



Big Data Science Certified Professional (BDSCP) Curriculum

Module 1 - Fundamental Big Data	(Exam B90.01)
Module 2 - Big Data Analysis & Technology Concepts	(Exam B90.02)
Module 3 - Big Data Analysis & Technology Lab	(Exam B90.03)
Module 4 - Fundamental Big Data Analysis & Science	(Exam B90.04)
Module 5 - Advanced Big Data Analysis & Science	(Exam B90.05)
Module 6 - Big Data Analysis & Science Lab	(Exam B90.06)
Module 7 - Fundamental Big Data Engineering	(Exam B90.07)
Module 8 - Advanced Big Data Engineering	(Exam B90.08)
Module 9 - Big Data Engineering Lab	(Exam B90.09)
Module 10 - Fundamental Big Data Architecture	(Exam B90.10)
Module 11 - Advanced Big Data Architecture	(Exam B90.11)
Module 12 - Big Data Architecture Lab	(Exam B90.12)
Module 13 - Fundamental Big Data Governance	(Exam B90.13)
Module 14 - Advanced Big Data Governance	(Exam B90.14)
Module 15 - Big Data Governance Lab	(Exam B90.15)



BDSCP Certification Requirements

Big Data Science School
CERTIFIED BIG DATA
Professional

Certified Big Data Science Professional
1, 2, 3

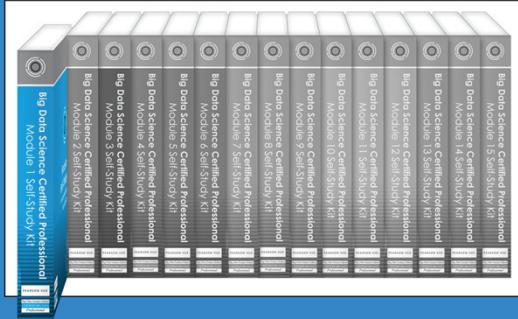
Certified Big Data Scientist
1, 2, 4, 5, 6

Certified Big Data Consultant
1, 2, 3, 4, 7

Certified Big Data Engineer
1, 2, 7, 8, 9

Certified Big Data Architect
1, 2, 10, 11, 12

Certified Big Data
Governance Specialist
1, 2, 13, 14, 15



For complete descriptions
of certifications, visit:
www.bigdatascienceschool.com

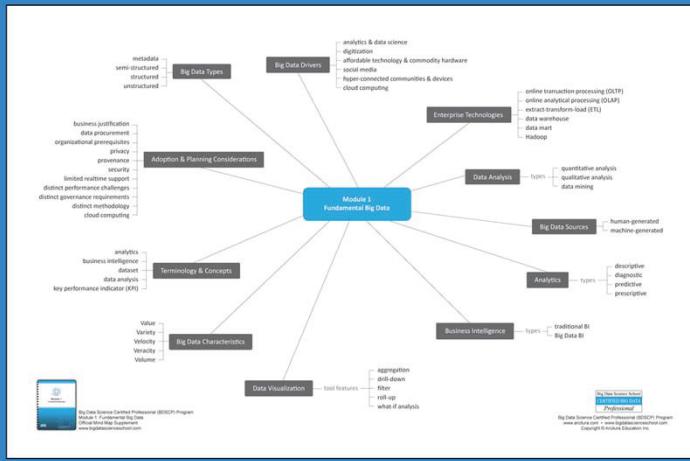
Introduction





Mind Map Poster

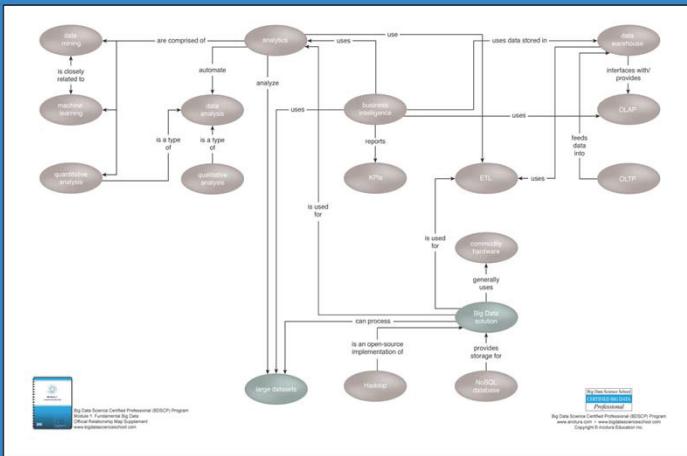
The BDSCP Module 1: Mind Map Poster that accompanies this course booklet provides an alternative visual representation of all primary topics covered in this course.





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.





Overview

- Understanding Big Data
- Fundamental 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

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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Introducing Big Data

- The qualities that distinguish data processed by Big Data solutions are commonly known as the “Five V’s” 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.



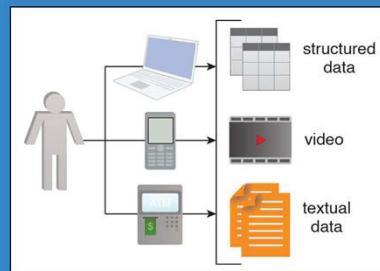
Introducing Big Data

- 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).



Introducing Big Data

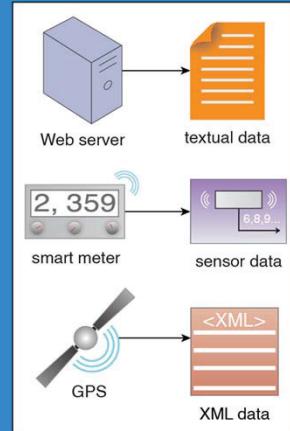
- The data processed by Big Data solutions can be human-generated or machine-generated, although it is ultimately the responsibility of machines to generate the processing results.
- **Human-generated data** is the result of human interaction with systems, such as online services and digital devices.
- Examples include social media, micro blogging, e-mails, photo sharing and messaging.





Introducing Big Data

- Machine-generated data is the result of the automated, event-driven generation of data by software programs or hardware devices.
- Examples include Web logs, sensor data, telemetry data, smart meter data and appliance usage data.



Copyright © Arcitura Education Inc. (www.arcitura.com)



Introducing Big 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.



Reading

“Big Data Fundamentals”

Chapter 1
Pages 20-24, Case Study Background

Copyright © Arcitura Education Inc. (www.arcitura.com)

Fundamental Terminology & Concepts





Overview

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.

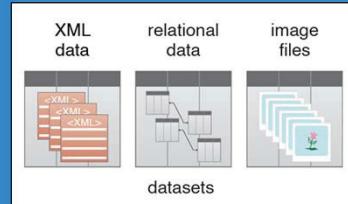


Datasets

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.

The diagram on the bottom right 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





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.





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.
- 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.





Analytics

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

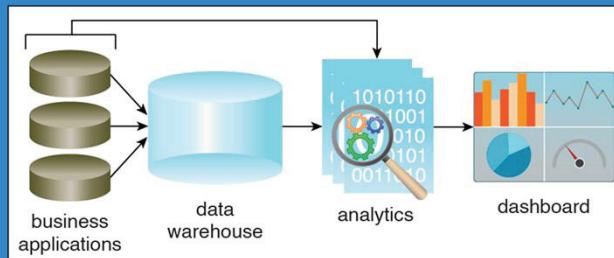
- In **business-oriented environments**, analytics results can lower operational costs and facilitate strategic decision-making.
- In the **scientific domain**, analytics can help identify the cause of a phenomenon to improve the accuracy of predictions.
- In **services-based environments** (such as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





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. Each KPI is based on a quantifiable indicator that is identified and agreed upon beforehand.

