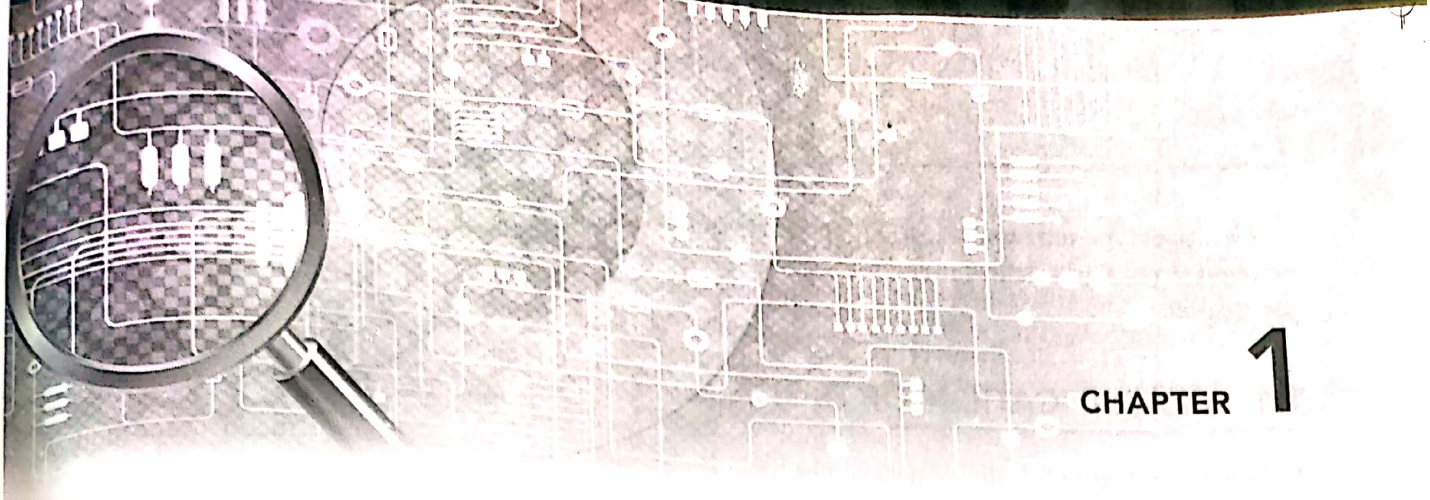


CHAPTER 1

# Types of Digital Data

## BRIEF CONTENTS


- What's in Store?
- Classification of Digital Data
- Structured Data
  - Sources of Structured Data
  - Ease of Working with Structured Data
- Semi-Structured Data
  - Sources of Semi-Structured Data
- Unstructured Data
  - Issues with "Unstructured" Data
  - How to Deal with Unstructured Data


*"In God we trust, all others must bring data."*

– W. Edwards Deming

## WHAT'S IN STORE?

Irrespective of the size of the enterprise (big or small), data continues to be a precious and irreplaceable asset. Data is present internal to the enterprise and also exists outside the four walls and firewalls of the enterprise. Data is present in homogeneous sources as well as in heterogeneous sources. The need of the hour is to understand, manage, process, and take the data for analysis to draw valuable insights.

Data → Information

Information → Insights

This chapter is a "must read" for first-time learners interested in understanding the role of data in business intelligence and business analysis and businesses at large. This chapter will introduce you to the various formats of digital data (structured, semi-structured, and unstructured data), the sources of each format of data, the issues with the terminology of unstructured data, etc.

Scanned by CamScanner

We suggest you refer to the learning resources suggested at the end of this chapter and also attempt all the exercises to get a grip on this topic. We suggest you make your own notes/bookmarks while reading through the chapter.

## 1.1 CLASSIFICATION OF DIGITAL DATA

As depicted in Figure 1.1, digital data can be broadly classified into structured, semi-structured, and unstructured data.

1. **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, researches, white papers, body of an email, etc.
2. **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. Metadata for this data is available but is not sufficient.
3. **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. Relationships exist between entities of data, such as classes and their objects. Data stored in databases is an example of structured data.

Ever since the 1980s most of the enterprise data has been stored in relational databases complete with rows/records/tuples, columns/attributes/fields, primary keys, foreign keys, etc. Over a period of time Relational Database Management System (RDBMS) matured and the RDBMS, as they are available today, have become more robust, cost-effective, and efficient. We have grown comfortable working with RDBMS – the storage, retrieval, and management of data has been immensely simplified. The data held in RDBMS is typically structured data. However, with the Internet connecting the world, data that existed beyond one's enterprise started to become an integral part of daily transactions. This data grew by leaps and bounds so much so that it became difficult for the enterprises to ignore it. All of this data was not structured. A lot of it was unstructured. In fact, Gartner estimates that almost 80% of data generated in any enterprise today is unstructured data. Roughly around 10% of data is in the structured and semi-structured category. Refer Figure 1.2.

### 1.1.1 Structured Data

Let us begin with a very basic question – When do we say that the data is structured? The simple answer is when data conforms to a pre-defined schema/structure we say it is structured data.

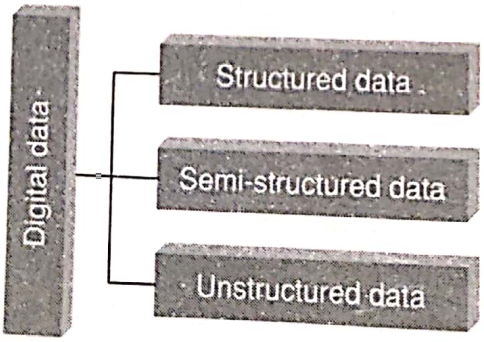



**Figure 1.1** Classification of digital data.

Scanned by CamScanner

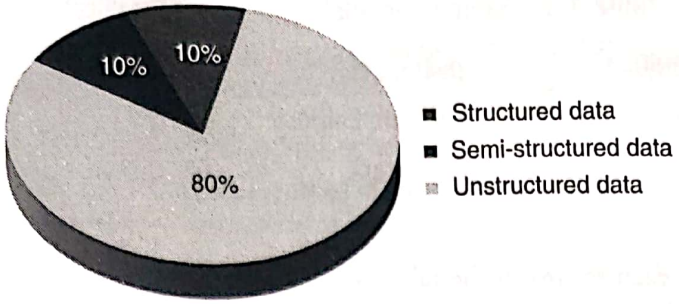



**Figure 1.2** Approximate percentage distribution of digital data.

Think structured data, and think data model – a model of the types of business data that we intend to store, process, and access. Let us discuss this in the context of an RDBMS. Most of the structured data is held in RDBMS. An RDBMS conforms to the relational data model wherein the data is stored in rows/columns. Refer Table 1.1.

The number of rows/records/tuples in a relation is called the *cardinality of a relation* and the number of columns is referred to as the *degree of a relation.*

The first step is the design of a relation/table, the fields/columns to store the data, the type of data that will be stored [number (integer or real), alphabets, date, Boolean, etc.]. Next we think of the constraints that we would like our data to conform to (constraints such as UNIQUE values in the column, NOT NULL values in the column, a business constraint such as the value held in the column should not drop below 50, the set of permissible values in the column such as the column should accept only "CS", "IS", "MS", etc., as input).

To explain further, let us design a table/relation structure to store the details of the employees of an enterprise. Table 1.2 shows the structure/schema of an "Employee" table in a RDBMS such as Oracle.

Table 1.2 is an example of a good structured table (complete with table name, meaningful column names with data types, data length, and the relevant constraints) with absolute adherence to relational data model.

**Table 1.1** A relation/table with rows and columns

| | Column 1 | Column 2 | Column 3 | Column 4 |
|---|---|---|---|---|
| Row 1 | | | | |

**Table 1.2** 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 |

Scanned by CamScanner

**Table 1.3** Sample records in the "Employee" table

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

It goes without saying that each record in the table will have exactly the same structure. Let us take a look at a few records in Table 1.3.

The tables in an RDBMS can also be related. For example, the above "Employee" table is related to the "Department" table on the basis of the common column, "DeptNo". It is not mandatory for the two tables that are related to have exactly the same name for the common column. On the contrary, the two tables are related on the basis of values held within the column, "DeptNo". Given in Figure 1.3 is a depiction of referential integrity constraint (primary – foreign key) with the "Department" table being the referenced table and "Employee" table being the referencing table.

#### 1.1.1.1 Sources of Structured Data

If your data is highly structured, one can look at leveraging any of the available RDBMS [Oracle Corp. – Oracle, IBM – DB2, Microsoft – Microsoft SQL Server, EMC – Greenplum, Teradata – Teradata, MySQL (open source), PostgreSQL (advanced open source), etc.] to house it. Refer Figure 1.4. These databases are typically used to hold transaction/operational data generated and collected by day-to-day business activities. In other words, the data of the On-Line Transaction Processing (OLTP) systems are generally quite structured.

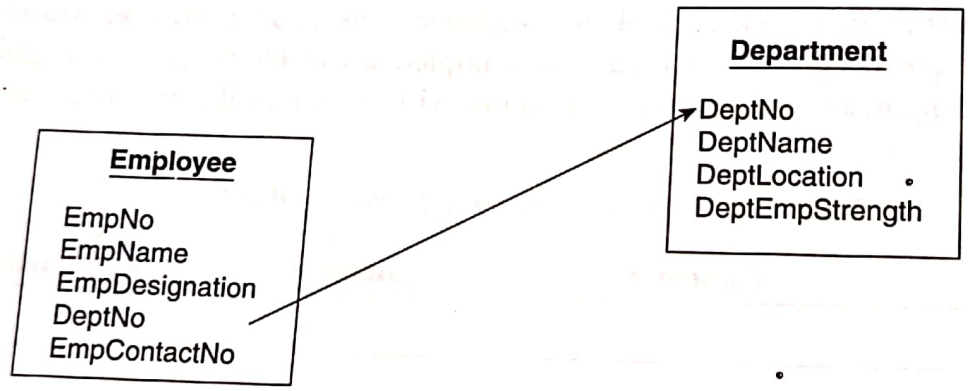



**Figure 1.3** Relationship between "Employee" and "Department" tables.

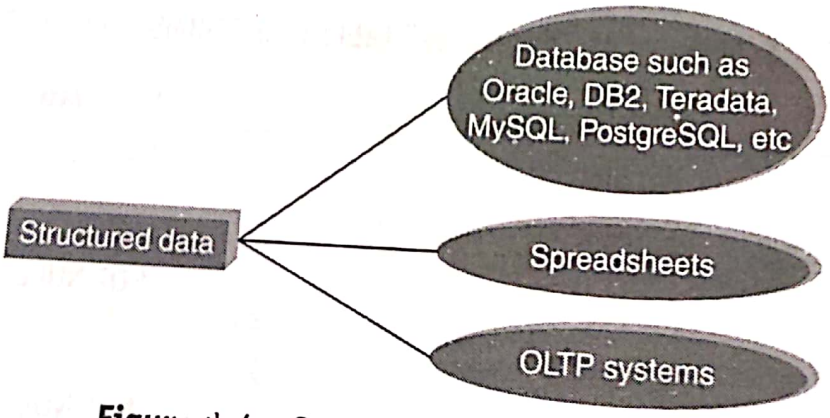



**Figure 1.4** Sources of structured data.

Scanned by CamScanner

#### 1.1.1.2 Ease of Working with Structured Data

Structured data provides the ease of working with it. Refer Figure 1.5. The ease is with respect to the following:

1. **Insert/update/delete:** The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis, etc.
2. **Security:** How does one ensure the security of information? There are available staunch encryption and tokenization solutions to warrant the security of information throughout its lifecycle. Organizations are able to retain control and maintain compliance adherence by ensuring that only authorized individuals are able to decrypt and view sensitive information.
3. **Indexing:** An index is a data structure that speeds up the data retrieval operations (primarily the SELECT DML statement) at the cost of additional writes and storage space, but the benefits that ensue in search operation are worth the additional writes and storage space.
4. **Scalability:** The storage and processing capabilities of the traditional RDBMS can be easily scaled up by increasing the horsepower of the database server (increasing the primary and secondary or peripheral storage capacity, processing capacity of the processor, etc.).
5. **Transaction processing:** RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction. Given next is a quick explanation of the ACID properties:
   - ***Atomicity:*** A transaction is atomic, means that either it happens in its entirety or none of it at all.
   - ***Consistency:*** The database moves from one consistent state to another consistent state. In other words, if the same piece of information is stored at two or more places, they are in complete agreement.
   - ***Isolation:*** The resource allocation to the transaction happens such that the transaction gets the impression that it is the only transaction happening in isolation.
   - ***Durability:*** All changes made to the database during a transaction are permanent and that accounts for the durability of the transaction.

### 1.1.2 Semi-Structured Data

Semi-structured data is also referred to as self-describing structure. Refer Figure 1.6. It has the following features:

1. It does not conform to the data models that one typically associates with relational databases or any other form of data tables.
2. It uses tags to segregate semantic elements.

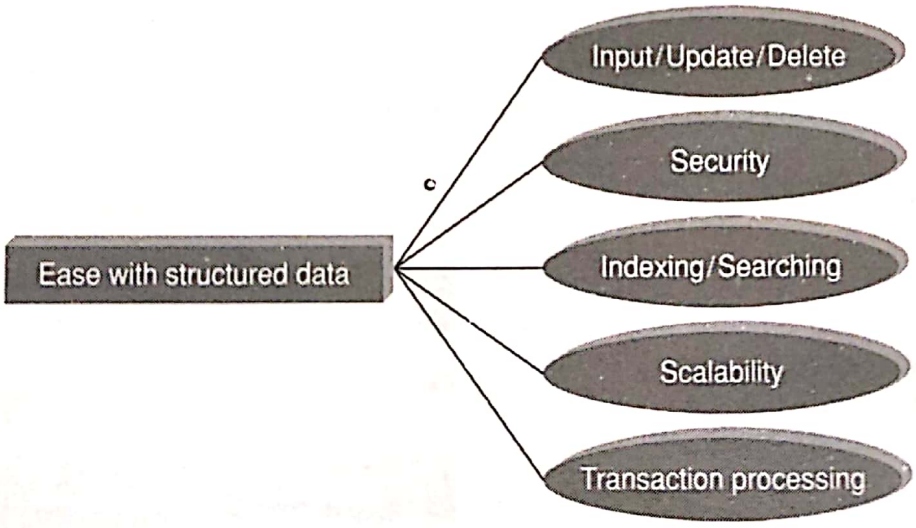



**Figure 1.5** Ease of working with structured data.

Scanned by CamScanner

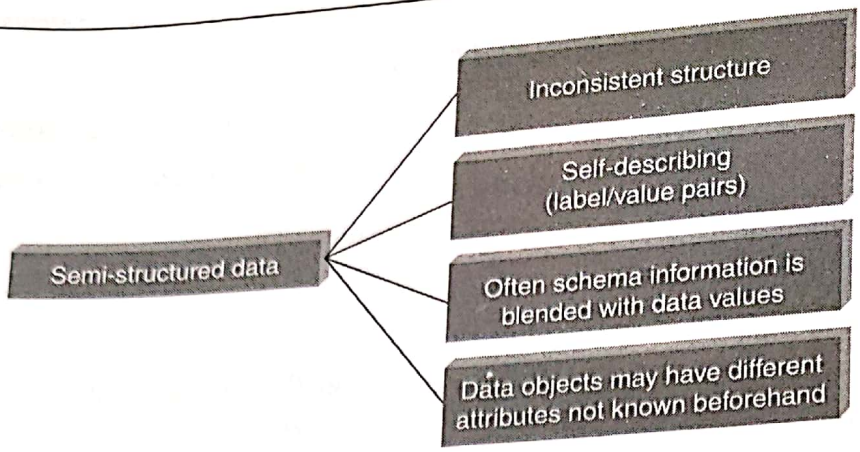



**Figure 1.6** Characteristics of semi-structured data.

3. Tags are also used to enforce hierarchies of records and fields within data.
4. There is no separation between the data and the schema. The amount of structure used is dictated by the purpose at hand.
5. In semi-structured data, entities belonging to the same class and also grouped together need not necessarily have the same set of attributes. And if at all, they have the same set of attributes, the order of attributes may not be similar and for all practical purposes it is not important as well.

