Module 1: Fundamental Big Data

Introduction

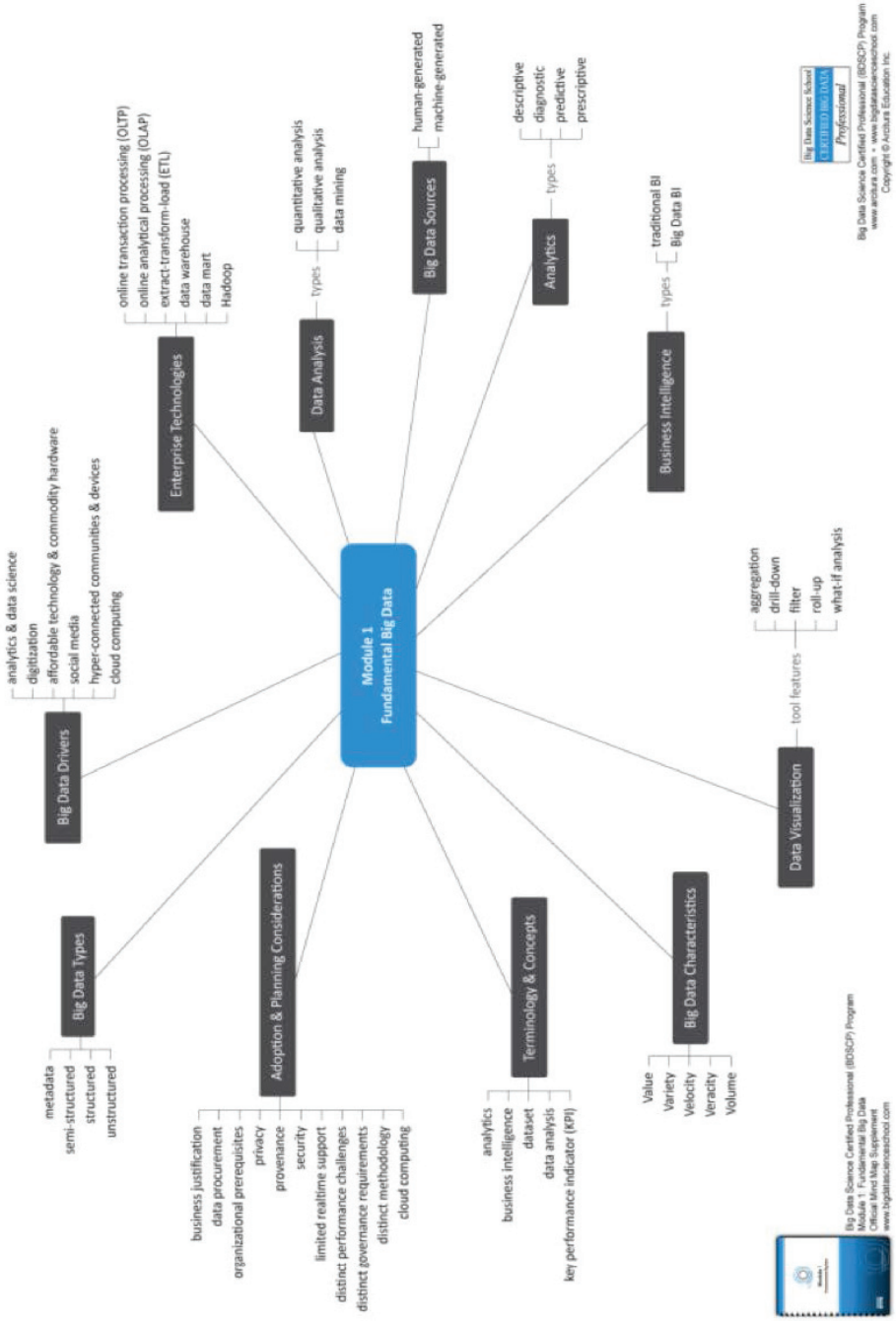
This is the official course booklet for the Big Data Science Certified Professional Course **Module 1: Fundamental Big Data** and the corresponding Pearson VUE **Exam B90.01**.

The purpose of this document is to establish an understanding of fundamental Big Data concepts, which include but are not limited to:

- Understanding Big Data
- Fundamental Big Data Terminology & Concepts
- Big Data Business & Technology Drivers
- Traditional Enterprise Technologies Related to Big Data
- Characteristics of Data in Big Data Environments
- Types of Data in Big Data Environments
- Fundamental Analysis, Analytics & Machine Learning Types
- Business Intelligence & Big Data
- Data Visualization & Big Data
- Big Data Adoption & Planning Considerations

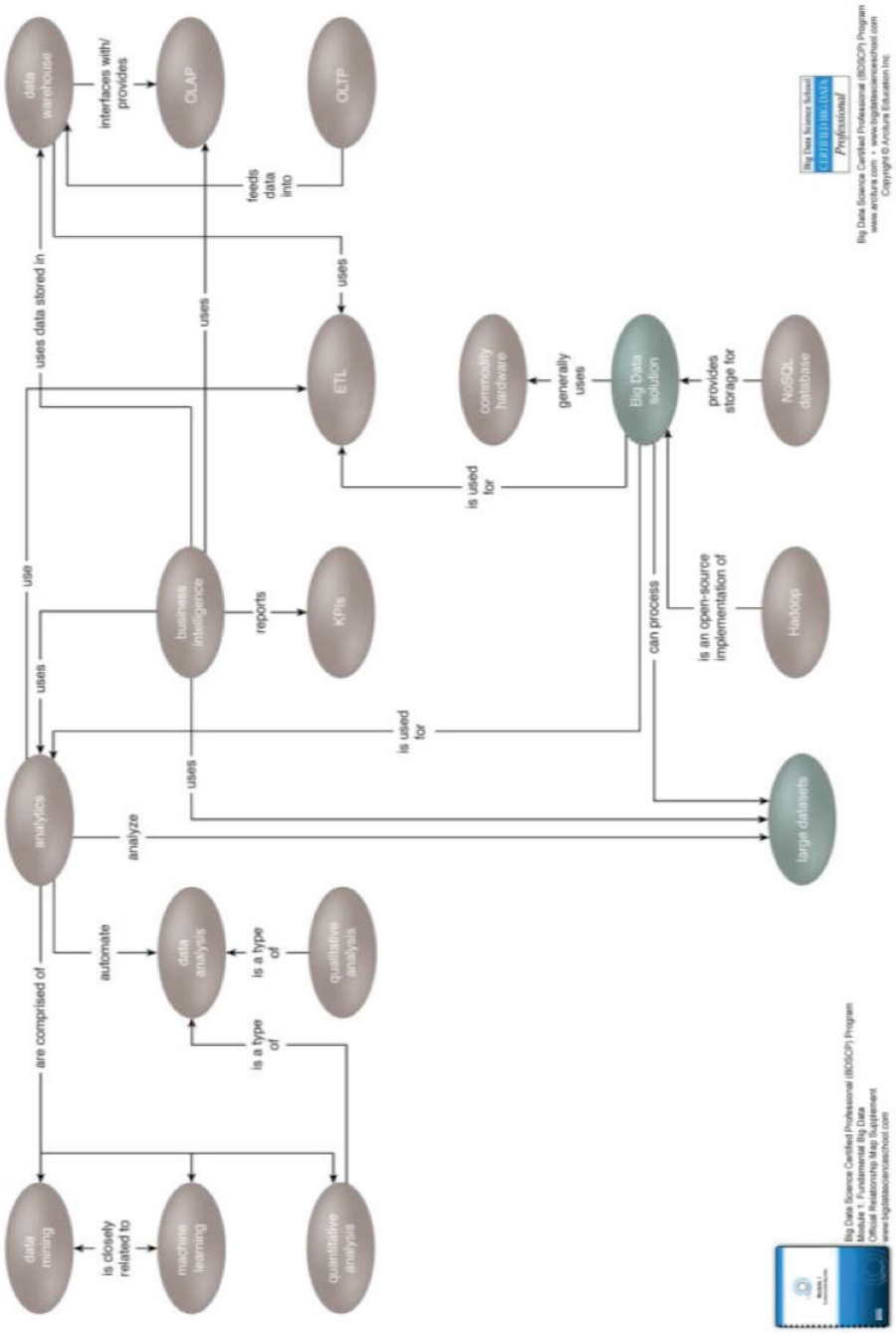
Mind Map Poster

The *BDSCP Module 1: Mind Map Poster* that accompanies this course booklet provides an alternative visual representation of all primary topics covered in this course.



Relationship Poster

The *BDSCP Module 1: Relationship Poster* that accompanies this course booklet provides an alternative visual representation of key topics covered in this course.



Understanding Big Data

Introducing Big Data

Big Data is a field dedicated to the analysis, processing, and storage of large collections of data that frequently originate from disparate sources. Big Data solutions and practices are typically required when traditional data analysis, processing, and storage technologies and techniques are insufficient. Specifically, Big Data addresses distinct requirements, such as the combining of multiple unrelated datasets, processing of large amounts of unstructured data, and harvesting of hidden information in a time-sensitive manner.

The qualities that distinguish data processed by Big Data solutions are commonly known as the "Five Vs" and are documented in the upcoming *Characteristics of Data in Big Data Environments* section. Using Big Data solutions, complex analysis tasks can be carried out to arrive at deeply meaningful and insightful analysis results for the benefit of the business. Big Data solutions can process massive quantities of data that arrive at varying speeds, may be of many different varieties, and have numerous incompatibilities.

Data within Big Data environments accumulates from being amassed within the enterprise via applications or from external sources that are then stored by the Big Data solution. Data processed by a Big Data solution can be used by enterprise applications directly, or fed into a data warehouse to enrich existing data. This data is typically analyzed and subjected to analytics.

Processed data and analysis results are commonly used for meaningful and complex reporting and assessment tasks, and can also be fed back into applications to enhance their behavior, such as when product recommendations are displayed online. The data processed by Big Data solutions can be human-generated or machine-generated, although it is ultimately the responsibility of machines to generate the processing results. Human-generated data is the result of human interaction with systems, such as online services and digital devices. Figure 1.1 shows examples of human-generated data, which can come in the form of structured data, video, and textual data.

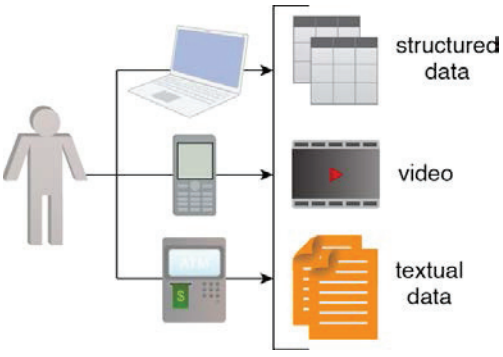


Figure 1.1 – Examples of human-generated data include social media, micro blogging, e-mails, photo sharing, and messaging.

Machine-generated data is the result of the automated, event-driven generation of data by software programs or hardware devices. Figure 1.2 provides a visual representation of examples of machine-generated data from Web servers, smart meters, and GPS devices.

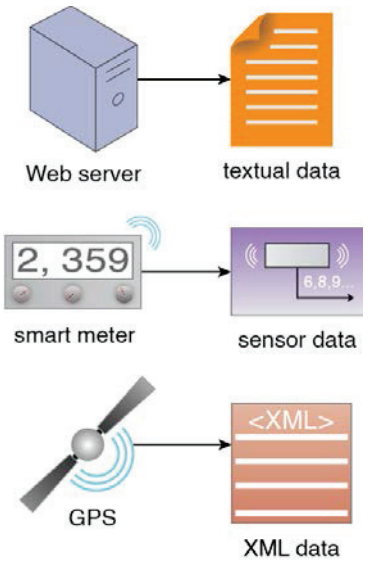


Figure 1.2 – Examples of machine-generated data include Web logs, sensor data, telemetry data, smart meter data, and appliance usage data.

Big Data solution processing results can lead to a wide range of insights and benefits, such as:

- operational optimization
- actionable intelligence
- identification of new markets
- accurate predictions
- fault and fraud detection
- more detailed records
- improved decision-making
- scientific discoveries

There are numerous concerns, limitations, and considerations that come with a Big Data adoption effort, all of which need to be understood and weighed against its anticipated benefits. Many of these are discussed separately in the *Big Data Adoption & Planning Considerations* section.

Optional Reading

The Big Data Analytics book that is included with Module 2: Big Data Analysis & Technology Concepts discusses case study examples with real-world perspectives in the Too Big to Ignore: The Business Case for Big Data section of Chapter 5.