Examples include calls handled per day and 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 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



Note

The primary types of data analysis relevant to Big Data are covered in the *Fundamental Analysis, Analytics & Machine Learning Types* section.

Business Intelligence and KPIs are further explored in the *Business Intelligence & Big Data* section.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Exercise

Exercise 1.1

Fill in the Blanks

Copyright © Arcitura Education Inc. (www.arcitura.com)

Big Data Business & Technology Drivers





Overview

Big Data emerged from a combination of business needs and technology innovations. The following are the 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



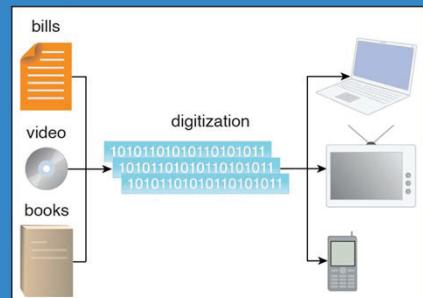
Analytics & Data Science

- 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. Mining this data may allow for customized marketing, automated recommendations, and the development of optimized product features.
- Examples include on-demand TV and streaming video.





Affordable Technology & Commodity Hardware

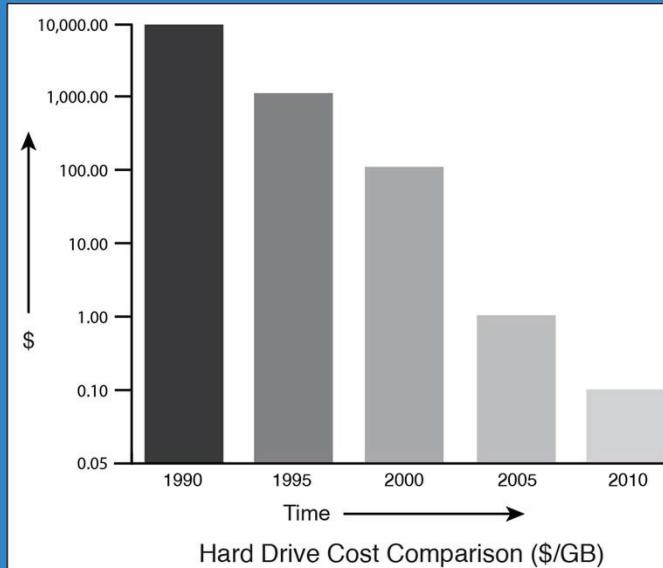
- Technology related to collecting and processing large quantities of diverse data has become increasingly affordable.
- Typical Big Data solutions are based on open-source software that requires little more than commodity hardware.
- The use of commodity hardware makes the adoption of Big Data solutions accessible to businesses without large capital investments.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Affordable Technology & Commodity Hardware

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





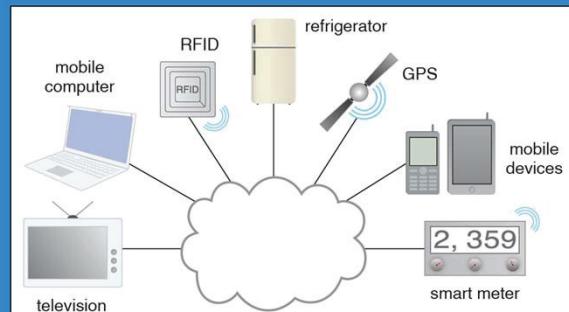
Social Media

- The emergence of social media has empowered customers to **provide feedback in near-realtime via open and public mediums**.
- This shift 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 interaction 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.
- This is either directly through online interaction or indirectly through the usage of connected devices.
- This has resulted in **massive data streams**.
- Some streams are public. Other streams go to vendors and businesses directly.





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.

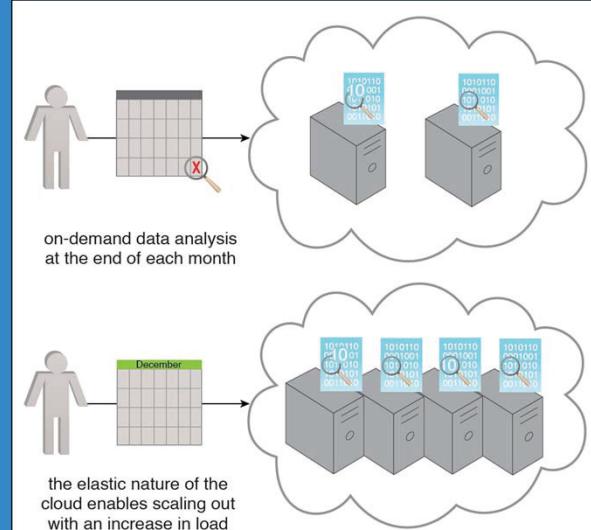
Copyright © Arcitura Education Inc. (www.arcitura.com)



Cloud Computing

The cloud 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.





Optional Reading

“Big Data Fundamentals”

Chapter 2

Pages 29-42, Marketplace Dynamics, Business Architecture, Business Process Management, Information and Communications Technology, Internet of Everything (IoE)

Copyright © Arcitura Education Inc. (www.arcitura.com)



Reading

“Big Data Fundamentals”

Chapter 2
Pages 43-45, Case Study Example

Copyright © Arcitura Education Inc. (www.arcitura.com)

Traditional Enterprise Technologies Related to Big Data





Overview

This section briefly describes the following technologies:

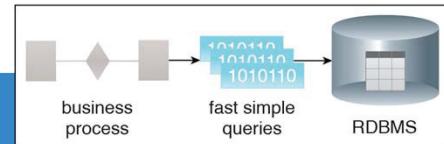
- Online Transaction Processing
- Online Analytical Processing
- Extract Transform Load
- Data Warehouses
- Data Marts
- Hadoop

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.**

If you are already familiar with these 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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



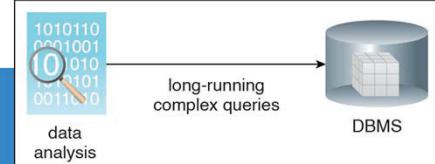
Online Transaction Processing (OLTP)

- OLTPs 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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



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.
- OLAPs 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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Online Analytical Processing (OLAP)

- OLAPs are used in **diagnostic**, **predictive** and **prescriptive** analytics (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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



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 the diagram on the following page, relational data from two OLTP systems is periodically imported via bulk data import tasks into an OLAP system.

