Big Data Analytics

Course Outcomes

At the end of the course students will be able to –

1. Explore the fundamental concepts of big data and its analytics
2. Analyze the big data using Hadoop and intelligent techniques
3. Apply NoSQL big data management
4. Recognize the suitable secure models for building competitive business decisions

Big Data Analytics

1. Importance of Big Data
2. Hadoop Architecture
3. Hadoop I/O
4. NoSQL Management
5. Analytics Framework
6. Securing Ecosystem

Text and Reference Books

Text Books

1. Seema Acharya, Subhasini Chellappan, “Big Data Analytics”, Wiley.
2. Tom White, “Hadoop: The Definitive Guide” (O’Reilly Media)
3. P. J. Sadalage, M. Flower, “NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence” (Addison-Wesley)
4. Sudeesh Narayanan, “Securing Hadoop” (O’Reilly Media)

Reference Books

1. Michael Mineli, Michele Chambers, Ambiga Dhiraj, “Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses” (Wiley Publication)
2. Chris Eaton,Dirk derooset al., “Understanding Big data”, McGraw Hill.
3. G James, D. Witten, T Hastie, R. Tibshirani, “An Introduction to Statistical Learning: with Applications in R”, Springer.
4. Douglas Eadline, “Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem”, Pearson Education.
5. E. Capriolo, D. Wampler, J. Rutherglen, “Programming Hive”, O’ Reilly.
6. Lars George, “HBase: The Definitive Guide”, O’ Reilly.
7. Alan Gates, "Programming Pig”, O’ Reilly

Big Data Analytics

- Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

- Big data analytics is an important aspect of Data Science.

Big Data Analytics

- Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency.
- Big data has one or more of the following characteristics:
  - high volume,
  - high velocity or
  - high variety

Big Data Analytics

- Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data.

- For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media — much of it generated in real time and at a very large scale.

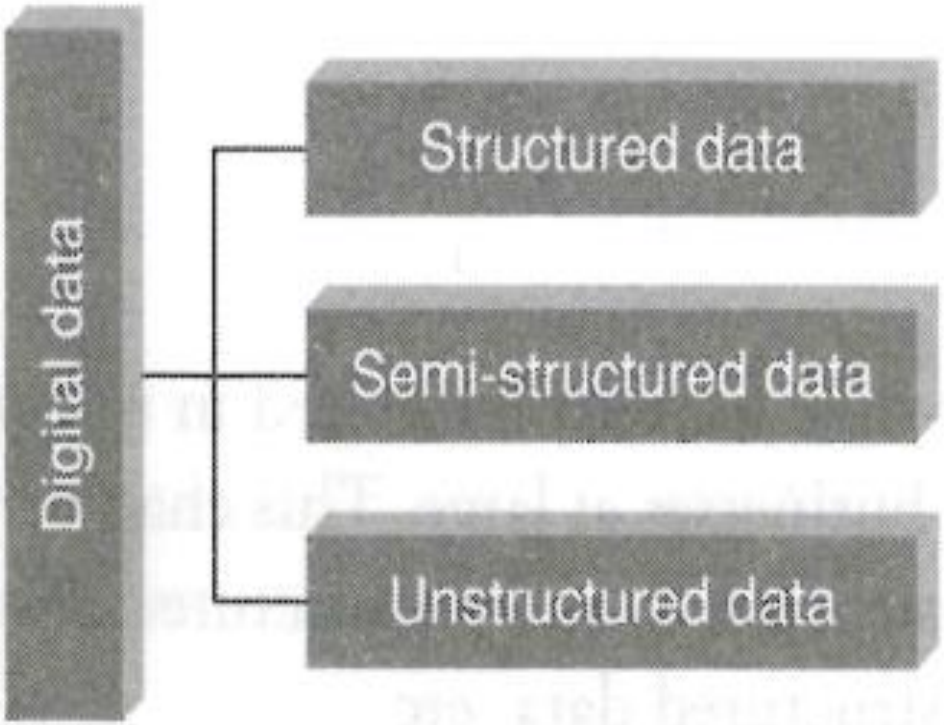
Why you should study Big Data

5 reasons to study Big Data

1. Data driven decisions provide a competitive advantage
2. Big Data provides a spring board for AI
3. Big Data skills are in high demand
4. Investments in Big Data keep growing
5. Studying Big Data will broaden your horizon

Unit-I
Importance of Big Data

CLASSIFICATION OF DIGITAL DATA



Classification of digital data.

CLASSIFICATION OF DIGITAL DATA

Unstructured data

- Does not conform to a data model
- Can not used easily by a computer program.
- About 80-90% data of an organization in unstructured form
- For example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

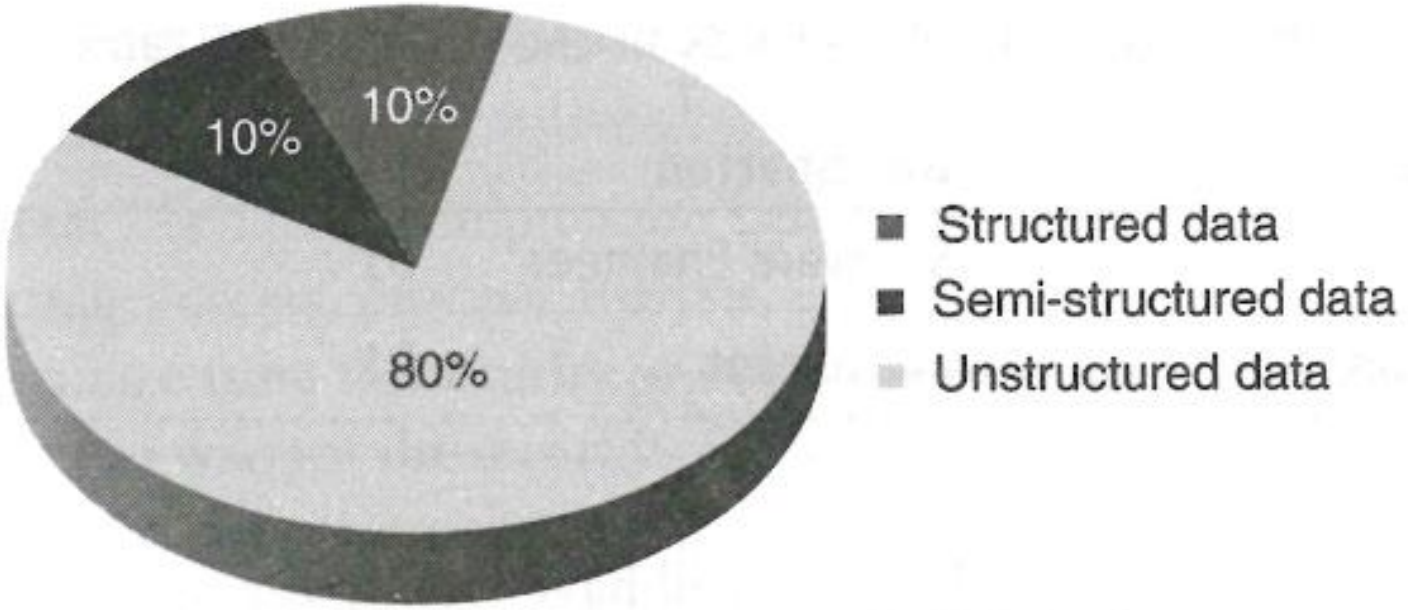
Semi-structured data

- Does not conform to a data model but has some structure
- Not in a form which can be used easily by a computer program;
- For example, emails, XML, markup languages like HTML, etc.
- Metadata for this data is available but is not sufficient.

Structured data

- Data in an organized form (e.g., in rows and columns)
- Can be easily used by a computer program
- Relationships exist between entities of data
- Data stored in databases is an example of structured data

CLASSIFICATION OF DIGITAL DATA



Approximate percentage distribution of digital data.

Structured data

- When do we say that the data is structured?
  The simple answer is when data conforms to a pre-defined schema/structure
- Think structured data, and think data model
- Most of the structured data is held in RDBMS

A relation/table with rows and columns

Column 1	Column 2	Column 3	Column 4
Row 1			

Schema of an "Employee" table in a RDBMS such as Oracle

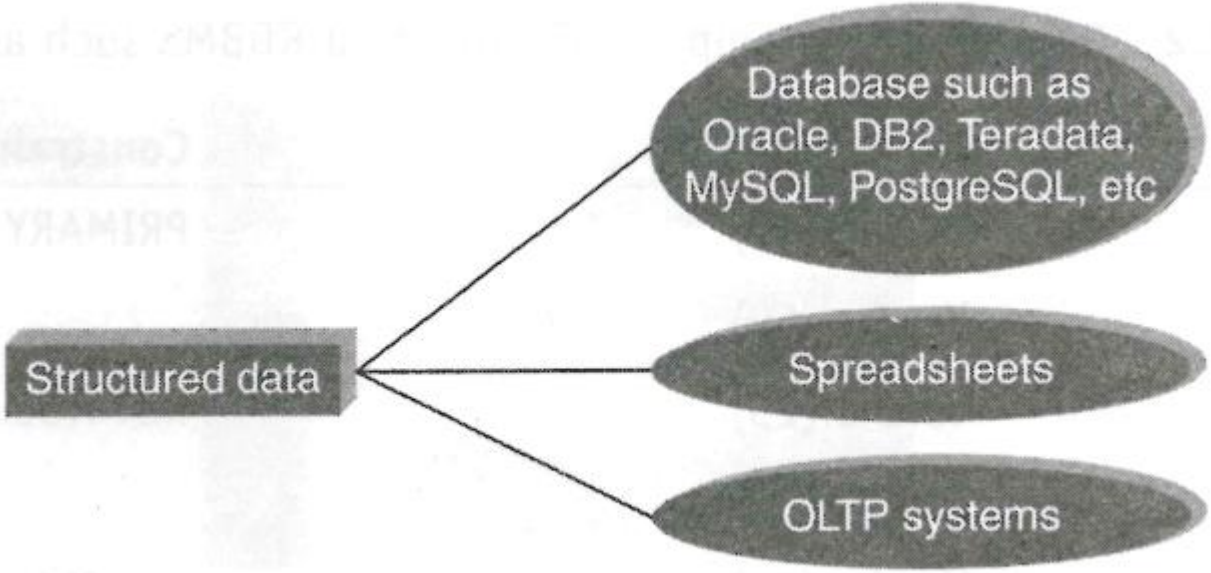
Column Name	Data Type	Constraints
EmpNo	Varchar(10)	PRIMARY KEY
EmpName	Varchar(50)	
Designation	Varchar(25)	NOT NULL
DeptNo	Varchar(5)	
ContactNo	Varchar(10)	NOT NULL

Sample records in the "Employee" table

EmpNo	EmpName	Designation	DeptNo	ContactNo
E101	Allen	Software Engineer	D1	0999999999
E102	Simon	Consultant	D1	0777777777

Structured data

Sources of Structured Data

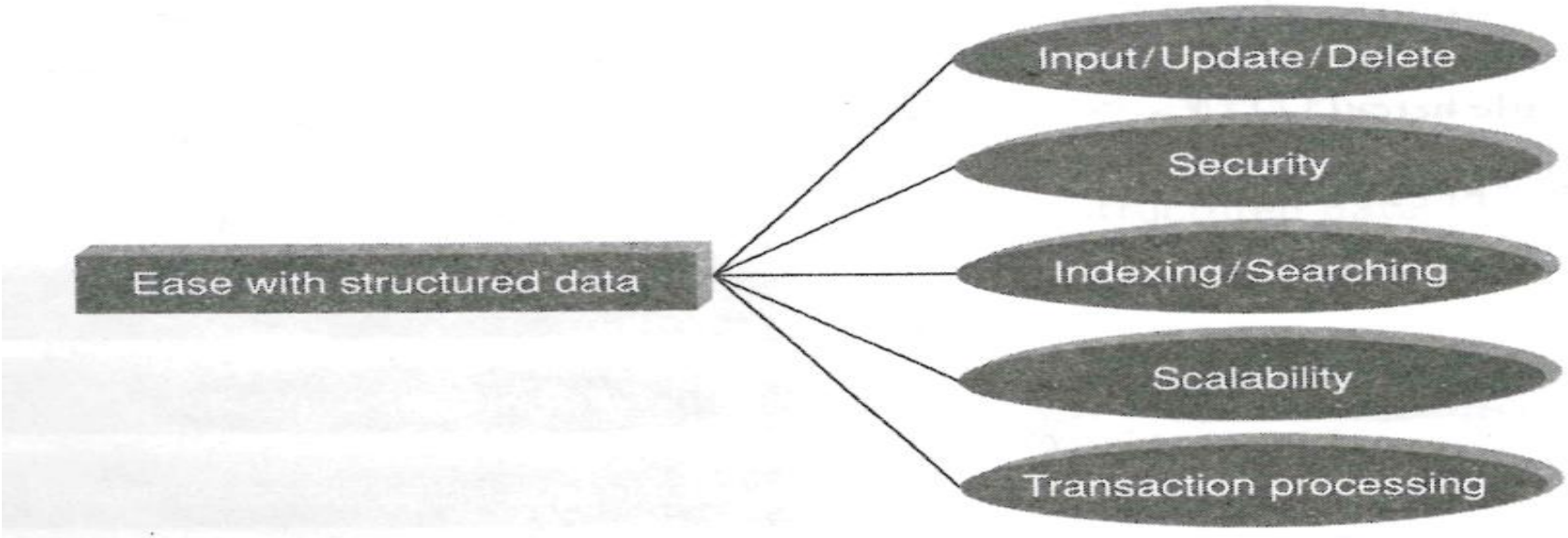


Sources of structured data.