#### *1.1.2.1 Sources of Semi-Structured Data*

Amongst the sources for semi-structured data, the front runners are "XML" and "JSON" as depicted in Figure 1.7.

1. **XML:** eXtensible Markup Language (XML) is hugely popularized by web services developed utilizing the Simple Object Access Protocol (SOAP) principles.
2. **JSON:** Java Script Object Notation (JSON) is used to transmit data between a server and a web application. JSON is popularized by web services developed utilizing the Representational State Transfer (REST) – an architecture style for creating scalable web services. MongoDB (open-source, distributed, NoSQL, documented-oriented database) and Couchbase (originally known as Membase, open-source, distributed, NoSQL, document-oriented database) store data natively in JSON format.

An example of HTML is as follows:

```
<HTML>
<HEAD>
<TITLE>Place your title here</TITLE>
</HEAD>
<BODY BGCOLOR="FFFFFF">
```

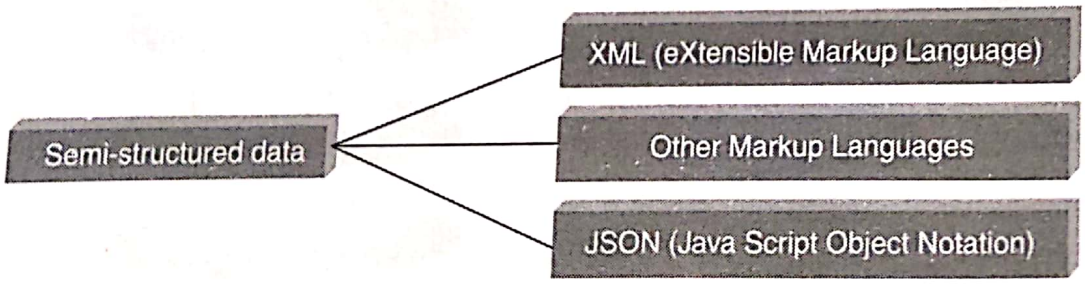



**Figure 1.7** Sources of semi-structured data.

Scanned by CamScanner

```
<CENTER><IMG SRC="clouds.jpg" ALIGN="BOTTOM"></CENTER>
<HR>
<a href="http://bigdatauniversity.com">Link Name</a>
<H1>this is a Header</H1>
<H2>this is a sub Header</H2>
Send me mail at <a href="mailto:support@yourcompany.com">
support@yourcompany.com</a>.
<P>a new paragraph!
<P><B>a new paragraph!</B>
<BR><B><I>this is a new sentence without a paragraph break, in bold italics.</I></B>
<HR>
</BODY>
</HTML>
```

***Sample JSON document***

```
{
_id:9,
BookTitle: "Fundamentals of Business Analytics",
AuthorName: "Seema Acharya",
Publisher: "Wiley India",
YearofPublication: "2011"
}
```

### 1.1.3 Unstructured Data

Unstructured data does not conform to any pre-defined data model. In fact, to explain things a little more, let us take a closer look at the various kinds of text available and the possible structure associated with it. As can be seen from the examples quoted in Table 1.4, the structure is quite unpredictable. In Figure 1.8 we look at the other sources of unstructured data.

#### *1.1.3.1 Issues with "Unstructured" Data*

Although unstructured data is known NOT to conform to a pre-defined data model or be organized in a pre-defined manner, there are incidents wherein the structure of the data (placed in the unstructured category) can still be implied. As mentioned in Figure 1.9, there could be few other reasons behind placing data in the unstructured category despite it having some structure or being highly structured.

There are situations where people argue that a text file should be in the category of semi-structured data and not unstructured data. Let us look at where they are coming from. Well, the text file does have a name,

**Table 1.4** Few examples of disparate unstructured data

| | |
|---|---|
| Twitter message | Feeling miffed ☹. Victim of twishing. |
| Facebook post | LOL. C ya. BFN |
| Log files | 127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I; Nav)" |
| Email | Hey Joan, possible to send across the first cut on the Hadoop chapter by Friday EOD or maybe we can meet up over a cup of coffee. Best regards, Tom |

Scanned by CamScanner

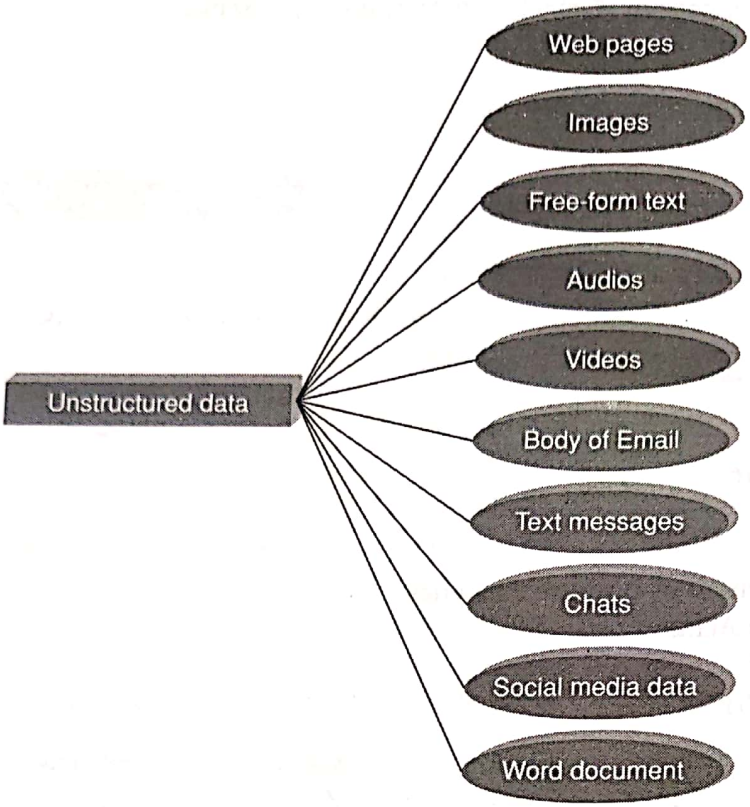



**Figure 1.8** Sources of unstructured data.

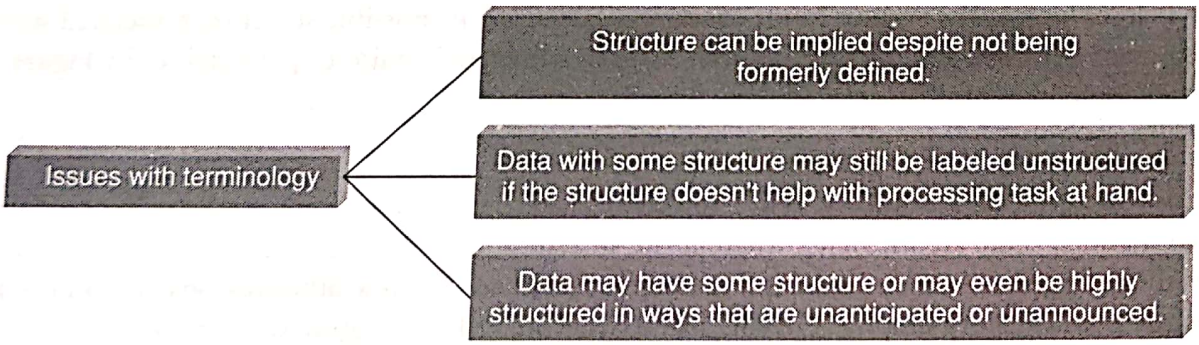



**Figure 1.9** Issues with terminology of unstructured data.

one can easily look at the properties to get information such as the owner of the file, the date on which the file was created, the size of the file, etc. Okay, we do have little metadata. But when it comes to analysis, we are more concerned with the content of the text file rather than the name or any of the other properties. In fact, the other properties may not in any way contribute to the processing/analysis task at hand. Therefore, it is fair to place it in the unstructured data category.

### *1.1.3.2 How to Deal with Unstructured Data?*

Today, unstructured data constitutes approximately 80% of the data that is being generated in any enterprise. The balance is clearly shifting in favor of unstructured data as shown in Figure 1.10. It is such a big percentage that it cannot be ignored. Figure 1.11 states a few ways of dealing with unstructured data.

Scanned by CamScanner

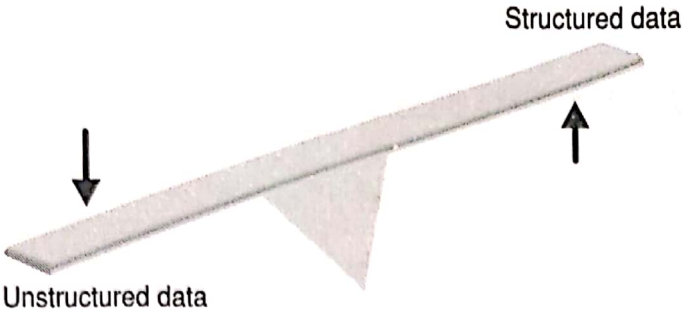



**Figure 1.10** Unstructured data clearly constitutes a major percentage of enterprise data.

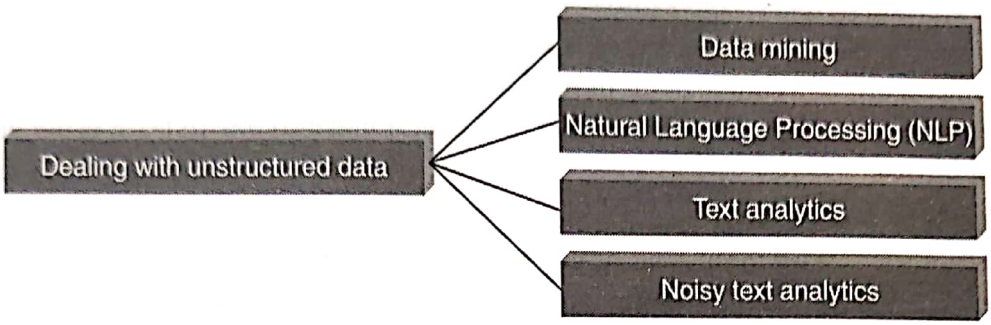



**Figure 1.11** Dealing with unstructured data.

The following techniques are used to find patterns in or interpret unstructured data:

1. **Data mining:** First, we deal with large data sets. Second, we use methods at the intersection of artificial intelligence, machine learning, statistics, and database systems to unearth consistent patterns in large data sets and/or systematic relationships between variables. It is the analysis step of the "knowledge discovery in databases" process.

   Few popular data mining algorithms are as follows:
   - ***Association rule mining:*** It is also called "market basket analysis" or "affinity analysis". It is used to determine "What goes with what?" It is about when you buy a product, what is the other product that you are likely to purchase with it. For example, if you pick up bread from the grocery, are you likely to pick eggs or cheese to go with it.
   - ***Regression analysis:*** It helps to predict the relationship between two variables. The variable whose value needs to be predicted is called the dependent variable and the variables which are used to predict the value are referred to as the independent variables.

**PICTURE THIS...**

You are interested in purchasing real estate. You have been looking at a few good sites. You have come to the conclusion that cost of the real estate depends on the location (outskirts or prime locale), the amenities provided by the builder (joggers track, senior citizen zone, gymnasium, swimming pools, etc.), the built up area, etc. The cost of the real estate is the dependent variable and the location, amenities, built-up area are called the independent variables.

Scanned by CamScanner

**Table 1.5** Sample records depicting learners' preferences for modes of learning

| | Learning using Audios | Learning using Videos | Textual Learners |
|---|---|---|---|
| User 1 | Yes | Yes | No |
| User 2 | Yes | Yes | Yes |
| User 3 | Yes | Yes | No |
| User 4 | Yes | ? | ? |

- ***Collaborative filtering:*** It is about predicting a user's preference or preferences based on the preferences of a group of users. For example, take a look at Table 1.5.
  We are looking at predicting whether User 4 will prefer to learn using videos or is a textual learner depending on one or a couple of his or her known preferences. We analyze the preferences of similar user profiles and on the basis of it, predict that User 4 will also like to learn using videos and is not a textual learner.

2. **Text analytics or text mining:** Compared to the structured data stored in relational databases, text is largely unstructured, amorphous, and difficult to deal with algorithmically. Text mining is the process of gleaning high quality and meaningful information (through devising of patterns and trends by means of statistical pattern learning) from text. It includes tasks such as text categorization, text clustering, sentiment analysis, concept/entity extraction, etc.
3. **Natural language processing (NLP):** It is related to the area of human computer interaction. It is about enabling computers to understand human or natural language input.
4. **Noisy text analytics:** It is the process of extracting structured or semi-structured information from noisy unstructured data such as chats, blogs, wikis, emails, message-boards, text messages, etc. The noisy unstructured data usually comprises one or more of the following: Spelling mistakes, abbreviations, acronyms, non-standard words, missing punctuation, missing letter case, filler words such as "uh", "um", etc.
5. **Manual tagging with metadata:** This is about tagging manually with adequate metadata to provide the requisite semantics to understand unstructured data.
6. **Part-of-speech tagging:** It is also called POS or POST or grammatical tagging. It is the process of reading text and tagging each word in the sentence as belonging to a particular part of speech such as "noun", "verb", "adjective", etc.
7. **Unstructured Information Management Architecture (UIMA):** It is an open source platform from IBM. It is used for real-time content analytics. It is about processing text and other unstructured data to find latent meaning and relevant relationship buried therein. Read up more on UIMA at the link: http://www.ibm.com/developerworks/data/downloads/uima/

## REMIND ME

- ***Structured data:*** It conforms to a data model. For example, RDBMS conforms to relational data model. It has a pre-defined schema.
- ***Semi-structured data:*** For this format of data, little metadata is available, but is insufficient. Semi-structured data have a self-describing structure. There is little or no separation between data and schema.

Scanned by CamScanner

- ***Unstructured data:*** This data is growing by the day and growing by leaps and bounds. It has innumerable sources such as human generated (social media data, emails, word documents, presentations, audio and video files that we create and share every day, etc.) and machine generated data (sensors, web server logs, call data records, etc.).

## POINT ME (BOOK)

- Chapter 2: Types of Digital Data, "Fundamentals of Business Analytics", Wiley India; Authors – RN Prasad and Seema Acharya, 2011.

## CONNECT ME (INTERNET RESOURCES)

- 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
- Matching unstructured data and structured data by Bill Inmon: http://www.tdan.com/view-articles/5009
- Semi-structured data analytics: Relational or Hadoop platform? – IBM: http://www.ibmbigdatahub.com/blog/semi-structured-data-analytics-relational-or-hadoop-platform-part-1

## TEST ME

### A. 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

Scanned by CamScanner

**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 & DBs |
| 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 |

2.

| Column A | Column B |
|---|---|
| JSON | SOAP |
| MongoDB | REST |
| XML | JSON |
| Flexible structure | CouchDB |
| JSON | XML |

Scanned by CamScanner

**Answer:**

| Column A | Column B |
|---|---|
| JSON | REST |
| MongoDB | JSON |
| XML | SOAP |
| Flexible structure | XML |
| JSON | CouchDB |

### C. Solve Me

You are a senior faculty at a premier engineering institute of the city. The Head of the Department has asked you to take a look at the institute's learning website and make a list of the unstructured data that gets generated on the website that can then be stored and analyzed to improve the website to facilitate and enhance the student's learning. You log into the institute's learning website and observe the following features on it:

- Presentation decks (.pdf files)
- Laboratory Manual (.doc files)
- Discussion forum
- Student's blog
- Link to Wikipedia
- A survey questionnaire for the students
- Student's performance sheet downloadable into an .xls sheet
- Student's performance sheet downloadable into a .txt file
- Audio/Video learning files (.wav files)
- .xls sheet having a compiled list of FAQs

From this list, you select the following as sources of unstructured data:

**1.**
**2.**
**3.**
**4.**
**5.**
**6.**

You have just finished making your list when your colleague comes in looking for you. Both of you decide to go away to the cafeteria in the vicinity of the institute's campus. You have forever liked this cafeteria. And you have reasons for the same. There are a couple of machines in the cafeteria's reception area that the customers can use to feed in their orders from a selection of menu items. Once the order is done, you are given a token number. Once your order is ready for serving, the display flashes your token number. It goes without saying

Scanned by CamScanner

that the billing is also automated. You being in the IT department cannot refrain from thinking about the d that gets collected by these automatic applications. Here's your list:

1.
2.
3.
4.
5.
6.