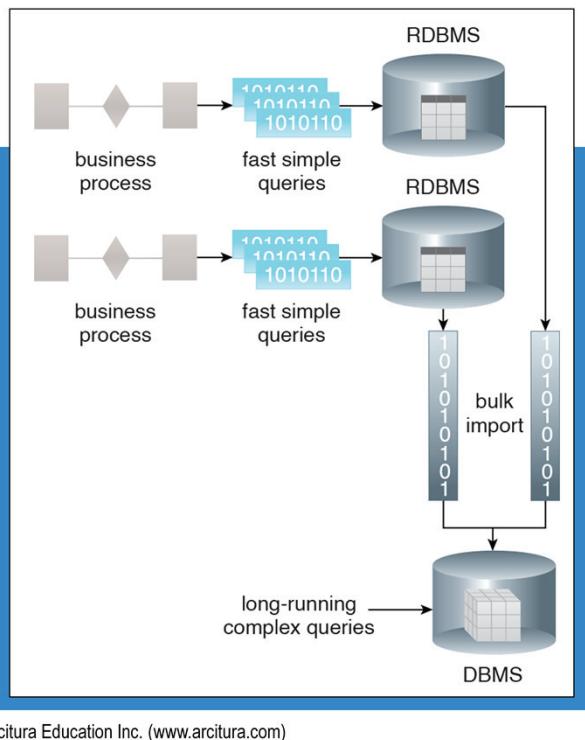
Copyright © Arcitura Education Inc. (www.arcitura.com)



OLTPs & OLAPs

The 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.



Copyright © Arcitura Education Inc. (www.arcitura.com)



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.

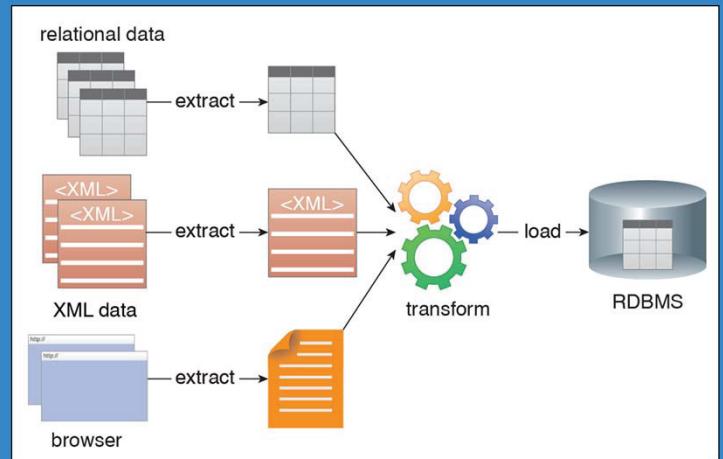
Copyright © Arcitura Education Inc. (www.arcitura.com)



Extract Transform Load (ETL)

The required data is first obtained (**extracted**) from the sources, after which the extracts are modified (**transformed**) by applying rules.

The data is finally inserted (**loaded**) into the target system.



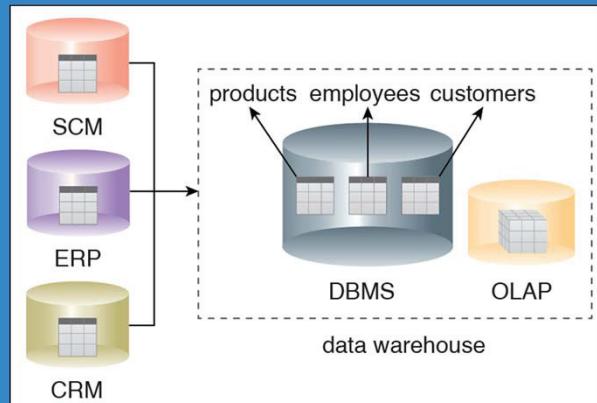


Data Warehouses

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.

Data warehouses usually interface with an OLAP system to support analytical queries.

It also periodically pulls data from other sources for consolidation into a dataset (such as from OLTP, ERP, CRM, and SCM systems).





Data Warehouses

- 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.

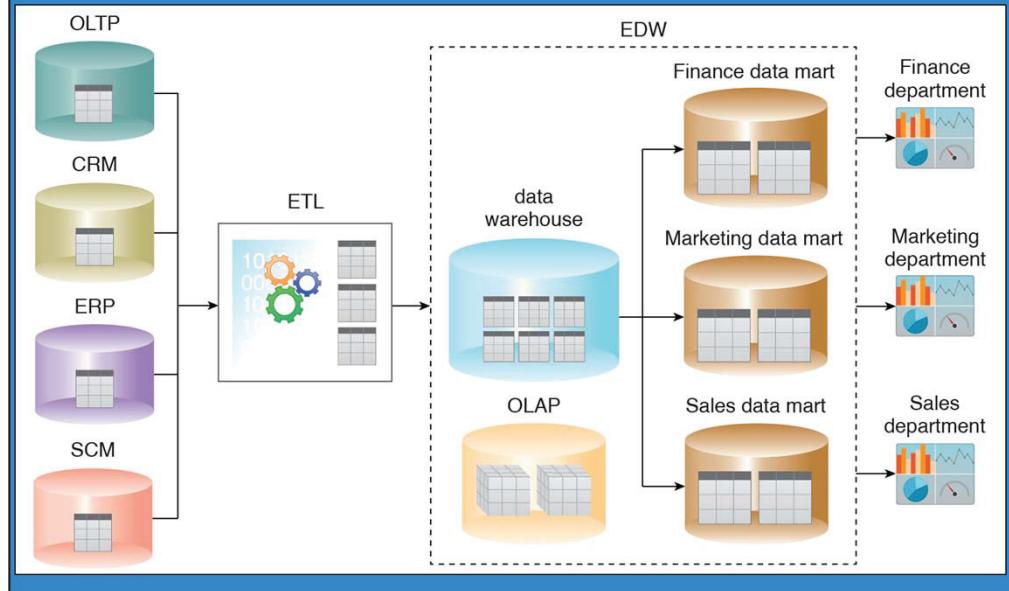


Data Warehouses & Data Marts

- 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 the diagram on the following page, enterprise-wide data is collected and business entities are then extracted.
- Domain-specific entities are persisted into the data warehouse via an ETL process.
- 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 side of the diagram).



Data Warehouses & Data Marts





Hadoop

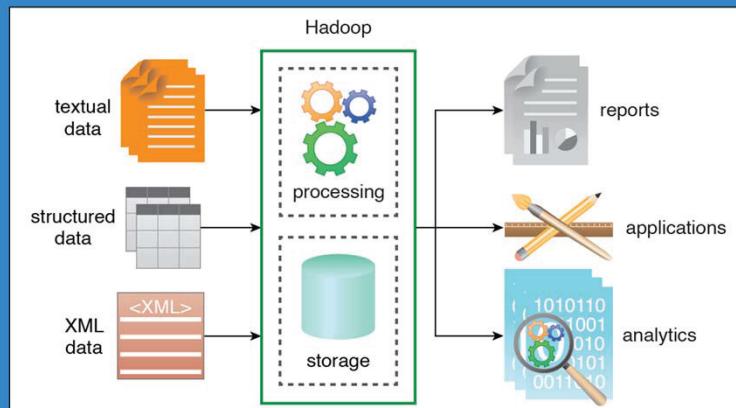
- Hadoop is an open-source framework for large-scale data storage and data processing that is more or less run on commodity hardware.
- The Hadoop framework has established itself as a de facto industry platform for contemporary Big Data solutions.
- Hadoop can be used as an ETL engine, or as an analytics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

Copyright © Arcitura Education Inc. (www.arcitura.com)



Hadoop

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



Characteristics of Data in Big Data Environments



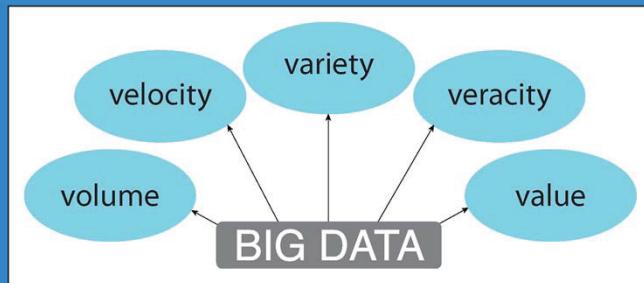


Data Characteristics

Below are the five main characteristics that differentiate data categorized as “Big” data from other forms of data.