Notes

Notes / Sketches

Fundamental Terminology & Concepts

In preparation for subsequent sections that cover the next set of introductory topics, the upcoming pages provide concise definitions of the following basic terms:

- Datasets
- Data Analysis
- Analytics
- Business Intelligence (BI)
- Key Performance Indicators (KPIs)

The section concludes with the terms and abbreviations used as part of data size terminology.

Datasets

Collections or groups of related data are generally referred to as datasets in these course modules. Each group or dataset member (datum) shares the same set of attributes as others in the same dataset.

Figure 1.3 shows three datasets based on three different formats. Examples can include:

- tweets stored in a flat file
- a collection of image files
- an extract of rows stored in a table
- historical weather observations that are stored as XML files

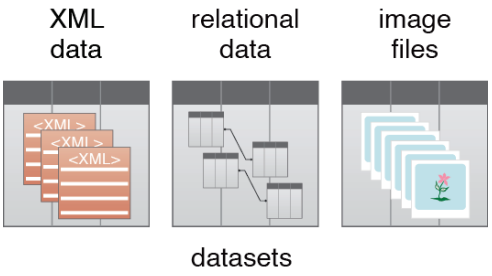


Figure 1.3 – Datasets can be based on XML data, relational data, and/or image files.

Data Analysis

Data analysis is the process of examining data to find facts, relationships, patterns, insights, and/or trends. The eventual goal of data analysis is to support decision-making. A simple data analysis example is the analysis of ice cream sales data in order to determine how the number of ice cream cones sold is related to the daily temperature. This supports decisions of how much ice cream and how many cones a store should order and stock in relation to weather forecast information. Carrying out data analysis helps establish patterns and relationships among the data being analyzed.

Figure 1.4 - The symbol used to represent data analysis.

Analytics

Analytics is the discipline of gaining an understanding of data by analyzing it via a multitude of scientific techniques and automated tools, with a focus on locating hidden patterns and correlations. In Big Data environments, analytics is usually applied using highly scalable distributed technologies and frameworks for analyzing large volumes of data from different sources.

Figure 1.5 - The symbol used to represent analytics.

The process of analytics generally involves sifting through large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data, or performing large-scale searches.

Different kinds of organization use analytics tools and techniques in different ways, such as these three sectors:

- In business-oriented environments, analytics results can lower operational costs and facilitate strategic decision-making.
- In the scientific domain, analytics can help identify the cause of a phenomenon to improve the accuracy of predictions.

- In services-based environments, such as in public sector organizations, analytics can help strengthen the focus on delivering high quality services by driving down costs.

In general, analytics enables data-driven decision-making with scientific backing, so that decisions can be based on factual data and not on past experience or intuition alone.

Business Intelligence (BI)

Business intelligence (BI) is the process of gaining insight into the workings of an enterprise to improve decision-making by analyzing external data and data generated by its business processes. BI applies analytics to large amounts of data across the enterprise. BI can further utilize the consolidated data contained in data warehouses to run analytical queries. As shown in Figure 1.6, BI can be used via a dashboard mechanism to access and analyze queries on this data.

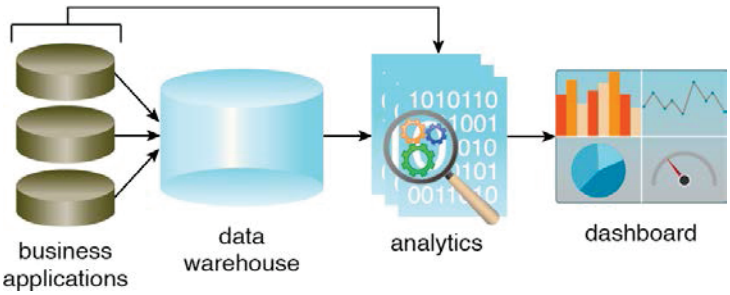


Figure 1.6 – BI can be used to improve business applications, consolidate data in data warehouses, and analyze queries via a dashboard.

Key Performance Indicators (KPIs)

A key performance indicator (KPI) is a measure for gauging success within a particular context. KPIs are closely linked with an enterprise's strategic objectives and are generally used to:

- identify problem areas in order to take corrective actions
- achieve regulatory compliance

KPIs act as quick reference points for measuring the overall performance of the business via KPI dashboards, shown in Figure 1.7. Each KPI is based on a quantifiable indicator that is identified and agreed upon beforehand.

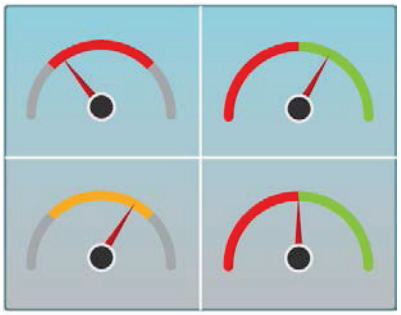


Figure 1.7 – A KPI dashboard can gauge calls handled per day and the number of units manufactured per month.

Data Size Units

When discussing ranges in data sizes, it is necessary to understand the corresponding ranges in units of data measurement. The following data sizes are listed in Table 1.1 using the byte as the fundamental unit of measurement with decimal, not binary, prefixes.

Data Size	Number of Bytes
Kilobyte (KB)	1,000
Megabyte (MB)	1,000,000
Gigabyte (GB)	1,000,000,000
Terabyte (TB)	1,000,000,000,000
Petabyte (PB)	1,000,000,000,000,000
Exabyte (EB)	1,000,000,000,000,000,000
Zettabyte (ZB)	1,000,000,000,000,000,000,000
Yottabyte (YB)	1,000,000,000,000,000,000,000,000

Table 1.1 – Data Size Units

NOTE
The primary types of data analysis relevant to Big Data are covered in the *Fundamental Analysis, Analytics & Machine Learning Types* section. BI and KPIs are further explored in the *Business Intelligence & Big Data* section.

Exercise 1.1: Fill in the Blanks

1. A _____ is a group of related data in which each member of the group possesses the same set of attributes.

2. The goal of _____ is to support decision-making by establishing patterns and relationships in the data being analyzed.

3. _____ are focused on sifting through large amounts of _____ data to extract meaningful information that may help enrich existing enterprise data.

4. The process of business intelligence can apply _____ to large amounts of data.

Exercise answers are provided at the end of this booklet.

Notes

Notes / Sketches

Big Data Business & Technology Drivers

Big Data emerged from a combination of business needs and technology innovations. This section will examine the following primary business and technology drivers that led to Big Data becoming its own field:

- Analytics & Data Science
- Digitization
- Affordable Technology & Commodity Hardware
- Social Media
- Hyper-Connected Communities & Devices
- Cloud Computing

Analytics & Data Science

As growing enterprises are collecting and storing more data to potentially find new insights and gain a competitive edge, the need for techniques and technologies that can extract meaningful information and insights has increased. Machine learning algorithms, statistical techniques, and data warehousing have advanced data science and analytics to such a point where they have emerged as individual disciplines, with specific techniques and tools to perform unique and complex analyses. The maturity of these fields of practice inspired and enabled much of the core functionality expected from contemporary Big Data solutions and tools.

Digitization

For many businesses, digital mediums have replaced physical mediums as the de facto communications and delivery mechanism. Digitized data leads to an opportunity to collect further “secondary” data, such as when individuals carry out searches or complete surveys. Collecting secondary data can be important for businesses, as mining this data may allow for customized marketing, automated recommendations, and the development of optimized product features. Figure 1.8 provides a visual representation of examples of digitization.

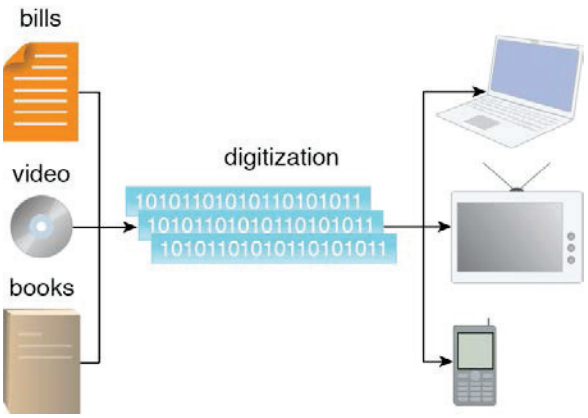


Figure 1.8 – Examples of digitization include online banking, on-demand television, and streaming video.

Affordable Technology & Commodity Hardware

Technology related to collecting and processing large quantities of diverse data has become increasingly affordable. Typical Big Data solutions are based on open-source software that requires little more than commodity hardware.

The use of commodity hardware makes the adoption of Big Data solutions accessible to businesses without large capital investments. Figure 1.9 provides an example of the cost savings associated with data storage prices.

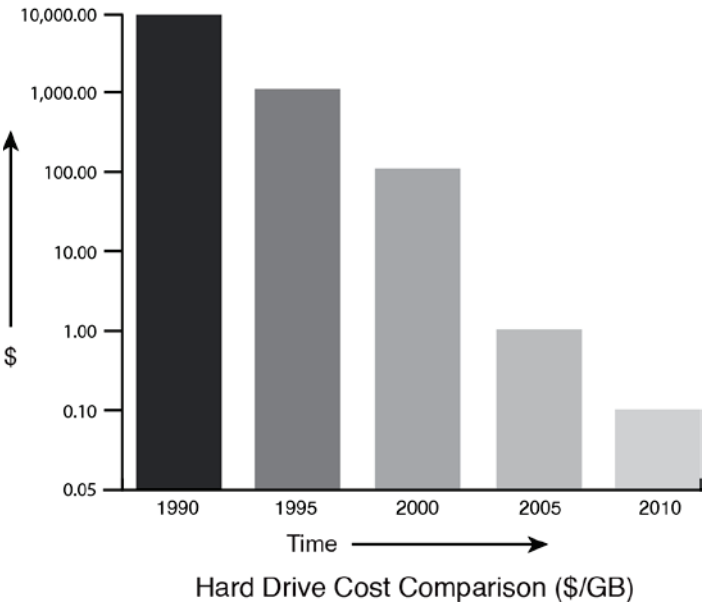


Figure 1.9 – Data storage prices have dropped dramatically from over $10,000 to less than $0.10 per GB over the decades.

Social Media

The emergence of social media has empowered customers to provide feedback in near-realtime via open and public mediums, a shift that has forced businesses to consider customer feedback on their offerings in their strategy planning. As a result, businesses are storing increasing amounts of data on customer interactions and from social media avenues in an attempt to harvest this data to increase sales, enable targeted marketing and create new products and services. Businesses are also increasingly interested in incorporating publicly available datasets from social media and other external data sources.

Hyper-Connected Communities & Devices

The broadening coverage of the Internet and the proliferation of cellular and Wi-Fi networks has enabled more people to be continuously active in virtual communities, either directly through online interaction or indirectly through the usage of connected devices. This has resulted in massive data streams. Some streams are public, while other streams go to vendors and businesses directly. The various types of hyper-connected communities and devices are shown in Figure 1.10.

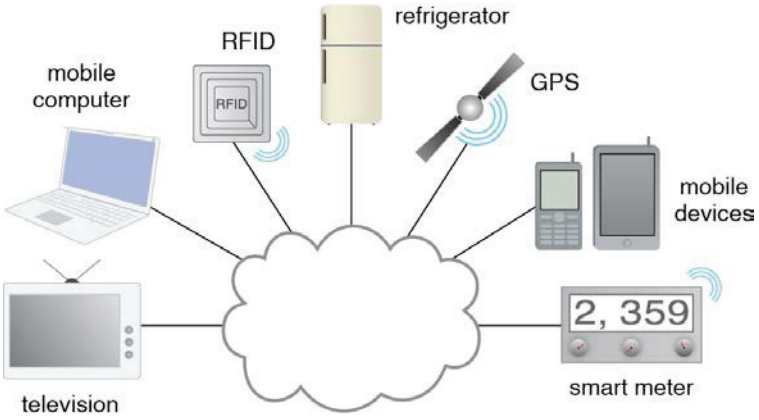


Figure 1.10 – Hyper-connected communities and devices include television, mobile computing, RFIDs, refrigerators, GPS devices, mobile devices, and smart meters.

Cloud Computing

Cloud computing technology advancements have led to the creation of remote environments referred to as “clouds.” These environments are capable of providing highly scalable, on-demand IT resources that can be leased via pay-as-you-go models. Businesses have the opportunity to leverage the infrastructure, storage, and processing capabilities provided by these environments in order to build large-scale Big Data solutions that can carry out large-scale processing tasks.