You are thinking of the analysis that you can perform on this data. Here's your list:

1.
2.
3.
4.
5.
6.

## D. Solved Exercises

1. Why is an email placed in the "unstructured category"?

   **Answer:** Let us take a look at what we can place in the body of the email. We can have any or more of the following:

   - Hyperlink
   - PDFs/DOCs/XLS/etc. attachments
   - Emoticons
   - Images
   - Audio/video attachments
   - Free flowing text, etc.

   The above are reasons behind placing the email in the "unstructured category".

2. What category will you place a CCTV footage into?

   **Answer:** Unstructured

3. You have just got a book issued from the library. What are the details about the book that can be place in an RDBMS table?

   **Answer:**

   - Title of the book
   - Author of the book
   - Publisher of the book
   - Year of Publication
   - No. of pages in the book
   - Type of book such as whether hardbound or paperback
   - Price of the book
   - ISBN No. of the book
   - Attachments such as With CD or Without CD, etc.

Scanned by CamScanner

4. Which category would you place the consumer complaints and feedback?
   **Answer:** Unstructured data

### E. Unsolved Exercises

1. Which category (structured, semi-structured, or unstructured) will you place a web page in?
2. What according to you are the challenges with unstructured data?
3. Which category (structured, semi-structured, or unstructured) will you place a PowerPoint presentation in?
4. Which category (structured, semi-structured, or unstructured) will you place a Word Document in?
5. State a few examples of human generated and machine-generated data.

## SCENARIO-BASED QUESTION

You are at the university library. You see a few students browsing through the library catalog on a kiosk. You observe the librarians busy at work issuing and returning books. You see a few students fill up the feedback form on the services offered by the library. Quite a few students are learning using the e-learning content.

Think for a while on the different types of data that are being generated in this scenario. Support your answer with logic.

Scanned by CamScanner

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?

Scanned by CamScanner

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, ronment 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 e of the session that was graded. There was a feedb form that was made available at the end of the se sion to hear back from the participants. They al provided additional reading contents in the fo of references to white papers/research papers. T lecture was recorded and made available for bet learning and comprehension of the participants.

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

There is no dearth of such virtual classroom se sions being conducted today. There is a huge learni community out there eager to learn. Just think on th volume of data that gets generated, and the varie (the list of attendees, their scores and grades, the chat conversations, their assignments, the pollin questions put forth by the instructor to gauge the leve of understanding and participation from the learner etc.) of data that we produce as well consume as w 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 d 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 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 d generated?" "How sensitive is this data?" "What are the events associated with this data?" and so o

Small data (data as it existed prior to the big data revolution) is about certainty. It is about fairly known d 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 genera exactly how we would like to use it. what questions will this data be able to answer, and so on. Big da

Scanned by CamScanner

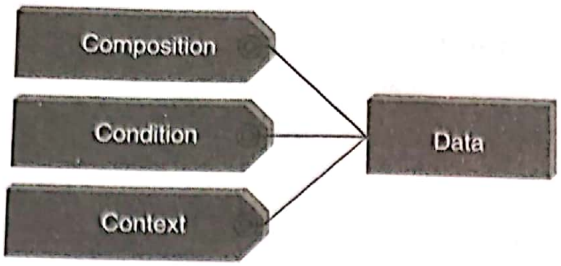



**Figure 2.1** Characteristics of data.

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.

Scanned by CamScanner

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 a 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.

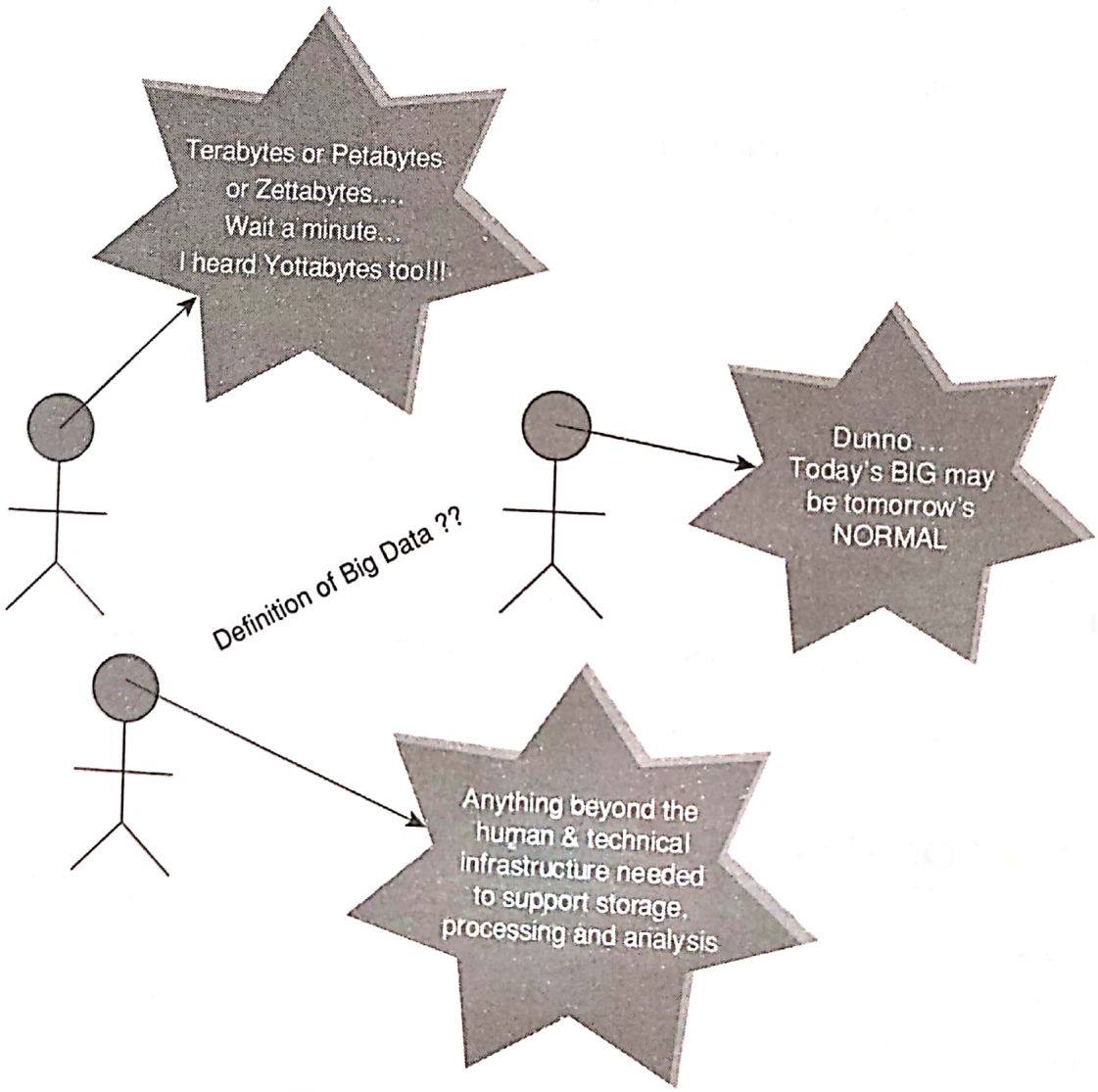



**Figure 2.2** Definition of big data.

Scanned by CamScanner

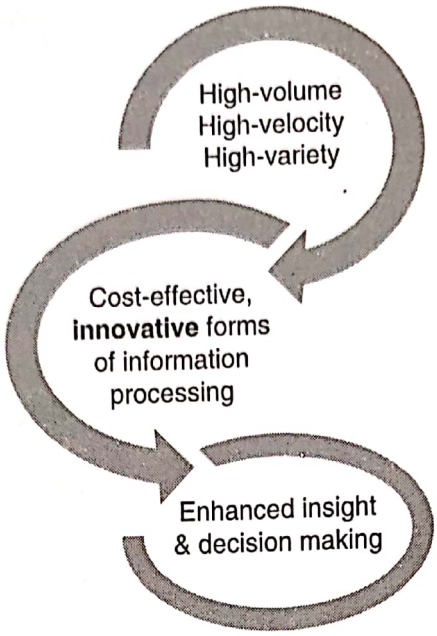



**Figure 2.3** Definition of big data – Gartner.

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.

Scanned by CamScanner

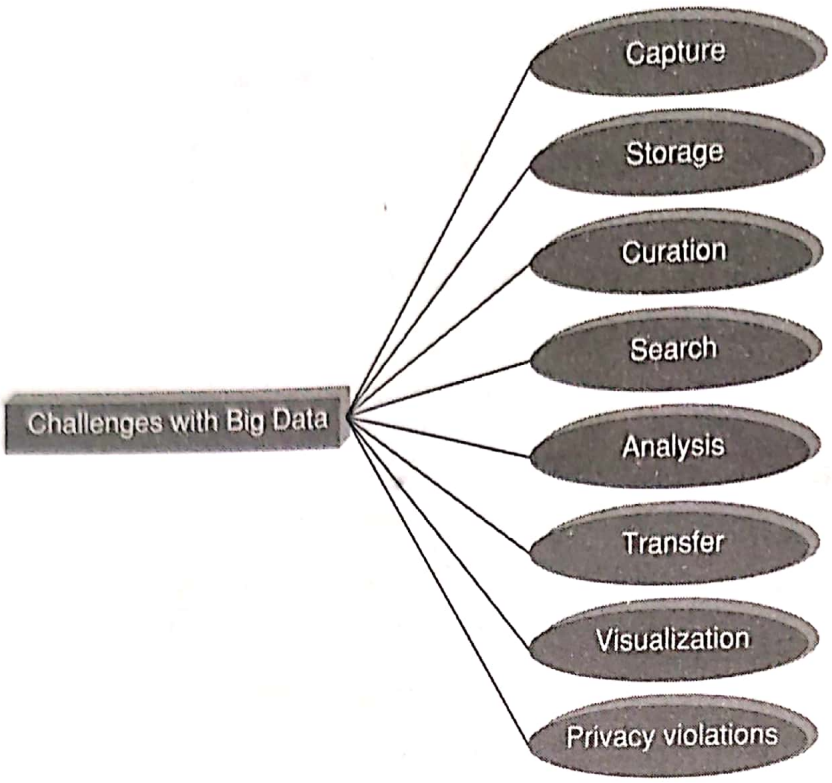



**Figure 2.4** Challenges with big data.

4. There is a dearth of skilled professionals who possess a high level of proficiency in data sciences that i vital in implementing big data solutions.
5. Then, of course, there are other challenges with respect to capture, storage, preparation, search, anal ysis, transfer, security, and visualization of big data. Big data refers to datasets whose size is typicall beyond the storage capacity of traditional database software tools. There is no explicit definition o how big the dataset should be for it to be considered "big data." Here we are to deal with data that i 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

Scanned by CamScanner

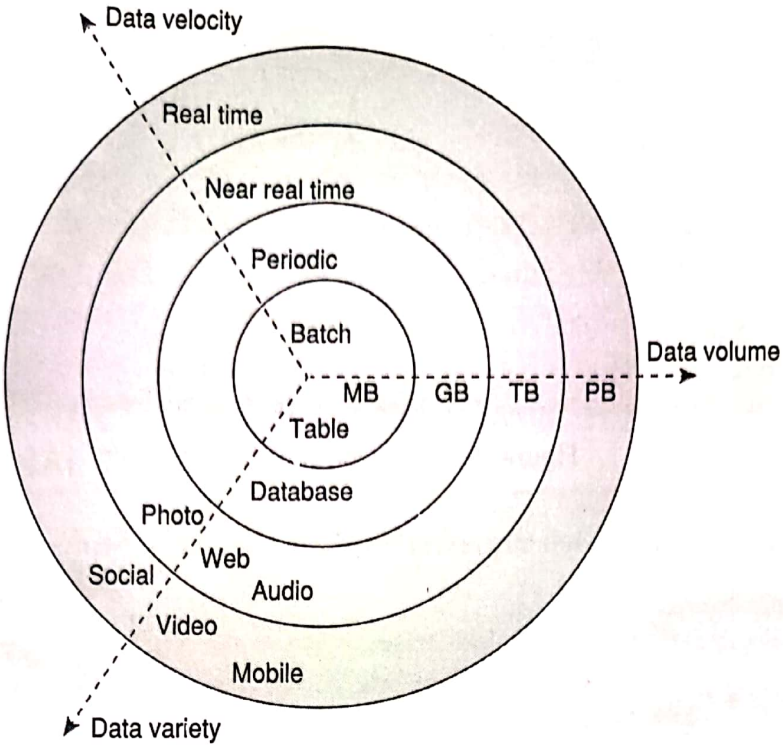



**Figure 2.5** Data: Big in volume, variety, and velocity.

**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.

Scanned by CamScanner

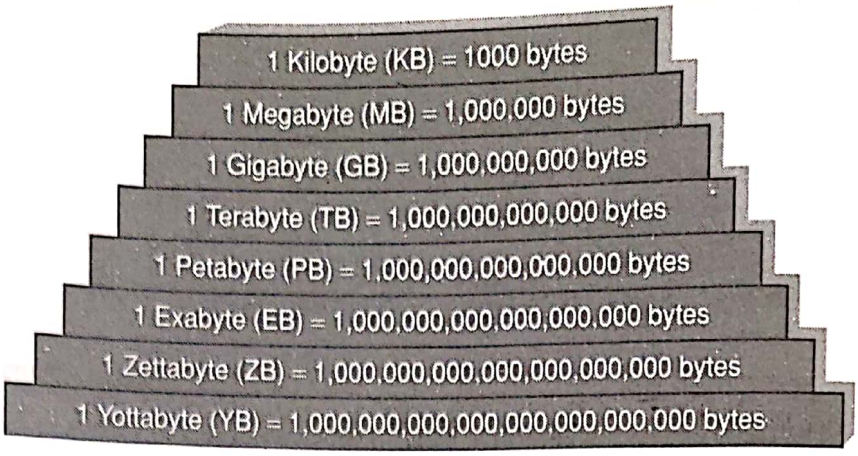



**Figure 2.6** A mountain of data.

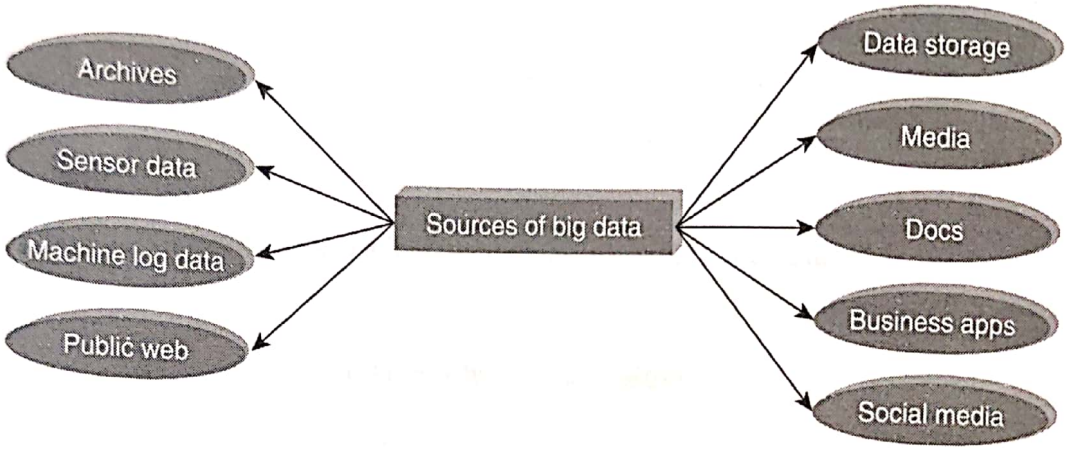



**Figure 2.7** Sources of big data.

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**

Scanned by CamScanner

### 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.**

Scanned by CamScanner

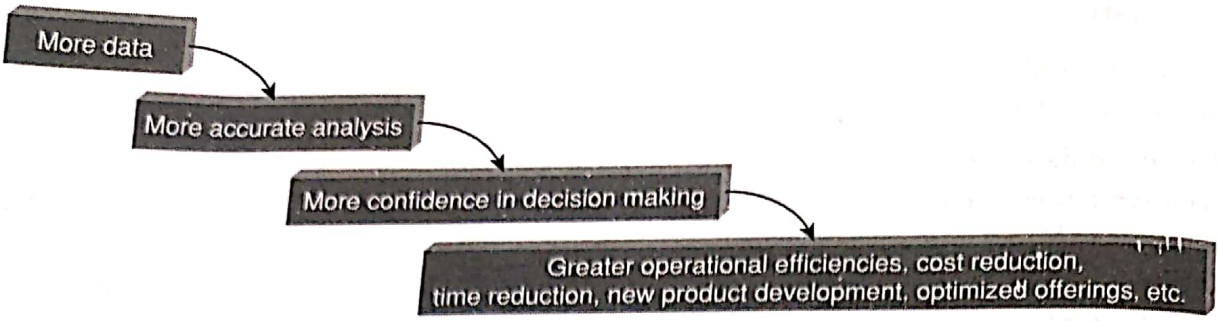



**Figure 2.8** Why big data?

## 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.*

Scanned by CamScanner

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.