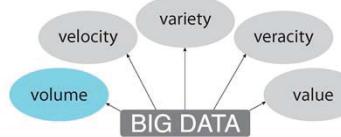
These traits are commonly referred to as the **Five Vs:**

- Volume
- Velocity
- Variety
- Veracity
- Value



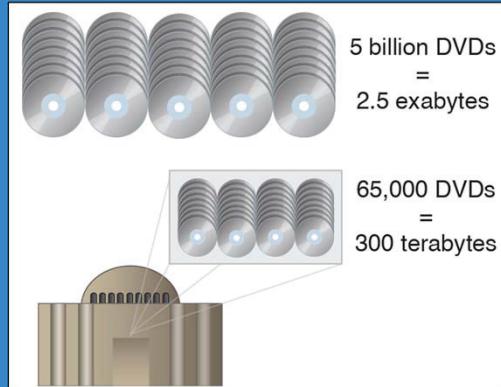


Volume



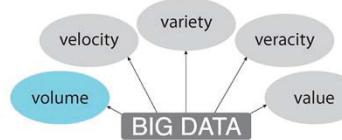
- The anticipated **volume of data that is processed by Big Data solutions** is substantial and usually ever-growing.
- High data volumes impose distinct data storage and processing demands, as well as management and access processes.

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





Volume



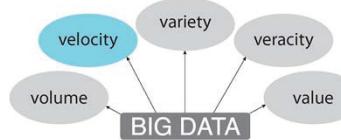
Typical data sources that are responsible for generating high data volumes can include:

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

Copyright © Arcitura Education Inc. (www.arcitura.com)



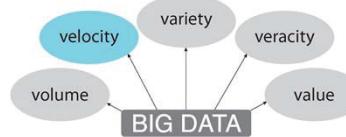
Velocity



- 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.

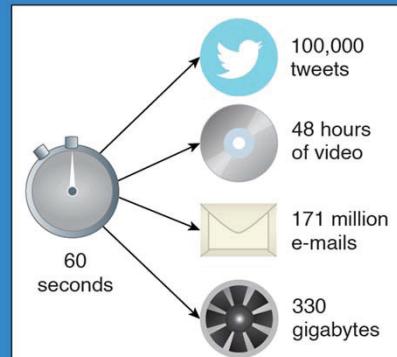


Velocity



To put potential data velocity into perspective, the following data is currently generated every minute:

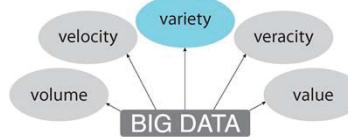
- 100,000 tweets
- 48 hours of video footage
- 171 million e-mails
- 330 GBs generated by the average jet engine



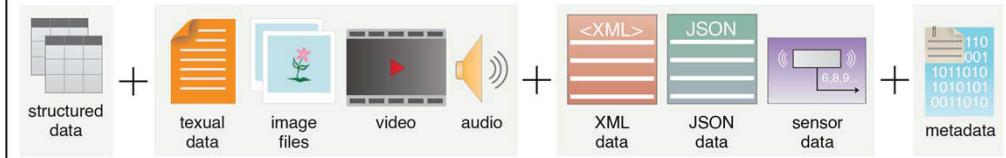
Copyright © Arcitura Education Inc. (www.arcitura.com)



Variety

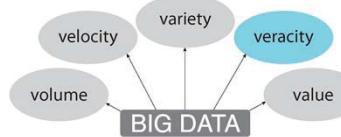


- Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions, such as structured, semi-structured and unstructured data (further described in the upcoming *Types of Data in Big Data Environments* section).
- Data variety brings challenges for enterprises in terms of data integration, transformation, processing and storage. Examples include financial transactions (structured), e-mails (semi-structured) and images (unstructured).





Veracity



Veracity refers to the quality or fidelity of data.

Data that exists in Big Data environments can be meaningful or just add clutter. Data assessed in relation to veracity are either:

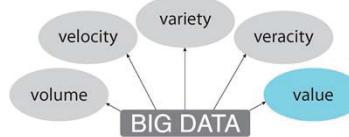
- **Noise** – data carrying no value
- **Signal** – data bearing value leading to meaningful information

Data acquired in a controlled manner, such as 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 noise-to-signal ratio varies depending on the type of data present.



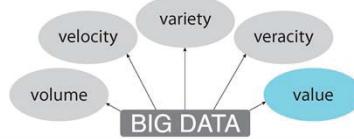
Value



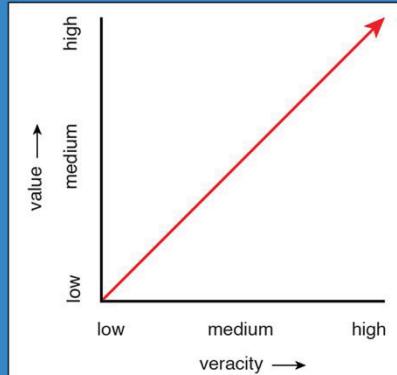
- Value is defined as the usefulness of data for an enterprise.
- The Value characteristic is directly related to the Veracity characteristic in that the higher the data fidelity, the more value it holds for the business.
- Value is also dependent on how long data processing takes, as value and time 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).



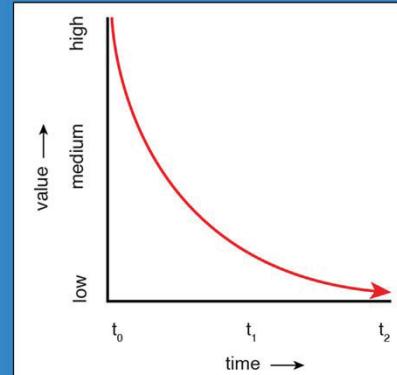
Value



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

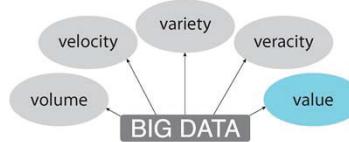


The longer it takes to analyze the data, the less potential value it has for the business.





Value



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?

Copyright © Arcitura Education Inc. (www.arcitura.com)

Types of Data in Big Data Environments





Data Types

The types of data processed by Big Data solutions can be divided into the following primary categories:

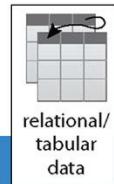
- structured data
- unstructured data
- semi-structured data

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, but rather another form of data that itself can vary in structure, **metadata** is briefly described at the end of this section.



