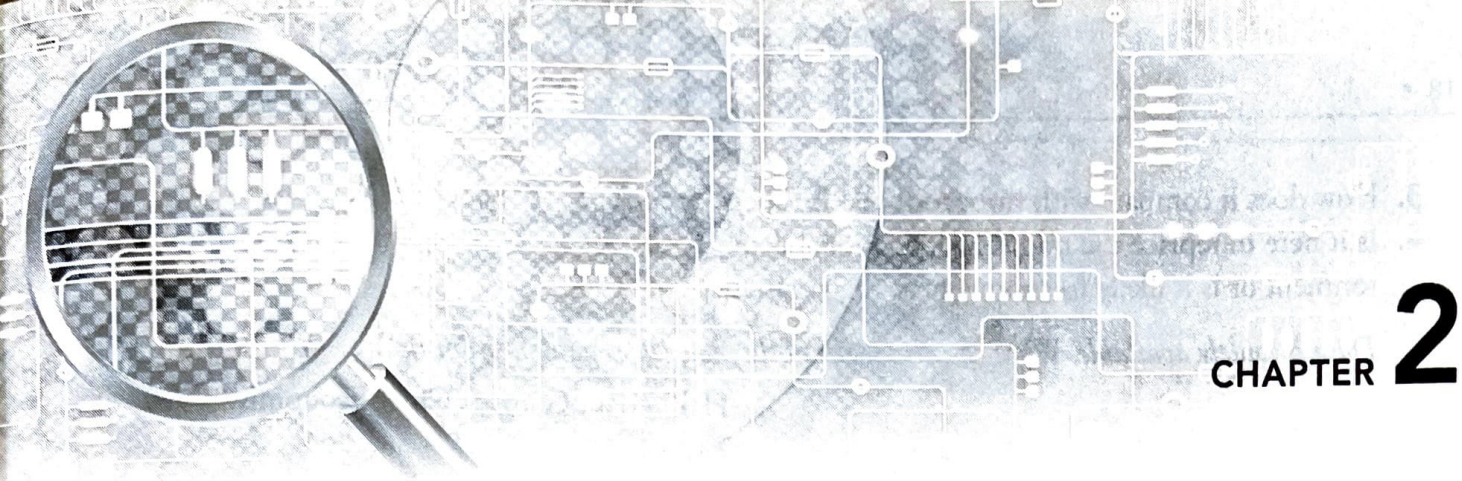


CHAPTER 2

Introduction to Big Data

BRIEF CONTENTS

- What's in Store?
- Characteristics of Data
- Evolution of Big Data
- Definition of Big Data
- Challenges with Big Data
- What is Big Data?
  - Volume
  - Velocity
  - Variety
- Other Characteristics of Data Which are Not Definitional Traits of Big Data
- Why Big Data?
- Are We Just an Information Consumer or Do We Also Produce Information?
- Traditional Business Intelligence (BI) versus Big Data
- A Typical Data Warehouse Environment
- A Typical Hadoop Environment
- What is New Today?
  - Coexistence of Big Data and Data Warehouse
- What is Changing in the Realms of Big Data?

"Data is the new science. Big Data holds the answers."

– Pat Gelsinger, the Chief Executive Officer of VMware, Inc. and former Chief Operating Officer of EMC Corporation

WHAT'S IN STORE?

This chapter focuses on defining and explaining big data. The "Internet of Things" and its widely ultra-connected nature are leading to a burgeoning rise in big data. There is no dearth of data for today's enterprise. On the contrary, they are mired in data and quite deep at that. That brings us to the following questions:

1. Why is it that we cannot forego big data?
2. How has it come to assume such magnanimous importance in running business?

3. How does it compare with the traditional Business Intelligence (BI) environment?
4. Is it here to replace the traditional, relational database management system and data warehouse environment or is it likely to complement their existence?"

Data is widely available. What is scarce is the ability to extract wisdom from it.

Hal Varian, Google's Chief Economist, 2010

PICTURE THIS...

You recently availed the opportunity to attend a virtual classroom session from a leading training institute. You are reflecting back on the experience. Since the session was on big data, it gets you thinking on the types and volume of data that was created before, during, and after the session. It all began with you registering online a week ago for the "Big Data" course. You remember having received an acknowledgment confirming your registration. They had also stated that they will send across some reading contents two days prior to the session. And true to their word, they did. When you logged into the session, you saw that there were 493 other participants. The presenter was introducing the process on smooth learning through the session. During the session, the participants could converse with the presenter as well as with other participants using the chat facility. They had also activated a discussion forum for participants to share their learnings/views/opinions/experiences, etc. There were assignments, which would have to be attempted and submitted on their site. There was an assessment towards the end of the session that was graded. There was a feedback form that was made available at the end of the session to hear back from the participants. They also provided additional reading contents in the form of references to white papers/research papers. The lecture was recorded and made available for better learning and comprehension of the participants.

It was a good experience and you are already thinking of being part of another such experience very soon.

There is no dearth of such virtual classroom sessions being conducted today. There is a huge learning community out there eager to learn. Just think on the volume of data that gets generated, and the variety (the list of attendees, their scores and grades, their chat conversations, their assignments, the polling questions put forth by the instructor to gauge the level of understanding and participation from the learners, etc.) of data that we produce as well consume as we become part of these virtual training sessions.

2.1 CHARACTERISTICS OF DATA

Let us start with the characteristics of data. As depicted in Figure 2.1, data has three key characteristics:

1. **Composition:** The composition of data deals with the structure of data, that is, the sources of data, the granularity, the types, and the nature of data as to whether it is static or real-time streaming.
2. **Condition:** The condition of data 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?"
3. **Context:** The context of data 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?" and so on.