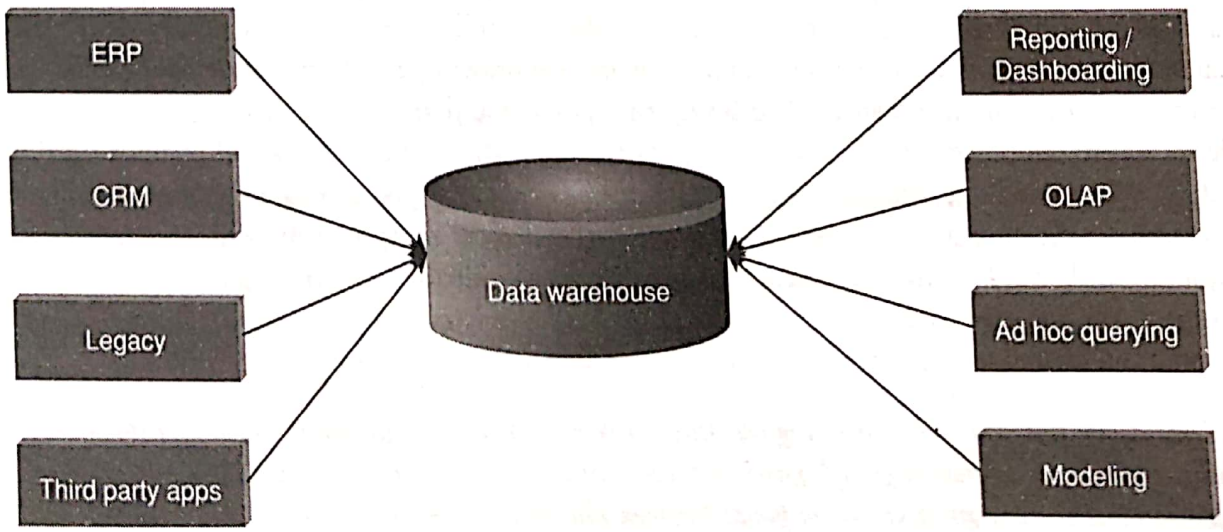



**Figure 2.9** A typical data warehouse environment.

Scanned by CamScanner

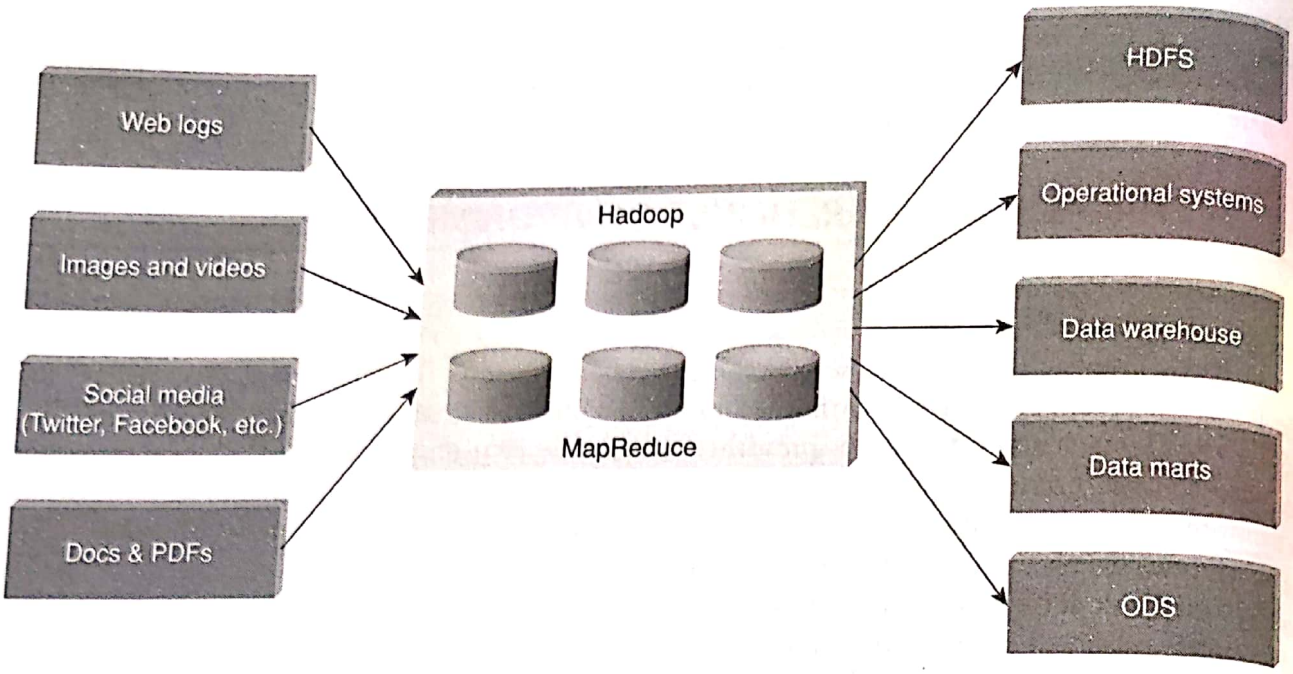



**Figure 2.10** A typical Hadoop environment.

## 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*

Scanned by CamScanner

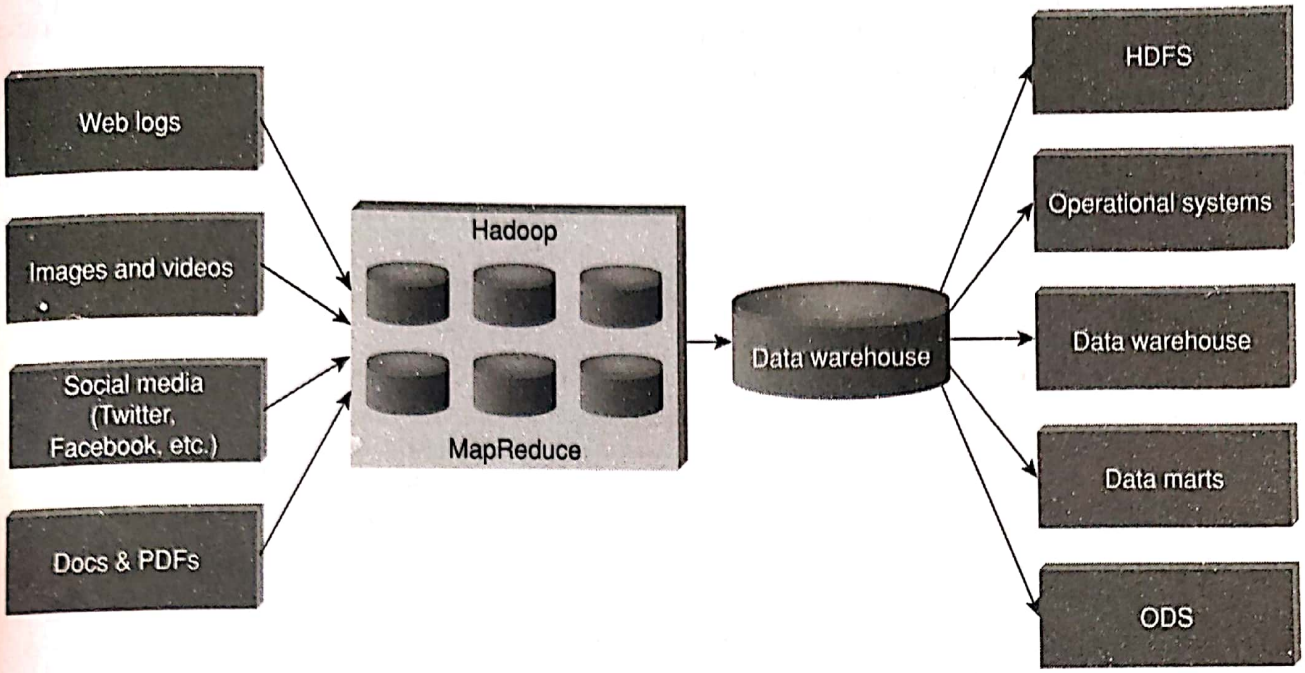



**Figure 2.11** Big data and data warehouse coexistence.

*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.

Scanned by CamScanner

3. **Value of data:** The value of data continues to see a steep rise. As the all-important resource, it is ti 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)

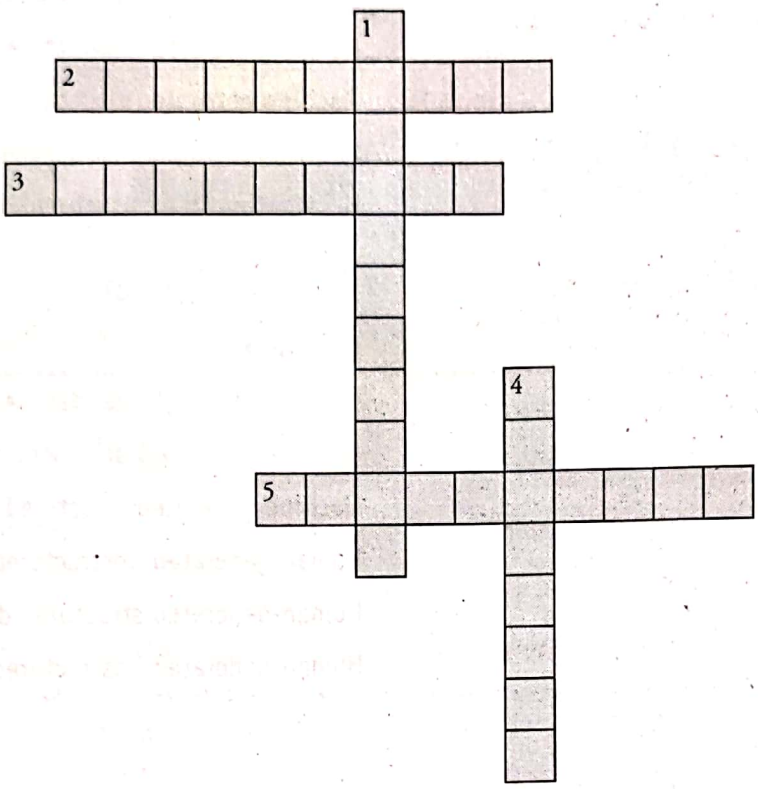
- http://en.wikipedia.org/wiki/Big_data
- http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
- https://www.oracle.com/bigdata/
- http://bigdatauniversity.com/
- http://www.sap.com/solution/big-data/software/overview.html
- http://www.ibm.com/software/data/bigdata/
- http://www.ibm.com/big-data/us/en/
- 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

Scanned by CamScanner

## 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

Scanned by CamScanner

### 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 created 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 data 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 warehouse

Scanned by CamScanner

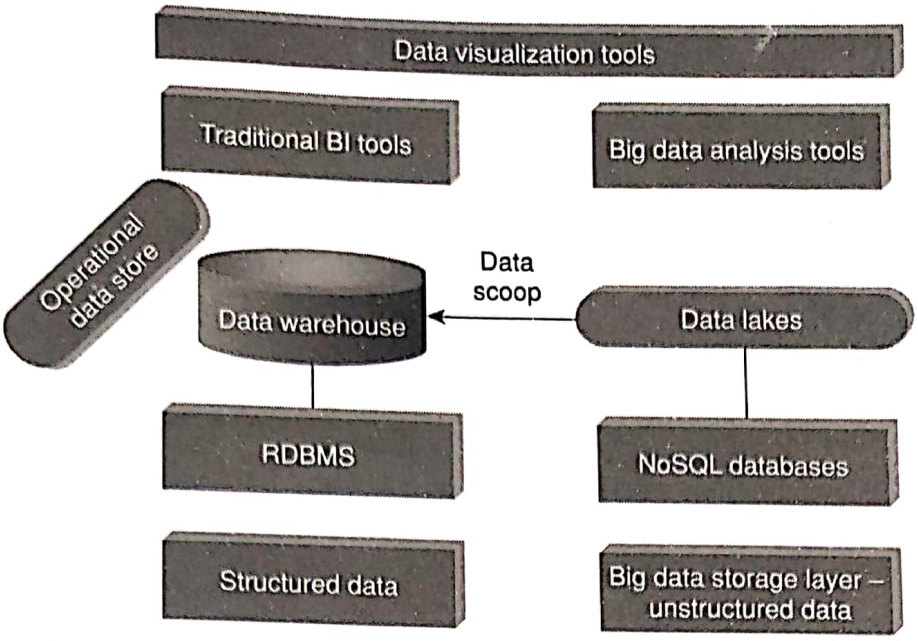



**Figure 2.12** Visualization tools for traditional BI and big data.

(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.

Scanned by CamScanner

CHAPTER 3

# Big Data Analytics

## BRIEF CONTENTS

- What's in Store?
- Where do we Begin?
- What is Big Data Analytics?
- What Big Data Analytics isn't?
- Why this Sudden Hype Around Big Data Analytics?
- Classification of Analytics
- Greatest Challenges that Prevent Businesses from Capitalizing on Big Data
- Top Challenges Facing Big Data
- Why is Big Data Analytics Important?
- What Kind of Technologies are we Looking Toward to Help Meet the Challenges Posed by Big Data?
- Data Science
- Data Scientist … Your New Best Friend!!!
- Terminologies Used in Big Data Environment
  - In-Memory Analytics
  - In-Database Processing
  - Symmetric Multiprocessor System
  - Massively Parallel Processing
  - Difference between Parallel and Distributed Systems
  - Shared Nothing Architecture
  - Consistency, Availability, Partition Tolerance (CAP) Theorem Explained
  - Basically Available Soft State Eventual Consistency (BASE)
- Few Top Analytics Tools

*"If you do not know how to ask the right question, you discover nothing."*

– W. Edwards Deming

## WHAT'S IN STORE?

This chapter is about understanding "Big Data Analytics." We have taken you through the comprehension of the term Big Data – datasets which are voluminous, rich in variety, and calls for processing at a great speed. Big data analytics is the process of examining these large datasets of big data – to unearth hidden

Scanned by CamScanner

patterns, decipher unknown correlations, understand the rationale behind market trends, and reco
customer preferences and other useful business information. The analytical findings can lead to more e
tive marketing, better customer service and satisfaction, newer products and services, improved operati
efficiency, reduced expenditure, competitive advantages over rival organizations, boosted business gains,

**PICTURE THIS...**

**Scenario 1:** You have heard a lot from your friends about the deals on offer on the Amazon site. You decide to register on www.amazon.co.in to avail their discount offers and bumper sales.