Structured Data



relational/
tabular
data

Structured data:

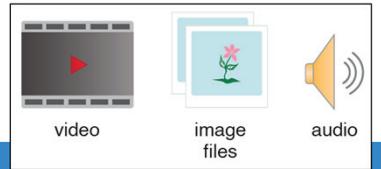
- conforms to a data model or schema
- is stored in a tabular form
- can be relational

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.

Structured data does not generally have any special pre-processing or storage requirements. Examples include banking transactions, OLTP system records and customer records.



Unstructured Data



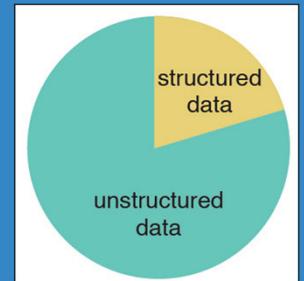
Unstructured data:

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

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.*





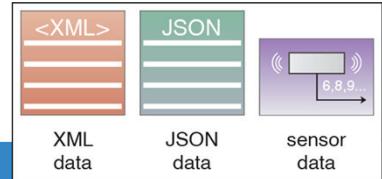
Unstructured Data

- Unlike structured data, unstructured data does generally require special or customized logic when it comes to pre-processing and storage.
- Unstructured data 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.

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



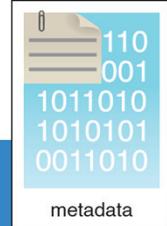
Semi-Structured Data



- Semi-structured data has a defined level of structure and consistency, but cannot be relational in nature.
- Semi-structured data 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 does often have special pre-processing and storage requirements, especially if the underlying format is not text-based.



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

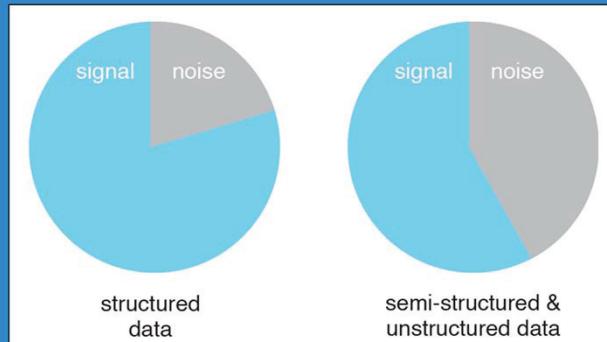
Big Data solutions particularly rely on metadata when processing semi-structured and unstructured data.



Data Types & Veracity

Semi-structured data and unstructured data have a greater noise-to-signal ratio than structured data.

The larger amounts of noise in semi-structured and unstructured data require automated data cleansing and data verification when carrying out extract-transform-load (ETL) processes.





Reading

“Big Data Fundamentals”

Chapter 1
Pages 24-27, Case Study Example

Copyright © Arcitura Education Inc. (www.arcitura.com)



Exercise

Exercise 1.2

Fill in the Blanks

Copyright © Arcitura Education Inc. (www.arcitura.com)

Fundamental Analysis, Analytics & Machine Learning Types





Types of Data Analysis

In the *Fundamental Terminology & Concepts* section we introduced the term “data analysis” and provided a simple example.



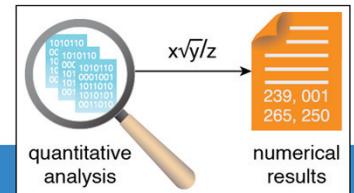
The upcoming pages further describe the following basic types of data analysis:

- quantitative analysis
- qualitative analysis
- data mining

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



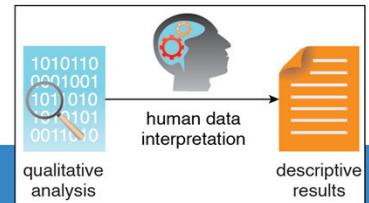
Quantitative Analysis



- 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%.



Qualitative Analysis



- Qualitative analysis is a data analysis technique that focuses on describing various data qualities using words.
- Qualitative analysis 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.



Data Mining



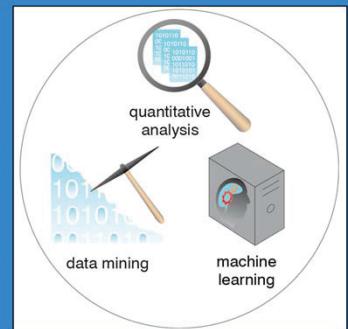
- Data mining, also known as data discovery, is a specialized form of data analysis that targets large datasets.
- Data mining in relation to Big Data analysis generally refers to automated, software-based techniques that sift through massive datasets to identify patterns and trends.
- Specifically, data mining 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).



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.



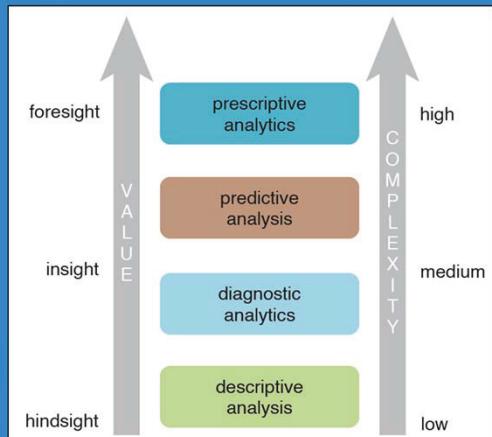


Types of Analytics

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

Value and complexity increase as we move from descriptive to prescriptive analytics.





Descriptive Analytics

Descriptive analytics is 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?*

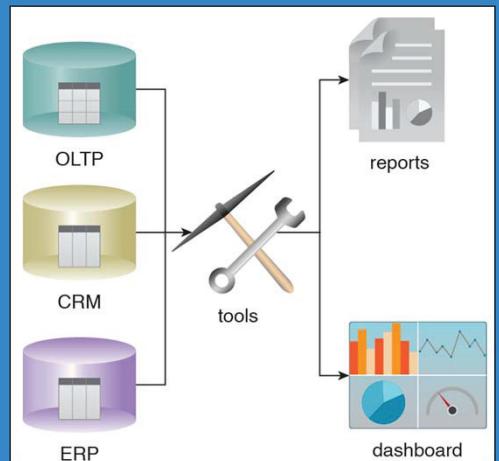
Around 80% of analytics are descriptive in nature. Value-wise, descriptive analytics provides the least value and requires a relatively basic skillset.



Descriptive Analytics

Descriptive analytics are often carried out via ad-hoc reporting or dashboards. 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 various other information systems, such as CRMs and ERPs.





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:

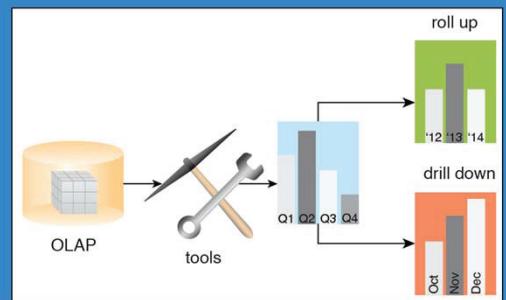
- Why were Q2 sales less than Q1 sales?
- Why have there been more support calls originating from the Eastern region than from the Western region?
- Why was there an increase in patient re-admission rates over the past three months?