Structured data

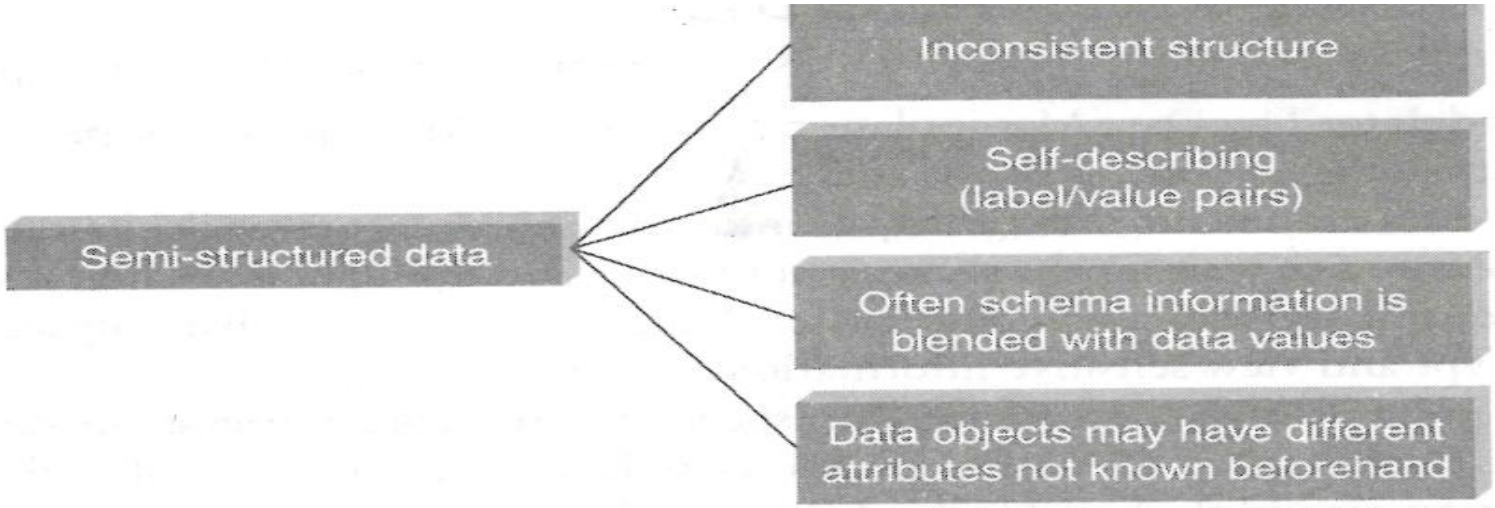
Ease of Working with Structured Data

- Insert/update/delete
- Security
- Indexing
- Scalability
- Transaction processing



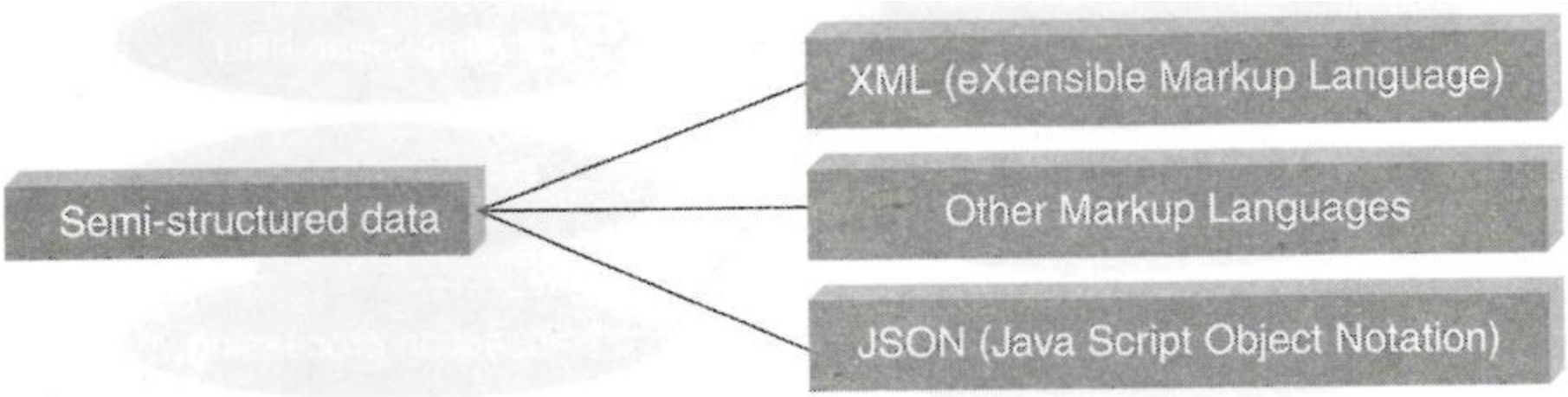
Semi-Structured Data

- It does not conform to the data models (relational databases or data tables)
- It uses tags to segregate semantic elements.
- Tags used to enforce hierarchies of records and fields within data.
- No separation between the data and the schema.
- The amount of structure used is dictated by the purpose at hand.
- In semi-structured data, entities belonging to the same class and also grouped together need not necessarily have the same set of attributes.



Characteristics of semi-structured data.

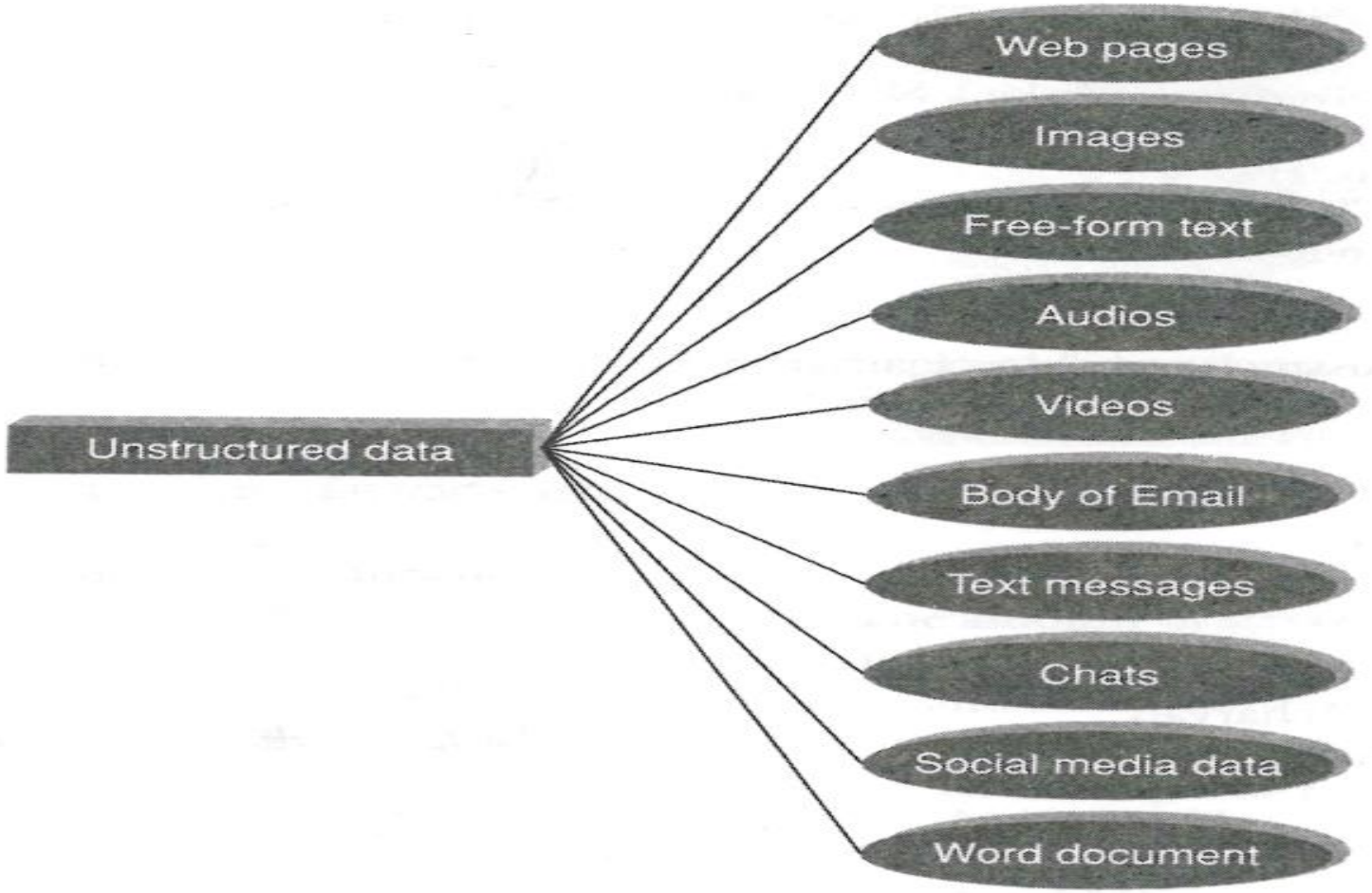
Semi-Structured Data



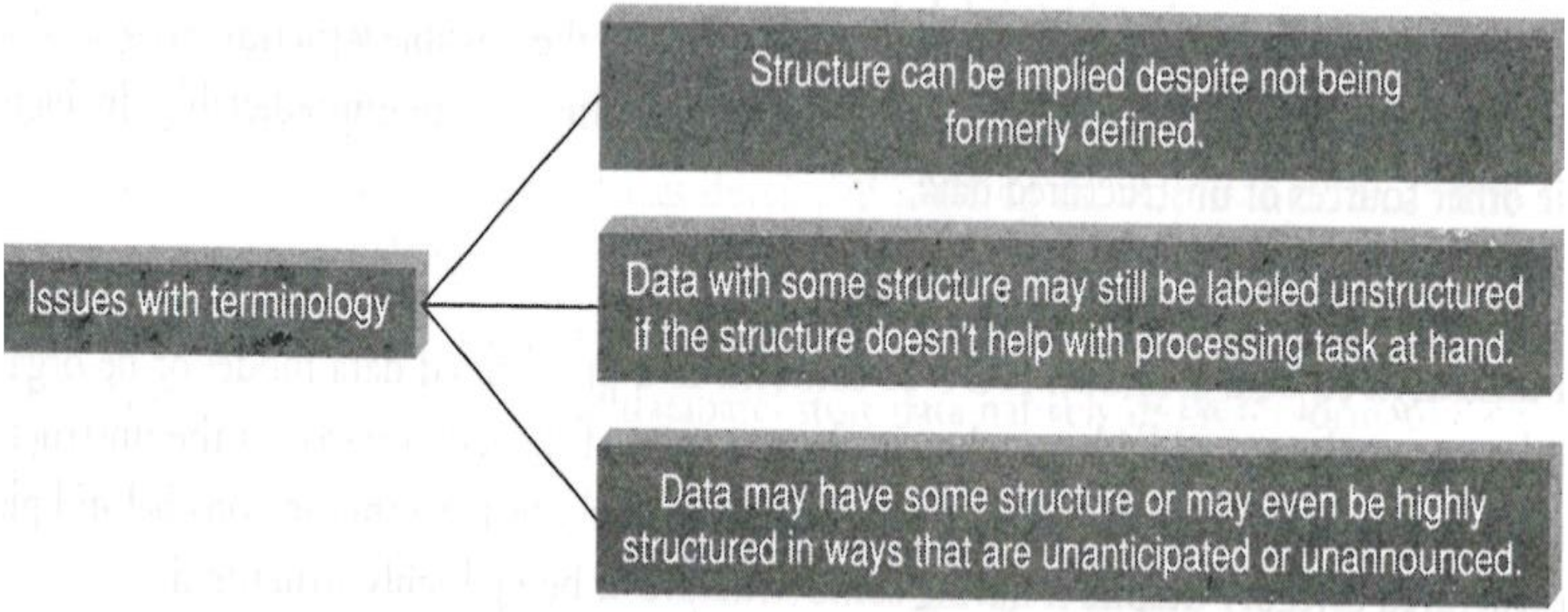
Sources of semi-structured data.

