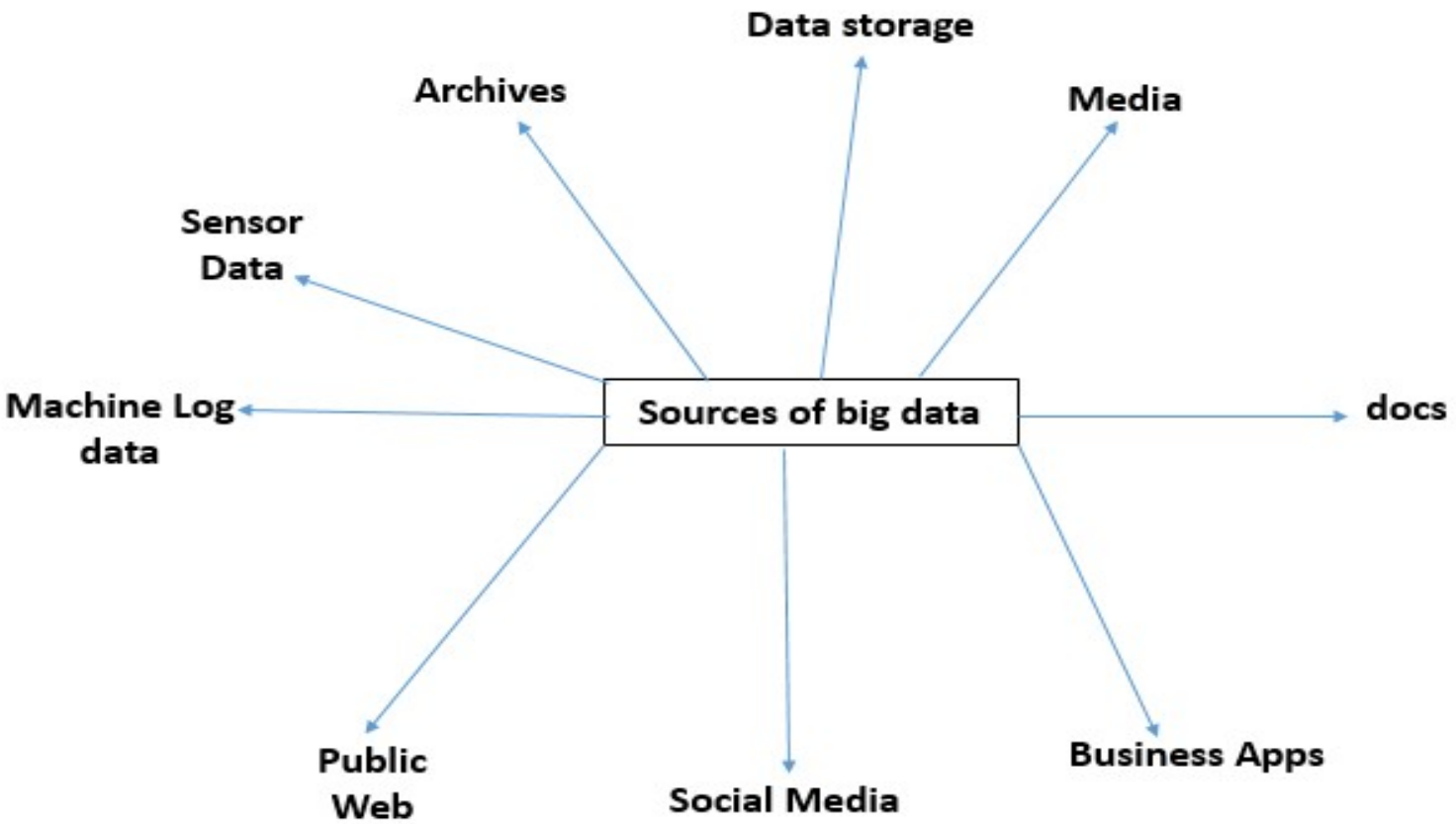


Volume

Where does this data get generated?