Diagnostic analytics are considered to provide more value than descriptive analytics, requiring a more advanced skillset.



Diagnostic Analytics

- Diagnostic analytics usually require collecting data from multiple sources and storing it in a structure that lends itself to performing drill-downs and roll-ups.
- 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.





Predictive Analytics

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

Questions are usually formulated using on 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

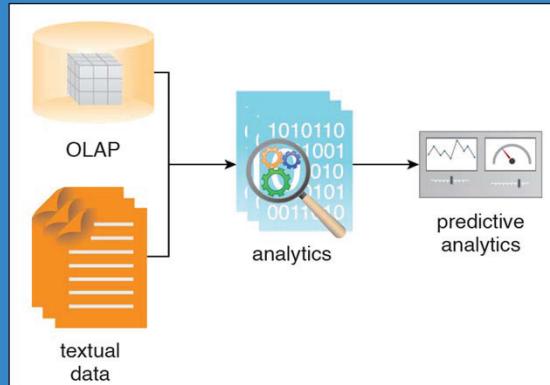
- Predictive analytics try to predict the event outcome and predictions are made based on patterns, trends and exceptions found in historical and current data.
- This can lead to the identification of risks and opportunities.
- Predictive analytics involve the use of large datasets (comprised of both internal and external data), statistical techniques, quantitative analysis, machine learning and data mining techniques.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Predictive Analytics

- This type of analytics is considered to provide more value and requires a more advanced skillset than both descriptive and diagnostic analytics.
- Predictive analytics tools generally abstract underlying statistical intricacies by providing 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, 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.



Prescriptive Analytics

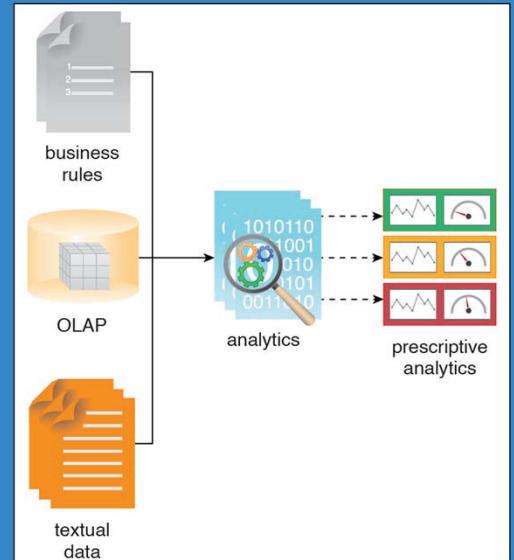
- Various outcomes are calculated, and the best course of action for each outcome is suggested.
- The approach shifts from explanatory to advisory and can include the simulation of various scenarios.
- Prescriptive analytics incorporate internal data (current and historical sales data, customer information, product data, business rules) and external data (social media data, weather data, demographic data).

Copyright © Arcitura Education Inc. (www.arcitura.com)



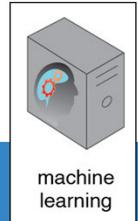
Prescriptive Analytics

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.





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



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.



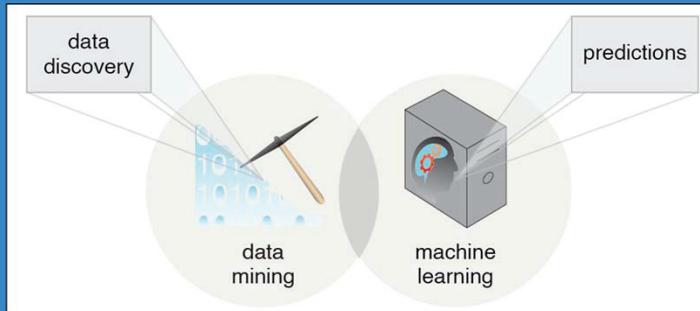
Machine Learning vs. Data Mining

- 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 vs. Data Mining

Machine learning can use the output from data mining (identified patterns) for further data classification through supervised learning.





Exercise

Exercise 1.3

Fill in the Blanks

Copyright © Arcitura Education Inc. (www.arcitura.com)

Business Intelligence & Big Data





Overview

- 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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Traditional BI

Traditional BI utilizes descriptive and diagnostic analytics to provide information on historical and current events.

Traditional BI is not “intelligent” as such 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:

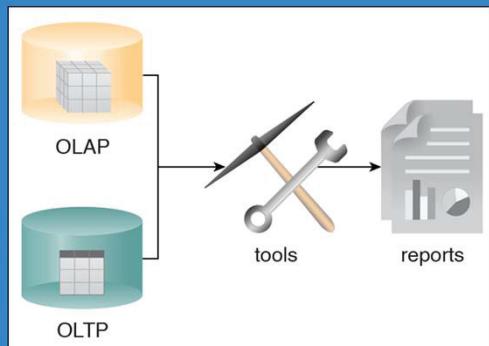
- ad-hoc reports
- dashboards



Traditional BI: Ad-Hoc Reporting

- Ad-hoc reporting is a process that involves manually processing data to produce custom-made reports.
- 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.

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

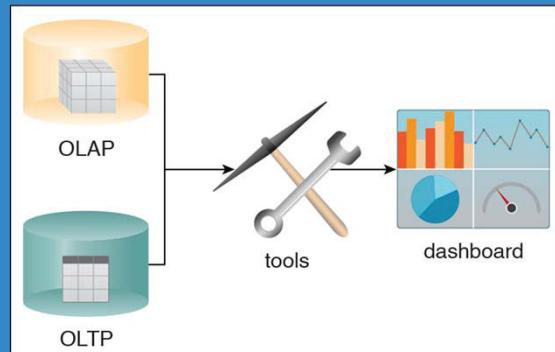




Traditional BI: Dashboards

- 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, such as column charts, pie charts and gauges.

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





Traditional BI

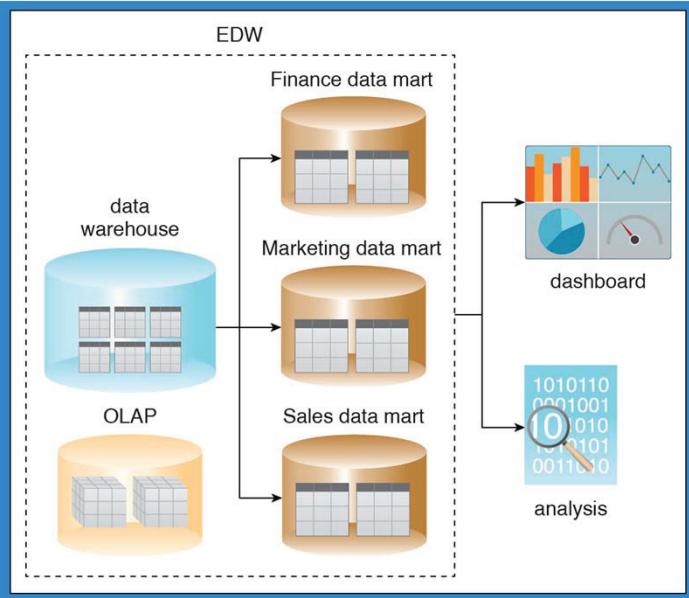
- As previously explained, data warehouses and data marts contain consolidated and validated information about enterprise-wide business entities.
- Traditional BI cannot function effectively without data marts because they contain the optimized and segregated data BI requires for reporting purposes.
- 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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Traditional BI

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.