Unstructured Data

- Unstructured data does not conform to any pre-defined data model



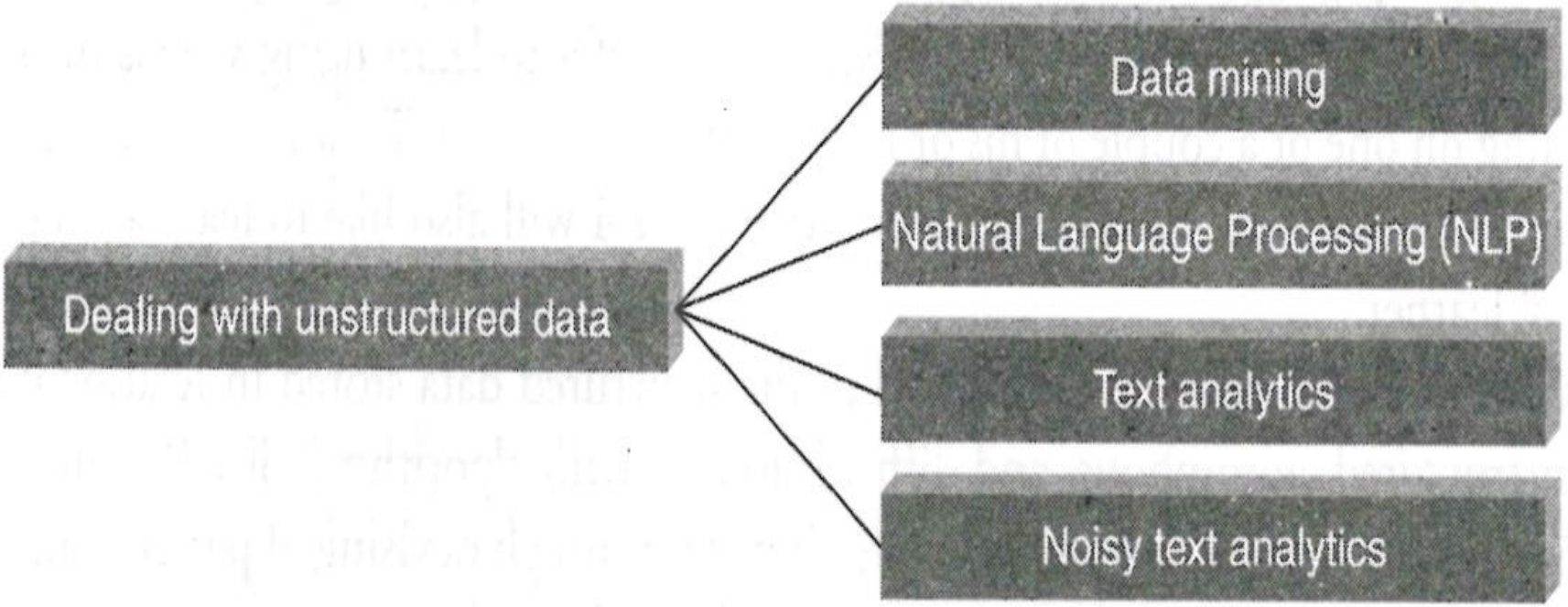
Issues with "Unstructured" Data



Issues with terminology of unstructured data.

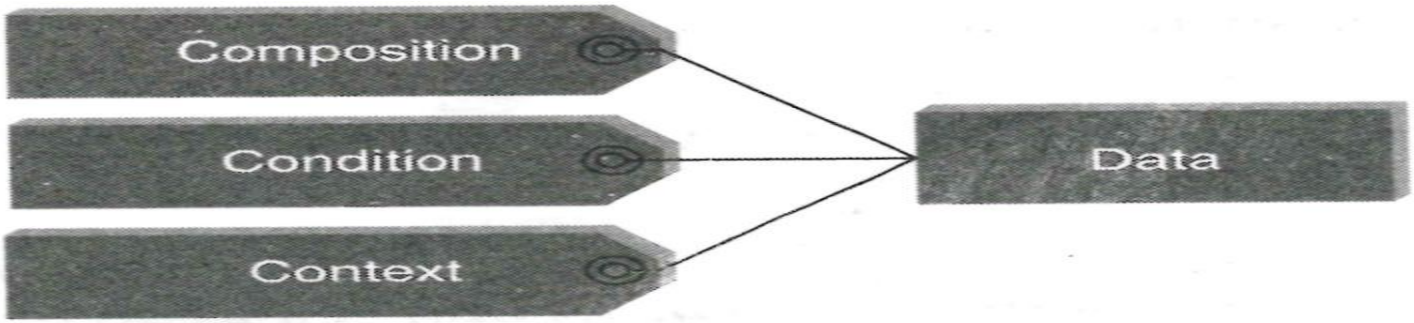
Unstructured Data

How to Deal with Unstructured Data?



Dealing with unstructured data.

CHARACTERISTICS OF DATA



Characteristics of data.

Composition:

- The composition of data deals with the structure of data, that is, the sources of data, granularity, types, and the nature (static or real-time)

Condition:

- Deals with the state of data, that is, "Can one use this data as is for analysis?" or "Does it require cleansing for further enhancement and enrichment?"

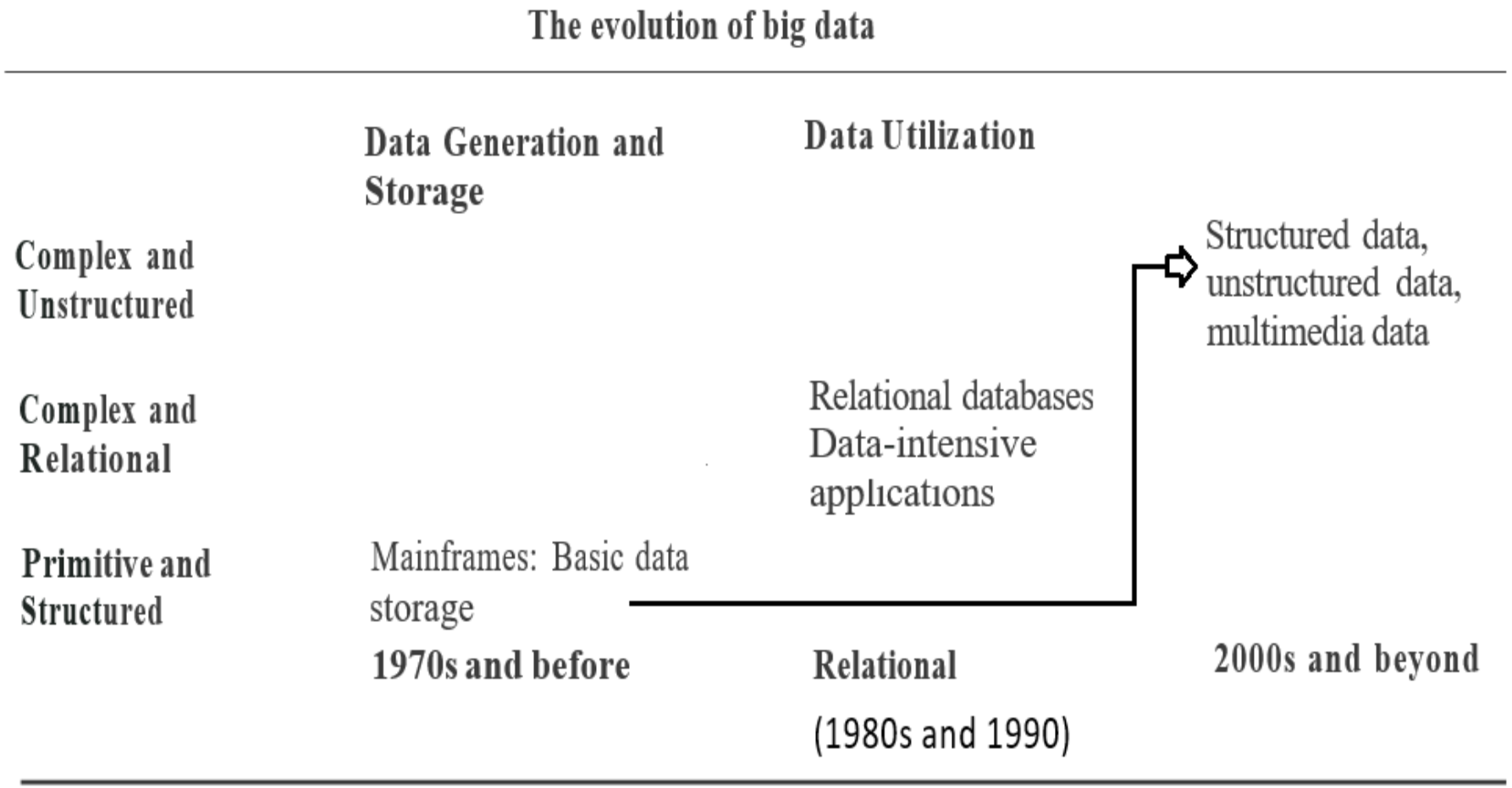
Context:

- Deals with
  - Where has this data been generated?
  - Why was this data generated?
  - How sensitive is this data?
  - What are the events associated with this data?

CHARACTERISTICS OF DATA

- Big Data is about complexity ...
  - complexity in terms of multiple and unknown datasets,
  - in terms of exploding volume,
  - in terms of speed at which the data is being generated and speed at which it needs to be processed, and
  - in terms of the variety of data (internal or external, behavioral or social) that is being generated .

EVOLUTION OF BIG DATA

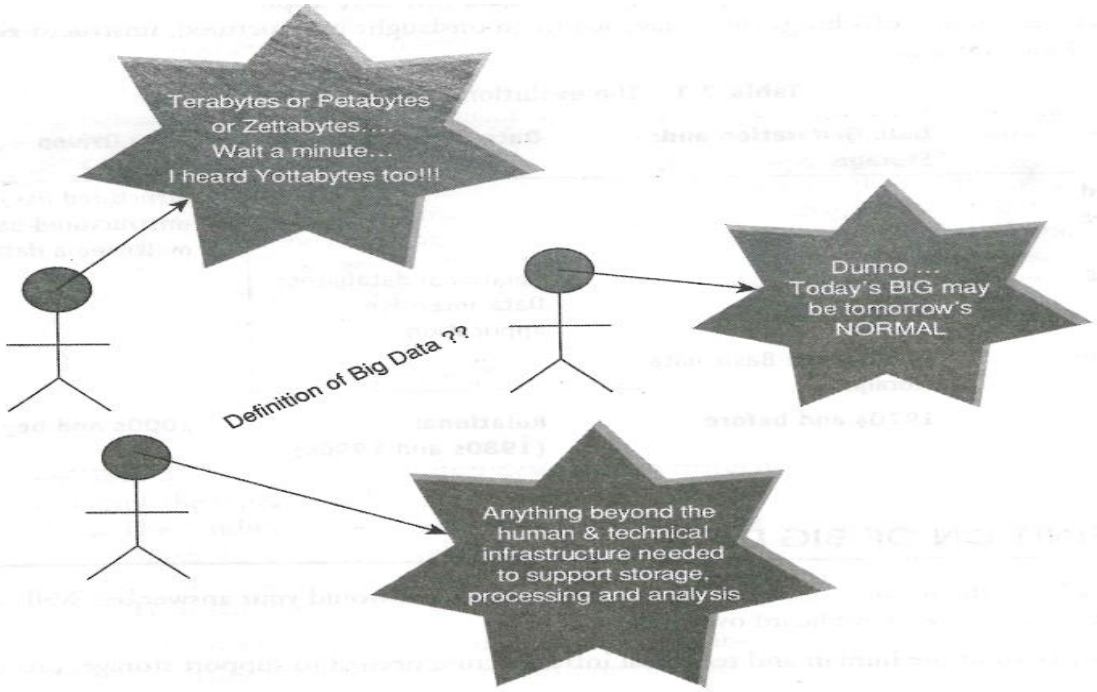


DEFINITION OF BIG DATA

Question: "Define Big Data", what would your answer be?