A couple of days later, you make a purchase on their site. You landed yourself a good deal by going for books by your favorite author. There is something that does not escape your attention. Amazon has made a few suggestions (of books on similar topics or books by the same author) to you to help with your next or future purchases. You wonder how Amazon's recommendation engine was able to do this for you. Is it something that they do for all their customers?

Well, Amazon's recommendation engine churns out these sort of good suggestions for customers like you, day in and day out. The company gathers all information about your past purchases together with what it knows about you, studies your buying patterns, and the buying patterns of customers like you to arrive at the recommendations that can help wit your future purchase. At the core they have big dat analytics working for them.

**Scenario 2:** You are the owner of a trucks transpor company. Your company has 500 trucks plying severa routes and carrying cargo from one place to another. It is one of those busy days where almost all the trucks are engaged in carrying cargo. You get a call to help with a cargo delivery. They are ready to pay double the charge. You do not want to miss this opportunity. But which truck should you engage. The one that is the nearest but is facing the heaviest traffic or the second nearest one but that is occupied to 75% and will not be able to take more load. There is a need to analyze the truck load, the fuel consumption, the traffic on various routes, etc. before deciding on which truck to select to pick up the new delivery.

## 3.1 WHERE DO WE BEGIN?

Raw data is collected, classified, and organized. Associating it with adequate metadata and laying bare the context converts this data into meaningful information. It is then aggregated and summarized so that it becomes easy to consume it for analysis. Gradual accumulation of such meaningful information builds a knowledge repository. This, in turn, helps with actionable insights which prove useful for decision making. Refer Figure 3.1.

Organizations have realized that they will not be able to ignore big data if they want to be competitive enough and make those timely decisions to make well of the fleeting opportunities. They will have to analyze

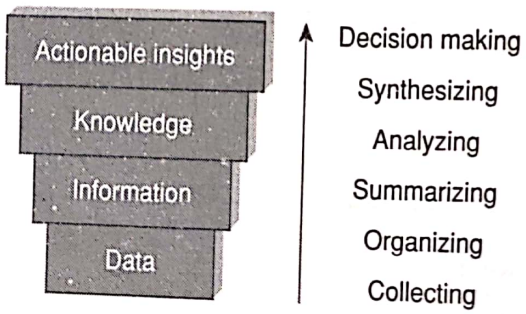



**Figure 3.1** Transformation of data to yield actionable insights.

Scanned by CamScanner

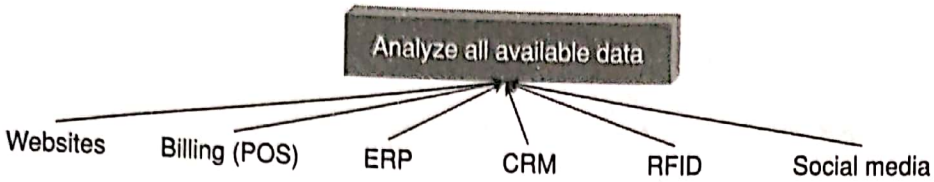



**Figure 3.2** Types of unstructured data available for analysis.

big time and also take into consideration big data that makes it to the organization at unprecedented level in terms of volume, velocity, and variety.

Big data analytics is the process of examining big data to uncover patterns, unearth trends, and find unknown correlations and other useful information to make faster and better decisions. Analytics begin with analyzing all available data. Refer Figure 3.2.

## 3.2 WHAT IS BIG DATA ANALYTICS?

***Big Data Analytics is...***

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.

   *Author's experience:* The other day I was pleasantly surprised to get a few recommendations via email from one of my frequently visited online retailers. They had recommended clothing line from my favorite brand and also the color suggested was one to my liking. How did they arrive at this? In the recent past, I had been buying clothing line of a particular brand and the color preference was pastel shades. They had it stored in their database and pulled it out while making recommendations to me.
3. About a competitive edge over your competitors by enabling you with findings that allow quicker and better decision-making.
4. A tight handshake between three communities: IT, business users, and data scientists. Refer Figure 3.3.
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. This makes perfect sense as the program for distributed processing is tiny (just a few KBs) compared to the data (Terabytes or Petabytes today and likely to be Exabytes or Zettabytes in the near future).

## 3.3 WHAT BIG DATA ANALYTICS ISN'T?

We have often asked participants of our learning programs as what comes to mind when you hear the term "Big Data." And we are not surprised by the answer... it is "Volume." But now that we have a clear understanding of big data, we know it isn't only about volume but the variety and velocity too are very important factors.

Scanned by CamScanner

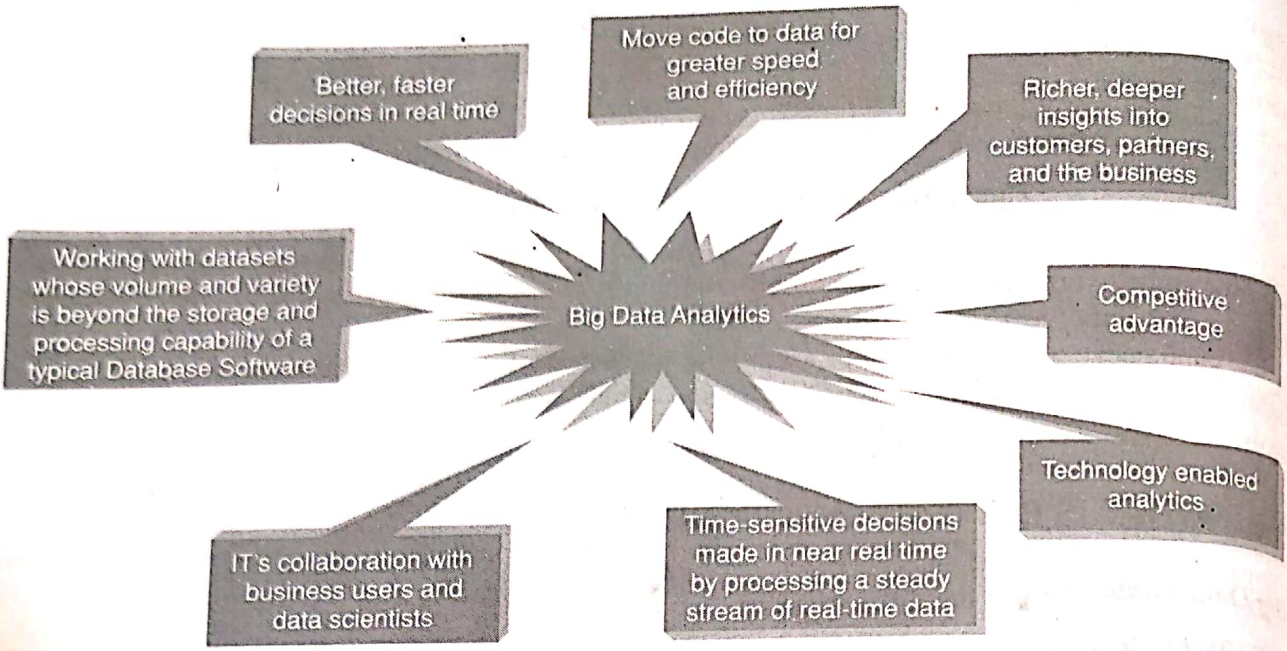



**Figure 3.3** What is big data analytics?

Refer Figure 3.4. Big data isn't just about technology. It is about understanding what the data is sayin to us. It is about understanding relationships that we thought never existed between datasets. It is abou patterns and trends waiting to be unveiled.

And of course, big data analytics is not here to replace our now very robust and powerful Relationa Database Management System (RDBMS) or our traditional Data Warehouse. It is here to coexist with bot RDBMS and Data Warehouse, leveraging the power of each to yield business value. Big data analytics is no "One-size fits all" traditional RDBMS built on shared disk and memory.

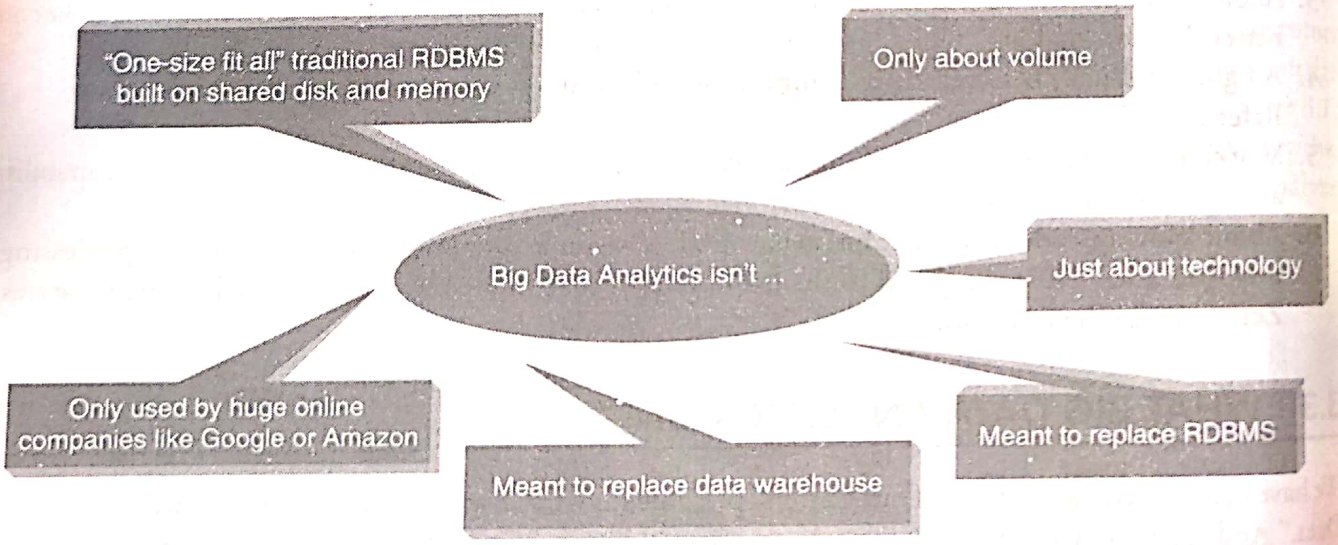



**Figure 3.4** What big data analytics isn't?

Scanned by CamScanner

And before we think it is only used by huge online companies like a Google or Amazon, let us clear the myth. It is for any business and any industry that needs actionable insights out of their data (both internal and external).

## 3.4 WHY THIS SUDDEN HYPE AROUND BIG DATA ANALYTICS?

If we go by the industry buzz, every place there seems to be talk about big data and big data analytics. Why this sudden hype? Refer Figure 3.5.

Let us put it down to three foremost reasons:

1. Data is growing at a 40% compound annual rate, reaching nearly 45 ZB by 2020. In 2010, almost about 1.2 trillion Gigabyte of data was generated. This amount doubled to 2.4 trillion Gigabyte in 2012 and to about 5 trillion Gigabytes in the year 2014. The volume of business data worldwide is expected to double every 1.2 years. Wal-Mart, the world retailer, processes one million customer transactions per hour. 500 million "tweets" are posted by Twitter users every day. 2.7 billion "Likes" and comments are posted by Facebook users in a day. Every day 2.5 quintillion bytes of data is created, with 90% of the world's data created in the past 2 years alone.

   *Source:*

   (a) http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

   (b) http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
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.

## 3.5 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.

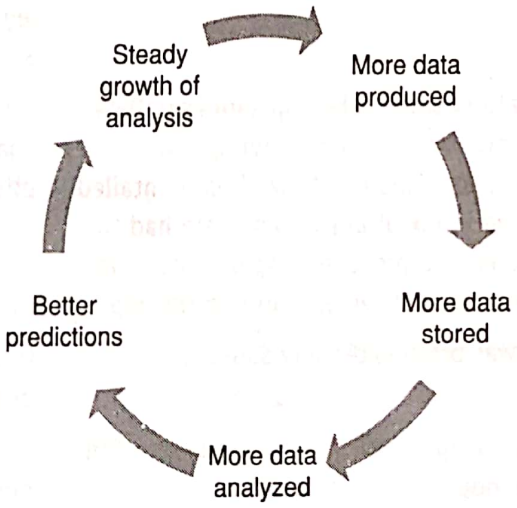



**Figure 3.5** What big data entails?

Scanned by CamScanner

### 3.5.1 First School of Thought

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

### 3.5.2 Second School of Thought

Let us take a closer look at analytics 1.0, analytics 2.0, and analytics 3.0. Refer Table 3.1.

**Table 3.1** 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 metho machine learning techniques, etc. |

Scanned by CamScanner

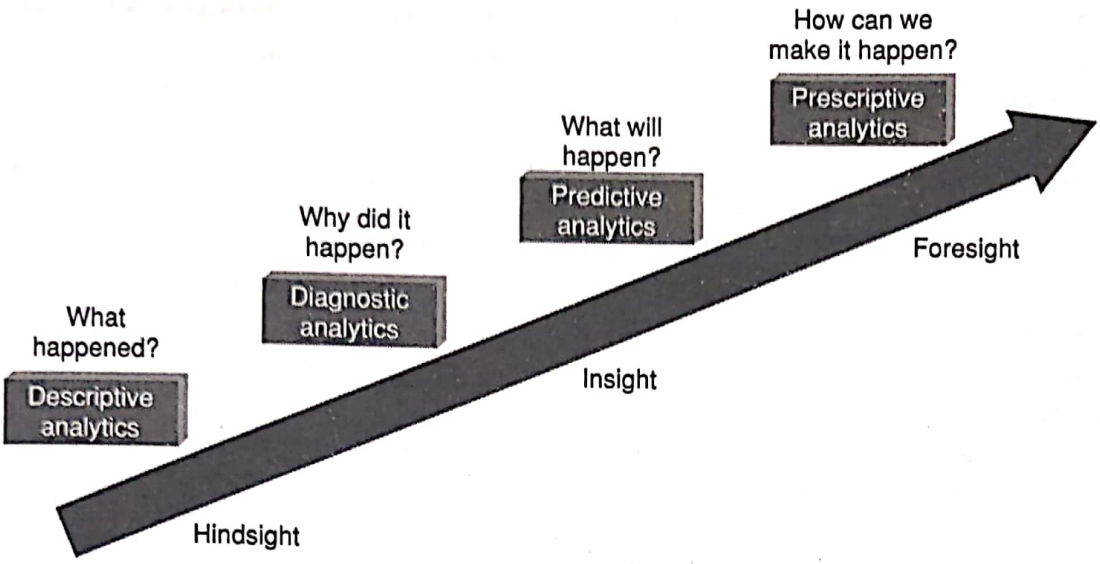



**Figure 3.6** Analytics 1.0, 2.0, and 3.0.

Figure 3.6 shows the subtle growth of analytics from Descriptive → Diagnostic → Predictive → Prescriptive analytics.

## 3.6 GREATEST 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. In other words, the 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.

## 3.7 TOP CHALLENGES FACING BIG DATA

1. **Scale:** Storage (RDBMS (Relational Database Management System) or NoSQL (Not only SQL)) is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically. The need of the hour is a storage that can best withstand the onslaught of large volume, velocity, and variety of big data? Should you scale vertically or should you scale horizontally?

Scanned by CamScanner

2. **Security:** Most of the NoSQL big data platforms have poor security mechanisms (lack of proper authentication and authorization mechanisms) when it comes to safeguarding big data. A spot that cannot be ignored given that big data carries credit card information, personal information, and other sensitive data.
3. **Schema:** Rigid schemas have no place. We want the technology to be able to fit our big data and not the other way around. The need of the hour is dynamic schema. Static (pre-defined schemas) are passé.
4. **Continuous availability:** The big question here is how to provide 24/7 support because almost all RDBMS and NoSQL big data platforms have a certain amount of downtime built in.
5. **Consistency:** Should one opt for consistency or eventual consistency?
6. **Partition tolerant:** How to build partition tolerant systems that can take care of both hardware and software failures?
7. **Data quality:** How to maintain data quality – data accuracy, completeness, timeliness, etc.? Do we have appropriate metadata in place?