Figure 1.11 shows an example of how a cloud environment can be leveraged for its scaling capabilities to perform Big Data processing tasks. The fact that cloud-based IT resources can be leased dramatically reduces the required up-front investment of Big Data projects.

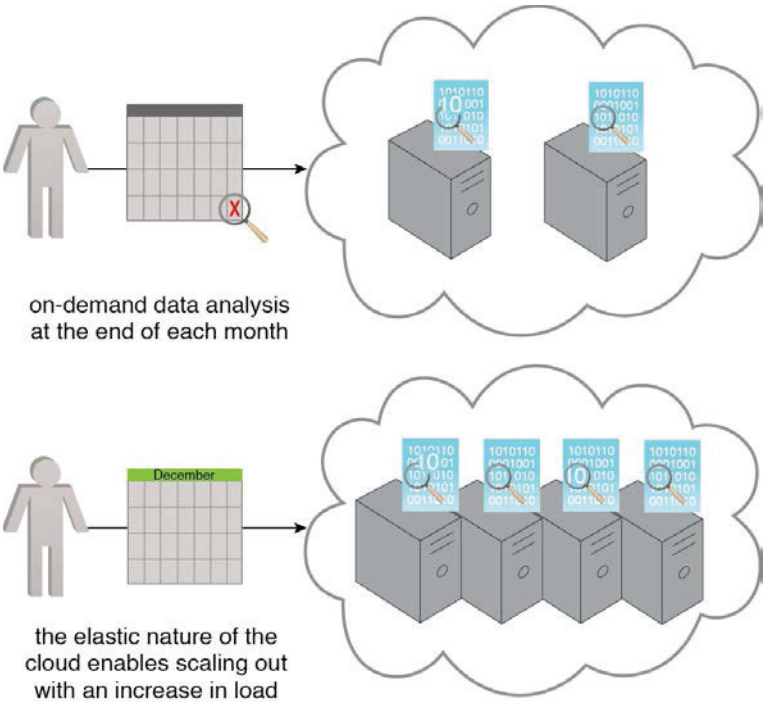


Figure 1.11 – The cloud can be used, for example, to complete on-demand data analysis at the end of each month or enable the scaling out of systems with an increase in load.

Notes

Notes / Sketches

Traditional Enterprise Technologies Related to Big Data

This section briefly describes the following technologies:

- Online Transaction Processing
- Online Analytical Processing
- Extract Transform Load
- Data Warehouses
- Data Marts
- Hadoop

Most of these technologies are well-established in the IT industry and pre-date the advent of Big Data. They are covered here because each technology is uniquely relevant to modern-day Big Data solutions and ecosystems.

NOTE
If you are already familiar with the listed technologies, feel free to skip ahead to the next section.

Online Transaction Processing (OLTP)

Online transaction processing (OLTP) is a software system that processes transaction-oriented data. The term “online transaction” refers to the completion of an activity in realtime and not batch-processed. OLTP systems store operational data that is fully normalized, and are important to Big Data in representing a common source of structured analytics input. Big Data analysis results can also be fed back into OLTPs.

The queries supported by OLTP systems are comprised of simple insert, delete, and update operations with sub-second response times. Examples include ticket reservation systems and banking and POS transactions.

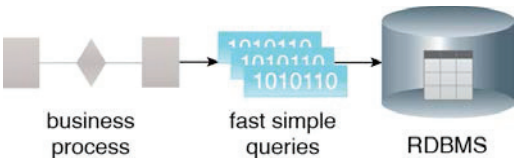


Figure 1.12 – The symbols used to represent OLTP.

Online Analytical Processing (OLAP)

Online analytical processing (OLAP) is a system used for processing data analysis queries. OLAPs form an integral part of business intelligence, data mining and machine learning processes. They are relevant to Big Data in that they can serve as both a data source as well as a data sink that is capable of receiving data. They are used in diagnostic, predictive, and prescriptive analytics, which are covered later in this course module.

OLAP systems store historical data that is aggregated and denormalized to support fast reporting capability. They further use databases that store historical data in multidimensional arrays and can answer complex queries based on multiple dimensions of the data.

Figure 1.13 – The symbols used to represent OLAP.

OLTPs & OLAPs

An OLAP system is always fed with data from multiple OLTP systems using regular batch processing jobs. Unlike OLTP systems, the response time of OLAP queries can take several minutes or even longer, depending on the complexity of the query and the number of records queried.

In Figure 1.14, relational data from two OLTP systems is periodically imported via bulk data import tasks into an OLAP system. Relational data is stored in the OLAP system as denormalized data in the form of cubes. This allows the data to be queried during any data analysis tasks that are performed later.

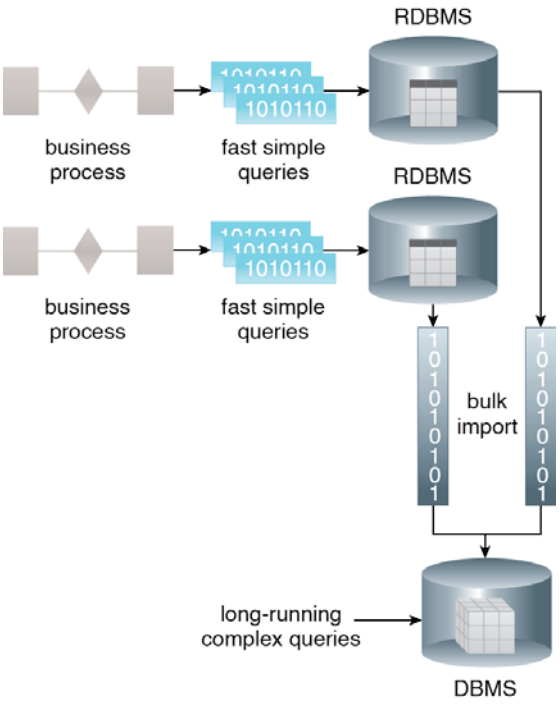


Figure 1.14 – The OLAP system stores relational data that is denormalized for future data analysis.

Extract Transform Load (ETL)

Extract-transform-load (ETL) is a process of loading data from a source system into a target system. The source system can be a database, a flat file, or an application. Similarly, the target system can be a database or some other information system.

ETL represents the main operation through which data warehouses are fed data. A Big Data solution encompasses the ETL feature-set for converting data of different types. Figure 1.15 shows that the required data is first obtained or extracted from the sources, after which the extracts are modified or transformed by applying rules. The data is finally inserted or loaded into the target system.

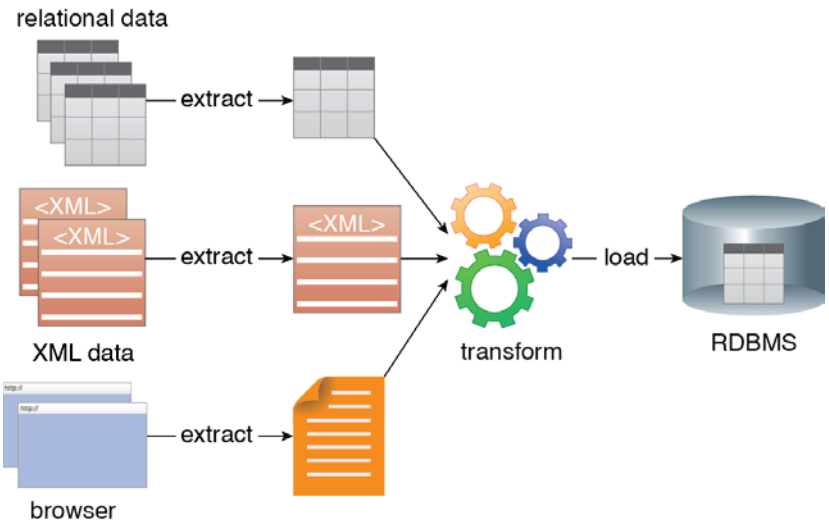


Figure 1.15 – The ETL process can extract browser data, XML data and relational data.

Data Warehouses

A data warehouse is a central, enterprise-wide repository consisting of historical and current data. Data warehouses are heavily used by BI to run various analytical queries, and usually interface with an OLAP system to support analytical queries, as shown in Figure 1.16.

Data pertaining to multiple business entities from different operational systems is periodically extracted, validated, transformed, and consolidated into a single database. With periodic data imports from across the enterprise, the amount of data contained in a given data warehouse will continue to increase. Query response times for data analysis tasks performed as part of BI can suffer as a result.

To resolve this shortcoming, data warehouses usually contain optimized databases, called analytical databases, to handle reporting and data analysis tasks. An analytical database can exist as a separate DBMS, as in the case of an OLAP database.

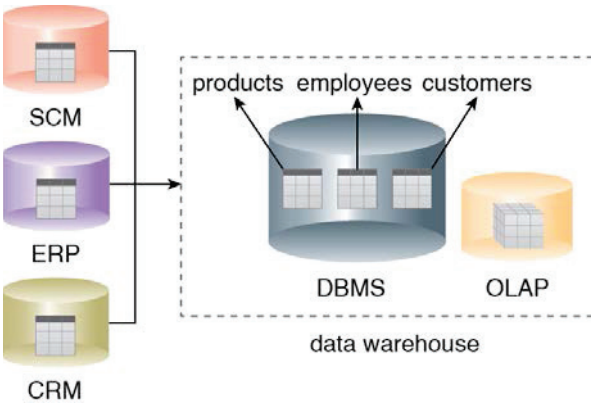


Figure 1.16 – A data warehouse periodically pulls data from other sources, such as OLTP, ERP, CRM, and SCM systems, for consolidation into a dataset.

Data Marts

A data mart is a subset of the data stored in a data warehouse that typically belongs to a department, division, or specific line of business. Data warehouses can have multiple data marts. As shown in Figure 1.17, enterprise-wide data is collected and business entities are then extracted. Domain-specific entities are persisted into the data warehouse via an ETL process.

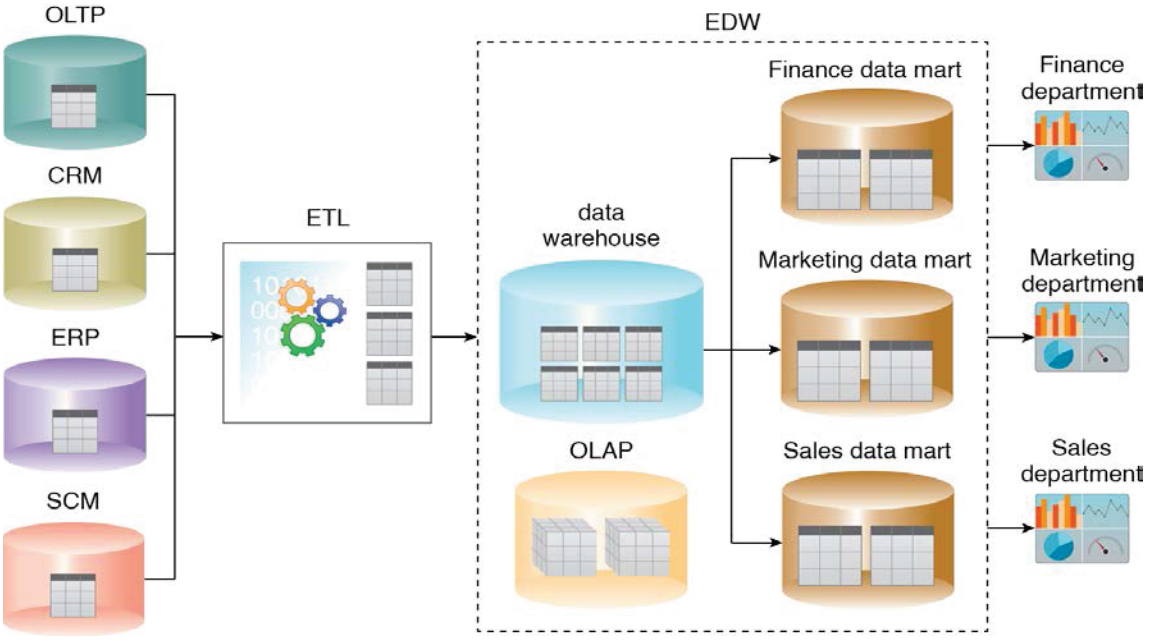


Figure 1.17 – A data warehouse's single version of "truth" is based on cleansed data, which is a prerequisite for accurate and error-free reports, as per the output shown on the right.