Few responses chat heard over time:

1. Anything beyond human and technical infrastructure needed to support storage, processing, and analysts.
2. Today's BIG may be tomorrow's NORMAL
3. Terabytes or petabytes or zettabytes of data
4. It is about **3 Vs** (Volume, Velocity, Variety)



Definition of big data.

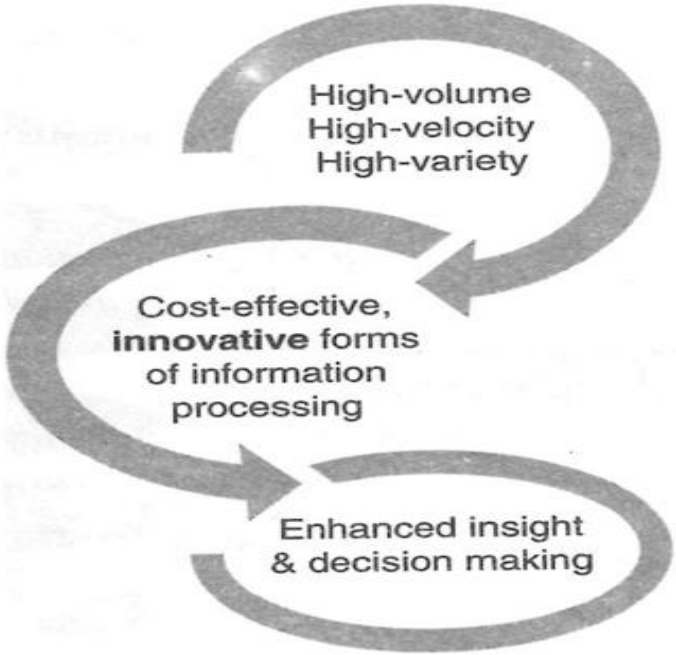
DEFINITION OF BIG DATA

- Wikipedia big data
  - An all-encompassing term for any collection of data sets so **large** and **complex** that it becomes **difficult** to process using on-hand data management tools or traditional data processing applications.

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



Definition of big data – Gartner.

Data → Information → Actionable intelligence → Better decisions → Enhanced business value

Who is generating Big Data?