Small data (data as it existed prior to the big data revolution) is about certainty. It is about fairly known data sources; it is about no major changes to the composition or context of data.

Most often we have answers to queries like why this data was generated, where and when it was generated, exactly how we would like to use it, what questions will this data be able to answer, and so on. Big data is

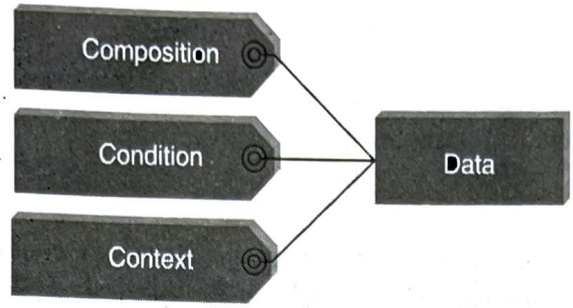


Figure 2.1 Characteristics of data.

about complexity... complexity in terms of multiple and unknown datasets, in terms of exploding volume, in terms of the speed at which the data is being generated and the 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.

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**

2.3 DEFINITION OF BIG DATA

If we were to ask you the simple question: "Define Big Data", what would your answer be? Well, we will give you a few responses that we have heard over time:

1. Anything beyond the human and technical infrastructure needed to support storage, processing, and analysis.
2. Today's BIG may be tomorrow's NORMAL.

3. Terabytes or petabytes or zettabytes of data.
4. I think it is about 3 Vs.

Refer Figure 2.2. Well, all of these responses are correct. But it is not just one of these; in fact, big data is all of the above and more.

> *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

The 3Vs concept was proposed by the Gartner analyst Doug Laney in a 2001 MetaGroup research publication, titled, *3D Data Management: Controlling Data Volume, Variety and Velocity.*

Source: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

For the sake of easy comprehension, we will look at the definition in three parts. Refer Figure 2.3.