## 3.8 WHY IS BIG DATA ANALYTICS IMPORTANT?

Let us study the various approaches to analysis of data and what it leads to.

1. **Reactive – Business Intelligence:** What does Business Intelligence (BI) help us with? 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.
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.
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.
4. **Proactive – Big Data Analytics:** This is sieving through terabytes, petabytes, exabytes 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.

## 3.9 WHAT KIND OF TECHNOLOGIES ARE WE LOOKING TOWARD TO HELP MEET THE CHALLENGES POSED BY BIG DATA?

1. The first requirement is of cheap and abundant storage.
2. We need faster processors to help with quicker processing of big data.
3. Affordable open-source, distributed big data platforms, such as Hadoop.
4. Parallel processing, clustering, virtualization, large grid environments (to distribute processing to a number of machines), high connectivity, and high throughputs rather than low latency.
5. Cloud computing and other flexible resource allocation arrangements.

Scanned by CamScanner

## 3.10 DATA SCIENCE

*Data science* is the science of extracting knowledge from data. In other words, it is a science of drawing out hidden patterns amongst data using statistical and mathematical techniques. It employs techniques and theories drawn from many fields from the broad areas of mathematics, statistics, information technology including machine learning, data engineering, probability models, statistical learning, pattern recognition and learning, etc.

Today we have a plethora of use-cases for "Data Science" that are already exploring massive datasets (Peta to Zetta bytes of Information) for weather predictions, oil drillings, seismic activities, financial frauds, terrorist network and activities, global economic impacts, sensor logs, social media analytics, and so many others beyond standard retail, manufacturing use-cases such as customer churn, market basket analytics (associative mining), collaborative filtering, regression analysis, etc. Data science is multi-disciplinary. Refer to Figure 3.7.

### 3.10.1 Business Acumen Skills

A data scientist should have the prowess to counter the pressures of business. A firm understanding of business domain further helps. The following is a list of traits that needs to be honed to play the role of data scientist.

1. Understanding of domain.
2. Business strategy.
3. Problem solving.
4. Communication.
5. Presentation.
6. Inquisitiveness.

### 3.10.2 Technology Expertise

It goes without saying that technology expertise will come in handy if one is to play the role of a data scientist. Cited below are few skills required as far as technical expertise is concerned.

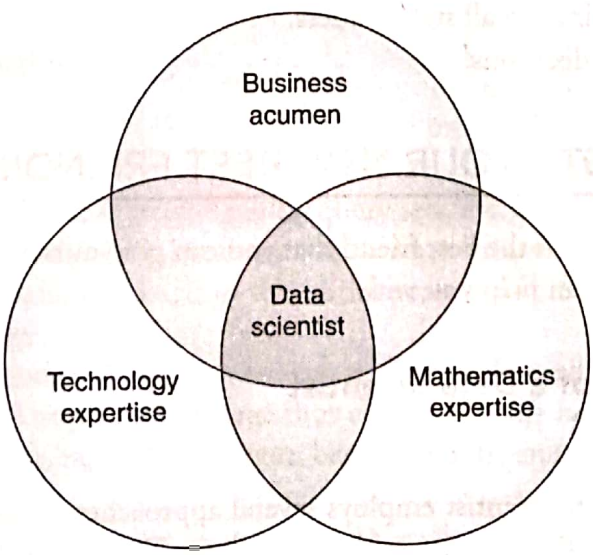



**Figure 3.7** Data scientist.

Scanned by CamScanner

1. Good database knowledge such as RDBMS.
2. Good NoSQL database knowledge such as MongoDB, Cassandra, HBase, etc.
3. Programming languages such as Java, Python, C++, etc.
4. Open-source tools such as Hadoop.
5. Data warehousing.
6. Data mining.
7. Visualization such as Tableau, Flare, Google visualization APIs, etc.

### 3.10.3 Mathematics Expertise

Since the core job of the data scientist will require him to comprehend data, interpret it, make sense of it, and analyze it, he/she will have to dabble in learning algorithms. The following are the key skills that a data scientist will have to have in his arsenal.

1. Mathematics.
2. Statistics.
3. Artificial Intelligence (AI).
4. Algorithms.
5. Machine learning.
6. Pattern recognition.
7. Natural Language Processing.

***To sum it up, the data science process is***

1. Collecting raw data from multiple disparate data sources.
2. Processing the data.
3. Integrating the data and preparing clean datasets.
4. Engaging in explorative data analysis using model and algorithms.
5. Preparing presentations using data visualizations (commonly called Infographics, or BizAnalytics, or VizAnalytics, etc.)
6. Communicating the findings to all stakeholders.
7. Making faster and better decisions.

## 3.11 DATA SCIENTIST...YOUR NEW BEST FRIEND!!!

In today's data age, a data scientist is the best friend that you can gift yourself. Refer Figure 3.8 to learn about the tasks that the data scientist can help you with.

### 3.11.1 Responsibilities of a Data Scientist

Refer Figure 3.8.

1. **Data Management:** A data scientist employs several approaches to develop the relevant datasets for analysis. Raw data is just "RAW," unsuitable for analysis. The data scientist works on it to prepare it to reflect the relationships and contexts. This data then becomes useful for processing and further analysis.

Scanned by CamScanner

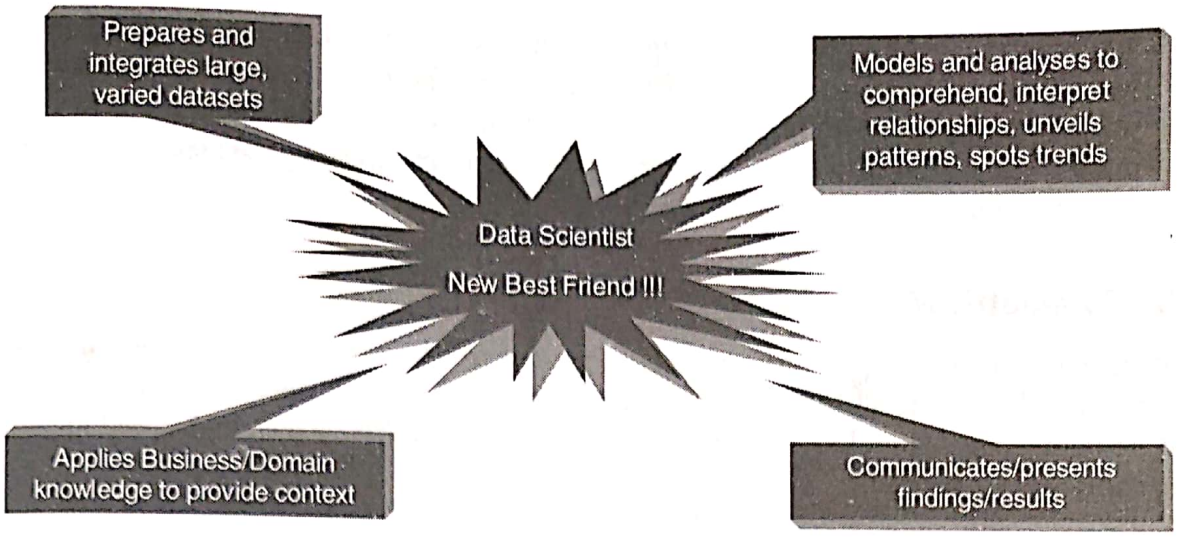



**Figure 3.8** Data scientist: your new best friend!!!

2. **Analytical Techniques:** Depending on the business questions which we are trying to find answers to and the type of data available at hand, the data scientist employs a blend of analytical techniques to develop models and algorithms to understand the data, interpret relationships, spot trends, and unveil patterns.
3. **Business Analysts:** A data scientist is a business analyst who distinguishes cool facts from insights and is able to apply his business acumen and domain knowledge to see the results in the business context. He is a good presenter and communicator who is able to communicate the results of his findings in a language that is understood by the different business stakeholders.

## 3.12 TERMINOLOGIES USED IN BIG DATA ENVIRONMENTS

In order to get a good handle on the big data environment, let us get familiar with a few key terminologies in this arena.

### 3.12.1 In-Memory Analytics

Data access from non-volatile storage such as hard disk is a slow process. The more the data is required to be fetched from hard disk or secondary storage, the slower the process gets. One way to combat this challenge is to pre-process and store data (cubes, aggregate tables, query sets, etc.) so that the CPU has to fetch a small subset of records. But this requires thinking in advance as to what data will be required for analysis. If there is a need for different or more data, it is back to the initial process of pre-computing and storing data or fetching it from secondary storage.

This problem has been addressed using in-memory analytics. Here all the relevant data is stored in Random Access Memory (RAM) or primary storage thus eliminating the need to access the data from hard disk. The advantage is faster access, rapid deployment, better insights, and minimal IT involvement.

### 3.12.2 In-Database Processing

In-database processing is also called as *in-database analytics*. It works by fusing data warehouses with analytical systems. Typically the data from various enterprise On Line Transaction Processing (OLTP) systems after

Scanned by CamScanner

cleaning up (de-duplication, scrubbing, etc.) through the process of ETL is stored in the Enterprise Data Warehouse (EDW) or data marts. The huge datasets are then exported to analytical programs for complex and extensive computations. With in-database processing, the database program itself can run the computations eliminating the need for export and thereby saving on time. Leading database vendors are offering this feature to large businesses.

### 3.12.3 Symmetric Multiprocessor System (SMP)

In SMP, there is a single common main memory that is shared by two or more identical processors. The processors have full access to all I/O devices and are controlled by a single operating system instance.

SMP are tightly coupled multiprocessor systems. Each processor has its own high-speed memory, called cache memory and are connected using a system bus. Refer Figure 3.9.

### 3.12.4 Massively Parallel Processing

*Massive Parallel Processing* (MPP) refers to the coordinated processing of programs by a number of processors working parallel. The processors, each have their own operating systems and dedicated memory. They work on different parts of the same program. The MPP processors communicate using some sort of messaging interface. The MPP systems are more difficult to program as the application must be divided in such a way that all the executing segments can communicate with each other. MPP is different from Symmetrically Multiprocessing (SMP) in that SMP works with the processors sharing the same operating system and same memory. SMP is also referred to as *tightly-coupled multiprocessing*.

### 3.12.5 Difference Between Parallel and Distributed Systems

The next two terms that we discuss are parallel and distributed systems.

As is evident from Figure 3.10, a parallel database system is a tightly coupled system. The processors co-operate for query processing. The user is unaware of the parallelism since he/she has no access to a specific

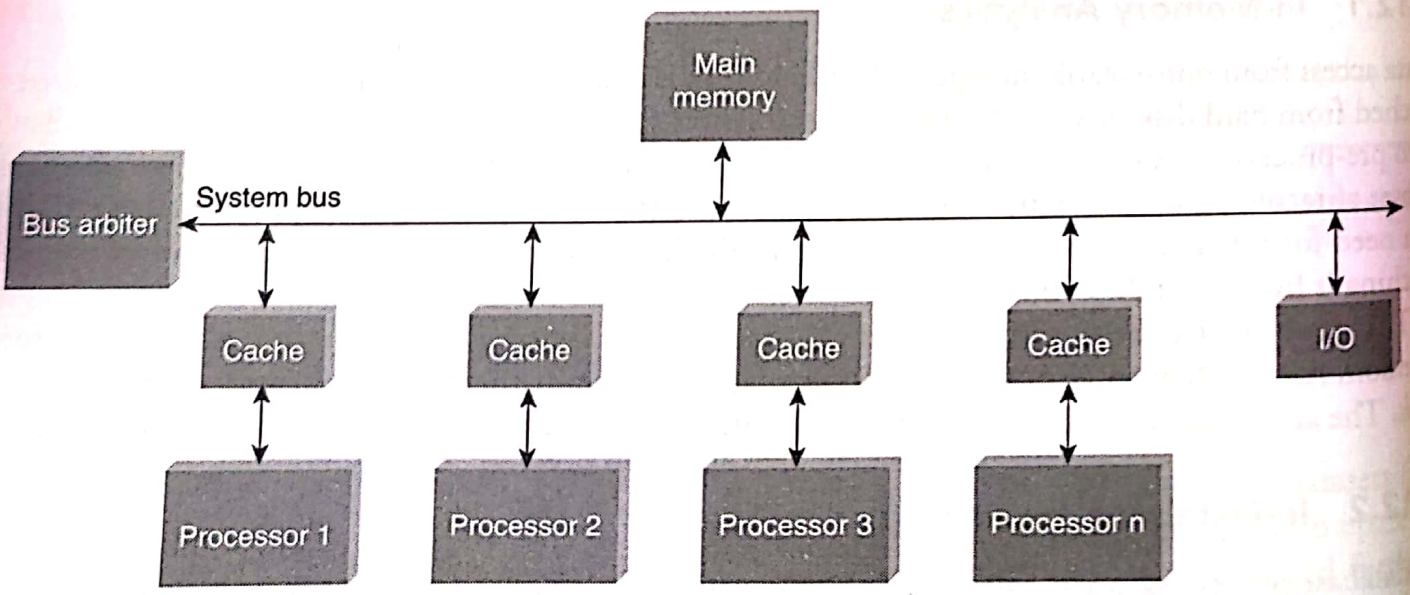



**Figure 3.9** Symmetric Multiprocessor System.

Scanned by CamScanner

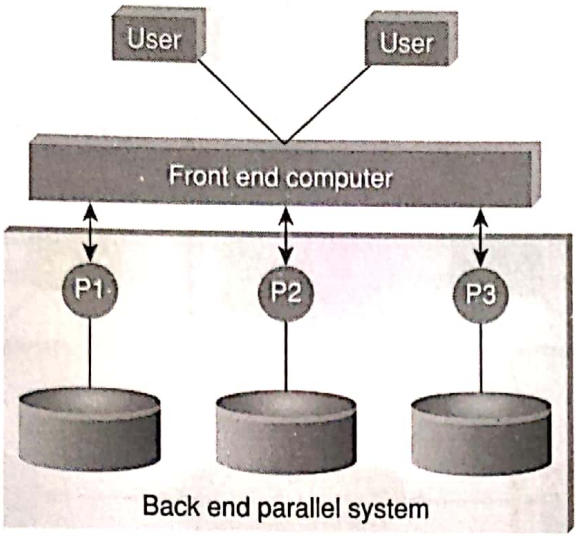



**Figure 3.10** Parallel system.

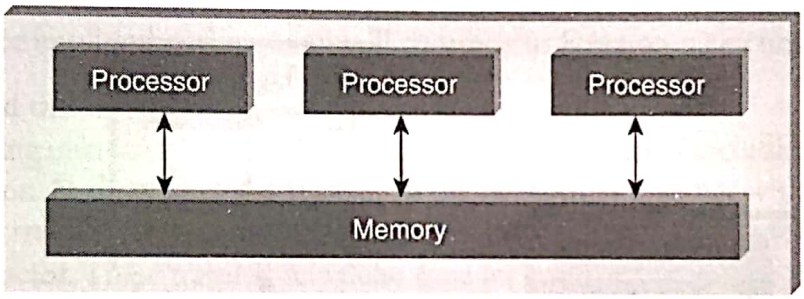



**Figure 3.11** Parallel system.

processor of the system. Either the processors have access to a common memory (Refer Fig 3.11) or make use of message passing for communication.

Distributed database systems are known to be loosely coupled and are composed by individual machines. Refer Figure 3.12. Each of the machines can run their individual application and serve their own respective user. The data is usually distributed across several machines, thereby necessitating quite a number of machines to be accessed to answer a user query. Refer Figure 3.13.

### 3.12.6 Shared Nothing Architecture

Let us look at the three most common types of architecture for multiprocessor high transaction rate systems. They are:

1. Shared Memory (SM).
2. Shared Disk (SD).
3. Shared Nothing (SN).

In shared memory architecture, a common central memory is shared by multiple processors. In shared disk architecture, multiple processors share a common collection of disks while having their own private memory. In shared nothing architecture, neither memory nor disk is shared among multiple processors.