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.
- Big Data BI 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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Big Data BI

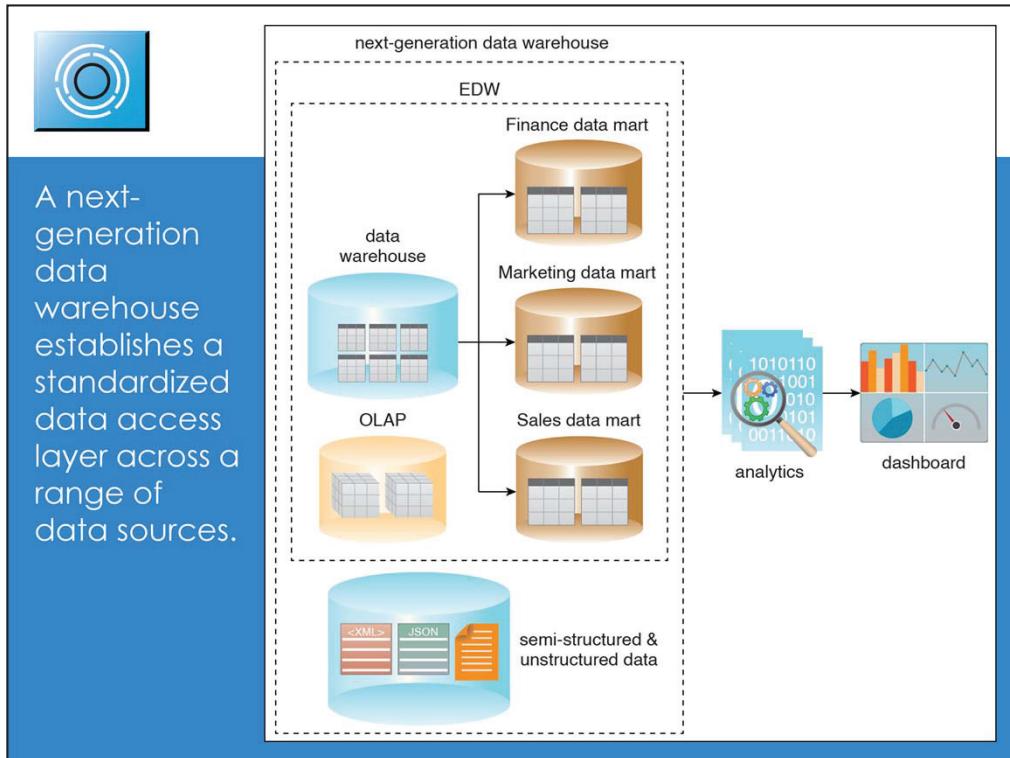
- This helps reveal patterns and anomalies across a broader scope within the enterprise. It also leads to data discovery by identifying insights and information that may have been previously absent or unknown.
- Big Data BI requires the analysis of unstructured, semi-structured and structured data residing 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.



Big Data BI

- Coupling a traditional data warehouse with these new technologies results in a hybrid data warehouse.
- This type of data warehouse 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 data they require.
- This eliminates the need for Big Data BI tools to have to connect to multiple data sources to retrieve or access data.

Copyright © Arcitura Education Inc. (www.arcitura.com)





Reading

“Big Data Fundamentals”

Chapter 4
Pages 86-88, Case Study Example

Copyright © Arcitura Education Inc. (www.arcitura.com)

Data Visualization & 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 provided 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

Exercise 1.4

Fill in the Blanks

Copyright © Arcitura Education Inc. (www.arcitura.com)

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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Business Justification

- 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.
- Also, **the quality of the data targeted for processing by Big Data solutions** needs to be assessed.

Copyright © Arcitura Education Inc. (www.arcitura.com)



Organizational Prerequisites

- 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.

Copyright © Arcitura Education Inc. (www.arcitura.com)



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 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.



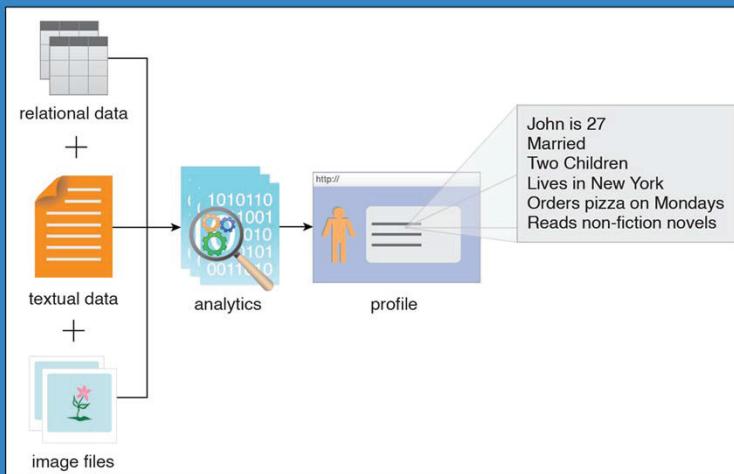
Privacy

- 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.



Privacy

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





Security

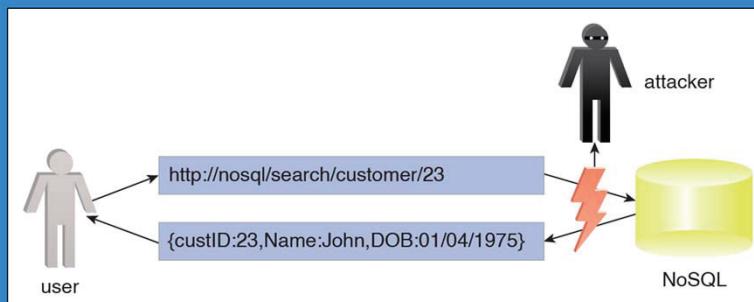
- Some of the components of Big Data solutions lack the robustness of traditional enterprise solution environments when it comes to access control and data security.
- Securing Big Data involves ensuring that data networks provide access to repositories that are sufficiently secured, via custom authentication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

Copyright © Arcitura Education Inc. (www.arcitura.com)



Security

For example, unlike traditional RDBMSs, 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.



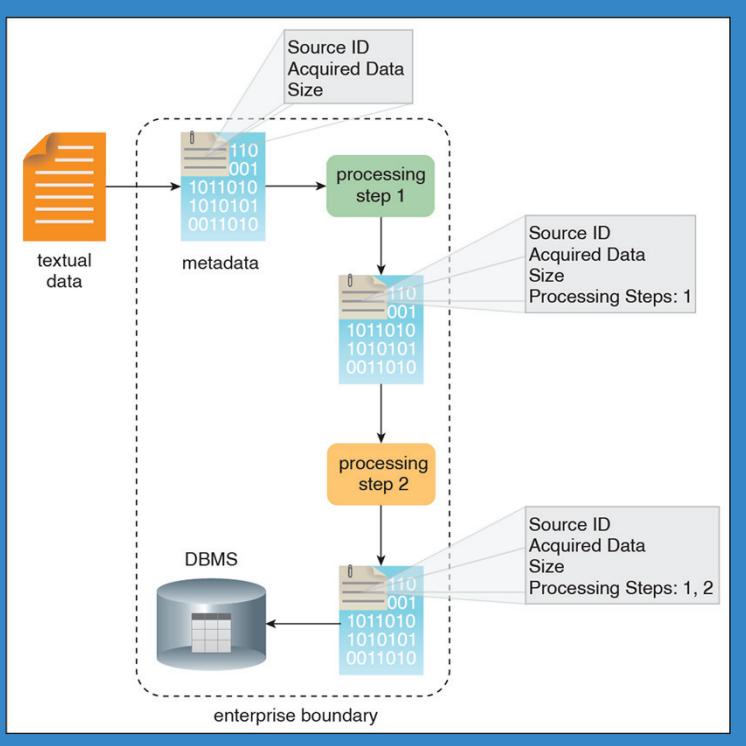


Provenance

- 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.