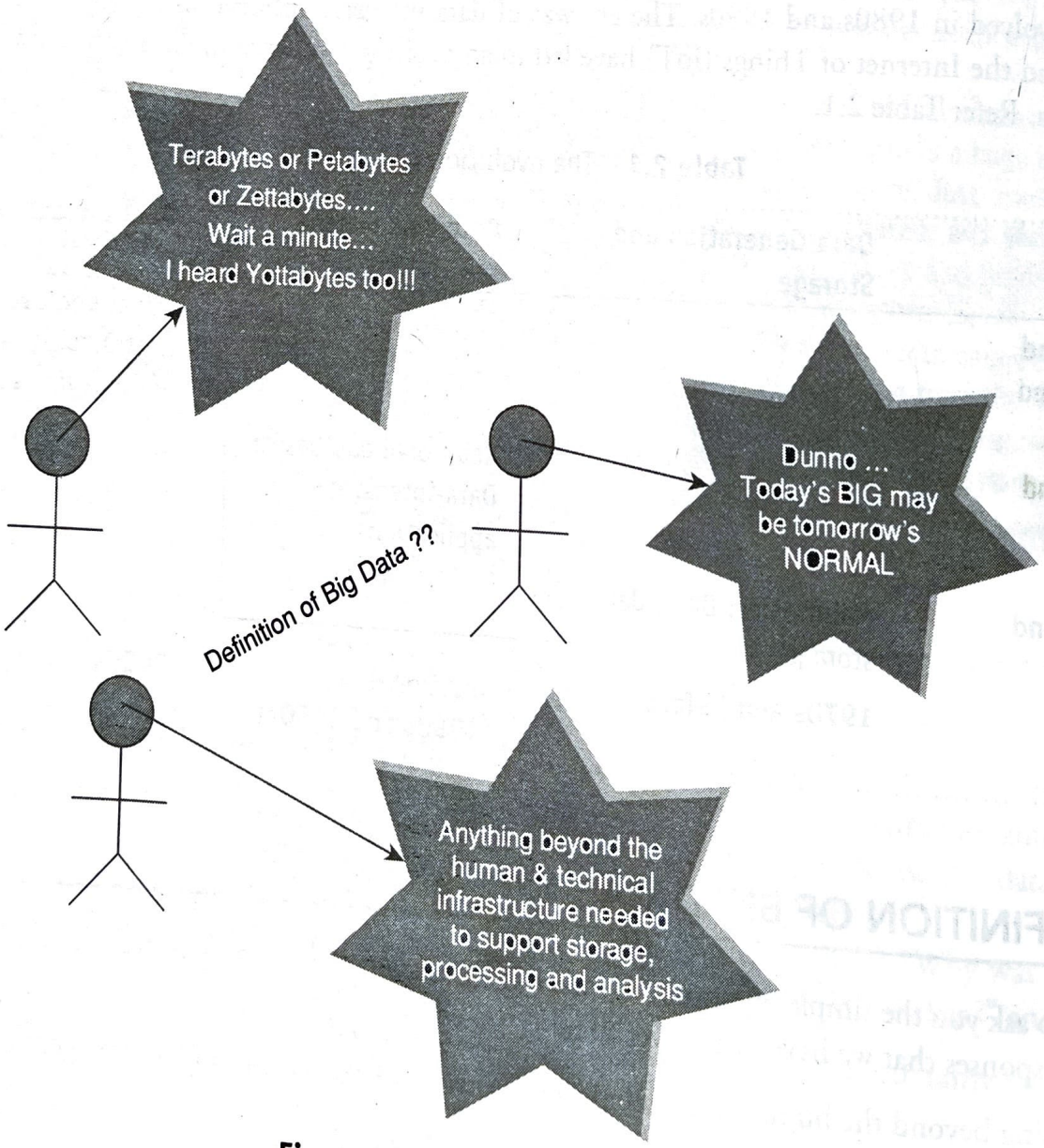


Figure 2.2 Definition of big data

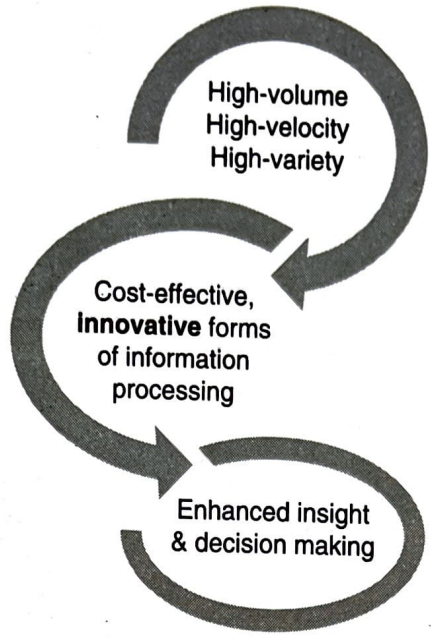


Figure 2.3 Definition of big data – Gartner.

Part I of the definition "big data is high-volume, high-velocity, and high-variety information assets" talks about voluminous data (humongous data) that may have great variety (a good mix of structured, semi-structured, and unstructured data) and will require a good speed/pace for storage, preparation, processing, and analysis.

Part II of the definition "cost effective, innovative forms of information processing" 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 "enhanced insight and decision making" 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

2.4 CHALLENGES WITH BIG DATA

Refer Figure 2.4. Following are a few 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.

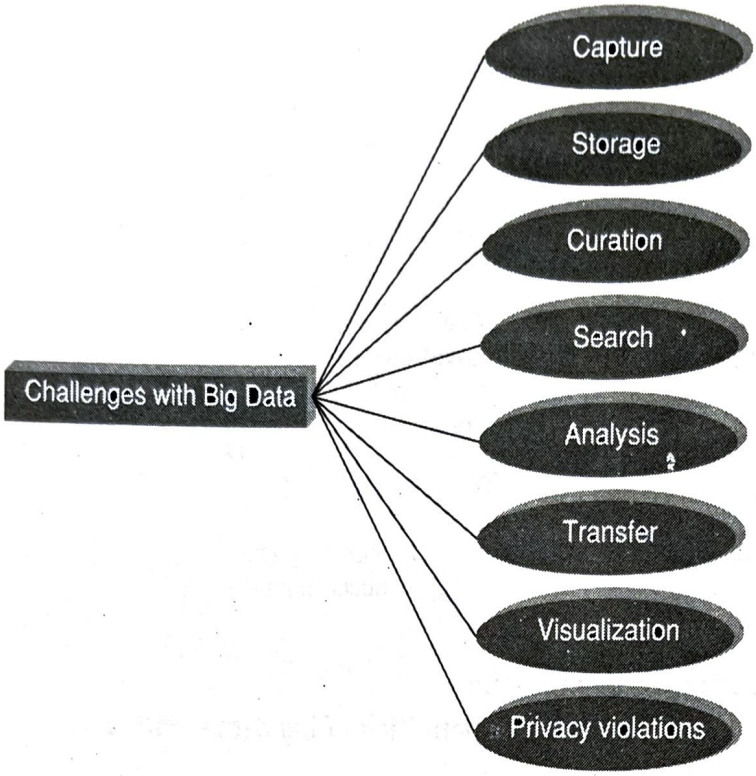


Figure 2.4 Challenges with big data.

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. 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.

2.5 WHAT IS BIG DATA?

Big data is data that is big in volume, velocity, and variety. Refer Figure 2.5.

2.5.1 Volume

We have seen it grow from bits to bytes to petabytes and exabytes. Refer Table 2.2 and Figure 2.6.

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

2.5.1.1 *Where Does This Data get Generated?*

There are a multitude of sources for big data. An XLS, a DOC, a PDF, etc. is unstructured data; a video on YouTube, a chat conversation on Internet Messenger, a customer feedback form on an online retail website