1. Typical internal sources:
    - **Data Storage-** File systems, SQL, NoSQL (MongoDB, Cassandra).
    - **Archives** – Archives of scanned documents, paper archives, customer records, patient health records etc,.
2. External data sources:
    - **public web -** Wikipedia, weather, regulatory, census etc.
3. Both (internal+external)
    - **Sensor data** – Car sensors, smart electric meters, office buildings etc,.
    - **Machine log data** – Event logs, application logs, Business process logs, audit logs 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** – CSV, Word Documents, PDF,XLS, PPT and so on.

Sources of Big Data



Velocity

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

Variety

- **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.

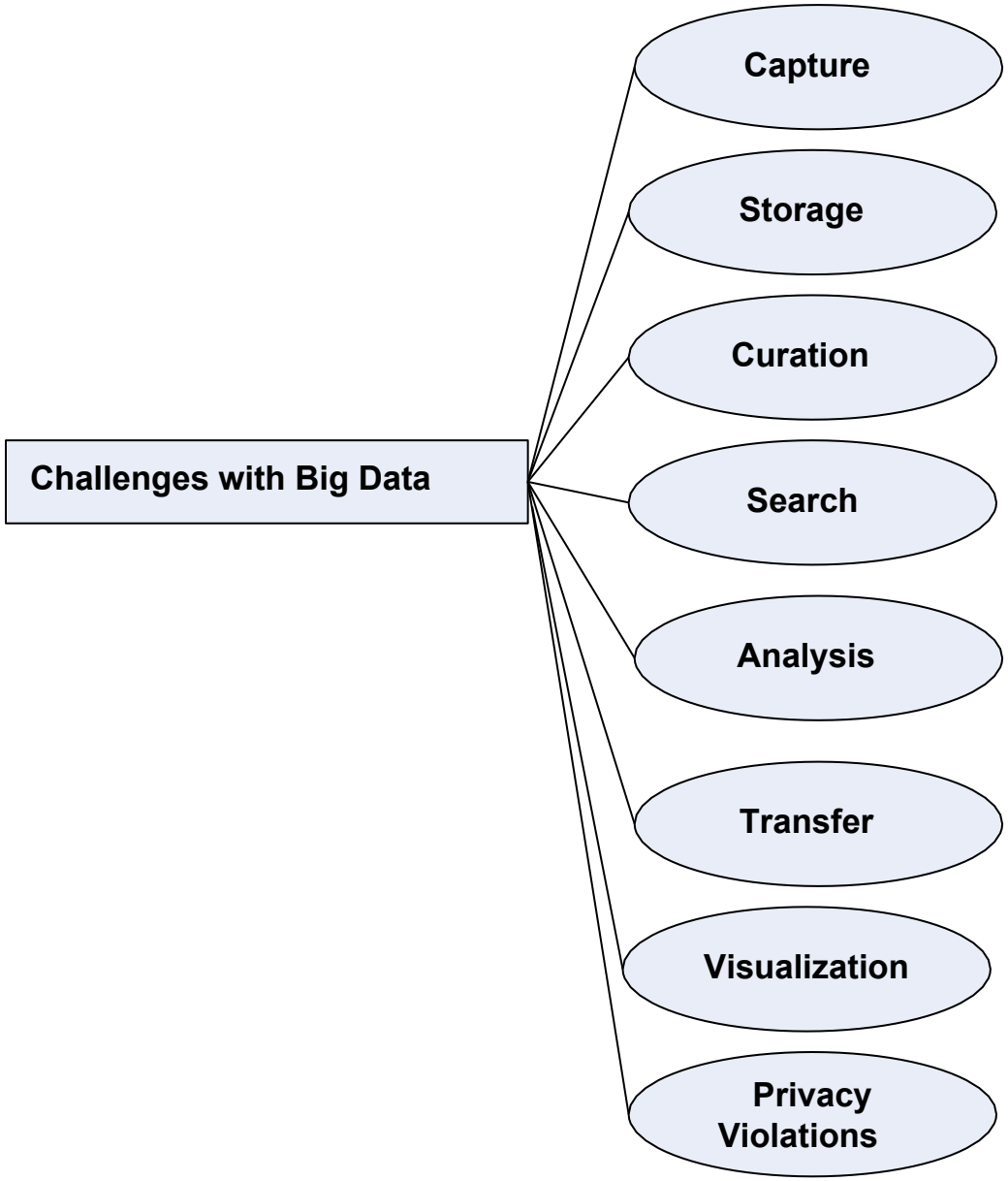
Other Characteristics of Data - Which are not Definitional Traits of Big Data