Social

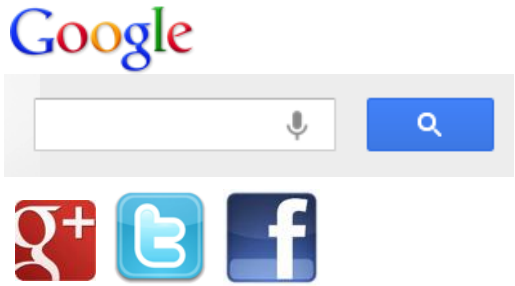
User Tracking & Engagement

Homeland Security

eCommerce

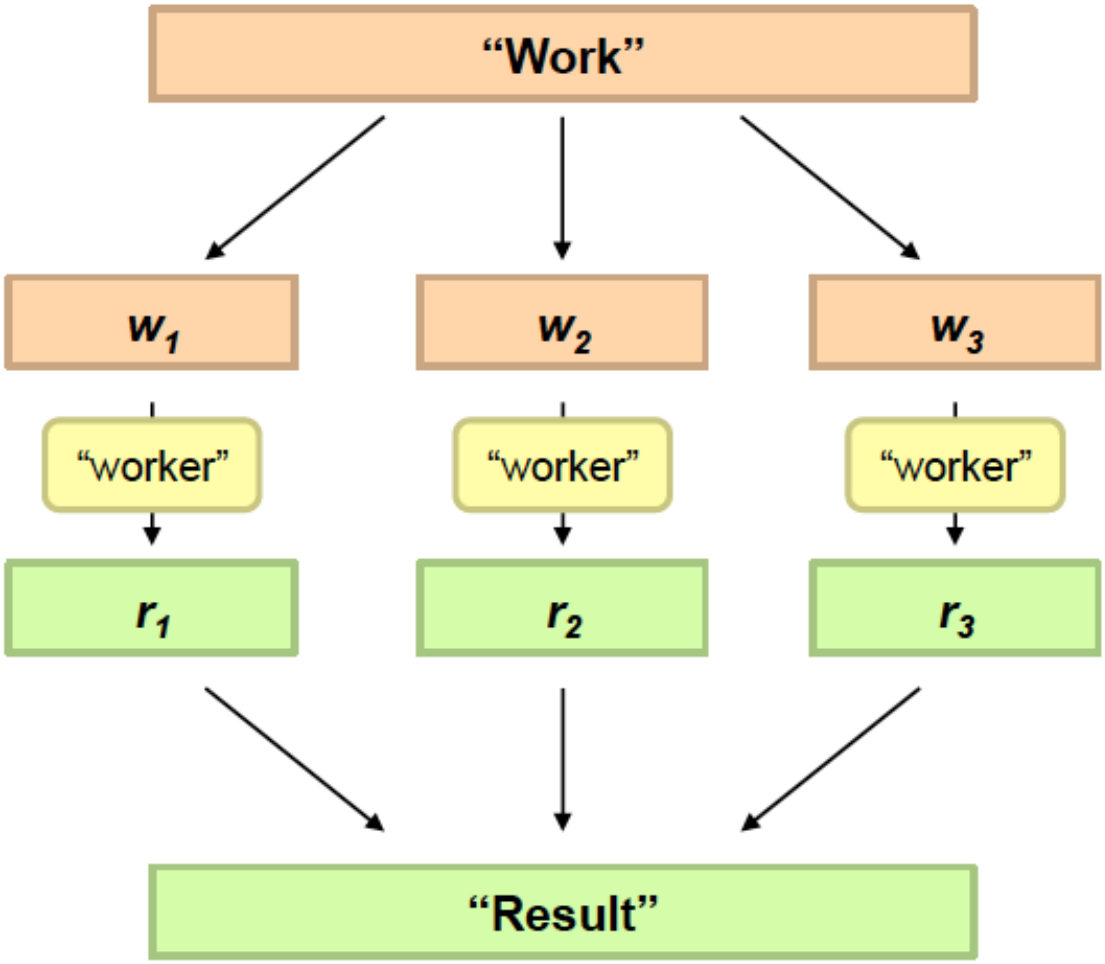
Financial Services

Real Time Search



Philosophy to Scale for Big Data

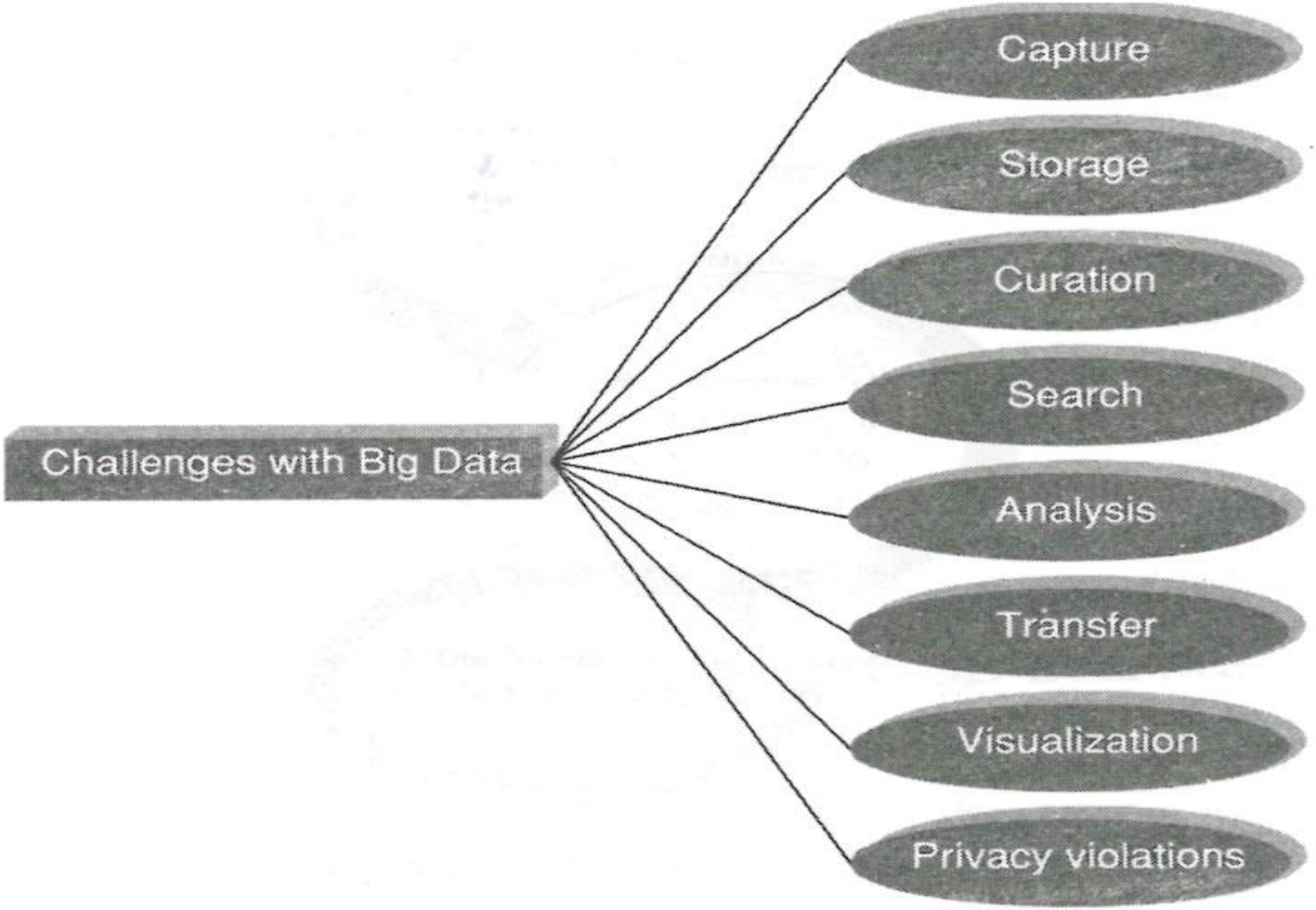
Divide and Conquer



Divide Work

Combine Results

CHALLENGES WITH BIG DATA

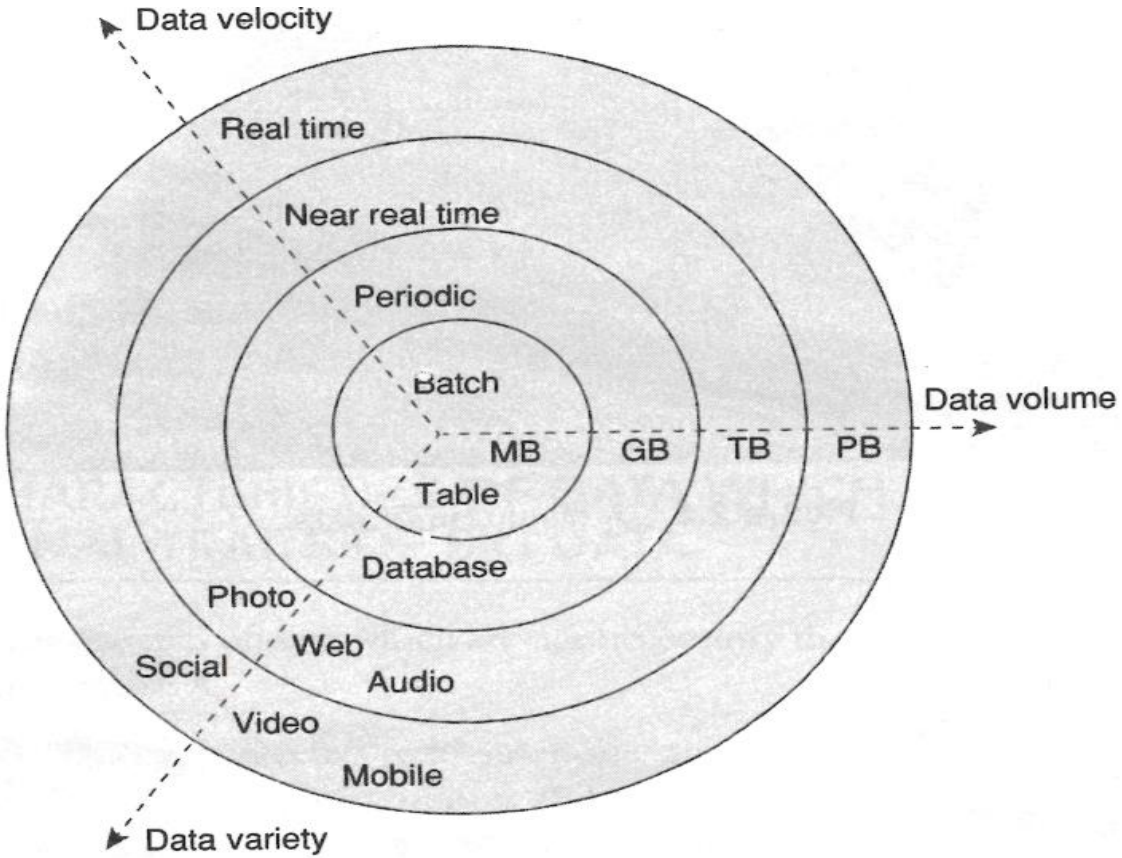


CHALLENGES WITH BIG DATA

1. Data today is growing at an exponential rate.
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.
3. The other challenge is to decide on the period of retention of big data
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. Challenges with respect to capture, storage, preparation, search, analysis, transfer, security, and visualization of big data.
6. Data visualization is becoming popular as a separate discipline.

WHAT IS BIG DATA?

Big data is data that is big in volume, velocity, and variety.



Data: Big in volume, variety, and velocity.

WHAT IS BIG DATA?

Volume

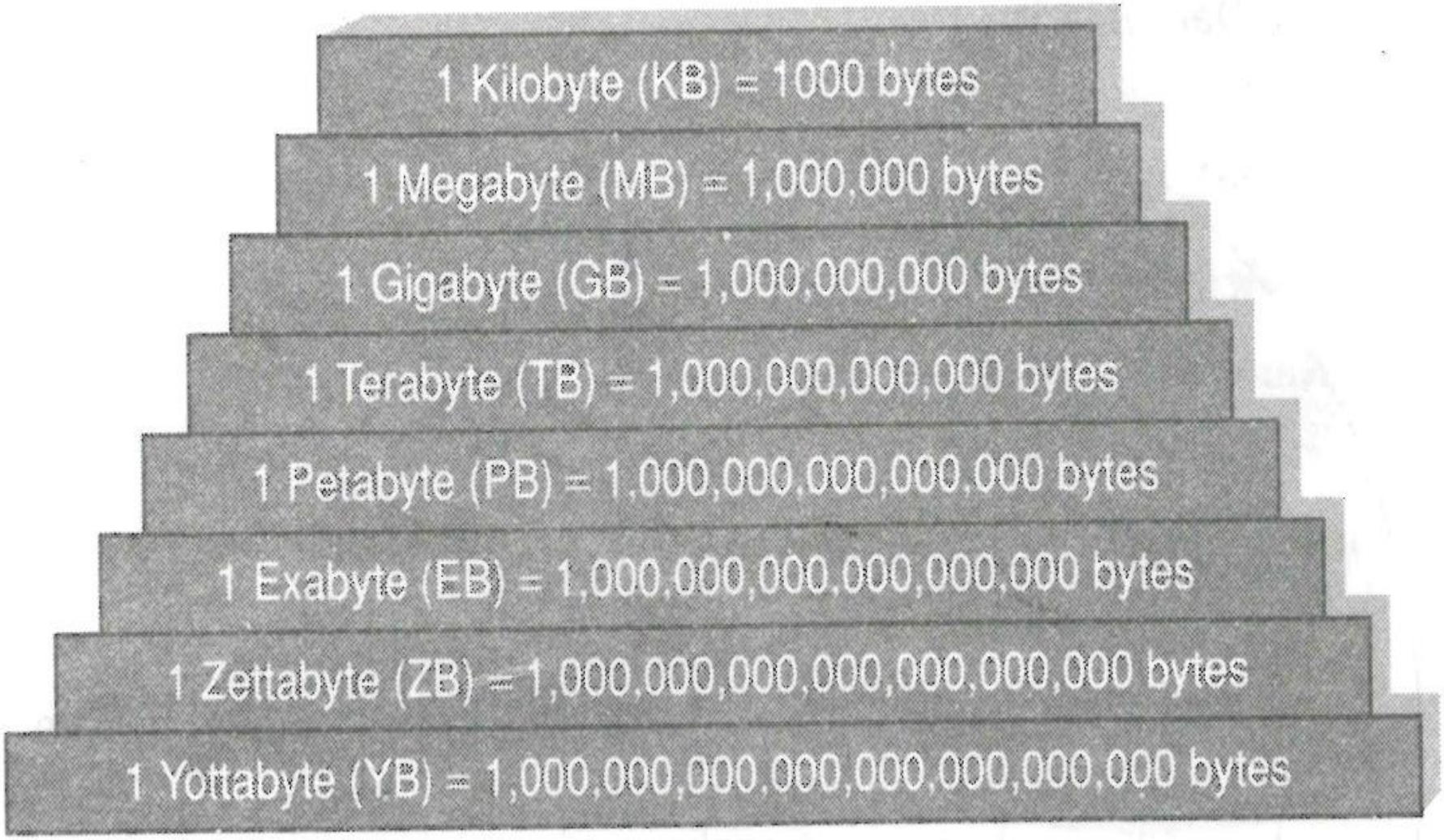
- Data grow from bits to bytes to petabytes and Exabytes.
- Where Does This Data get Generated?
- Sources of big data.

1. Typical internal data sources: Data storage, Archives: Archives of scanned documents, paper archives, customer correspondence records,
2. External data sources: Public web: Wikipedia, weather, regulatory, compliance, census, etc.
3. Both (internal + external data sources) : Sensor data, Machine log data, Social media, Business apps, Media, Docs

Growth of data

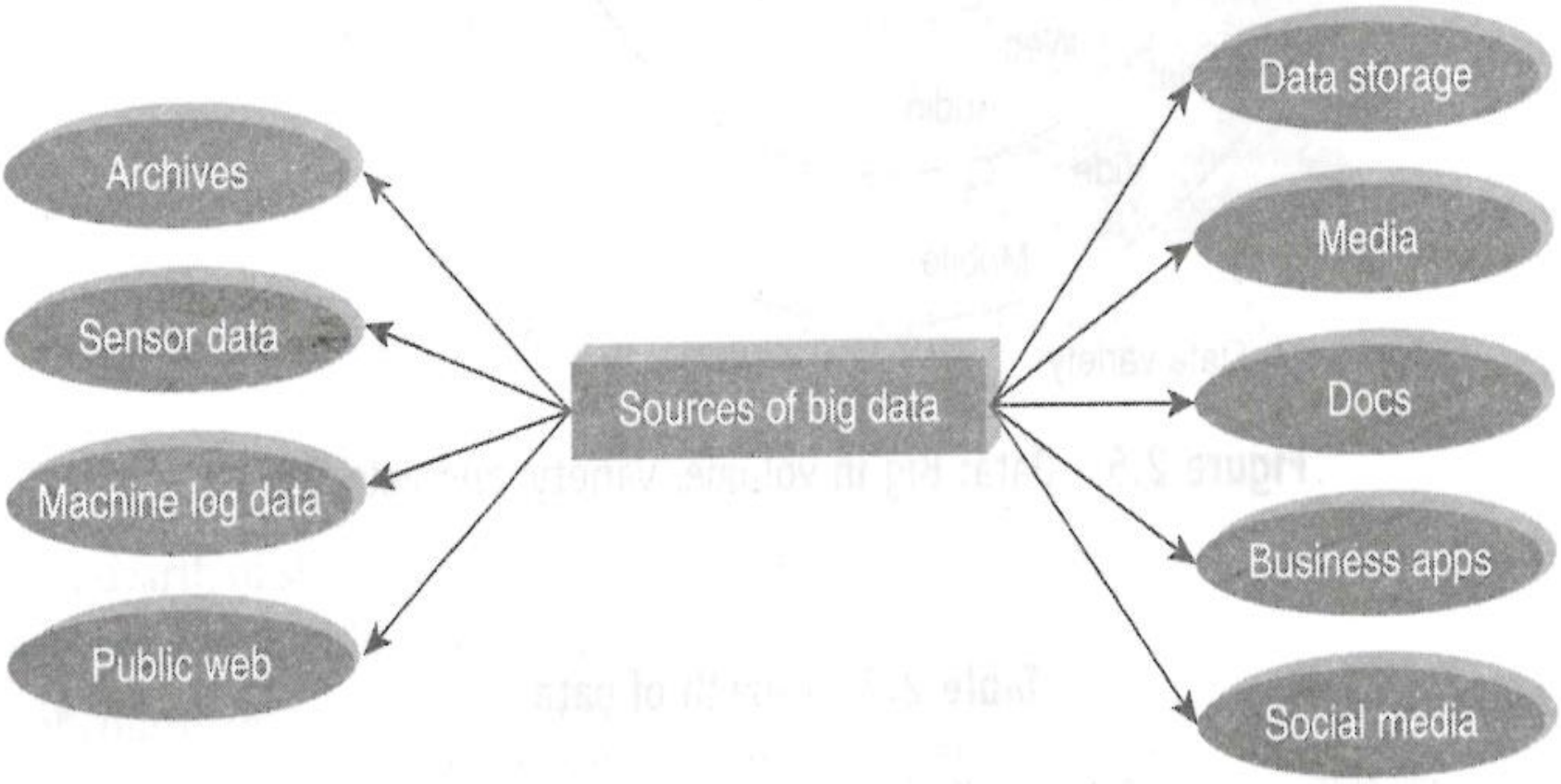
Bits	0 or 1
Bytes	8 bits
Kilobytes	1024 bytes
Megabytes	1024^2 bytes
Gigabytes	1024^3 bytes
Terabytes	1024^4 bytes
Petabytes	1024^5 bytes
Exabytes	1024^6 bytes
Zettabytes	1024^7 bytes
Yottabytes	1024^8 bytes

WHAT IS BIG DATA?



A mountain of data.

WHAT IS BIG DATA?



Sources of big data.

WHAT IS BIG DATA?

Velocity

- Have moved from the days of batch processing (remember our payroll applications) to real-time processing.

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

WHAT IS BIG DATA?

Variety

- Variety deals with a wide range of data types and sources of data.
- **Three categories:** Structured data, semi-structured data and unstructured data.