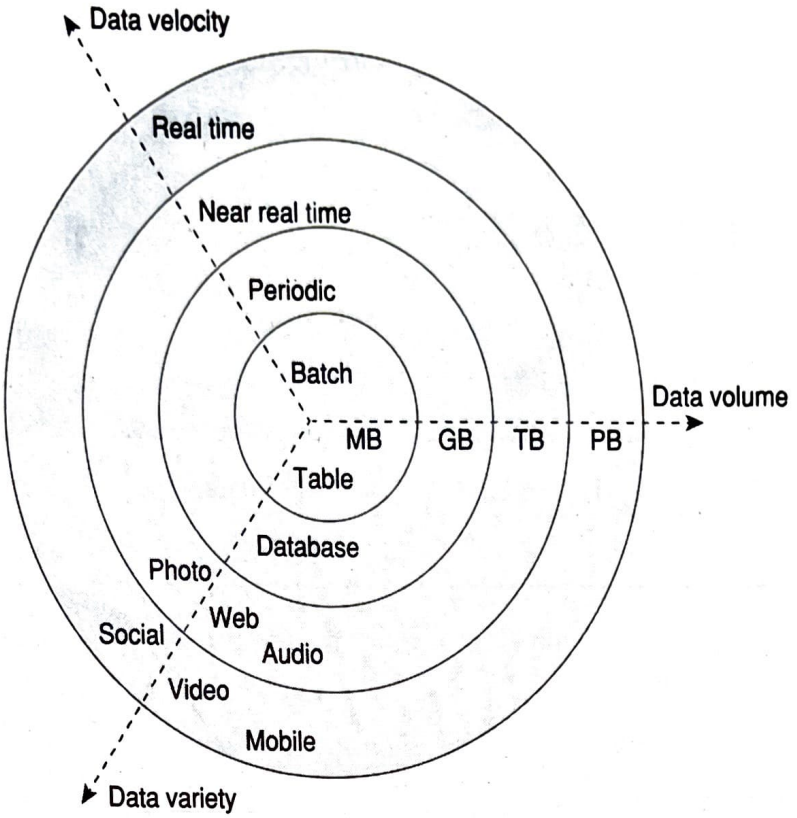


Figure 2.5 Data: Big in volume, variety, and velocity.

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

is unstructured data; a CCTV coverage, a weather forecast report is unstructured data too. Refer Figure 2.7 for the sources of big data.

1. **Typical internal data sources:** Data present within an organization's firewall. It is as follows:
   - ***Data storage:*** File systems, SQL (RDBMSs – Oracle, MS SQL Server, DB2, MySQL, PostgreSQL, etc.), NoSQL (MongoDB, Cassandra, etc.), and so on.
   - ***Archives:*** Archives of scanned documents, paper archives, customer correspondence records, patients' health records, students' admission records, students' assessment records, and so on.

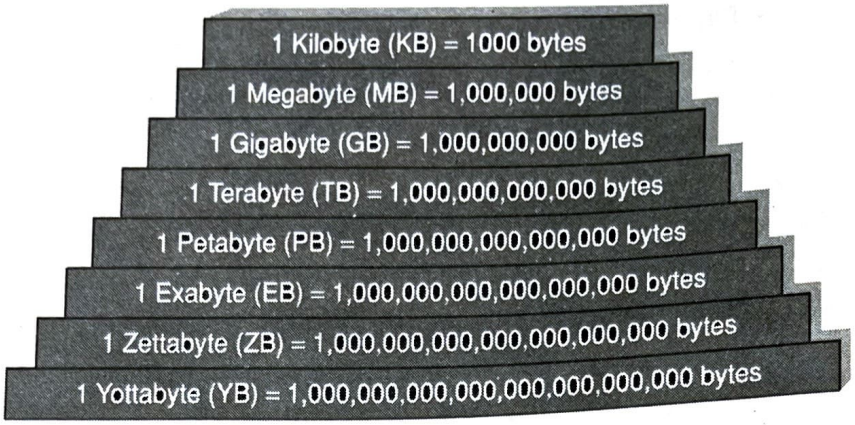


Figure 2.6 A mountain of data.

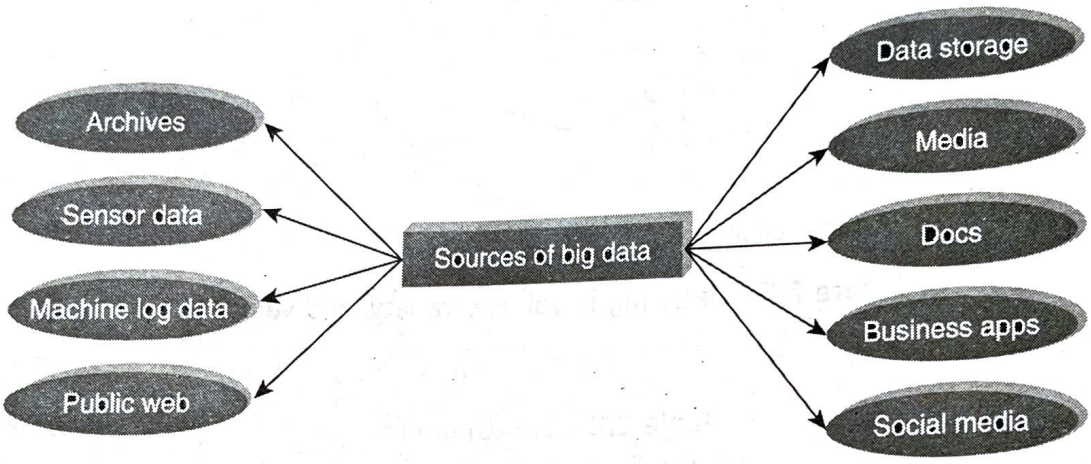


Figure 2.7 Sources of big data.

2. **External data sources:** Data residing outside an organization's firewall. It is as follows:
   - ***Public Web:*** Wikipedia, weather, regulatory, compliance, census, etc.
3. **Both (internal + 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.

2.5.2 Velocity

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

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

2.5.3 Variety

Variety deals with a wide range of data types and sources of data. We will study this under 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.

2.6 OTHER CHARACTERISTICS OF DATA WHICH ARE NOT DEFINITIONAL TRAITS OF BIG DATA

There are yet other characteristics of data which are not necessarily the definitional traits of big data. Few of these are listed as follows:

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. Any data that is picked up for analysis needs to be accurate. It is not just true about big data alone.
2. **Volatility:** Volatility of data deals with, how long is the data valid? And how long should it be stored? There is some data that is required for long-term decisions and remains valid for longer periods of time. However, there are also pieces of data that quickly become obsolete minutes after their generation.
3. **Variability:** Data flows can be highly inconsistent with periodic peaks.

PICTURE THIS...