- Veracity and Validity-Veracity refers to biases, noises and abnormality in data.
  Validity refers to the accuracy and correctness of the data.
- Volatility-Deals with, how long is the data valid? And how long should it be stored?
- Variability- Data flows can be highly inconsistent with periodic peaks.

Challenges with Big Data

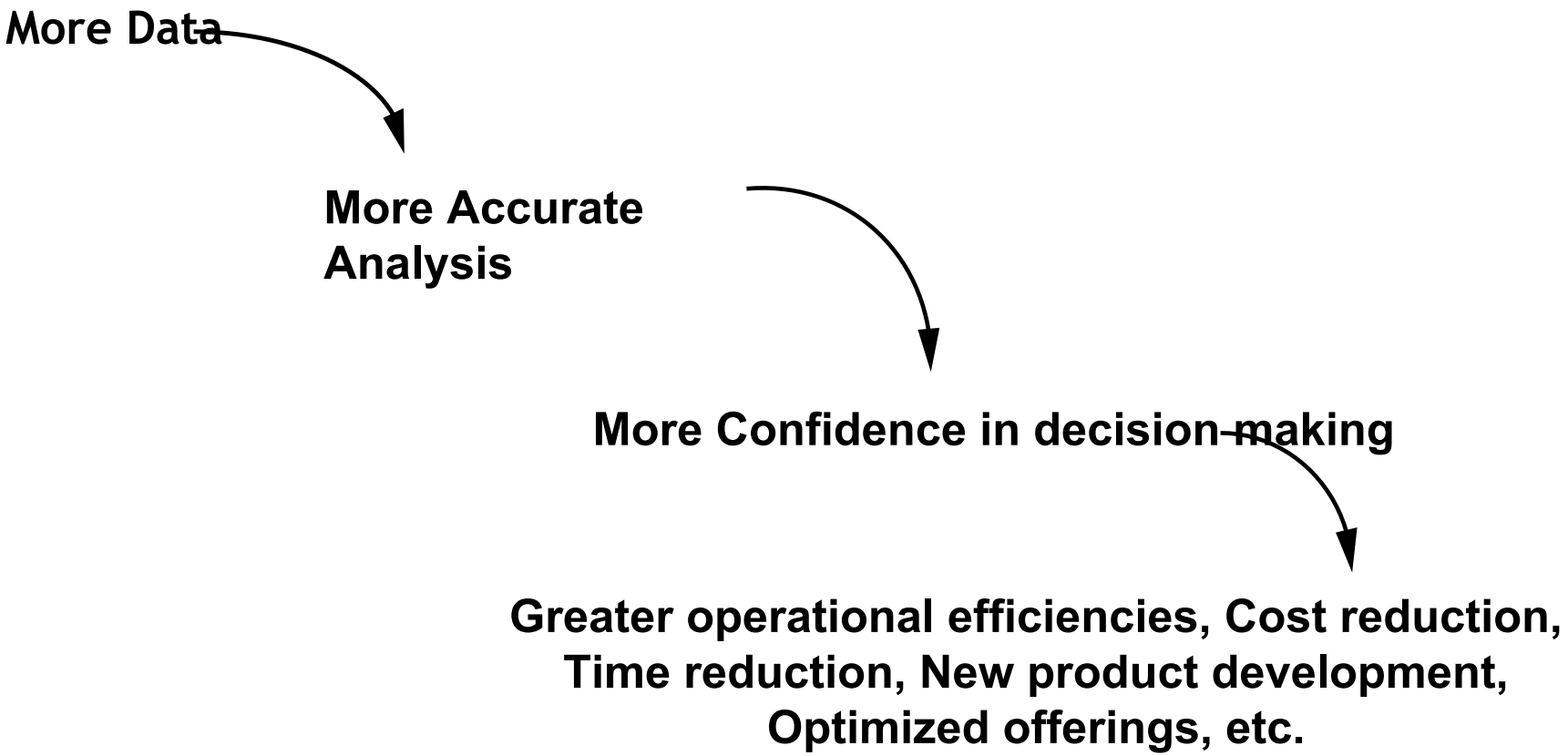
1. Data today is growing at an exponential rate. Most of the data that we have today has been generated in the last 2–3 years. This high tide of data will continue to rise incessantly. The key questions here are: "Will all this data be useful for analysis?", "Do we work with all this data or a subset of it?", "How will we separate the knowledge from the noise?", etc.
2. Cloud computing and virtualization are here to stay. Cloud computing is the answer to managing infrastructure for big data as far as cost-efficiency, elasticity, and easy upgrading/downgrading is concerned. This further complicates the decision to host big data solutions outside the enterprise.
3. The other challenge is to decide on the period of retention of big data. Just how long should one retain this data? A tricky question indeed as some data is useful for making long-term decisions, whereas in few cases, the data may quickly become irrelevant and obsolete just a few hours after having being generated.
4. There is a dearth of skilled professionals who possess a high level of proficiency in data sciences that is vital in implementing big data solutions.
5. Then, of course, there are other challenges with respect to capture, storage, preparation, search, analysis, transfer, security, and visualization of big data. Big data refers to datasets whose size is typically beyond the storage capacity of traditional database software tools. There is no explicit definition of how big the dataset should be for it to be considered "big data." Here we are to deal with data that is just too big, moves way to fast, and does not fit the structures of typical database systems. The data changes are highly dynamic and therefore there is a need to ingest this as quickly as possible.
6. Data visualization is becoming popular as a separate discipline. We are short by quite a number, as far as business visualization experts are concerned.