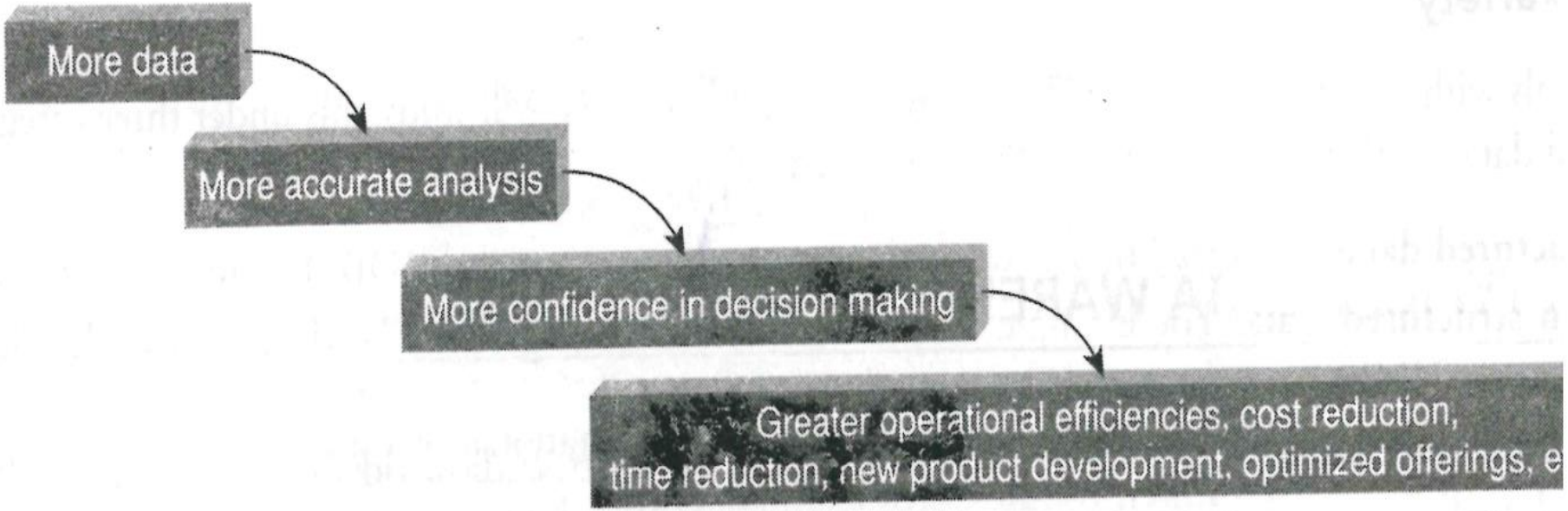
1. Structured data: From traditional transaction processing systems and RDBMS, etc.
2. Semi-structured data: For example Hyper Text Markup Language (HTML), eXtensible Markup Language (XML).
3. Unstructured data: For example unstructured text documents, audios, videos, emails, photos, PDFs, social media, etc.

There are yet other characteristics of data which are not necessarily the definitional traits of big data

1. **Veracity and validity:**
   - Veracity refers to biases, noise, and abnormality in data.
   - The key question here is: "Is all the data that is being stored, mined, and analyzed meaningful and pertinent to the problem under consideration?“
   - Validity refers to the accuracy and correctness of the data.
2. **Volatility:**
   - Volatility of data deals with, how long is the data valid?
   - And how long should it be stored?
3. **Variability:**
   - Data flows can be highly inconsistent with periodic peaks

WHY BIG DATA

More data → More accurate analysis → Greater confidence in decision making → Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offerings, etc.



Why big data?

WHY BIG DATA

ARE WE JUST AN INFORMATION CONSUMER OR DO WE ALSO PRODUCE INFORMATION?

Scenario

You have been invited to your friend's promotion party. You are happy and excited to join your friend at this important milestone in her career. You send in your confirmation through a text message. You get ready and leave for your friend's residence. On the way, you stop at a gas station to refuel. You pay using your credit card. You stop at an upmarket Archie's store to pick a good greeting card and a gift. You get the items billed at the Point of Sale system and pay cash at the counter. While at the party, you click photographs and post it on Facebook, Flickr, and the likes. Within minutes, you start to get likes and comments on your posts.

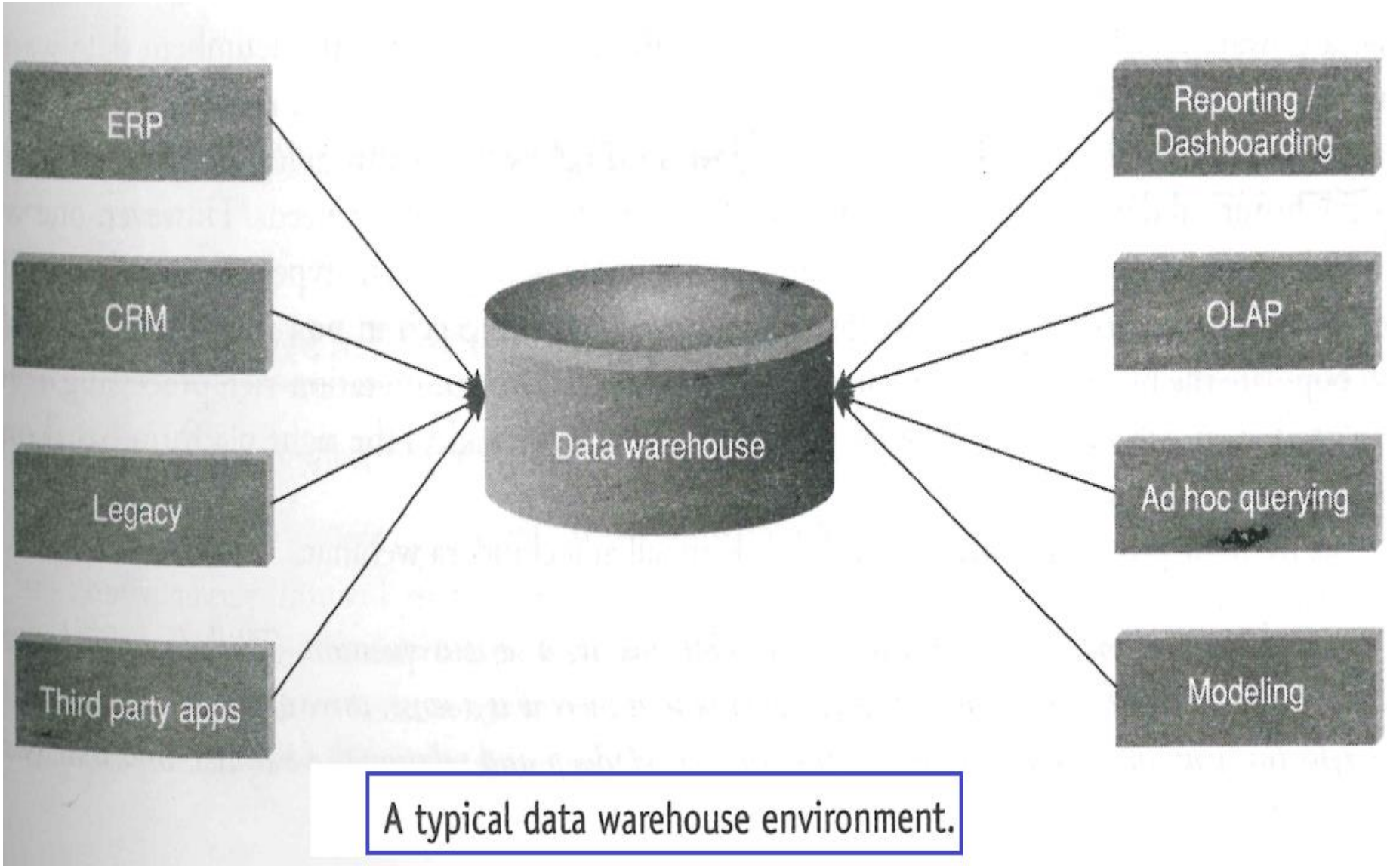
Places in this scenario where data was generated:

1. Text message to send in the confirmation to attend the promotion bash
2. Use of credit card to pay for gas/fuel at the gas station.
3. Point of Sale system at Archie's where your transaction gets recorded.
4. Photographs and posts on social networking sites.
5. Likes and comments to your post.

TRADITIONAL BUSINESS INTELLIGENCE (BI) VERSUS BIG DATA

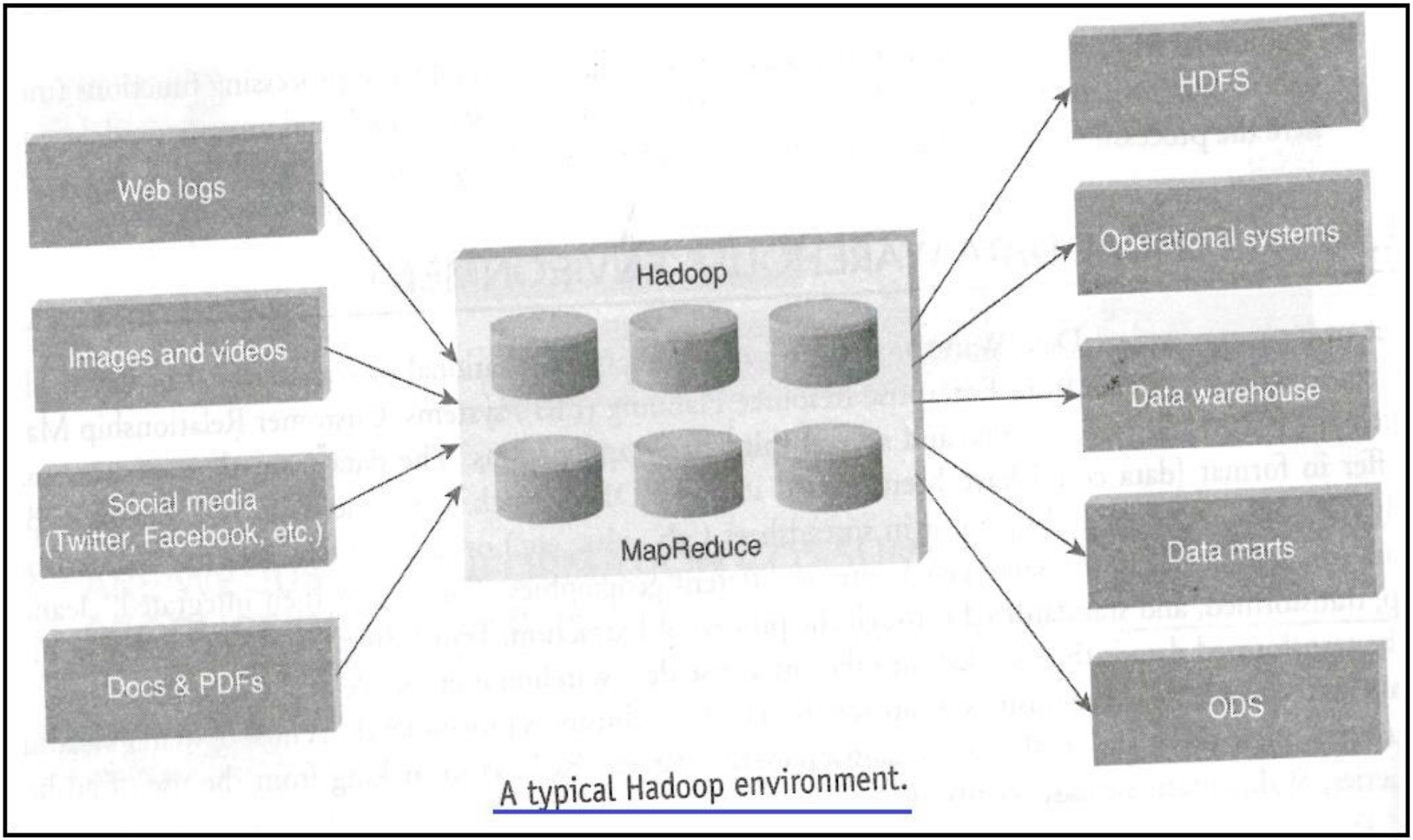
1. In traditional BI environment, all the enterprise's data is housed in a central server whereas in a big data environment data resides in a distributed file system. The distributed file system scales by scaling in or out horizontally as compared to typical database server that scales vertically. .
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 (move data to code) whereas big data is about variety: Structured, semi-structured, and unstructured data and here the processing functions are taken to the data (move code to data).

A TYPICAL DATA WAREHOUSE ENVIRONMENT



A typical data warehouse environment.