An online retailer announces the "big sale day" for a particular week. The retailer is likely to experience an upsurge in customer traffic to the website during this week. In the same way, he/she might experience a slump in his/her business immediately after the festival season. This reemphasizes the point that one might witness spikes in data at some point in time and at other times, the data flow can go flat.

2.7 WHY BIG DATA?

The more data we have for analysis, the greater will be the analytical accuracy and also the greater would be the confidence in our decisions based on these analytical findings. This will entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services, and optimizing existing services. Refer Figure 2.8.

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

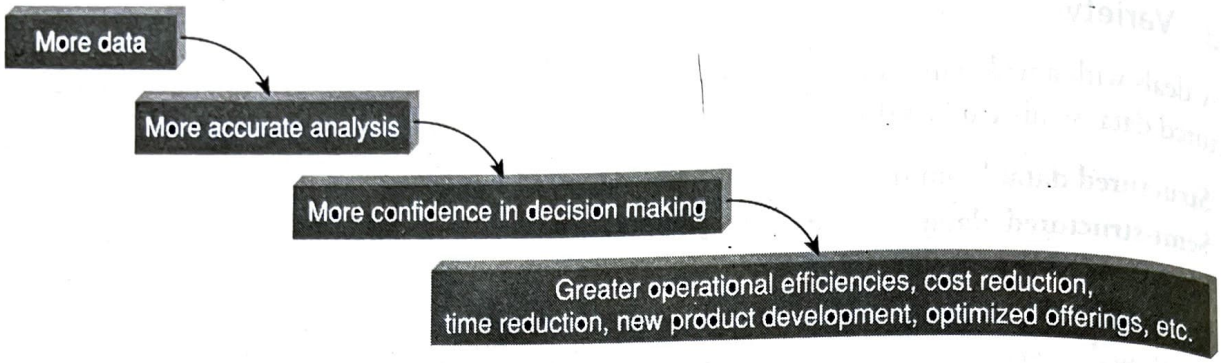


Figure 2.8 Why big data?

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

PICTURE THIS...

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.

Mention the 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.

Likewise, there are several instances everyday where you generate data. Think about cases where you are a consumer of information.

2.9 TRADITIONAL BUSINESS INTELLIGENCE (BI) VERSUS BIG DATA

Let us take a sneak peek into some of the differences that one encounters dealing with traditional BI and 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).

2.10 A TYPICAL DATA WAREHOUSE ENVIRONMENT

Let us look at a typical Data Warehouse (DW) environment. Operational or transactional or day-to-day business data is gathered from Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM), legacy systems, and several third party applications. The data from these sources may differ in format [data could have been housed in any RDBMS such as Oracle, MS SQL Server, DB2, MySQL, and Teradata, and so on or in spreadsheet (.xls, .xlsx, etc.) or .csv or txt]. Data may come from data sources located in the same geography or different geographies. This data is then integrated, cleaned up, transformed, and standardized through the process of Extraction, Transformation, and Loading (ETL). The transformed data is then loaded into the enterprise data warehouse (available at the enterprise level) or data marts (available at the business unit/ functional unit or business process level). A host of market leading business intelligence and analytics tools are then used to enable decision making from the use of ad-hoc queries, SQL, enterprise dashboards, data mining, etc. Refer Figure 2.9.

2.11 A TYPICAL HADOOP ENVIRONMENT

Let us now study the Hadoop environment. Is it very different from the data warehouse environment and where exactly is this difference?

As is fairly obvious from Figure 2.10, the data sources are quite disparate from web logs to images, audios, and videos to social media data to the various docs, pdfs, etc. Here the data in focus is not just the data within the company's firewall but also data residing outside the company's firewall. This data is placed in Hadoop Distributed File System (HDFS). If need be, this can be repopulated back to operational systems or fed to the enterprise data warehouse or data marts or Operational Data Store (ODS) to be picked for further processing and analysis.

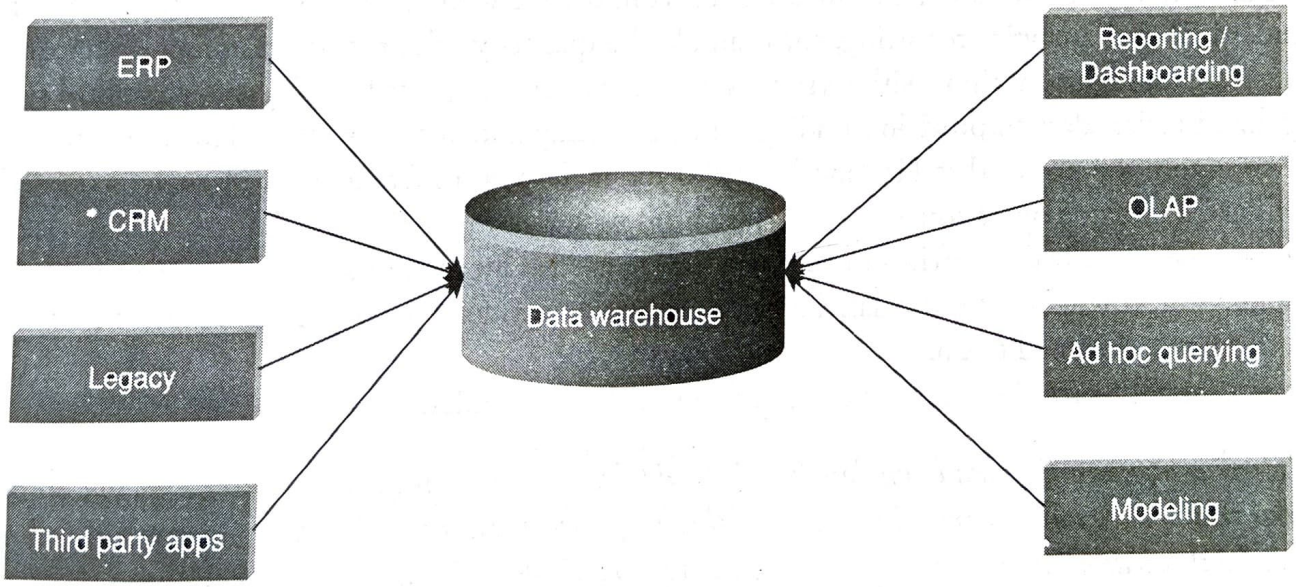


Figure 2.9 A typical data warehouse environment.