Hadoop

Hadoop is an open-source framework for large-scale data storage and data processing that is more or less run on commodity hardware. The Hadoop framework has established itself as a de facto industry platform for contemporary Big Data solutions. It can be used as an ETL engine or as an analytics engine for processing large amounts of structured, semi-structured, and unstructured data. Figure 1.18 illustrates some of Hadoop's features.

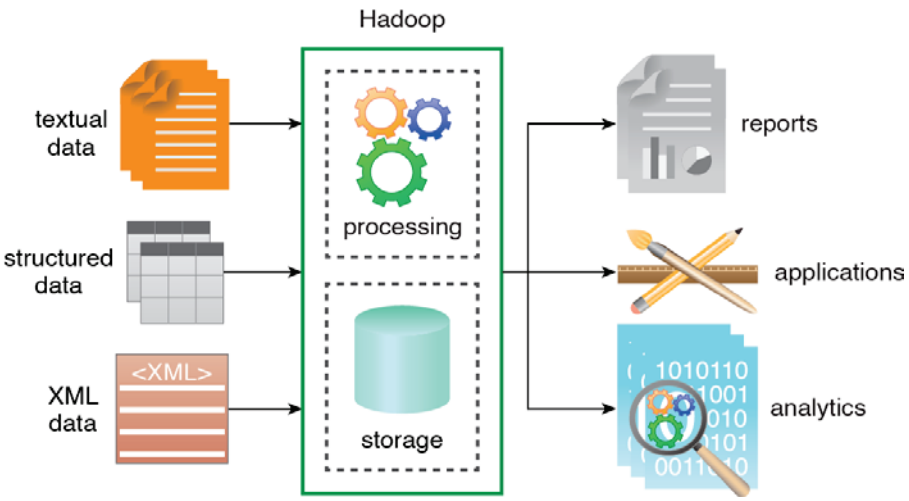


Figure 1.18 – The symbols used to represent Hadoop and its features.

NOTE
Module 2: Big Data Analysis & Technology Concepts provides more information regarding Hadoop features and related mechanisms.

Optional Reading

The *Big Data Analytics* book that is included with *Module 2: Big Data Analysis & Technology Concepts* discusses the relationship between Big Data and data warehouses in the *Big Data Analytics* section of *Chapter 3: The Rise of Big Data Options*.

Notes

Notes / Sketches

Characteristics of Data in Big Data Environments

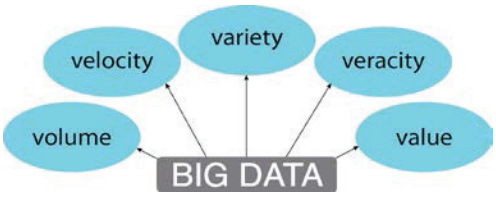
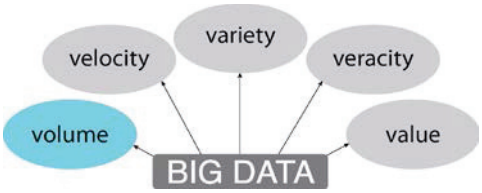


Figure 1.19 – The Five Vs of Big Data.

This section explores the five Big Data characteristics that can be used to help differentiate data categorized as "Big" data from other forms of data. The five Big Data traits are commonly referred to as the Five Vs:

- Volume
- Velocity
- Variety
- Veracity
- Value

Volume



The anticipated volume of data that is processed by Big Data solutions is substantial and usually ever-growing. High data volumes impose distinct data storage and processing demands, as well as management and access processes. Figure 1.20 provides a visual representation of the large volume of data used by organizations and users world-wide.

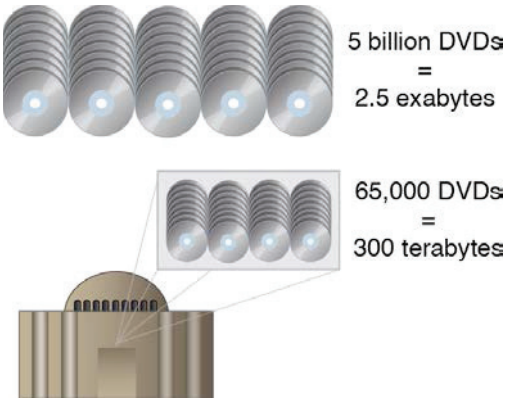
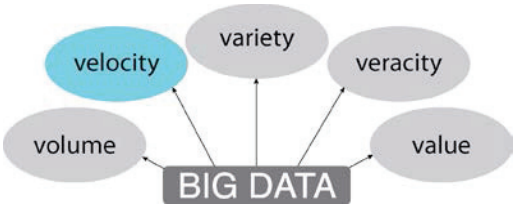


Figure 1.20 – Organizations and users world-wide create 2.5 EBs of data a day, while the Library of Congress currently holds over 300 TBs.

Typical data sources that are responsible for generating high data volumes can include:

- online transactions (point-of-sale, banking)
- scientific and research data (Large Hadron Collider, Atacama Large Millimeter/Submillimeter Array telescope)
- sensor data (RFIDs, Smart meters, GPS sensors)
- social media (Facebook, Twitter)

Velocity



Big Data arrives at such fast speeds that enormous datasets can accumulate within very short periods of time. From an enterprise's point of view, the velocity of data translates into the amount of time it takes for the data to be processed once it enters the enterprise's perimeter. Coping with the fast inflow of data requires the enterprise to design highly elastic and available processing solutions and corresponding data storage capabilities.

Depending on the data source, velocity may not always be high. For example, MRI scan images are usually not generated as frequently as log entries from a high-traffic Web server. As illustrated in Figure 1.21, potential data velocity is put into perspective when considering that the following data is currently generated every minute: 100,000 tweets, 48 hours of video footage, 171 million e-mails, and 330 GBs generated by the average jet engine.

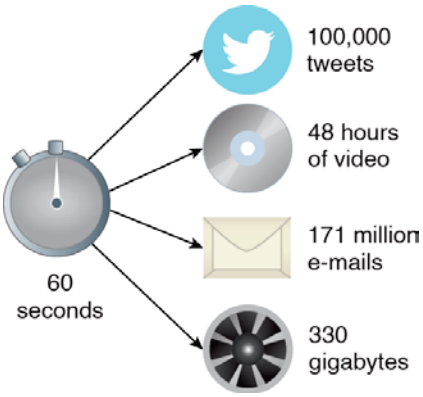
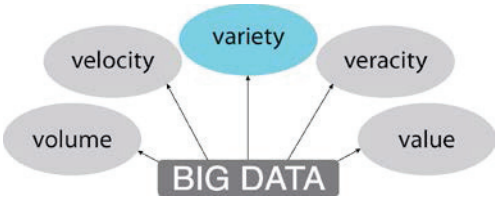


Figure 1.21 – Examples of high-velocity Big Data datasets produced every minute include tweets, video, e-mails, and GBs generated by the average jet engine.

Variety



Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions, such as structured, semi-structured, and unstructured data, which are further described in the upcoming *Types of Data in Big Data Environments* section. Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage. Figure 1.22 provides a visual representation of data variety, which includes structured data in the form of financial transactions, semi-structured data in the form of e-mails, and unstructured data in the form of images.

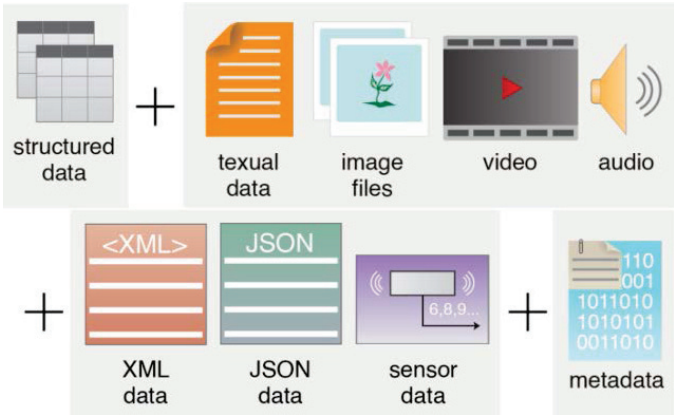
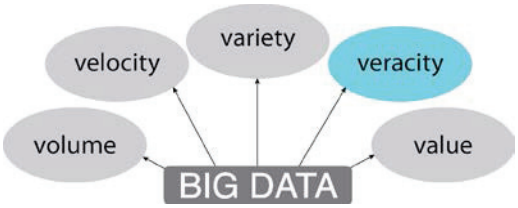


Figure 1.22 – Examples of high-variety Big Data datasets include structured, textual, image, video, audio, XML, JSON, sensor data, or metadata.

Veracity

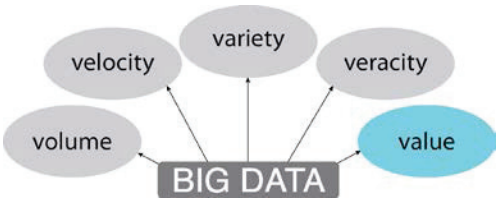


Veracity refers to the quality or fidelity of data. Data that exists in Big Data environments can be meaningful or it can just add clutter. Data assessed in relation to veracity are either:

- Noise – data carrying no value
- Signal – data bearing value leading to meaningful information

Data acquired in a controlled manner, for example via online customer registrations, usually contains less noise than data acquired via uncontrolled sources, such as blog postings. The degree of noise (bad data) or the noise-to-signal ratio varies depending on the type of data present.

Value



Value is defined as the usefulness of data for an enterprise. The value characteristic is directly related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business. Value is also dependent on how long data processing takes, as value and time are inversely proportional to each other. The longer it takes for data to be turned into meaningful information, the less value it may have for the business, because it inhibits the speed at which it can make informed decisions. Figures 1.23 and 1.24 provide a comparison of the potential value data may have against the time it takes to analyze data.

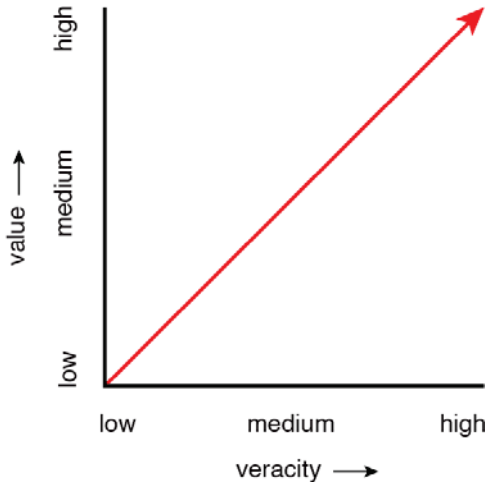


Figure 1.23 – The more confidence instilled in the data, the more potential value it has for the business.

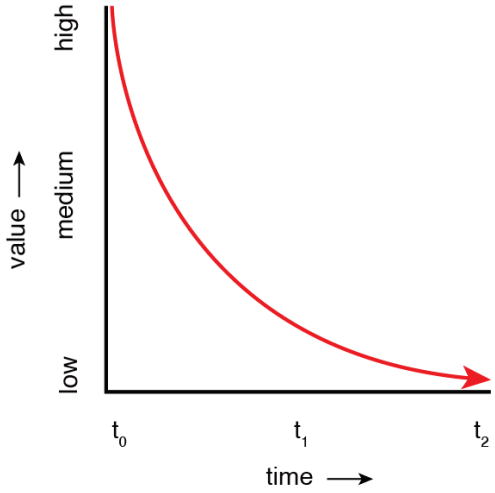


Figure 1.24 – The longer it takes to analyze the data, the less potential value it has for the business.

Apart from veracity and time, value is also determined by the following considerations:

- How well has the data been stored?
- Has the data been stripped of any valuable attributes?
- Are the right types of questions being asked during data analysis?
- Are the data analysis results being accurately communicated to the appropriate decision-makers?

Notes

Notes / Sketches

Types of Data in Big Data Environments

This section examines the types of data that are processed by Big Data solutions. These types of data can be divided into the following primary categories:

- structured data
- unstructured data
- semi-structured data

These data types refer to the internal organization of data and can also be referred to as data formats. Although technically not a data type, rather another form of data that itself can vary in structure, metadata is briefly described at the end of this section.

Structured Data

Structured data:

- conforms to a data model or schema
- is stored in a tabular form
- can be relational

Structured data is typically stored in relational databases and frequently generated by custom enterprise applications, enterprise resource planning (ERP) systems and customer relationship management (CRM) systems. It does not generally have any special pre-processing or storage requirements. Examples include banking transactions, OLTP system records, and customer records.

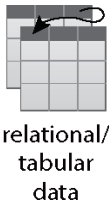


Figure 1.25 - The symbol used to represent relational structured data stored in a tabular form.