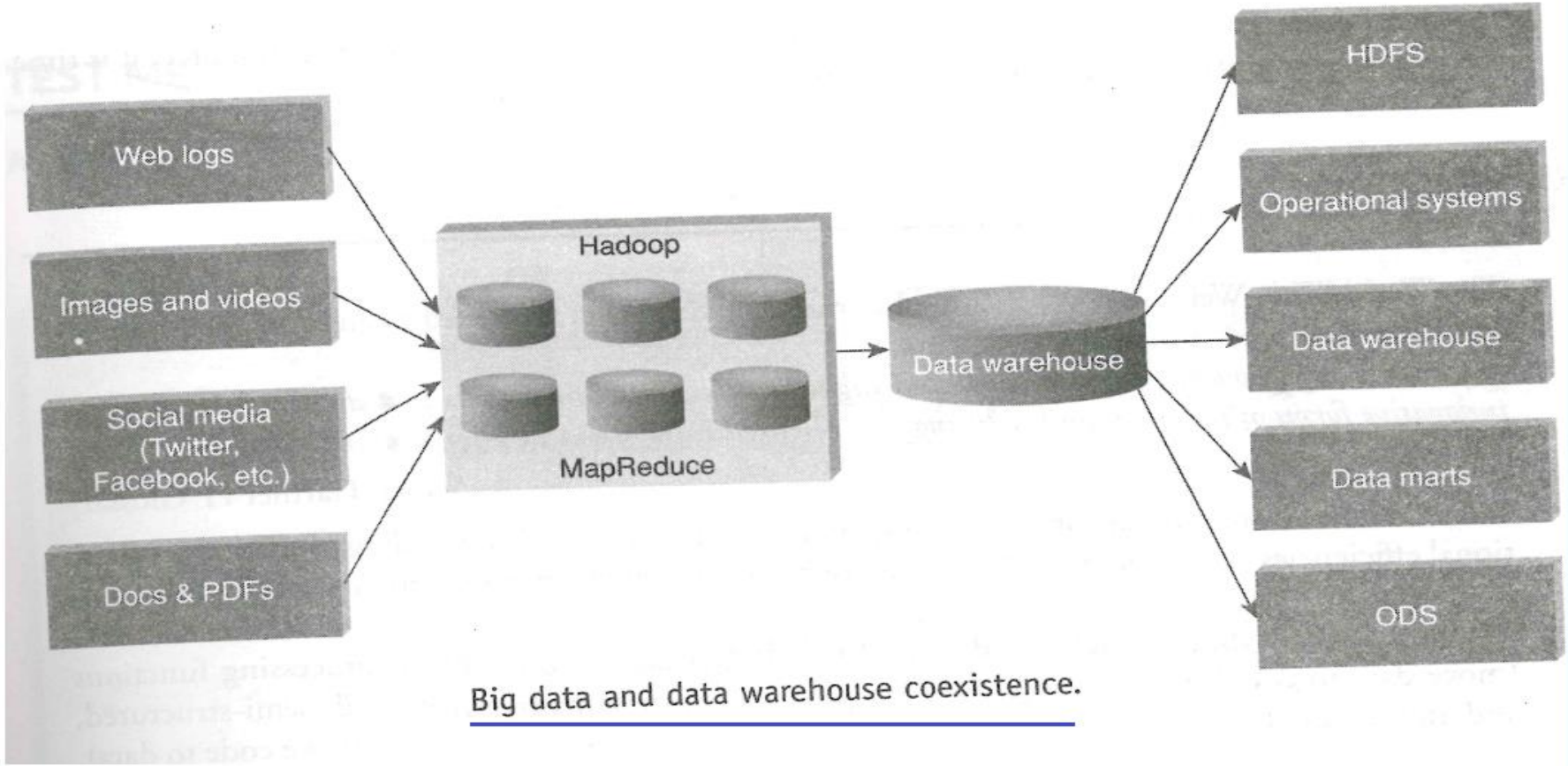
A TYPICAL HADOOP ENVIRONMENT



A typical Hadoop environment.

WHAT IS NEW TODAY?

A coexistence strategy that combines the best of legacy data warehouse and analytics environment with the new power of big data solutions is the best of both the worlds



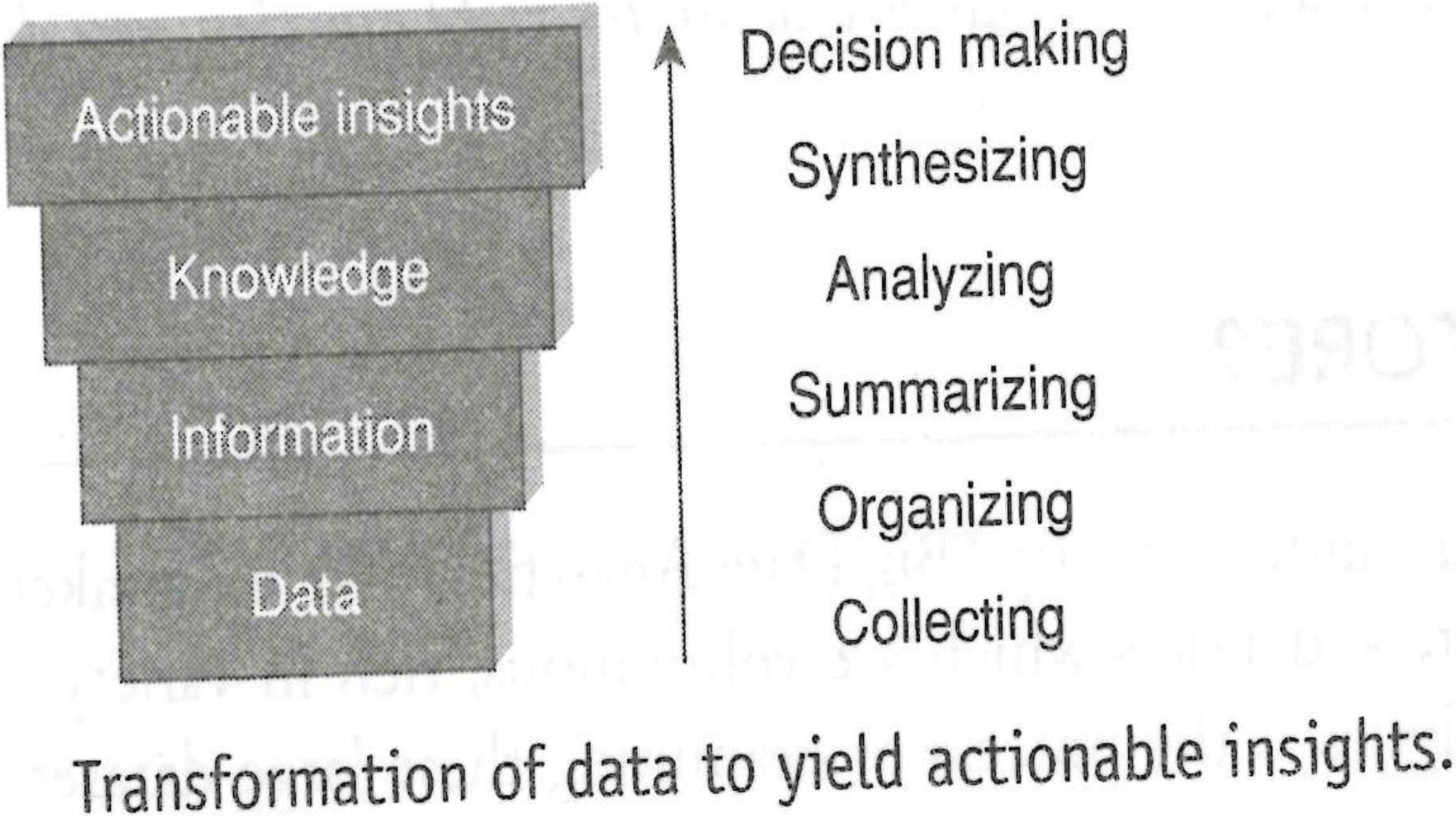
Big data and data warehouse coexistence.

WHAT IS CHANGING IN THE REALMS OF BIG DATA?

Three very important reasons why companies should compulsorily consider leveraging big data:

1. **Competitive advantage:**
   - The most important resource with any organization today is their data.
   - What they do with it will determine their fate in the market.
2. **Decision making:**
   - Decision making has shifted from the hands of the elite few to the empowered many.
   - Good decisions play a significant role in furthering customer engagement, reducing operating margins in retail, cutting cost and other expenditures in the health sector
3. **Value of data:**
   1. The value of data continues to see a steep rise.
   2. As the all-important resource, it is time to look at newer architecture, tools, and practices to leverage this.

Transformation of data to yield actionable insights.

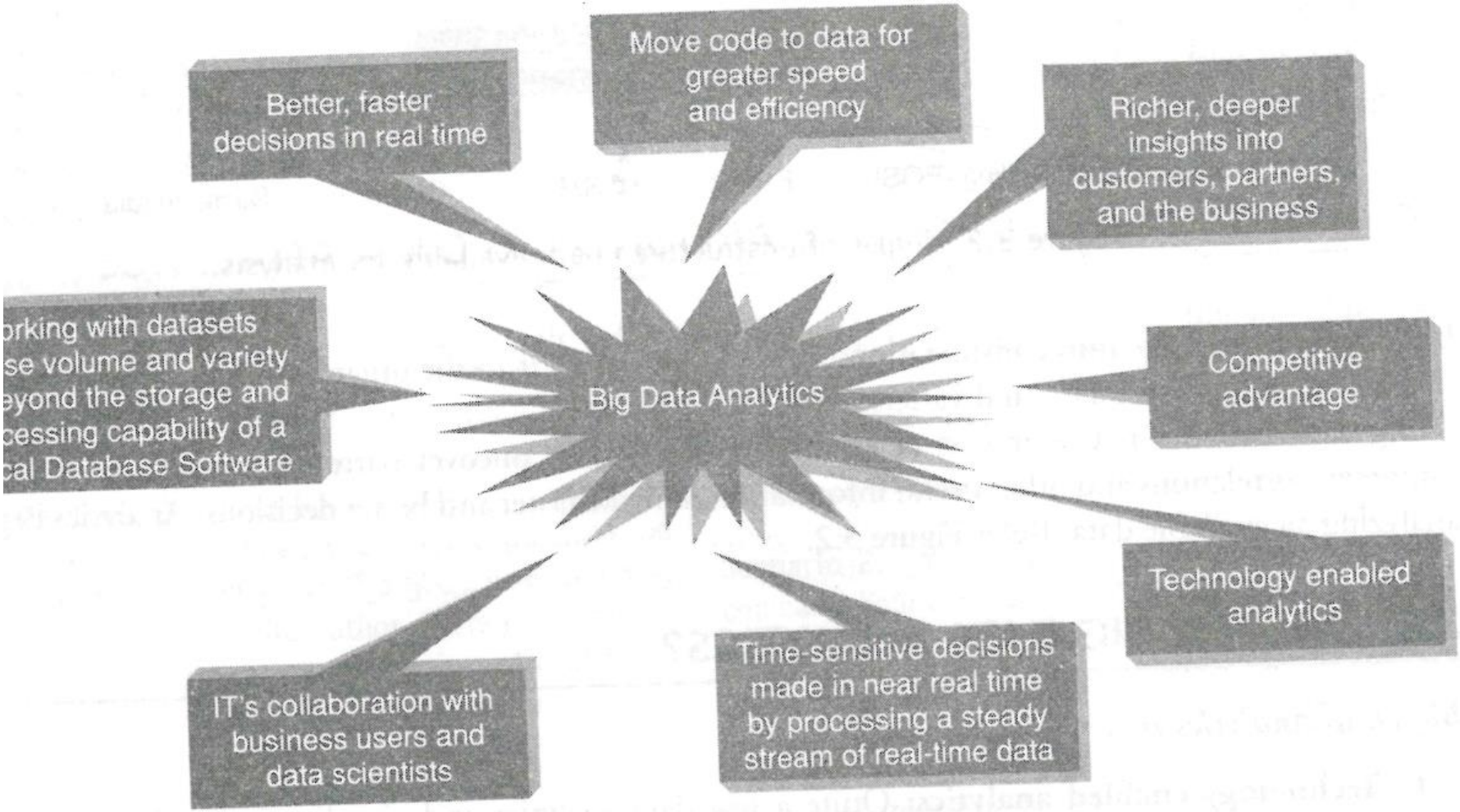


Transformation of data to yield actionable insights.