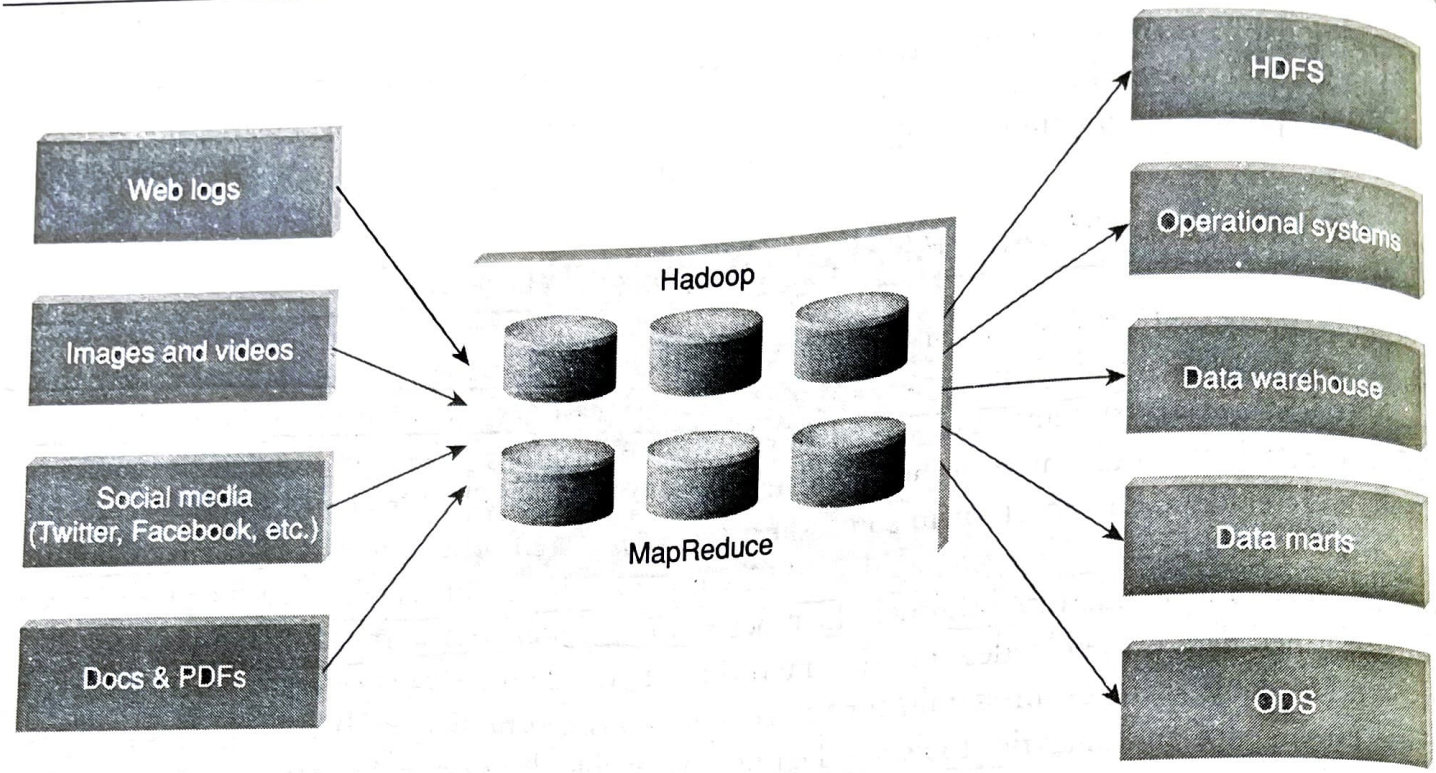


Figure 2.10 A typical Hadoop environment.

2.12 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. Refer Figure 2.11.

2.12.1 Coexistence of Big Data and Data Warehouse

It is NOT about rip and replace. It will not be possible to get rid of RDBMS or massively parallel processing (MPP), but instead use the right tool for the right job.

As we are aware that few companies are a wee bit comfortable working with incumbent data warehouse for standard BI and analytics reporting, for example the quarterly sales report, customer dashboard, etc. The data warehouse can continue with its standard workload drawing data from legacy operational systems, storing the historical data to provision traditional BI reporting and analytics needs. However, one will not be able to ignore the power that Hadoop brings to the table with different types of analysis on different types of data. The same operational systems, which till now was engaged in powering the data warehouse, can also populate the big data environment when they're needed for computation-rich processing or for raw data exploration. It will be a tight balancing act to steer the workload to the right platform based on what that platform was designed to do.