Scanned by CamScanner

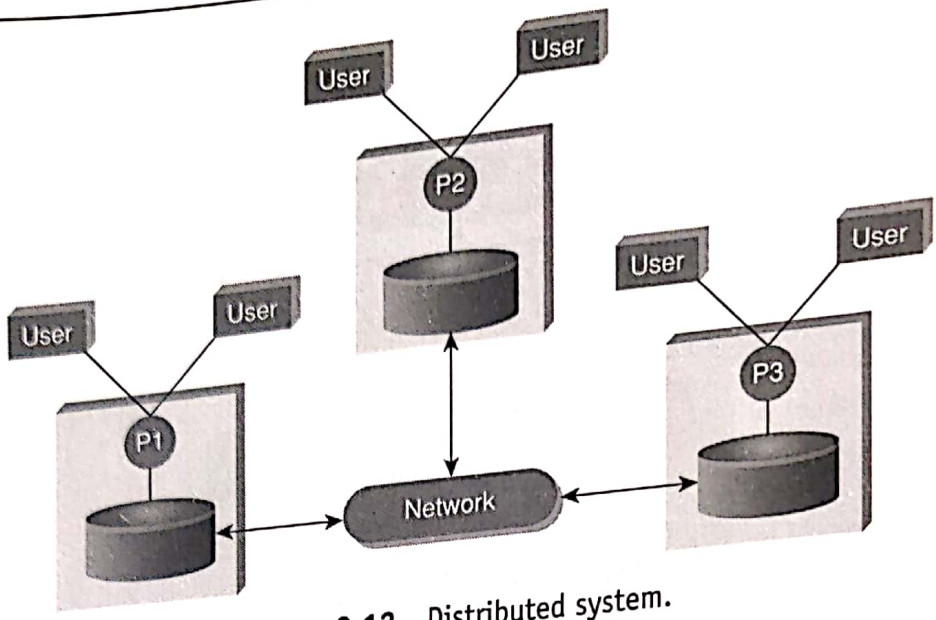



**Figure 3.12** Distributed system.

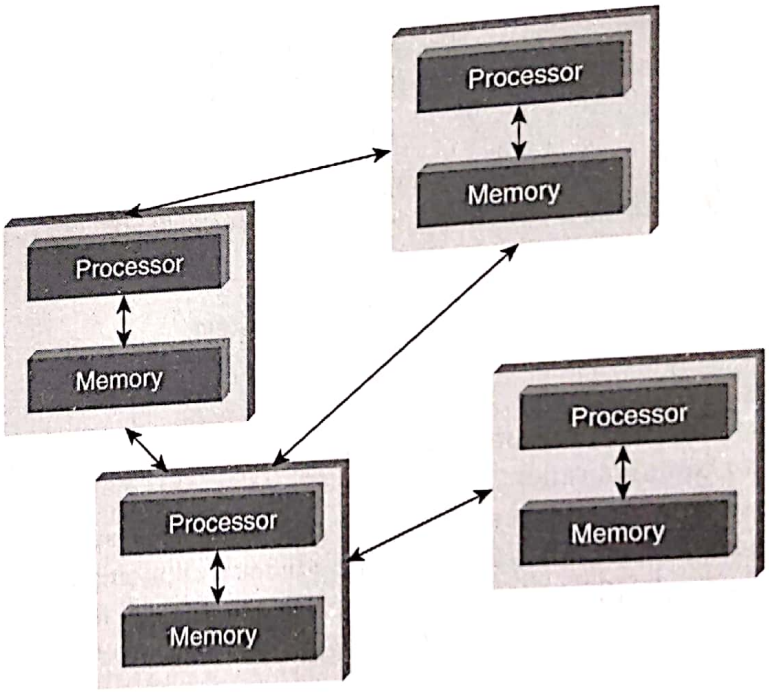



**Figure 3.13** Distributed system.

### *3.12.6.1 Advantages of a "Shared Nothing Architecture"*

1. **Fault Isolation:** A "Shared Nothing Architecture" provides the benefit of isolating fault. A fault in a single node is contained and confined to that node exclusively and exposed only through messages (or lack of it).
2. **Scalability:** Assume that the disk is a shared resource. It implies that the controller and the disk bandwidth are also shared. Synchronization will have to be implemented to maintain a consistent shared state. This would mean that different nodes will have to take turns to access the critical data. This

Scanned by CamScanner

imposes a limit on how many nodes can be added to the distributed shared disk system, thus compromising on scalability.

### 3.12.7 CAP Theorem Explained

The CAP theorem is also called the *Brewer's Theorem*. It states that in a distributed computing environment (a collection of interconnected nodes that share data), it is impossible to provide the following guarantees. Refer Figure 3.14. At best you can have two of the following three – one must be sacrificed.

1. Consistency
2. Availability
3. Partition tolerance

#### 3.12.7.1 CAP Theorem

Let us spend some time understanding the earlier mentioned terms.

1. Consistency implies that every read fetches the last write.
2. Availability implies that reads and writes always succeed. In other words, each non-failing node will return a response in a reasonable amount of time.
3. Partition tolerance implies that the system will continue to function when network partition occurs.

Let us try to understand this using a real-life situation.

You work for a training institute, "XYZ." The institute has 50 instructors including you. All of you report to a training coordinator. At the end of the month, all the instructors together with the training coordinator peruse through the training requests received from the various corporate houses and prepare a training schedule for each instructor. These training schedules (one for each instructor) are shared with "Amey," the office administrator. Each morning, you either call the office helpdesk (essentially Amey's desk) or check in-person with Amey for your schedule for the day. In case a training request has been cancelled or updated (updates can be in the form of change in course, change in duration, change of the training timings, etc.), Amey is informed of the updates and the schedules are subsequently updated by him.

Things were good until now. Few corporate houses were your clients and the schedules of each instructor could be smoothly managed without any major hiccups. But your training institute has been implementing promotion campaigns to expand the business. As a result of advertising in the media and word of mouth publicity by your existing clients, you suddenly see an upsurge in training requests from existing and new clients. In consequence of that, more instructors have been recruited. Few trainers/consultants have also been roped in from other training institutes to help tackle the load.

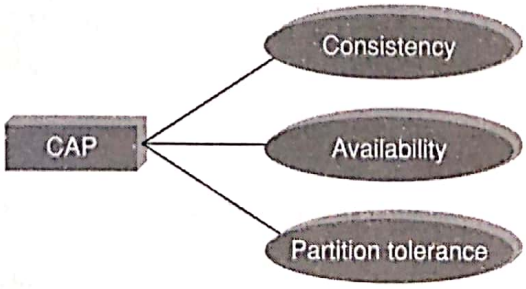



**Figure 3.14** Brewer's CAP.

Scanned by CamScanner

Now when you go to Amey to check your schedule or call in at the helpdesk, you are prepared for a wait in the queue. Looking at the current state of affairs, the training coordinator decides to recruit an additional office administrator "Joey." The helpdesk number will remain the same and will be shared by both the office administrators.

This arrangement works well for a couple of days. Then one day…

*You*: Hey Amey!

*Amey*: Hi! How can I help?

*You*: I think I am scheduled to anchor a training at 3:00 pm today. Can I please have the details?

*Amey*: Sure! Just a minute.

Amey browses through the file where he maintains the schedules. He does not see a training scheduled against your name at 3:00 pm today and responds back, "You do not have any training to conduct at 3:00 pm."

*You*: How is that possible? The training coordinator called up yesterday evening to inform of the same and said he has updated the office administrators of the same.

*Amey*: Oh! Did he say which office administrator? It could have been Joey. Please check with Joey.

*Amey*: Hey Joey! Please check the schedule for Paul here… Do you see something scheduled at 3:00 pm today?

*Joey*: Sure enough! He is anchoring the training for client "Z" today at 3:00 pm.

***A clear case of inconsistent system!!!*** The updates in the schedule were shared by the training coordinator with Joey and you were checking for your schedule with Amey.

You share this incident with the training coordinator and that gets him thinking. The issue has to be addressed immediately otherwise it will be difficult to avoid a chaotic situation. He comes up with a plan and shares it with both the office administrators the following day.

*Training Coordinator*: Folks, each time that either an instructor or me calls any one of you to update a schedule, make sure that both of you update it in your respective files. This way the instructor will always get the most recent and consistent information irrespective of whom amongst the two of you he/she speaks to.

*Joey*: But that could mean a delay in answering either a phone call or sharing the schedule with the instructor waiting in queue.

*Training Coordinator*: Yes, I understand. But there is no way that we can give incorrect information.

*Amey*: There is this other problem as well. Suppose one of us is on leave on a particular day. That would mean that we cannot take any update related calls as we will not be able to simultaneously update both the files (my file and Joey's).

*Training Coordinator*: Well, good point! *That's the availability problem!!!* But I have thought about that as well. Here is the plan:

1. If one of you receives the update call (any updates to any schedule), ensure that you inform the other person if he is available.
2. In case the other person is not available, ensure that you inform him of all the updates to all schedules via email. It is a must!!!
3. When the other person resumes duty, the first thing he will do is update his file with all the updates to all schedules that he has received via email.

Scanned by CamScanner

Wow!!! That is sure a Consistent and Available system!!!

Looks like everything is in control. Wait a minute! There is a tiff that has taken place between the office administrators. The two are pretty much available but are not talking to each other which, in other words, means that the updates are not flowing from one to the other. *We have to be partition tolerant!!!* As a training coordinator, you instruct them saying that none of you are taking any calls requesting for schedules or updates to schedules till you patch up. This implies that the system is partition tolerant but not available at that time.

In summary, one can at most decide to go with two of the three.

1. **Consistent:** The instructors or the training coordinator, once they have updated information with you, will always get the most updated information when they call subsequently.
2. **Availability:** The instructors or the training coordinators will always get the schedule if any or both of the office administrators have reported to work.
3. **Partition Tolerance:** Work will go on as usual even if there is communication loss between the office administrators owing to a spat or a tiff!

**When to choose consistency over availability and vice-versa...**

1. Choose availability over consistency when your business requirements allow some flexibility around when the data in the system synchronizes.
2. Choose consistency over availability when your business requirements demand atomic reads and writes.

**Examples of databases that follow one of the possible three combinations:**

1. Availability and Partition Tolerance (AP)
2. Consistency and Partition Tolerance (CP)
3. Consistency and Availability (CA)

Refer Figure 3.15 to get a glimpse of databases that adhere to two of the three characteristics of CAP theorem.

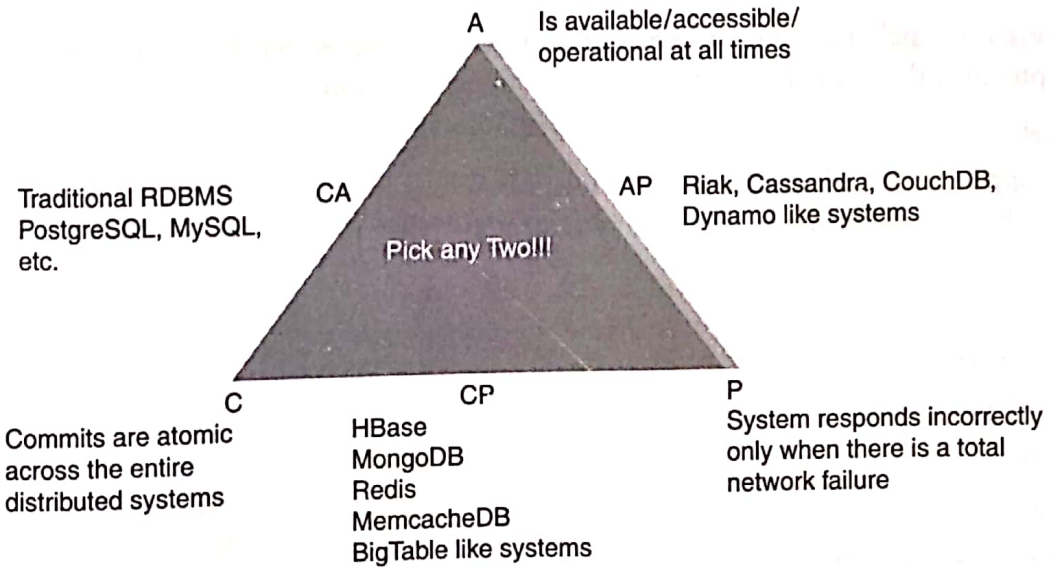



**Figure 3.15** Databases and CAP.

Scanned by CamScanner

## 3.13 BASICALLY AVAILABLE SOFT STATE EVENTUAL CONSISTENCY (BASE)

A few basic questions to start with:

1. ***Where is it used?***
   In distributed computing.

2. ***Why is it used?***
   To achieve high availability.

3. ***How is it achieved?***
   Assume a given data item. If no new updates are made to this given data item for a stipulated period of time, eventually all accesses to this data item will return the updated value. In other words, if no new updates are made to a given data item for a stipulated period of time, all updates that were made in the past and not applied to this given data item and the several replicas of it will percolate to this data item so that it stays as current/recent as is possible.

4. ***What is replica convergence?***
   A system that has achieved eventual consistency is said to have converged or achieved *replica convergence.*

5. ***Conflict resolution*: How is the conflict resolved?**
   (a) **Read repair:** If the read leads to discrepancy or inconsistency, a correction is initiated. It slows down the read operation.
   (b) **Write repair:** If the write leads to discrepancy or inconsistency, a correction is initiated. This will cause the write operation to slow down.
   (c) **Asynchronous repair:** Here, the correction is not part of a read or write operation.

## 3.14 FEW TOP ANALYTICS TOOLS

There is no dearth of analytical tools in the market. Please find below our list of few top analytics tools. We have also provided the links after each tool for you to explore more…

1. MS Excel
   https://support.office.microsoft.com/en-in/article/Whats-new-in-Excel-2013-1cbc42cd-bfaf-43d7-9031-5688ef1392fd?CorrelationId=1a2171cc-191f-47de-8a55-08a5f2e9c739&ui=en-US&rs=en-IN&ad=IN
2. SAS
   http://www.sas.com/en_us/home.html
3. IBM SPSS Modeler
   http://www-01.ibm.com/software/analytics/spss/products/modeler/
4. Statistica
   http://www.statsoft.com/

Scanned by CamScanner

5. Salford systems (World Programming Systems)
   http://www.salford-systems.com/
6. WPS
   http://www.teamwpc.co.uk/products/wps

### 3.14.1 Open Source Analytics Tools

Let us look at a couple of open source analytics tools. We have also provided the links after each tool for you to explore more…

1. R analytics
   http://www.revolutionanalytics.com/
2. Weka
   http://www.cs.waikato.ac.nz/ml/weka/

## REMIND ME

- 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.
- Big data analytics is a about a tight handshake between three communities: IT, business users, and data scientists.
- *Data science* is the science of extracting knowledge from data.
- The CAP theorem is also called the Brewer's Theorem. It states that in a distributed computing environment (a collection of interconnected nodes that share data), it is impossible to provide the following guarantees. At best you can have two of the following three – one must be sacrificed.
  - Consistency
  - Availability
  - Partition tolerance

## CONNECT ME (INTERNET RESOURCES)

- http://en.wikipedia.org/wiki/Data_science
- http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/
- http://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html
- http://spotfire.tibco.com/blog/?p=4240
- http://reports.informationweek.com/abstract/106/1255/Financial/tech-center-taking-advantage-of-in-memory-analytics.html
- http://www.informationweek.com/software/information-management/oracle-analytics-package-expands-in-database-processing-options/d/d-id/1102712?

Scanned by CamScanner

## TEST ME

### A. Fill Me

1. The _____ technology helps query data that resides in a computer's random access memory (RAM) rather than data stored on physical disks.
2. Eventual consistency is a consistency model used in distributed computing to achieve high _____.
3. A coordinated processing of a program by multiple processors, each working on different parts of the program and using its own operating system and memory is called _____.
4. A collection of independent computers that appear to its users as a single coherent system is _____.