WHAT IS BIG DATA ANALYTICS?

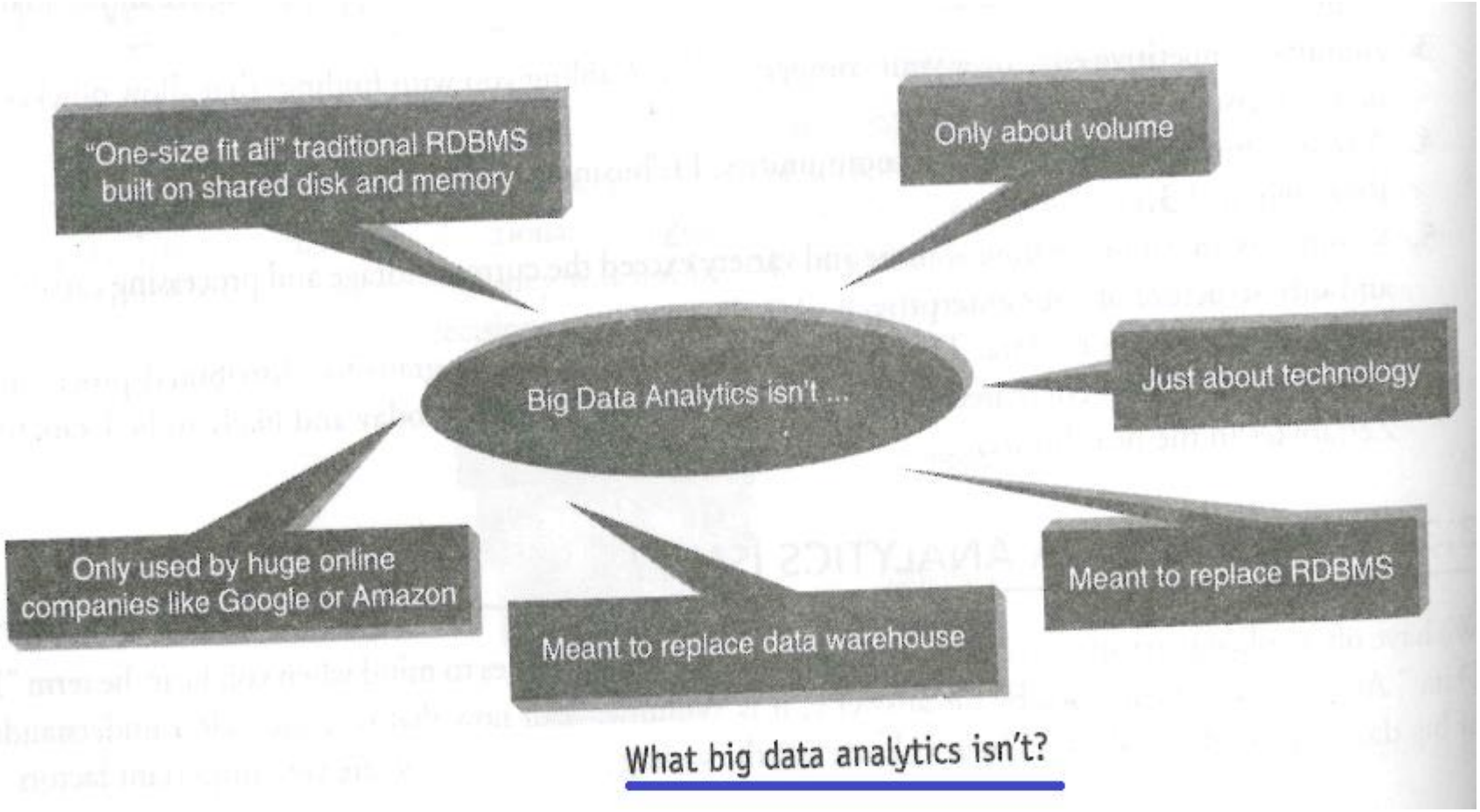
1. Technology-enabled analytics: Quite a few data analytics and visualization tools are available in the market today from leading vendors such as IBM, Tableau, SAS, R Analytics, Statistica, World Programming Systems (WPS), etc. to help process and analyze your big data.
2. About gaining a meaningful, deeper, and richer insight into your business to steer it in the right direction, understanding the customer's demographics to cross-sell and up-sell to them, better leveraging the services of your vendors and suppliers, etc.
3. About a competitive edge over your competitors by enabling you with findings that allow quicker and better decision-making.
4. A right handshake between three communities : IT, business users, and data scientists.
5. Working with datasets whose volume and variety exceed the current storage and processing capabilities and infrastructure of your enterprise.
6. About moving code to data.

WHAT IS BIG DATA ANALYTICS?



What is big data analytics?

WHAT BIG DATA ANALYTICS ISN'T?

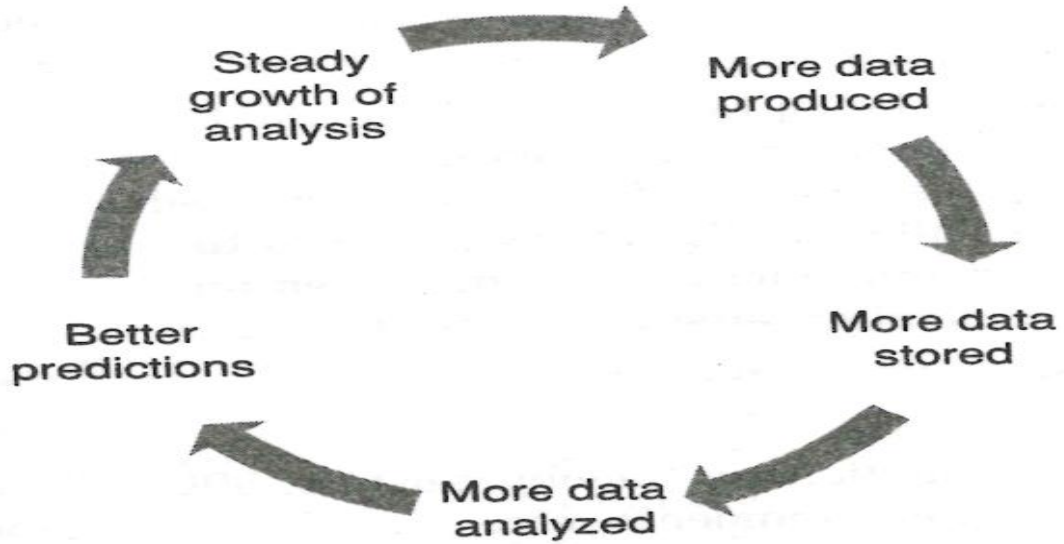


What big data analytics isn't?

WHY THIS SUDDEN HYPE AROUND BIG DATA ANALYTICS?

Three foremost reasons:

1. Data is growing at a 40% compound annual rate. 90% of the world's data created in the past 2 years alone.
2. Cost per gigabyte of storage has hugely dropped.
3. There are an overwhelming number of user-friendly analytics tools available in the market today.



What big data entails?

CLASSIFICATION OF ANALYTICS

There are basically two schools of thought:

1. Those that classify analytics into basic, operationalized, advanced , and monetized.
2. Those that classify analytics into analytics 1.0, analytics 2.0, and analytics 3.0.

CLASSIFICATION OF ANALYTICS

First School of Thought

1. Basic analytics:
   - This primarily is slicing and dicing of data to help with basic business insights.
   - This is about reporting on historical data, basic visualization, etc.
2. Operationalized analytics:
   - It is operationalized analytics if it gets woven into the enterprise's business processes.
3. Advanced analytics:
   - This largely is about forecasting for the future by way of predictive and prescriptive modeling.
4. Monetized analytics:
   - This is analytics in use to derive direct business revenue.

CLASSIFICATION OF ANALYTICS

Analytics 1.0, 2.0, and 3.0

Analytics 1.0	Analytics 2.0	Analytics 3.0
Era: mid 1950s to 2009	**2005 to 2012**	**2012 to present**
Descriptive statistics (report on events, occurrences, etc. of the past)	Descriptive statistics + predictive statistics (use data from the past to make predictions for the future)	Descriptive + predictive + prescriptive statistics (use data from the past to make prophecies for the future and at the same time make recommendations to leverage the situation to one's advantage)
Key questions asked: What happened? Why did it happen?	Key questions asked: What will happen? Why will it happen?	Key questions asked: What will happen? When will it happen? Why will it happen? What should be the action taken to take advantage of what will happen?
Data from legacy systems, ERP, CRM, and 3rd party applications.	Big data	A blend of big data and data from legacy systems, ERP, CRM, and 3rd party applications.
Small and structured data sources. Data stored in enterprise data warehouses or data marts.	Big data is being taken up seriously. Data is mainly unstructured, arriving at a much higher pace. This fast flow of data entailed that the influx of big volume data had to be stored and processed rapidly, often on massive parallel servers running Hadoop.	A blend of big data and traditional analytics to yield insights and offerings with speed and impact.
Data was internally sourced.	Data was often externally sourced.	Data is both being internally and externally sourced.
Relational databases	Database appliances, Hadoop clusters, SQL to Hadoop environments, etc.	In memory analytics, in database processing, agile analytical methods, machine learning techniques, etc.

CHALLENGES THAT PREVENT BUSINESSES FROM CAPITALIZING ON BIG DATA

1. Obtaining executive sponsorships for investments in big data and its related activities (such as training, etc.).
2. Getting the business units to share information across organizational silos.
3. Finding the right skills (business analysts and data scientists) that can manage large amounts of structured, semi-structured, and unstructured data and create insights from it.
4. Determining the approach to scale rapidly and elastically. Need to address the storage and processing of large volume, velocity, and variety of big data.
5. Deciding whether to use structured or unstructured, internal or external data to make business decisions.
6. Choosing the optimal way to report findings and analysis of big data (visual presentation and analytics) for the presentations to make the most sense.
7. Determining what to do with the insights created from big data.

TOP CHALLENGES FACING BIG DATA

1. Scale
2. Security
3. Schema
4. Continuous availability
5. Consistency
6. Partition tolerant
7. Data quality

WHY IS BIG DATA ANALYTICS IMPORTANT?

- Big data analytics helps organizations harness their data and use it to identify new opportunities.
- That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers.

1. Cost reduction

- Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data
- They can identify more efficient ways of doing business.

2. Faster, better decision making.

- With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.

3. New products and services.

- With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want.
- Thomas H. Davenport (Fellow at the MIT) points out that with big data analytics, more companies are creating new products to meet customers' needs.

WHY IS BIG DATA ANALYTICS IMPORTANT?

Various approaches to analysis of data and what it leads to:

1. Reactive - Business Intelligence
2. Reactive - Big Data Analytics
3. Proactive – Analytics
4. Proactive - Big Data Analytics

WHY IS BIG DATA ANALYTICS IMPORTANT?

1. Reactive - Business Intelligence:

- It allows the businesses to make faster and better decisions by providing the right information to the right person at the right time in the right format.
- It is about analysis of the past or historical data and then displaying the findings of the analysis or reports in the form of enterprise dashboards, alerts, notifications, etc.
- It has support for both pre-specified reports as well as ad hoc querying.

WHY IS BIG DATA ANALYTICS IMPORTANT?

2. Reactive - Big Data Analytics:

- Here the analysis is done on huge datasets but the approach is still reactive as it is still based on static data.

WHY IS BIG DATA ANALYTICS IMPORTANT?

3. Proactive - Analytics:

- This is to support futuristic decision making by the use of data mining, predictive modeling, text mining, and statistical analysis.
- This analysis is not on big data as it still uses the traditional database management practices on big data and therefore has severe limitations on the storage capacity and the processing capability.

WHY IS BIG DATA ANALYTICS IMPORTANT?

4. Proactive - Big Data Analytics:

- This is sieving through terabytes, petabytes, Exabyte's of information to filter out the relevant data to analyze.
- This also includes high performance analytics to gain rapid insights from big data and the ability to solve complex problems using more data.

Thank You !!!