Unstructured Data

Unstructured data:

- does not conform to a data model or data schema
- is generally inconsistent and non-relational

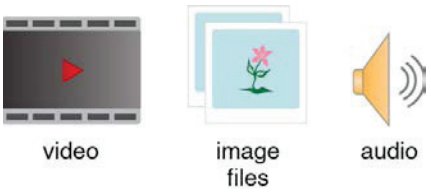


Figure 1.26 - The symbols used to represent video, image files, and audio.

Unstructured data either exists in textual or binary form. Examples include image, audio, and video files. Technically, both text and binary files have a structure defined by the file format itself. This is being disregarded to focus on the format of the data contained in the file only. Unstructured data generally makes up 80% of the data within an enterprise, and has a faster growth rate than structured data. Figure 1.27 provides a pie chart of the general proportions of unstructured and structured data within an enterprise.

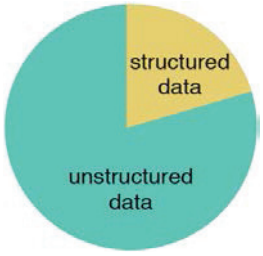


Figure 1.27 – Within an enterprise, unstructured data generally makes up 80% of the data, while structured data makes up the remaining 20% of the data

Unlike structured data, unstructured data generally requires special or customized logic when it comes to pre-processing and storage. It cannot be inherently processed or queried using SQL or traditional programming features, and is usually an awkward fit with relational databases. A Not-only SQL (NoSQL) database is a non-relational database that can be used to store unstructured data alongside structured data.

NOTE
NoSQL is covered further in *Module 2: Big Data Analysis & Technology Concepts.*

Semi-Structured Data

Semi-structured data has a defined level of structure and consistency, but cannot be relational in nature. It mostly exists in textual formats, such as XML or JSON files, and can generally be more easily processed than unstructured data.

Examples of common sources of semi-structured data include electronic data interchanges (EDI), e-mails, spreadsheets, RSS feeds, and sensor data. Semi-structured data often has

special pre-processing and storage requirements, especially if the underlying format is not text-based.

Figure 1.28 - The symbols used to represent XML, JSON, and sensor data.

Metadata

Metadata provides information about a dataset's characteristics and structure. This type of data is mostly machine-generated and automatically appended to the data. It is crucial to Big Data processing, storage, and analysis. Examples of metadata include:

- XML tags providing the author and creation date of a document
- attributes providing the file size and resolution of a digital photograph

Big Data solutions rely on metadata, particularly when processing semi-structured and unstructured data.

Figure 1.29 - The symbol used to represent metadata.

Data Types & Veracity

Semi-structured data and unstructured data have a greater noise-to-signal ratio than structured data. This larger amount of noise requires automated data cleansing and data verification when carrying out ETL processes. Figure 1.30 illustrates the noise-to-signal ratio of the different data types.

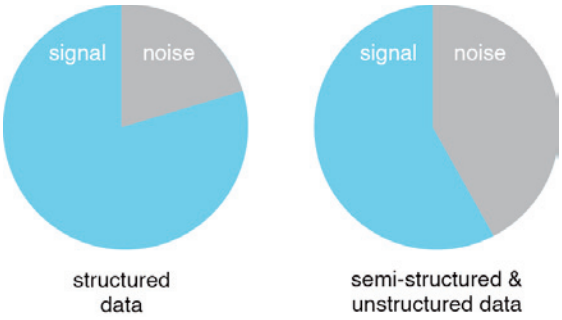


Figure 1.30 – A summary of the noise-to-signal ratio of structured, semi-structured, and unstructured data.

Exercise 1.2: Fill in the Blanks

1. A _____ can contain analytical databases that can improve query response times.

2. _____ is used to load data from a source system into a target system, and is the main operation through which data warehouses are fed data.

3. Volume, velocity, _____, _____ and _____ are the five primary Big Data characteristics that differentiate it from traditional data.

4. The value characteristic of Big Data is _____ on how long data processing takes.

5. In general, the data processed by Big Data solutions can be found in the following data types or formats: _____, _____, _____ and _____.

Exercise answers are provided at the end of this booklet.

Notes

Notes / Sketches

Fundamental Analysis, Analytics & Machine Learning Types

Types of Data Analysis

The *Fundamental Terminology & Concepts* section introduced the term “data analysis” and provided a simple example. The upcoming sections further describe the following basic types of data analysis:

- quantitative analysis
- qualitative analysis
- data mining

Each description includes a simple example based on the ice cream cone sales scenario used in the initial data analysis description.

Figure 1.31 - The symbol used to represent data analysis.

Quantitative Analysis

Quantitative analysis is a data analysis technique that focuses on quantifying the patterns and correlations found in the data. Based on statistical practices, this technique involves analyzing a large number of observations from a dataset. Since the sample size is large, the results can be applied in a generalized manner to the entire dataset.

Quantitative analysis results are absolute in nature and can therefore be used for numerical comparisons. For example, a quantitative analysis of ice cream sales may discover that a 5 degree increase in temperature increases ice cream sales by 15%.

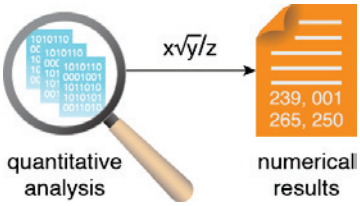


Figure 1.32 - The symbols used to represent quantitative analysis and numerical results.

Qualitative Analysis

Qualitative analysis is a data analysis technique that focuses on describing various data qualities using words. It involves analyzing a smaller sample in greater depth compared to quantitative data analysis.

These analysis results cannot be generalized to an entire dataset due to the small sample size. They also cannot be measured numerically or used for numerical comparisons. For example, an analysis of ice cream cone sales may reveal that May's sales figures were not as high as June's. The analysis results state only that the figures were "not as high as," and do not provide a numerical difference.

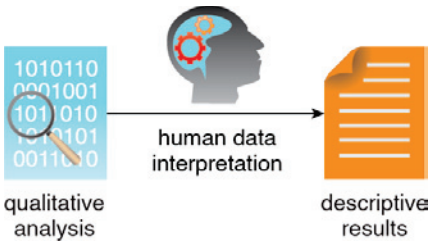


Figure 1.33 - The symbols used to represent qualitative analysis, human data interpretation, and descriptive results.

Data Mining

Data mining, also known as data discovery, is a specialized form of data analysis that targets large datasets. In relation to Big Data analysis, data mining generally refers to automated, software-based techniques that sift through massive datasets to identify patterns and trends. Specifically, it involves extracting hidden or unknown patterns in the data with the intention of identifying previously unknown patterns. Data mining forms the basis for predictive analytics and business intelligence (BI).

Figure 1.34 - The symbol used to represent data mining.

Analysis & Analytics

The time and effort required to carry out manual analysis is greatly magnified with Big Data. These techniques may not provide accurate findings in a timely manner because of the data's volume, velocity, and/or variety. These potential inefficiencies are further exacerbated if analysis needs to be repeated.

Analytics tools can automate data analyses through the use of highly scalable computational technologies that apply automated statistical quantitative analysis, data mining, and machine learning techniques, as shown in Figure 1.35.

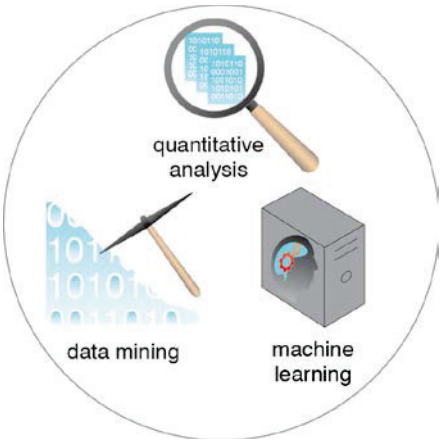


Figure 1.35 – Automating quantitative analysis, data mining, and machine learning can minimize the inefficiencies in repeat analysis.

Types of Analytics

The term "analytics" was initially described in the *Fundamental Terminology & Concepts* section. This section further explores analytics by describing the following four common analytics types:

- descriptive analytics
- diagnostic analytics
- predictive analytics
- prescriptive analytics

Figure 1.36 illustrates these types of analytics by value and complexity.

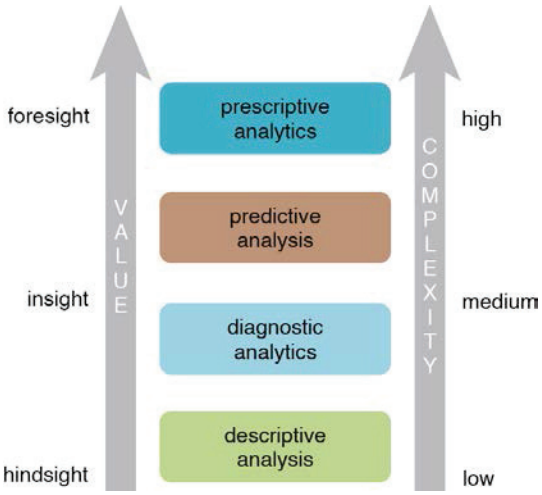


Figure 1.36 – Value and complexity increase from descriptive to prescriptive analytics.

Descriptive Analytics

Descriptive analytics are carried out to answer questions about events that have already occurred.

Sample questions can include:

- *What is the sales data over the past 12 months?*
- *What is the number of support calls received as categorized by severity and geographic location?*
- *What is the monthly commission earned by each sales agent?*

Around 80% of analytics are descriptive in nature. Value-wise, descriptive analytics provides the least value and requires a relatively basic skillset.

Descriptive analytics are often carried out via ad-hoc reporting or dashboards, as shown in Figure 1.37. The reports are generally static in nature and display historical data that is presented in the form of data grids or charts. Queries are executed on the OLTP systems or data obtained from a variety of other information systems, such as CRMs and ERPs.

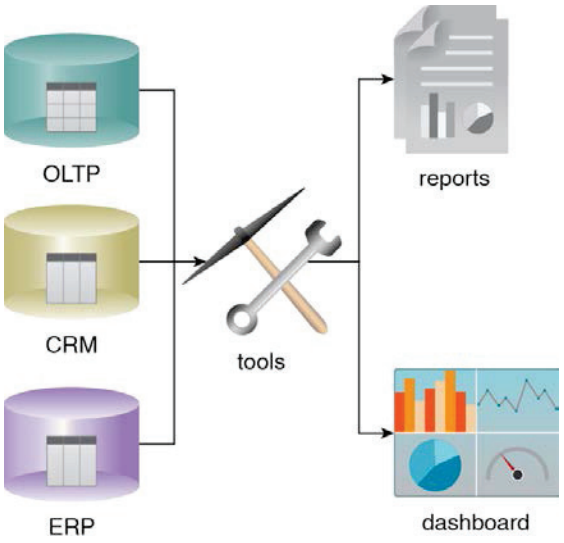


Figure 1.37 – Queries from the systems, pictured left, can be obtained via descriptive analytics tools that are communicated via reports or dashboards, pictured right.

Diagnostic Analytics

Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past, using questions that focus on the reason behind the event.

Sample questions can include:

- *Why were Q2 sales less than Q1 sales?*

- *Why have there been more support calls originating from the Eastern region than from the Western region?*
- *Why was there an increase in patient re-admission rates over the past three months?*

Diagnostic analytics are considered to provide more value than descriptive analytics, requiring a more advanced skillset, and usually require collecting data from multiple sources and storing it in a structure that lends itself to performing drill-downs and roll-ups, as shown in Figure 1.38. Analytics results are viewed via interactive visualization tools that enable users to identify trends and patterns. The executed queries are more complex compared to descriptive analytics, and are performed on multi-dimensional data held in OLAP systems.

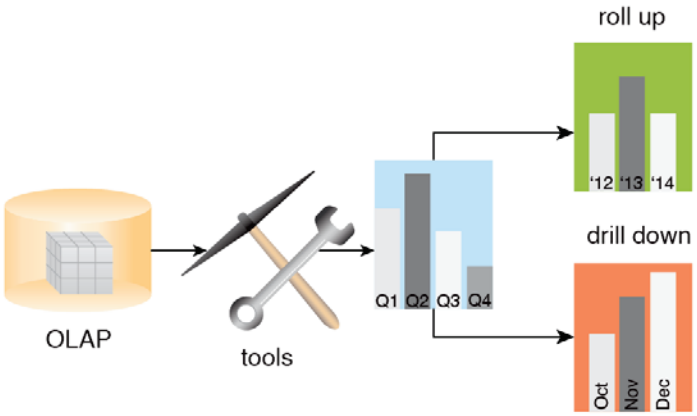


Figure 1.38 – Diagnostic analytics can result in data that is suitable for performing drill-downs and roll-ups.