Here is a thought-provoking piece from Ralph Kimball at a cloudera webinar:

"Here's a question that made me laugh a little bit, but it's a serious question: 'Well does this mean that relational databases are going to die?'. I think that there was a sense, three or four years ago, that maybe this was all a giant zero sum game between Hadoop and relational databases, and that has

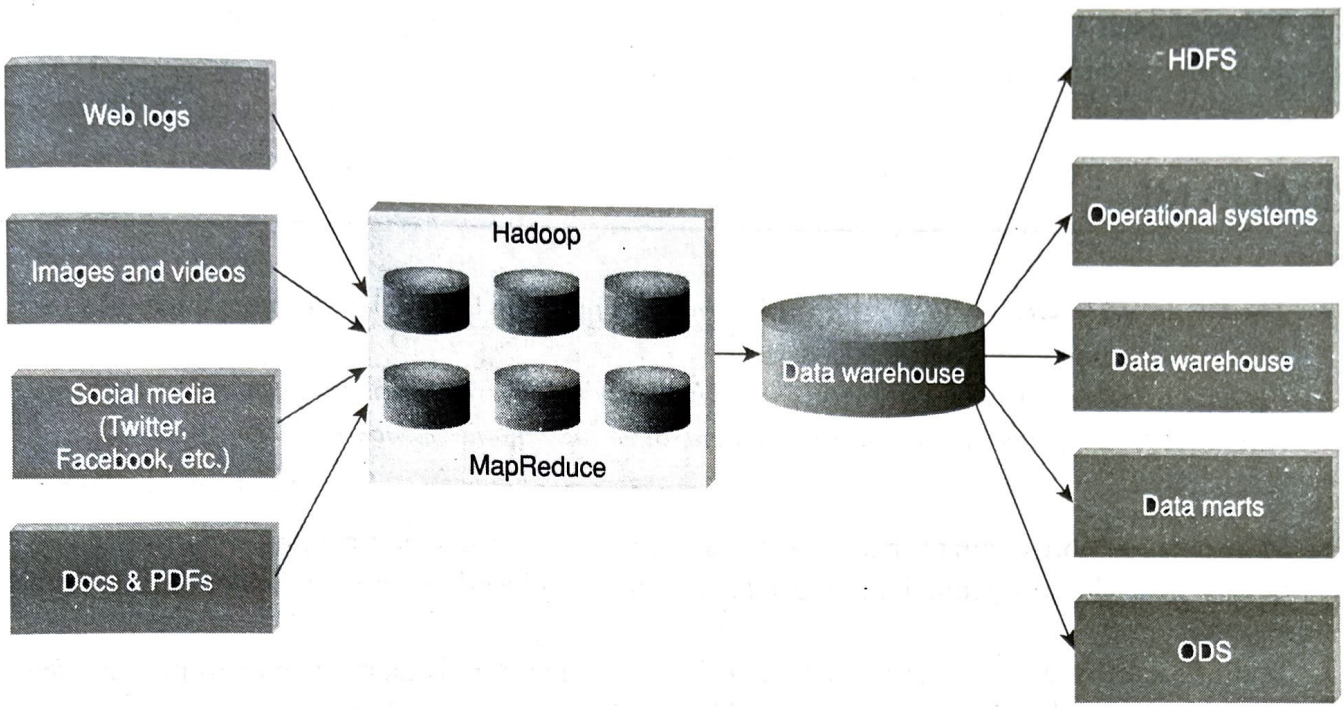


Figure 2.11 Big data and data warehouse coexistence.

simply gone away. Everyone has now realized that there's a huge legacy value in relational databases for the purposes they are used for. Not only transaction processing, but for all the much focused, index-oriented queries on that kind of data, and that will continue in a very robust way forever. Hadoop, therefore, will present this alternative kind of environment for different types of analysis for different kinds of data, and the two of them will coexist. And they will call each other. There may be points at which the business user isn't actually quite sure which one of them they are touching at any point of time."

Just as one cannot ignore the powerful analytics capability of Hadoop, one will not be able to ignore the revolutionary developments in RDBMS such as in-memory processing, etc. The need of the hour is to have both data warehouse and Hadoop co-exist in today's environment.

2.13 WHAT IS CHANGING IN THE REALMS OF BIG DATA?

Gone are the days when IT and business could work in silos and still see the business through. Today, it is an era of a tight handshake between business, IT, and yet another class called *Data Scientists* (more on it in Chapter 3 on "*Big Data Analytics*"). We are citing 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:** The value of data continues to see a steep rise. As the all-important resource, it is tim to look at newer architecture, tools, and practices to leverage this.

REMIND ME

- The World Wide Web (WWW) and the Internet of Things (IoT) have led to an onslaught of structured, unstructured, and multimedia 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

- More data → More accurate analysis → Greater confidence in decision making → Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offerings, etc.
- Traditional BI is about structured data and it is here that data is taken to processing functions (move data to code). On the other hand, 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).

POINT ME (BOOK)

- Big Data for Dummies – Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman, Wiley India Pvt. Ltd.