Challenges with Big Data



Why Big Data?

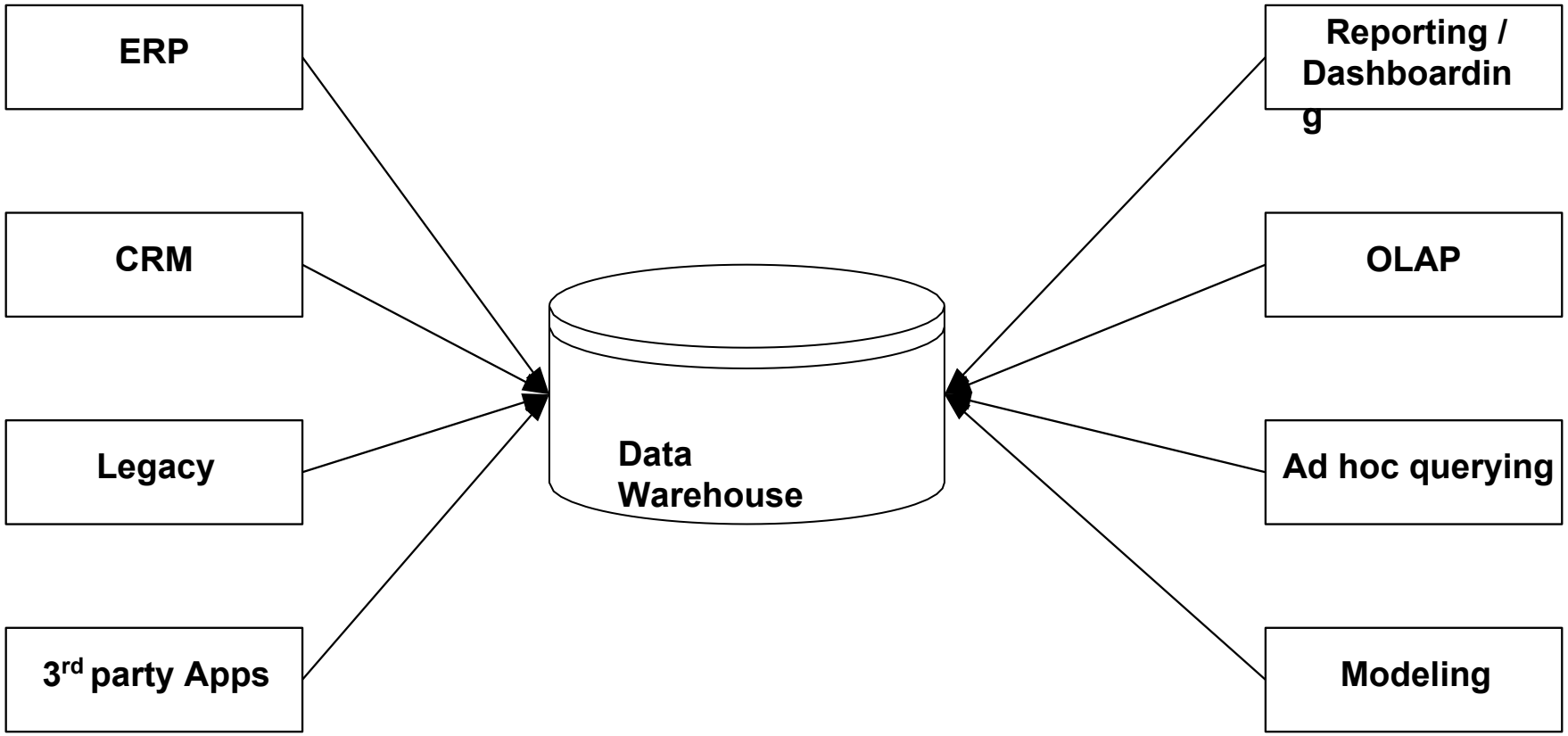
Why Big Data?