Copyright © Arcitura Education Inc. (www.arcitura.com)

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

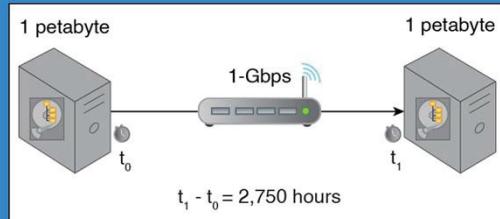
- Dashboards and other applications that require streaming data and alerts often demand realtime or near-realtime data transmissions.
- Many contemporary open-source Big Data solutions and tools are batch-oriented, meaning **support for streaming data analysis may either be limited or non-existent**.
- 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.

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 would encompass include:

- standardizing 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 data sources and analysis results
- policies for data cleansing and filtering



Distinct Methodology

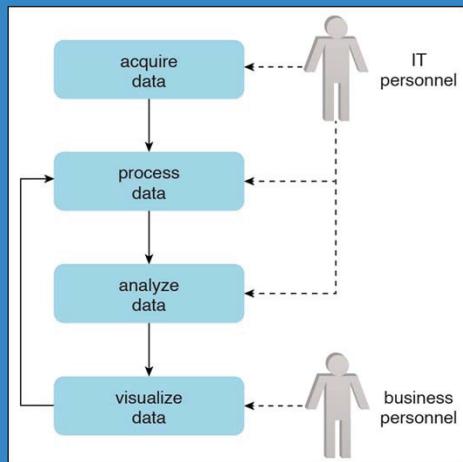
- A methodology will be required to control how data flows in and out of Big Data solutions and how feedback loops can be established to enable the processed data to undergo repeated refinements.
- For example, an iterative approach may be used to enable business personnel to provide IT personnel with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

Copyright © Arcitura Education Inc. (www.arcitura.com)



Distinct Methodology

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



Copyright © Arcitura Education Inc. (www.arcitura.com)



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.**

Copyright © Arcitura Education Inc. (www.arcitura.com)



Cloud Computing

- 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.

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

Copyright © Arcitura Education Inc. (www.arcitura.com)



Cloud Computing

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

- in-house hardware resources are inadequate
- 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



Conclusion and Q&A

Copyright © Arcitura Education Inc. (www.arcitura.com)



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.

PEARSON VUE

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)

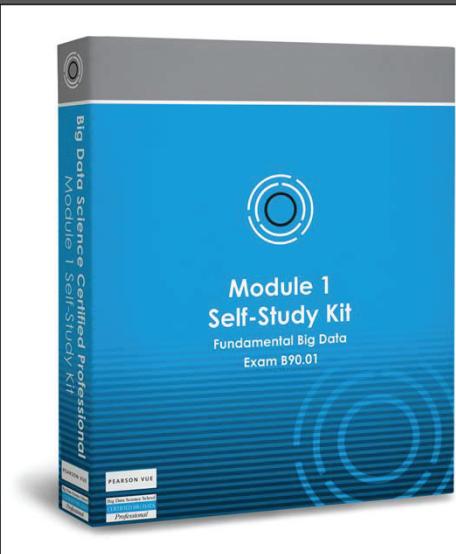


Module 1 Self-Study Kit

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





AITCP Community

Join the growing international Arcitura IT Certified Professional (AITCP) community by connecting with official social media channels available via LinkedIn, Twitter, Facebook and YouTube.

Social media and community links are accessible at:

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





Contact & Resources

Arcitura Education Inc.

www.arcitura.com

Big Data School

www.bigdatascienceschool.com

Workshops

www.bigdataworkshops.com

Self-Study Kits

www.bigdataselfstudy.com

Service Technology Book Series

www.servicetechbooks.com

Service Technology Standards

www.servicetechspecs.com

Big Data Patterns

www.bigdatapatterns.org

General Inquiries

info@arcitura.com

Automatic Updates

notify@arcitura.com

Becoming a Trainer/Partner

partners@arcitura.com



Cloud Certified Professional (CCP) Program

CloudSchool.com™
CLOUD CERTIFIED
Professional

PEARSON VUE

The Cloud Certified Professional (CCP) program from Cloud School™ provides a comprehensive vendor-neutral curriculum of 21 course modules and exams for a series of industry certifications dedicated to areas of specialization in the field of cloud computing.

www.cloudschool.com • www.cloudselfstudy.com • www.cloudworkshops.com





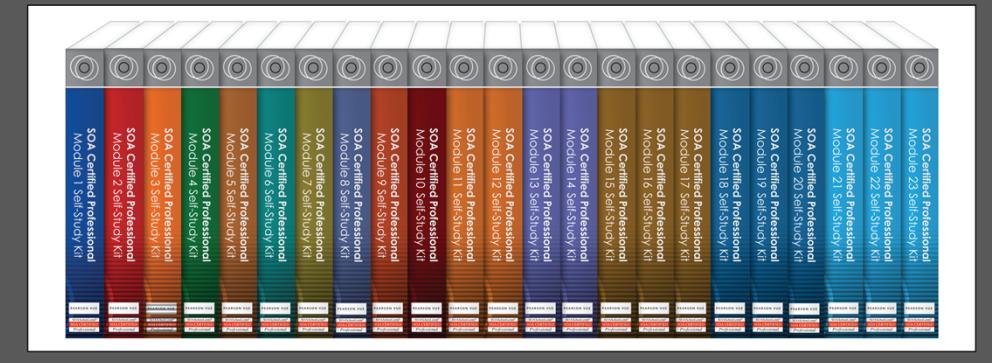
SOA Certified Professional (SOACP) Program

SOASchool.com®
SOA CERTIFIED
Professional

PEARSON VUE

The SOA Certified Professional (SOACP) program from SOA School® provides a comprehensive vendor-neutral curriculum of 23 course modules and exams for a series of industry certifications dedicated to areas of specialization in the fields of SOA and service-oriented computing.

www.soaschool.com • www.soaselfstudy.com • www.soaworkshops.com



Arcitura Education Inc. is a leading global provider of progressive, vendor-neutral training and certification programs. With a worldwide network of certified instructors, licensed training partners and testing centers, Arcitura schools and accreditation programs have become internationally established.

If you don't already have an online AITCP profile, visit

www.arcitura.com/aitcp
for more information.

Arcitura
the IT education company



www.arcitura.com