Predictive Analytics

Predictive analytics are carried out in attempt to determine the outcome of an event that might occur in the future.

Questions are usually formulated using a what-if rationale, such as the following:

- *What are the chances that a customer will default on a loan if he has missed a monthly payment?*
- *What will be the patient survival rate if Drug B is administered instead of Drug A?*
- *If a customer has purchased Products A and B, what are the chances that she will also purchase Product C?*

Predictive analytics try to predict the event outcome, and predictions are made based on patterns, trends, and exceptions found in historical and current data. This can lead to the identification of risks and opportunities.

Predictive analytics involve the use of large datasets comprised of both internal and external data, statistical techniques, quantitative analysis, machine learning, and data mining techniques. This type of analytics is considered to provide more value and requires

a more advanced skillset than both descriptive and diagnostic analytics. The tools used generally abstract underlying statistical intricacies by providing user-friendly front-end interfaces, as shown in Figure 1.39.

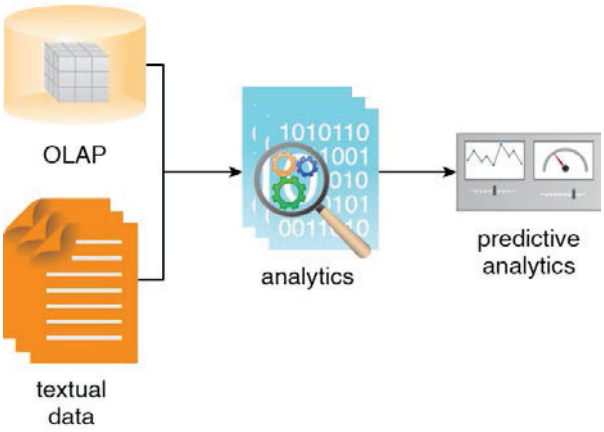


Figure 1.39 – Predictive analytics tools can provide user-friendly front-end interfaces.

Prescriptive Analytics

Prescriptive analytics build upon the results of predictive analytics by prescribing actions that should be taken. The focus is on which prescribed option to follow and why and when it should be followed in order to gain an advantage or mitigate a risk.

Sample questions can include:

- *With a choice of three drugs, which one provides the best results?*
- *When is the best time to trade a particular stock?*

Prescriptive analytics provide more value than any other type of analytics and correspondingly require the most advanced skillset, as well as specialized software and tools. Various outcomes are calculated, and the best course of action for each outcome is suggested. The approach shifts from explanatory to advisory and can include the simulation of various scenarios.

Prescriptive analytics incorporate internal data, which includes current and historical sales data, customer information, product data, and business rules, as well as external data, which includes social media data, weather data, and demographic data. Prescriptive analytics involve the use of business rules and large amounts of internal and/or external data to simulate outcomes and prescribe the best course of action, as shown in Figure 1.40.

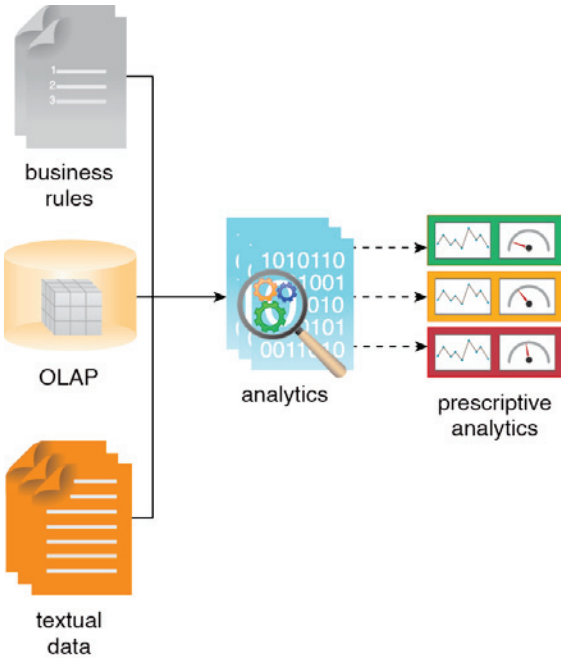


Figure 1.40 – Prescriptive analytics involve the use of business rules and internal or external data in order to perform an in-depth analysis.

Machine Learning

Machine learning is the process of teaching computers to learn from existing data and apply the acquired knowledge to formulate predictions about unknown data. This involves identifying patterns in the training data and classifying new or unseen data based on known patterns. Machine learning algorithms also have the ability to adjust behavior using a feedback loop as they work with new datasets. These algorithms can generally be grouped into the following two types:

- supervised learning
- unsupervised learning

Figure 1.41 – The symbol used to represent machine learning.

Machine Learning Types

A supervised learning algorithm is first fed sample data where the data categories are already known. Based on the input data, the algorithm develops an understanding of

which data belongs to which category. Having developed an understanding, the algorithm can then apply the learned behavior to categorize unknown data.

With an unsupervised learning algorithm, data categories are unknown and no sample data is fed. Instead, the algorithm attempts to categorize data by grouping data with similar attributes together.

Machine Learning vs. Data Mining

Although data mining and machine learning are closely related, they have notable differences. Whereas data mining unearths hidden patterns and relationships based on previously unknown attributes of data, machine learning makes predictions by categorizing data based on known patterns.

Data mining may employ machine learning algorithms, such as unsupervised learning, to extract previously unknown attributes. This is accomplished by categorizing data, which leads to the identification of patterns. Machine learning can use the output from data mining or identified patterns for further data classification through supervised learning, as shown in Figure 1.42.

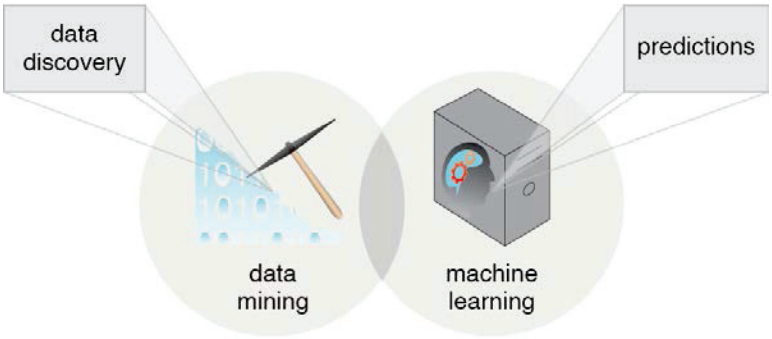


Figure 1.42 – Machine learning can make predictions for data classification, where data mining discovers the patterns used to automate the machine learning.

Exercise 1.3: Fill in the Blanks

1. Analytics can be categorized into four types according to their value attribute:

   _____, _____,

   _____ and _____.

2. _____ analytics results are viewed via interactive visualization tools that enable trends and patterns to be easily spotted.

3. _____ analytics are most valuable for enterprises because this technique offers a suggested course of action that can be taken.

4. Machine learning algorithms can be categorized into _____ and _____ learning.

Exercise answers are provided at the end of this booklet.

Notes

Notes / Sketches

Business Intelligence & Big Data

Contemporary enterprise Big Data solutions rely on BI and data warehouses as core components of Big Data environments and ecosystems. Conversely, the advent of Big Data has advanced BI and data warehouse technologies and practices to a point where a new generation of these platforms has emerged.

This section compares traditional and next-generation BI and data warehouse environments, and further defines their relationships to Big Data solutions.

Traditional BI

Traditional BI utilizes descriptive and diagnostic analytics to provide information on historical and current events. It is not "intelligent" because it only provides answers to correctly formulated questions. Correctly formulating questions requires an understanding of business problems and issues and of the data itself. BI reports on different KPIs through:

- ad-hoc reports
- dashboards

Traditional BI: Ad-Hoc Reporting

Ad-hoc reporting is a process that involves manually processing data to produce custom-made reports, as shown in Figure 1.43. The focus of an ad-hoc report is usually on a specific area of the business, such as its marketing or supply chain management. The generated custom reports are detailed and often tabular in nature.

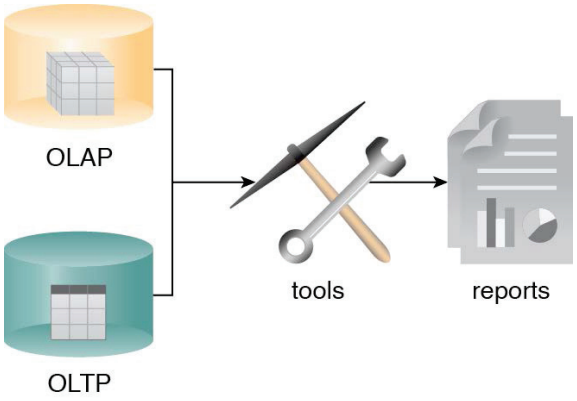


Figure 1.43 – OLAP and OLTP data sources can be used by BI tools for both ad-hoc reporting and dashboards.

Traditional BI: Dashboards

Dashboards provide a holistic view of key business areas. The information displayed on dashboards is generated at periodic intervals in realtime or near-realtime. The presentation of data on dashboards is graphical in nature, using column charts, pie charts, and gauges, as shown in Figure 1.44.

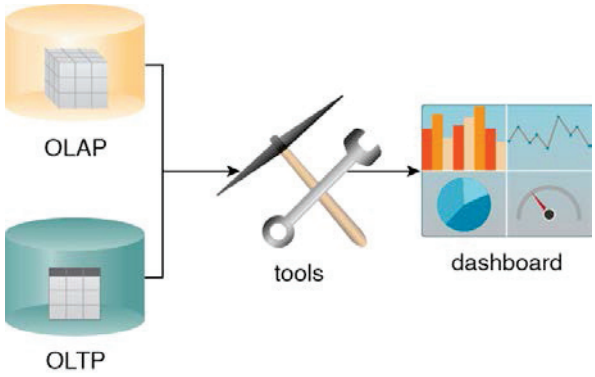


Figure 1.44 – BI tools use both OLAP and OLTP to display the information on dashboards.

As previously explained, data warehouses and data marts contain consolidated and validated information about enterprise-wide business entities. Traditional BI cannot function effectively without data marts because they contain the optimized and segregated data that BI requires for reporting purposes. Without data marts, data needs to be extracted from the data warehouse via an ETL process on an ad-hoc basis whenever a query needs to be run. This increases the time and effort to execute queries and generate reports.

Traditional BI uses data warehouses and data marts for reporting and data analysis because they allow complex data analysis queries with multiple joins and aggregations to be issued, as shown in Figure 1.45.

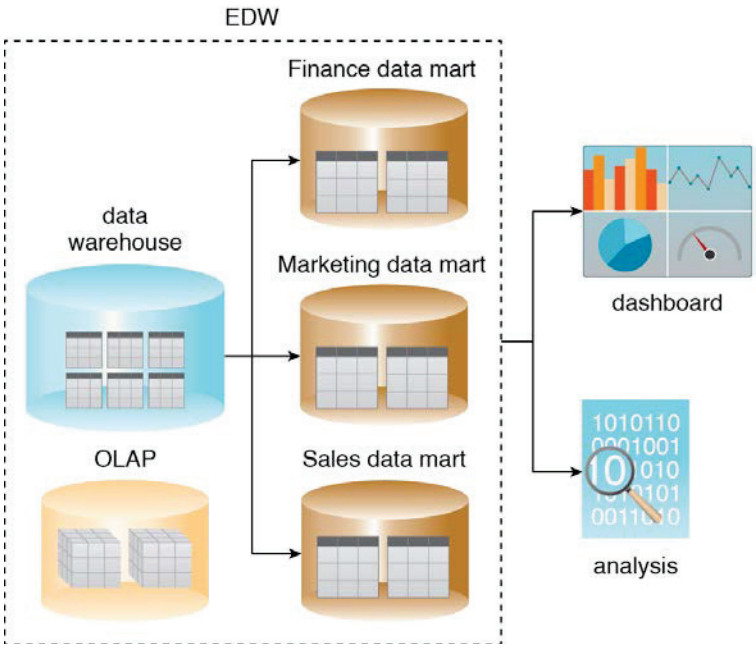


Figure 1.45 – An example of traditional BI.