More Data
More Accurate Analysis
More Confidence in decision-making
Greater operational efficiencies, Cost reduction, Time reduction, New product development, Optimized offerings, etc.

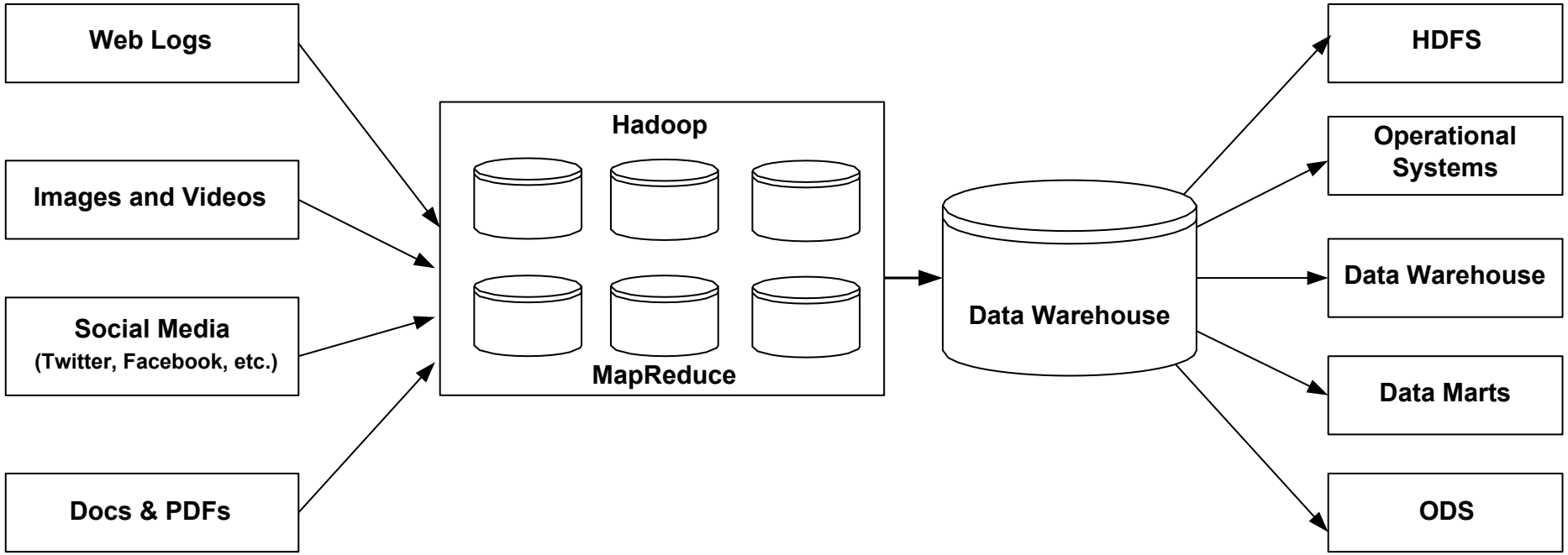
Traditional Business Intelligence (BI) versus Big Data