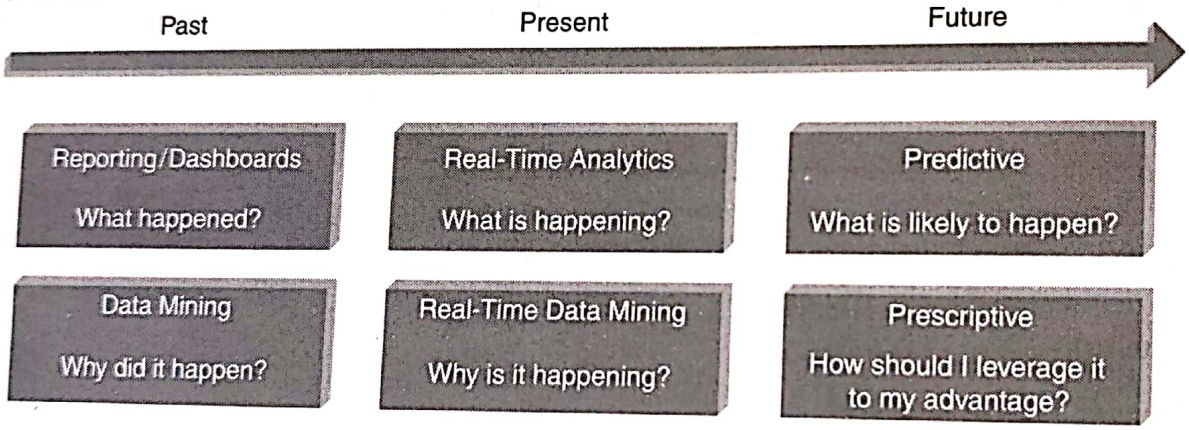
**Answers:**

1. In-memory analytics
2. Availability
3. Massively parallel processing
4. Distributed systems

### B. Answer Me

1. What are the various types of analytics?

**Answer:**




2. What are the key questions to be answered by all organizations stepping into analytics?

**Answer:** The key questions for any organization stepping into analytics are:

- Should you be storing all of your big data? If "Yes", where are you going to store it? If "No", how will you know what to store and what to discard?
- How will you sieve through your massive data to filter out the relevant from the irrelevant?
- How long will you store this data?
- How will you accommodate the peaks (variability in terms of data influx) in your data?
- How will you analyze? Will you analyze all the data that is stored or analyze a sample?
- What will you do with the insights generated from this analysis?

3. What can one expect from analytics 3.0?

**Answer:**

- In-memory analytics.
- In-database processing.
- Leveraging analytics to improve operational, tactical, and strategic decision making.

Scanned by CamScanner

- Coupling the in-memory analytics and in-database processing with agile analytical methods and machine learning techniques.
- Appropriate tools to effectively support decision-making at the front lines, such as mobile and self-serve analytical applications.

4. Which industries will be affected most by analytics 3.0? Who will benefit the most?

**Answer:** Almost all the firms in all the industries and not just online firms will be affected by analytics 3.0. A lot of analytics have already been done in the Transport, Retail, and Banking sector. Telecom, entertainment, and health sectors have a bit of catching up to do.

5. What is predictive and prescriptive analytics?

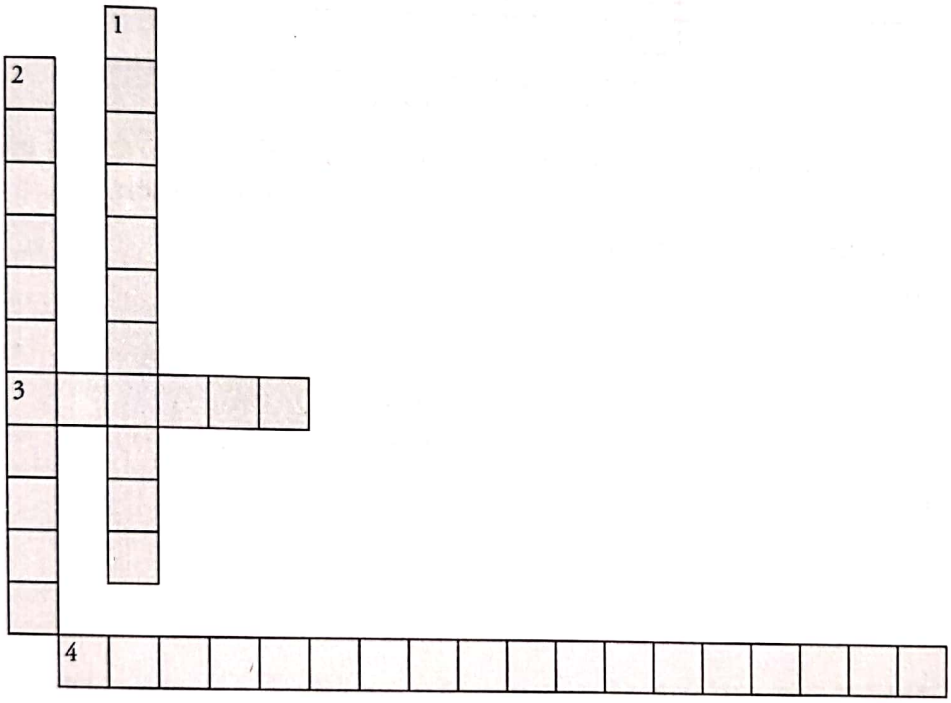
**Answer:**

Predictive analytics helps you answer the questions: "What will happen?" and "Why will it happen"?

Prescriptive analytics goes beyond "What will happen?" "Why will it happen?" and "When will it happen?" to answer "What should be the action taken to take advantage of what will happen"?

## C. Crossword

### 1. Puzzle on CAP Theorem




**Across**

3. CAP theorem is also called as _____ theorem.
4. System will continue to function even when network partition occurs.

**Down**

1. Every read fetches the most recent write.
2. A non-failing node will return a reasonable response within a reasonable amount of time.

**Solution:**

**Across**

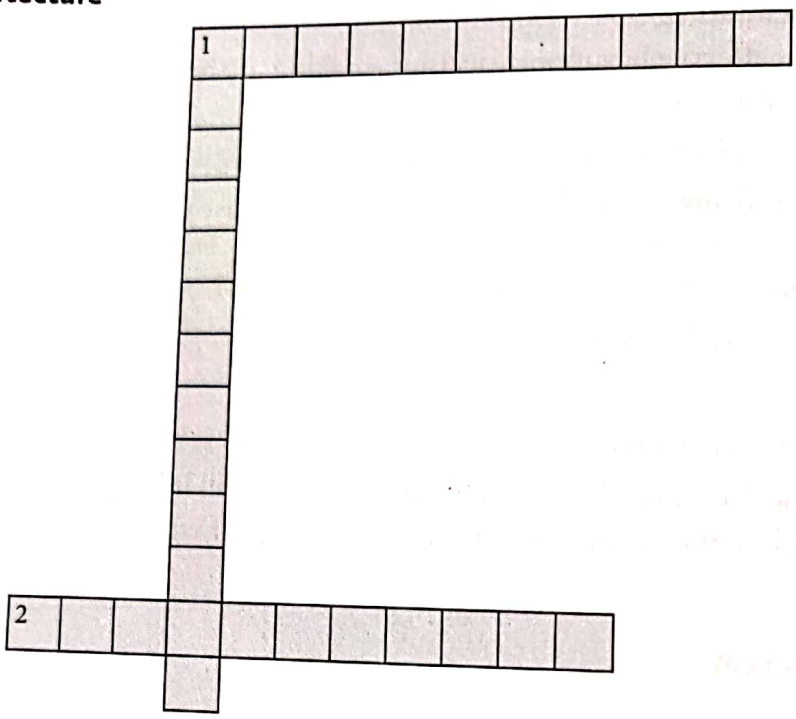
3. Brewer
4. Partition Tolerant

**Down**

1. Consistency
2. Availability

Scanned by CamScanner

## 2. Puzzle on Architecture




**Across**

1. _____ is an important advantage of shared nothing architecture.
2. In this architecture, multiple processors have their own private memory.

**Down**

1. In this architecture, central memory is share by multiple processors.

**Answer:**

**Across**

1. Scalability
2. Shared Disk

**Down**

1. Shared Memory

Scanned by CamScanner

CHAPTER 4

# The Big Data Technology Landscape

## BRIEF CONTENTS

- What's in Store?
- NoSQL (Not Only SQL)
  - Where is it used?
  - What is it?
  - Types of NoSQL Databases
  - Why NoSQL?
  - Advantages of NoSQL
  - What we miss with NoSQL?
  - Use of NoSQL in Industry
  - NoSQL Vendors
  - SQL versus NoSQL
  - NewSQL
  - Comparison of SQL, NoSQL, and NewSQL
- Hadoop
  - Features of Hadoop
  - Key Advantages of Hadoop
  - Versions of Hadoop
    - Hadoop 1.0
    - Hadoop 2.0
- Overview of Hadoop Ecosystems
- Hadoop Distributions
- Hadoop versus SQL
- Integrated Hadoop Systems Offered by Leading Market Vendors
- Cloud-Based Hadoop Solutions

*"The goal is to turn data into information, and information into insight."*

– Carly Fiorina, former CEO, Hewlett-Packard Co

## WHAT'S IN STORE?

The focus of this chapter is on understanding "big data technology landscape". This chapter is an overview on NoSQL and Hadoop. There are separate chapters on NoSQL (MongoDB and Cassandra) as well as Hadoop in the book.

The big data technology landscape can be majorly studied under two important technologies:

1. NoSQL
2. Hadoop

Scanned by CamScanner

## 4.1 NoSQL (NOT ONLY SQL)

The term NoSQL was first coined by Carlo Strozzi in 1998 to name his lightweight, open-source, relational database that did not expose the standard SQL interface. Johan Oskarsson, who was then a developer at last.fm, in 2009 reintroduced the term NoSQL at an event called to discuss open-source distributed network. The #NoSQL was coined by Eric Evans and few other database people at the event found it suitable to describe these non-relational databases.

Few features of NoSQL databases are as follows:

1. They are open source.
2. They are non-relational.
3. They are distributed.
4. They are schema-less.
5. They are cluster friendly.
6. They are born out of 21st century web applications.

### 4.1.1 Where is it Used?

NoSQL databases are widely used in big data and other real-time web applications. Refer Figure 4.1. NoSQL databases is used to stock log data which can then be pulled for analysis. Likewise it is used to store social media data and all such data which cannot be stored and analyzed comfortably in RDBMS.

### 4.1.2 What is it?

**NoSQL** stands for Not Only SQL. These are non-relational, open source, distributed databases. They are hugely popular today owing to their ability to scale out or scale horizontally and the adeptness at dealing with a rich variety of data: structured, semi-structured and unstructured data. Refer Figure 4.2 for additional features of NoSQL. NoSQL databases.

1. **Are non-relational:** They do not adhere to relational data model. In fact, they are either key–value pairs or document-oriented or column-oriented or graph-based databases.

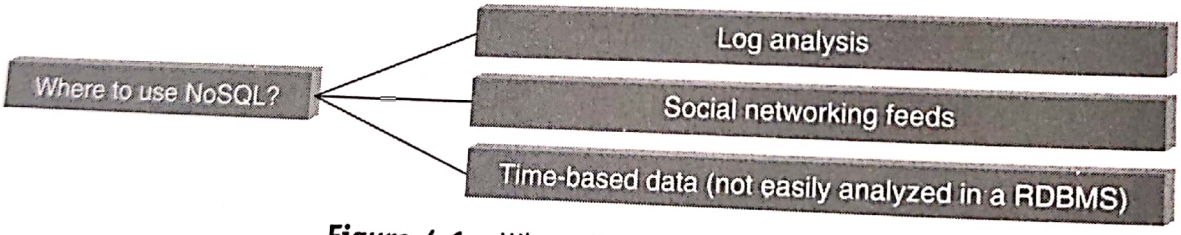



**Figure 4.1** Where to use NoSQL?

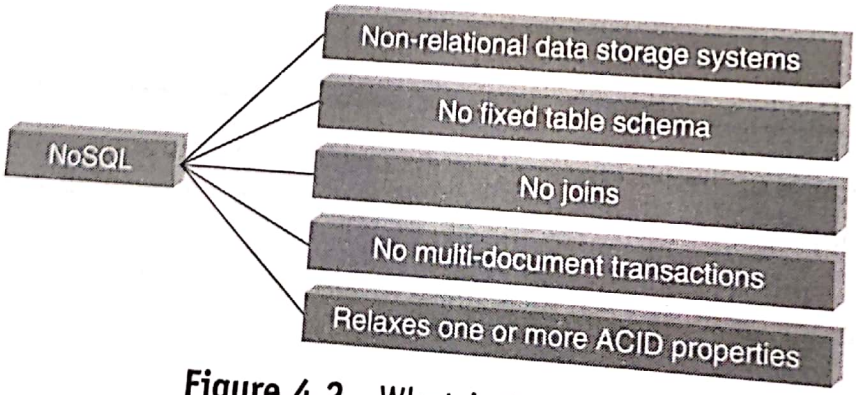



**Figure 4.2** What is NoSQL?

Scanned by CamScanner

2. **Are distributed:** They are distributed meaning the data is distributed across several nodes in a cluster constituted of low-cost commodity hardware.
3. **Offer no support for ACID properties (Atomicity, Consistency, Isolation, and Durability):** They do not offer support for ACID properties of transactions. On the contrary, they have adherence to Brewer's CAP (Consistency, Availability, and Partition tolerance) theorem and are often seen compromising on consistency in favor of availability and partition tolerance.
4. **Provide no fixed table schema:** NoSQL databases are becoming increasing popular owing to their support for flexibility to the schema. They do not mandate for the data to strictly adhere to any schema structure at the time of storage.

### 4.1.3 Types of NoSQL Databases

We have already stated that NoSQL databases are non-relational. They can be broadly classified into the following:

1. Key–value or the big hash table.
2. Schema-less.

Refer Figure 4.3. Let us take a closer look at key–value and few other types of schema-less databases:

1. **Key–value:** It maintains a big hash table of keys and values. For example, Dynamo, Redis, Riak, etc.
   ***Sample Key–Value Pair in Key–Value Database***

| Key | Value |
|---|---|
| First Name | Simmonds |
| Last Name | David |

2. **Document:** It maintains data in collections constituted of documents. For example, MongoDB, Apache CouchDB, Couchbase, MarkLogic, etc.
   ***Sample Document in Document Database***

```
{
"Book Name":    "Fundamentals of Business Analytics",
"Publisher":    "Wiley India",
"Year of Publication":    "2011"
}
```

. **Column:** Each storage block has data from only one column. For example: Cassandra, HBase, etc.

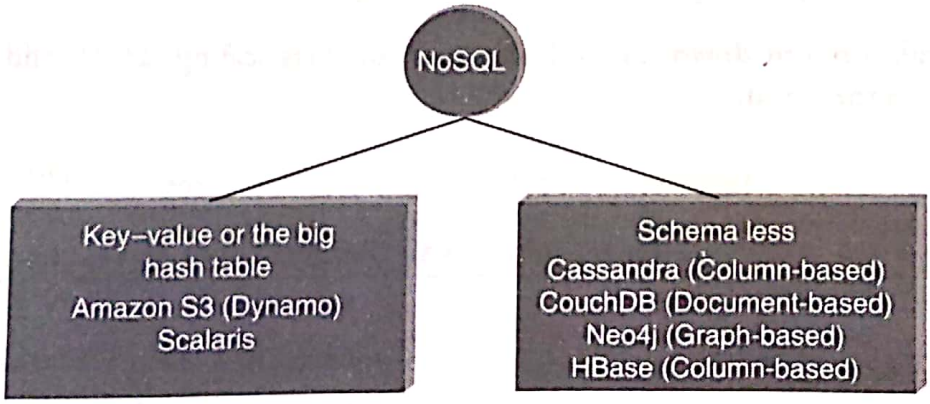



**Figure 4.3** Types of NoSQL databases.

Scanned by CamScanner