Big Data BI

Big Data BI builds upon traditional BI by acting on the cleansed, consolidated enterprise-wide data in the data warehouse and combining it with semi-structured and unstructured data sources. It comprises both predictive and prescriptive analytics to facilitate the development of an enterprise-wide understanding of the way a business works.

While traditional BI analyses generally focus on individual business processes, Big Data BI analyses focus on multiple business processes simultaneously. This helps reveal patterns and anomalies across a broader scope within the enterprise. It also leads to data discovery by identifying insights and information that may have been previously absent or unknown.

Big Data BI requires the analysis of unstructured, semi-structured and structured data residing in the enterprise data warehouse. This requires a "next-generation" data warehouse that uses new features and technologies to store cleansed data originating from a variety of sources in a single uniform data format. Coupling a traditional data warehouse with these new technologies results in a hybrid data warehouse which acts as a uniform and central repository of structured, semi-structured, and unstructured data that can provide Big Data BI tools with all of the required data. This eliminates the need for Big Data BI tools to have to connect to multiple data sources to retrieve or access data. In Figure 1.46, a next-generation data warehouse establishes a standardized data access layer across a range of data sources.

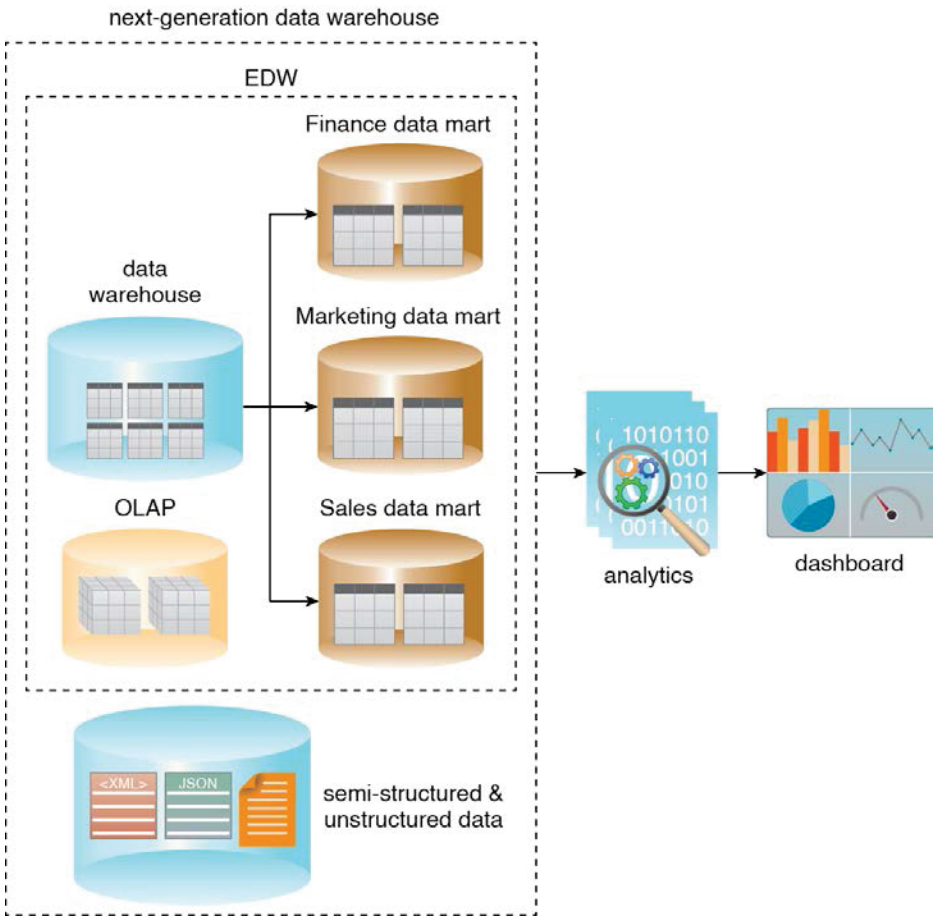


Figure 1.46 – A next-generation data warehouse.

Notes

Notes / Sketches

Data Visualization & Big Data

This section provides a closer look at data visualization, particularly in relation to Big Data.

Data Visualization

Data visualization is a technique whereby analytical results are graphically communicated using elements like charts, maps, data grids, infographics, and alerts. Graphically representing data can make it easier to understand reports, view trends, and identify patterns.

Traditional data visualization provides mostly static charts and graphs in reports and dashboards, whereas contemporary data visualization tools are interactive and can provide both summarized and detailed views of data. They are designed to help people who lack statistical and/or mathematical skills to better understand analytical results without having to resort to spreadsheets.

Data Visualization Tools

Traditional data visualization tools query data from relational databases, OLAP systems, data warehouses, and spreadsheets to present both descriptive and diagnostic analytics results. Big Data solutions require data visualization tools that can seamlessly connect to structured, semi-structured, and unstructured data sources, and are further capable of handling millions of data records. Data visualization tools for Big Data solutions generally use in-memory analytical technologies that reduce the latency normally attributed to traditional, disk-based data visualization tools.

Data Visualization Features

Common features of visualization tools used in Big Data:

- Aggregation – provides a holistic and summarized view of data across multiple contexts
- Drill-Down – enables a detailed view of the data of interest by focusing in on a data subset from the summarized view
- Filtering – helps focus on a particular set of data by filtering away the data that is not of immediate interest
- Roll-Up – groups data across multiple categories to show subtotals and totals
- What-If Analysis – enables multiple outcomes to be visualized by enabling related factors to be dynamically changed

Advanced Visualization Tools

Advanced data visualization tools for Big Data solutions incorporate predictive and prescriptive data analytics and data transformation features. These tools eliminate the need for data pre-processing methods, such as ETL, and provide the ability to directly connect to structured, semi-structured and unstructured data sources. As part of Big Data solutions, advanced data visualization tools can join structured and unstructured data that is kept in memory for fast data access. Queries and statistical formulae can then be applied as part of various data analysis tasks for viewing data in a user-friendly format, such as on a dashboard.

Exercise 1.4: Fill in the Blanks

1. Traditional BI uses _____ and

   _____ analytics.

2. _____ communicates analytical results using a variety of

   graphic, interactive tools.

3. Traditional data visualization tools present both _____ and

   _____ data analytics results.

4. Big Data BI adds value to traditional BI by using _____ and

   _____ analytics.

5. Advanced Big Data visualization tools use _____ and

   _____ data analytics tools.

Exercise answers are provided at the end of this booklet.

Notes

Notes / Sketches

Big Data Adoption & Planning Considerations

Business Justification

As Big Data initiatives are inherently business-driven, there needs to be a clear business case for adopting a Big Data solution to ensure that it is justified and that expectations are met. Clear goals regarding the measurable business value of an enterprise's Big Data solution need to be set. Anticipated benefits need to be weighed against risks and investments.

For example, a goal can be to build a 360-degree view of a company's customer base. This goal may require all in-house customer data to be consolidated from numerous systems. Risks associated with collecting accurate and relevant data, and with integrating the Big Data environment itself, need to be identified and quantified. It is important to accept that Big Data solutions are not necessary for all businesses. For example, some companies may simply not generate enough data to warrant a Big Data environment.

Organizational Prerequisites

Big Data frameworks are not turn-key solutions. In order for data analysis and analytics to be successful and offer value, enterprises need to have data management and Big Data governance frameworks. Sound processes and sufficient skillsets for those who will be responsible for implementing, customizing, populating, and using Big Data solutions are also necessary. Additionally, the quality of the data targeted for processing by Big Data solutions needs to be assessed.

Outdated, invalid, or poorly identified data will result in low-quality input which, regardless of how good the Big Data solution is, will continue to produce low-quality output. The longevity of the Big Data environment also needs to be planned for. A roadmap needs to be defined to ensure that any necessary expansion or augmentation of the environment is planned out to stay in sync with the requirements of the enterprise.

Data Procurement

The acquisition of Big Data solutions themselves can be economical, due to open-source platform availability and opportunities to leverage commodity hardware. However, a substantial budget may still be required to obtain external data. The nature of the business may make external data very valuable. The greater the volume and variety of data, the higher the chances are of finding hidden insights from patterns.

External data sources include data markets and the government. Government-provided data, like geo-spatial data, may be free. However, most commercially relevant data will need to be purchased. Such an investment may be on-going in order to obtain updated versions of the datasets.

Privacy

Performing analytics on datasets can reveal confidential information about organizations or individuals. Even analyzing separate datasets that contain seemingly benign data can reveal private information when the datasets are analyzed jointly. This can lead to intentional or inadvertent breaches of privacy.

Addressing these privacy concerns requires an understanding of the nature of data being accumulated and relevant data privacy regulations, as well as special techniques for data tagging and anonymization. For example, telemetry data, such as a car's GPS log or smart meter data readings, collected over an extended period of time can reveal an individual's location and behavior, as shown in Figure 1.47.

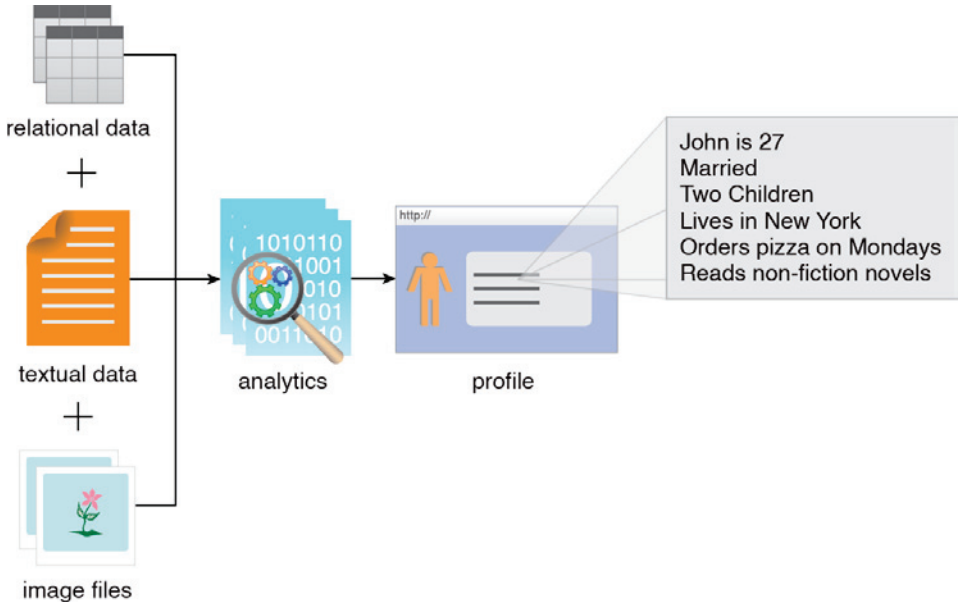


Figure 1.47 – Information gathered from running analytics on image files and relational and textual data is used to create John's profile.

Security

Some of the components of Big Data solutions lack the robustness of traditional enterprise solution environments when it comes to access control and data security. Securing Big Data involves ensuring that data networks provide access to repositories that are sufficiently secured, via custom authentication and authorization mechanisms.

Big Data security further involves establishing data access levels for different categories of users. For example, unlike traditional relational database management systems, NoSQL databases generally do not provide robust built-in security mechanisms. They instead rely on simple HTTP-based APIs where data is exchanged in plaintext, making the data prone to network-based attacks, as shown in Figure 1.48.

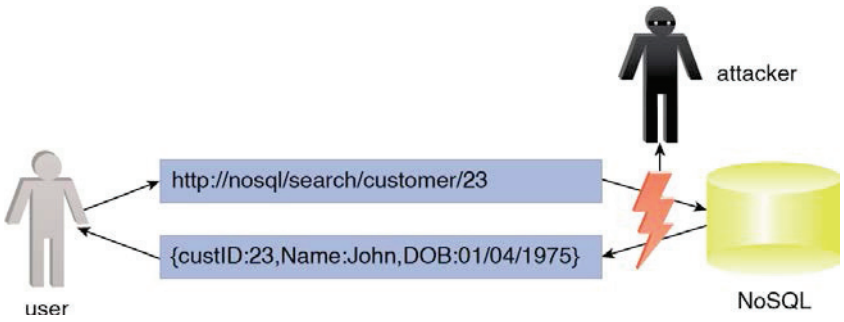


Figure 1.48 – Traditional databases can be susceptible to network-based attacks.

Provenance

Provenance refers to information about the source of the data that helps determine its authenticity and quality. It is also used for auditing purposes. Maintaining provenance as large volumes of data are acquired, combined, and put through multiple processing stages can be a complex task. Addressing provenance concerns can require the annotation of data with source information and other metadata, when it is generated or as it arrives, as shown in Figure 1.49.