1. In traditional BI environment, all the enterprise's data is housed in a central server whereas in a big data environment data resides in a distributed file system. The distributed file system scales by scaling in or out horizontally as compared to typical database server that scales vertically.
2. In traditional BI, data is generally analyzed in an offline mode whereas in big data, it is analyzed in both real time as well as in offline mode.
3. Traditional BI is about structured data and it is here that data is taken to processing functions whereas big data is about variety and here the processing functions are taken to the data.

A Typical Data Warehouse Environment



Co-existence of Big Data and Data Warehouse



What is changing in the realms of Big data

- Competitive Advantage
- Decision Making
- Value of Data

Its time for Activity…

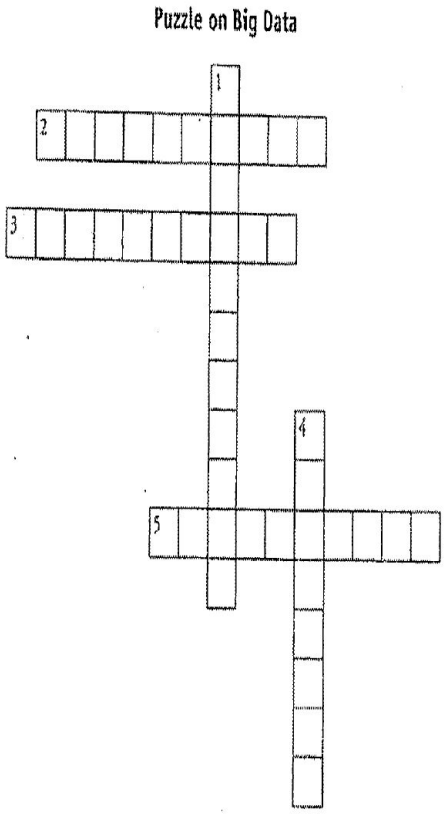
C. Match the Following

Column A	Column B
PostgreSql	Machine generated unstructured data
Scientific data	Open source relational database
Point-of-sale	Human-generated unstructured data
Social Media data	Machine-generated structured data
Gaming-related data	Human-generated unstructured data
Mobile data	Human-generated structured data

Answer:

Column A	Column B
PostgreSql	Open source relational database
Scientific data	Machine-generated unstructured data
Point-of-sale	Machine-generated structured data
Social Media data	Human-generated unstructured data
Gaming-related data	Human-generated structured data
Mobile data	Human-generated unstructured data

Teams Games Tournaments



Across

2. _____, a Gartner analyst coined the term, 'Big Data'
3. _____, is the characteristic of data dealing with its retention.
5. _____, is a large data repository that stores data in its native format until it is needed.

Down

1. _____ characteristic of data explains the spikes in data.
4. Near real time processing or real time processing deals with _____ characteristic of data.

Answer:

Across

2. Doug Laney
3. Volatility
5. Data Lakes

Down

1. Variability
4. Velocity

Answer Me

- Share your understanding of Big Data.

- How is traditional BI environment different from the Big Data environment?

- Share your experience as a customer on an e-commerce site. Comment on the big data that gets created on a typical e-commerce site.

Summary please…

Ask a few participants of the learning program to summarize the lecture.

References …

Further Readings

- Big data for dummies - Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman
- http://en.wikipedia.org/wiki/Big_data
- http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
- https://www.oracle.com/bigdata/
- http://bigdatauniversity.com/

THANK YOU