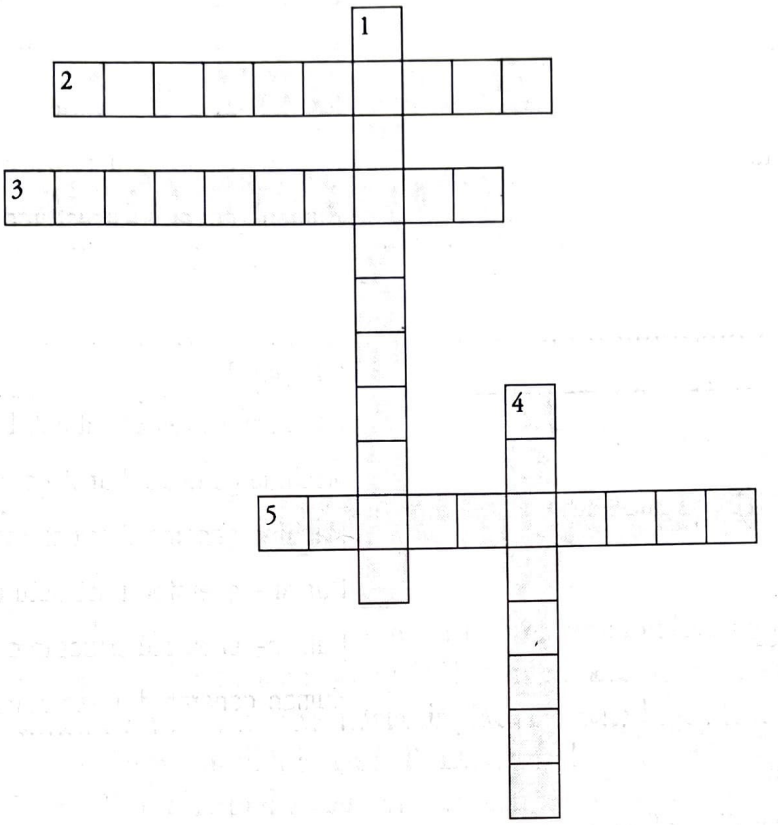
CONNECT ME (INTERNET RESOURCES)

- 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/
- http://www.sap.com/solution/big-data/software/overview.html
- http://www.ibm.com/software/data/bigdata/
- http://www.ibm.com/big-data/us/en/
- http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
- http://timoelliott.com/blog/2014/04/no-hadoop-isnt-going-to-replace-your-data-warehouse.html

TEST ME

A. Crossword

Puzzle on Big Data



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

B. Fill Me

1. Big data is high-volume, high-velocity, and high-variety information assets that demand _____, _____ forms of information processing for enhanced _____ and _____.

Answer: Cost-effective, Innovative, Insight, Decision making

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

D. Unsolved Exercises

1. Share your understanding of big data.
2. How is traditional BI environment different from the big data environment?
3. *Big data (Hadoop) will replace the traditional RDBMS and data warehouse.* Comment.
4. Share your experience as a customer on an e-commerce site. Comment on the big data that gets create on a typical e-commerce site.
5. What is your understanding of "Big Data Analytics"?

CHALLENGE ME

1. What is Internet of Things and why does it matter?

   Answer: See http://www.sas.com/en_us/insights/big-data/internet-of-things.html

2. Can the same visualization tool that we run over conventional data warehouse, be used in big dat environment?

   Answer: Let us look at Figure 2.12 to understand the solution:

   As per Figure 2.12, structured data is stored in Relational Database Management System (RDBMS whereas big data (largely unstructured data) is stored in NoSQL databases. Structured data after cleansing transforming, and converting to a uniform standard format are placed in the enterprise data warehous

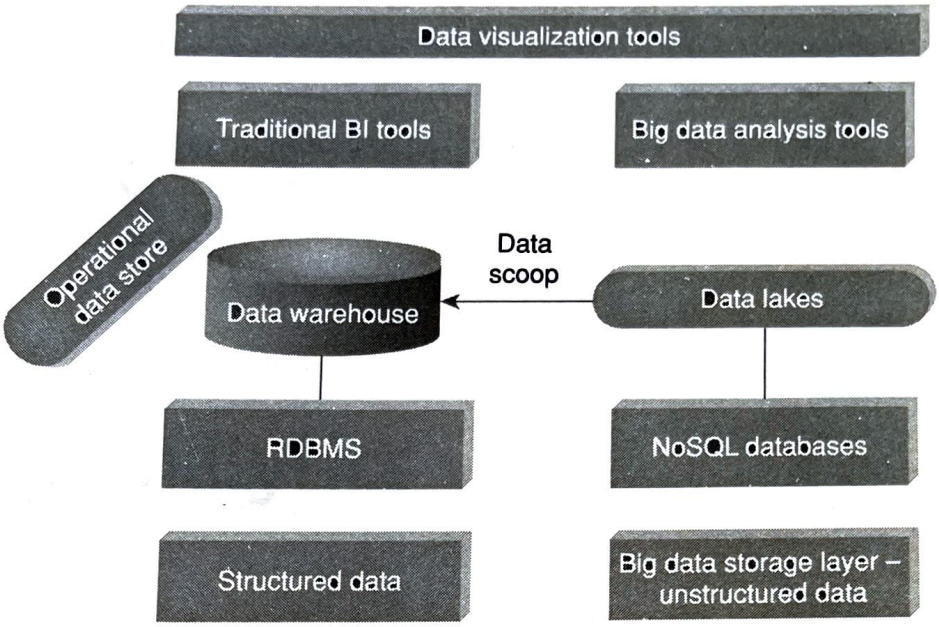


Figure 2.12 Visualization tools for traditional BI and big data.

(at the enterprise level) or the data marts (at the business unit or function level) or operational data stores (almost the complete operational data of an enterprise is housed here) whereas the good variety of data (structured, semi-structured, and unstructured data) is placed in data lakes (a large data repository that stores raw data in its native format until it is needed). Data can then be scooped from data lakes to data warehouses and traditional BI tools can then be run over them. A common set of data visualization tools can then be used to present results after analysis. This goes to emphasize the point, that it makes sense to use the tool that is a specialist for a particular function for example RDBMS for structured data and NoSQL for voluminous data that may be schema less.