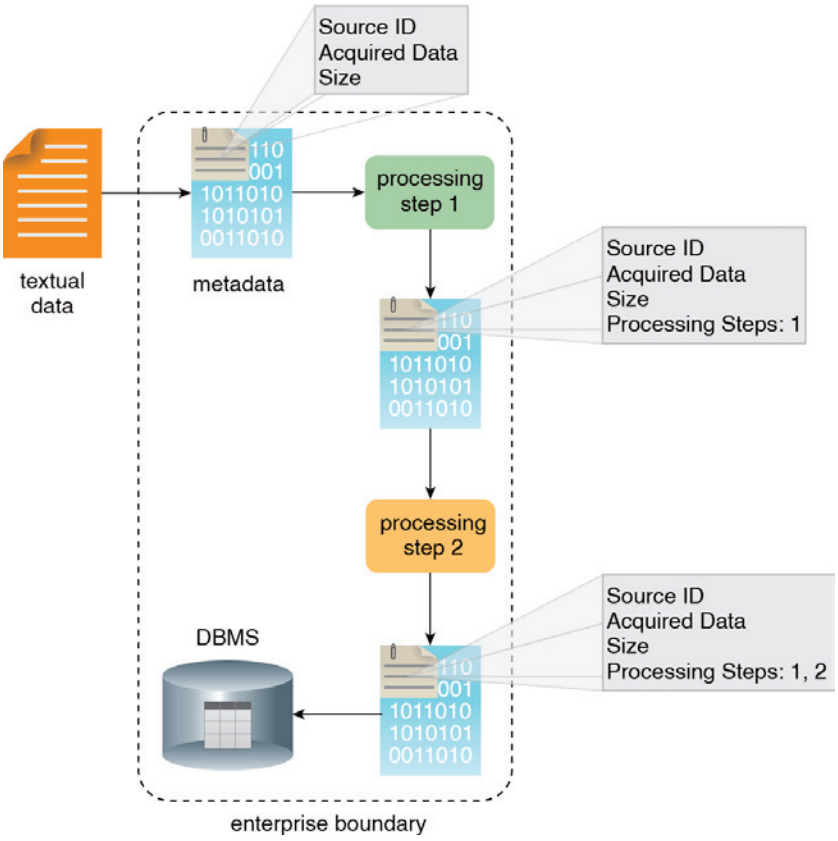


Figure 1.49 – Data may also need to be annotated with the source dataset attributes and processing step details as it passes through the data transformation steps.

Limited Realtime Support

Dashboards and other applications that require streaming data and alerts often demand realtime or near-realtime data transmissions. Many contemporary open-source Big Data solutions and tools are batch-oriented, meaning support for streaming data analysis may either be limited or non-existent. Some realtime data analysis solutions that do exist are proprietary. Near-realtime data processing can be achieved by processing transactional data as it arrives and combining it with already summarized batch-processed data.

Distinct Performance Challenges

Due to the volumes of data that some Big Data solutions are required to process, performance can sometimes become a concern. For example, large datasets coupled with complex search algorithms can lead to longer query times. Another example pertains to available bandwidth. With increasing data volumes, the time to transfer a unit of data can exceed its actual data processing time, as shown in Figure 1.50.

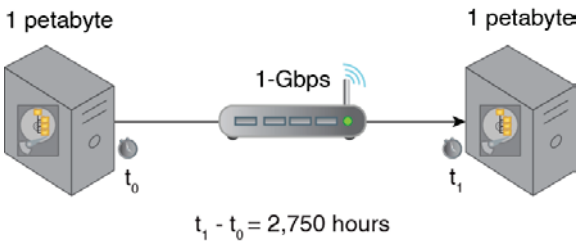


Figure 1.50 – Transferring 1 PB of data via a 1-Gigabit LAN connection at 80% throughput will take approximately 2,750 hours.

Distinct Governance Requirements

Big Data solutions access data and generate data, all of which become assets of the business. A governance framework is required to ensure that the data and the solution environment itself are regulated, standardized, and evolved in a controlled manner.

Examples of what a Big Data governance framework can encompass include:

- standardization of how data is tagged and the metadata used for tagging
- policies that regulate the kind of external data that can be acquired
- policies for data privacy and data anonymization
- policies for data archiving of data sources and analysis results
- policies for data cleansing and filtering

Distinct Methodology

A methodology will be required to control how data flows in and out of Big Data solutions and how feedback loops can be established to enable the processed data to undergo repeated refinements, as shown in Figure 1.51. For example, an iterative approach may be used to enable business personnel to provide IT personnel with feedback for system refinement on an iterative basis. Each feedback cycle may reveal the need for existing steps to be modified, or for new steps, such as pre-processing for data cleansing, to be added.

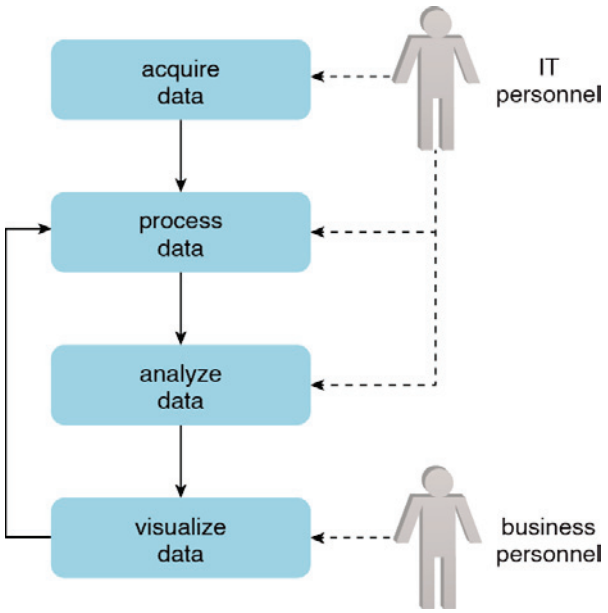


Figure 1.51 – Each iteration can help fine-tune processing steps, algorithms, and data models to improve the accuracy of the results and deliver greater value to the business.

Cloud Computing

As first mentioned in the preceding Big Data Business & Technology Drivers section, cloud computing introduces remote environments that can host IT infrastructure for, among other things, large-scale storage and processing. Regardless of whether an organization is already cloud-enabled, the adoption of a Big Data environment may necessitate that some or all of that environment be hosted within a cloud. For example, an enterprise that runs its CRM system in a cloud decides to add a Big Data solution in the same cloud environment in order to run analytics on its CRM data. This data can then be shared with its primary Big Data environment that resides within the enterprise boundaries.

NOTE
The topic of cloud computing in relation to Big Data is explored further in *Module 2: Big Data Analysis & Technology Concepts*.

Common justifications for incorporating a cloud environment in support of a Big Data solution include:

- inadequate in-house hardware resources
- upfront capital investment is not available
- the project is to be isolated from the rest of the business so that existing business processes are not impacted
- the Big Data initiative is a proof of concept
- datasets that need to be processed reside in a cloud
- the limits of available computing and storage resources used by an in-house Big Data solution are being reached

Optional Reading

Big Data adoption and planning considerations are discussed further in Chapter 3 of the *Big Data Analytics* book that is included with *Module 2: Big Data Analysis & Technology Concepts* and in Chapters 1, 2, 3, 5, 6, and 7 of *Too Big to Ignore: The Business Case for Big Data*.

Notes

Notes / Sketches

Exercise Answers

Exercise 1.1: Answers

1. A **dataset** is a group of related data in which each member of the group possesses the same set of attributes.
2. The goal of **data analysis** is to support decision-making by establishing patterns and relationships in the data being analyzed.
3. **Analytics** are focused on sifting through large amounts of **raw** or **unstructured** data to extract meaningful information that may help enrich existing enterprise data.
4. The process of business intelligence can apply **analytics** to large amounts of data.

Exercise 1.2: Answers

1. A **data warehouse** can contain analytical databases that can improve query response times.
2. **Extract-transform-load** or **ETL** is used to load data from a source system into a target system, and is the main operation through which data warehouses are fed data.
3. Volume, velocity, **variety**, **value** and **veracity** are the five primary Big Data characteristics that differentiate it from traditional data.
4. The value characteristic of Big Data is **dependent** on how long data processing takes.
5. In general, the data processed by Big Data solutions can be found in the following data types or formats: **structured**, **unstructured**, **semi-structured** and **metadata**.

Exercise 1.3: Answers

1. Analytics can be categorized into four types according to their value attribute: **descriptive**, **diagnostic**, **predictive** and **prescriptive**.
2. **Diagnostic** analytics results are viewed via interactive visualization tools that enable trends and patterns to be easily spotted.
3. **Prescriptive** analytics are most valuable for enterprises because this technique offers a suggested course of action that can be taken.
4. Machine learning algorithms can be categorized into **supervised** and **unsupervised** learning.

Exercise 1.4: Answers

1. Traditional BI uses **prescriptive** and **diagnostic** analytics.
2. **Data visualization** communicates analytical results using a variety of graphic, interactive tools.
3. Traditional data visualization tools present both **descriptive** and **diagnostic** data analytics results.
4. Big Data BI adds value to traditional BI by using **predictive** and **prescriptive** analytics.
5. Advanced Big Data visualization tools use **prescriptive** and **predictive** data analytics tools.

Exam B90.01

The course you just completed corresponds to Exam B90.01, which is an official exam that is part of the Big Data Science Certified Professional (BDSCP) program.

This exam can be taken at Pearson VUE testing centers worldwide or via Pearson VUE Online Proctoring, which enables you to take exams from your home or office workstation with a live proctor. For more information, visit:

www.bigdatascienceschool.com/exams/

www.pearsonvue.com/arcitura/

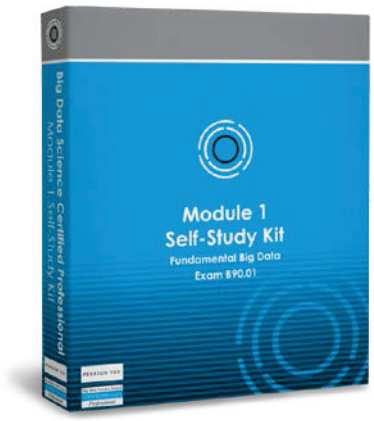
www.pearsonvue.com/arcitura/op/ (Online Proctoring)

Module 1 Self-Study Kit

An official BDSCP Self-Study Kit is available for this module, providing additional study aids and resources, including a separate self-study guide, Audio Tutor CDs and flash cards.

Note that versions of this self-study kit are available with and without a Pearson VUE exam voucher for Exam B90.01.

For more information, visit:
www.bigdataselfstudy.com



Contact Information and Resources

AITCP Community

Join the growing international Arcitura IT Certified Professional (AITCP) community by connecting on official social media platforms: LinkedIn, Twitter, Facebook, and YouTube.

Social media and community links are accessible at:

- www.arcitura.com/community
- www.servicetechbooks.com/community

General Program Information

For general information about the BDSCP program and Certification requirements, visit:

www.bigdatascienceschool.com and www.bigdatascienceschool.com/matrix/

General Information about Course Modules and Self-Study Kits

For general information about BDSCP Course Modules and Self-Study Kits, visit:

www.bigdatascienceschool.com and www.bigdataselfstudy.com

Pearson VUE Exam Inquiries

For general information about taking BDSCP Exams at Pearson VUE testing centers or via Pearson VUE Online Proctoring, visit:

www.pearsonvue.com/arcitura/
www.pearsonvue.com/arcitura/op/ (Online Proctoring)

Public Instructor-Led Workshop Schedule

For the latest schedule of instructor-led BDSCP workshops open for public registration, visit:

www.bigdatascienceschool.com/workshops

Private Instructor-Led Workshops

Certified trainers can deliver workshops on-site at your location with optional on-site proctored exams. To learn about options and pricing, contact:

info@arcitura.com

or

1-800-579-6582

Becoming a Certified Trainer

If you are interested in attaining the Certified Trainer status for this or any other Arcitura courses or programs, learn more by visiting:

www.arcitura.com/trainerdevelopment/

General BDSCP Inquiries

For any other questions relating to this Course or any Module, Exam, or Certification that is part of the BDSCP program, contact:

info@arcitura.com

or

1-800-579-6582

Automatic Notification

To be automatically notified of changes or updates to the BDSCP program and related resource sites, send a blank message to:

notify@arcitura.com

Feedback and Comments

Help us improve this course. Send your feedback or comments to:

info@arcitura.com