ICT 700
Introduction To
Business Information Systems

LECTURE 6

Business Intelligence and IS for decision making: Big Data and Analytics

Unit Coordinator:
Sajad Ghatrehsamani

King's Own Institute

Reading Chapters:

Chapter 6 - Baltzan (2019)

and

Chapter 6 – Stair & Reynolds (2020)

Learning Objectives

1. Understanding the modern history of data science.
2. Identify Data Types: Transactional and Analytical .
3. Identify five key characteristics associated with big data.
4. Identify five key challenges associated with big data.
5. Distinguish between the terms data warehouse, data mart, and data lake.
6. Identify three key organizational components that must be in place for an organization to get real value from its BI/analytics efforts.
7. Identify five broad categories of business intelligence/analytics techniques including the specific techniques used in each.

The Modern History of Data Science

- 2001: William S. Cleveland called for the new field and term data science. It became more widely used in the next few years.

- 2002: Data Science Journal launches.

- 2002: Torch machine learning library is created.

- 2003: Columbia University launches the Journal of Data Science.

- 2004: MapReduce algorithm is created.

- 2006: The Netflix prize is established.

- 2007: Machine learning library Scikit-learn is created.

The Modern History of Data Science

- 2008: The data scientist role emerged due to DJ Patil and Jeff Hammerbacher. Data scientists collect large amounts of data, transform it into a more usable format, and solve business-related problems using data-driven techniques and tools.

- 2009: ImageNet, a large data collection used for computer vision research, spawns the AI boom.

- 2010: The Kaggle machine learning competition launches.

- 2010: The Economist declares a new kind of professional: the data scientist.

- 2011: Jeff Dean and Andrew Ng build a neural net that sees cats, which marks the start of the Google Brain project.

- 2012: Geoffrey Hinton unleashes deep neural networks.

- 2012: Snowflake launches a cloud computing–based data warehousing company.

The Modern History of Data Science

- 2015: Google open sources TensorFlow, its AI engine.
- 2016: PyTorch is released.
- 2016: AlphaGo defeats the human Go champion.
- 2017: Amazon Web Services SageMaker launches.
- 2018: The BERT language model is created by Jacob Devlin and his colleagues from Google.

Value of data

1. Data Types
2. Data Governance
3. Data Quality
4. Data Timeliness

Case Study

.

Data Types

1. Transactional

2. Analytical

Transactional Data Types

1. Transactional data encompasses all of the data contained within a single business process or unit of work, and its primary purpose is to support daily operational tasks.

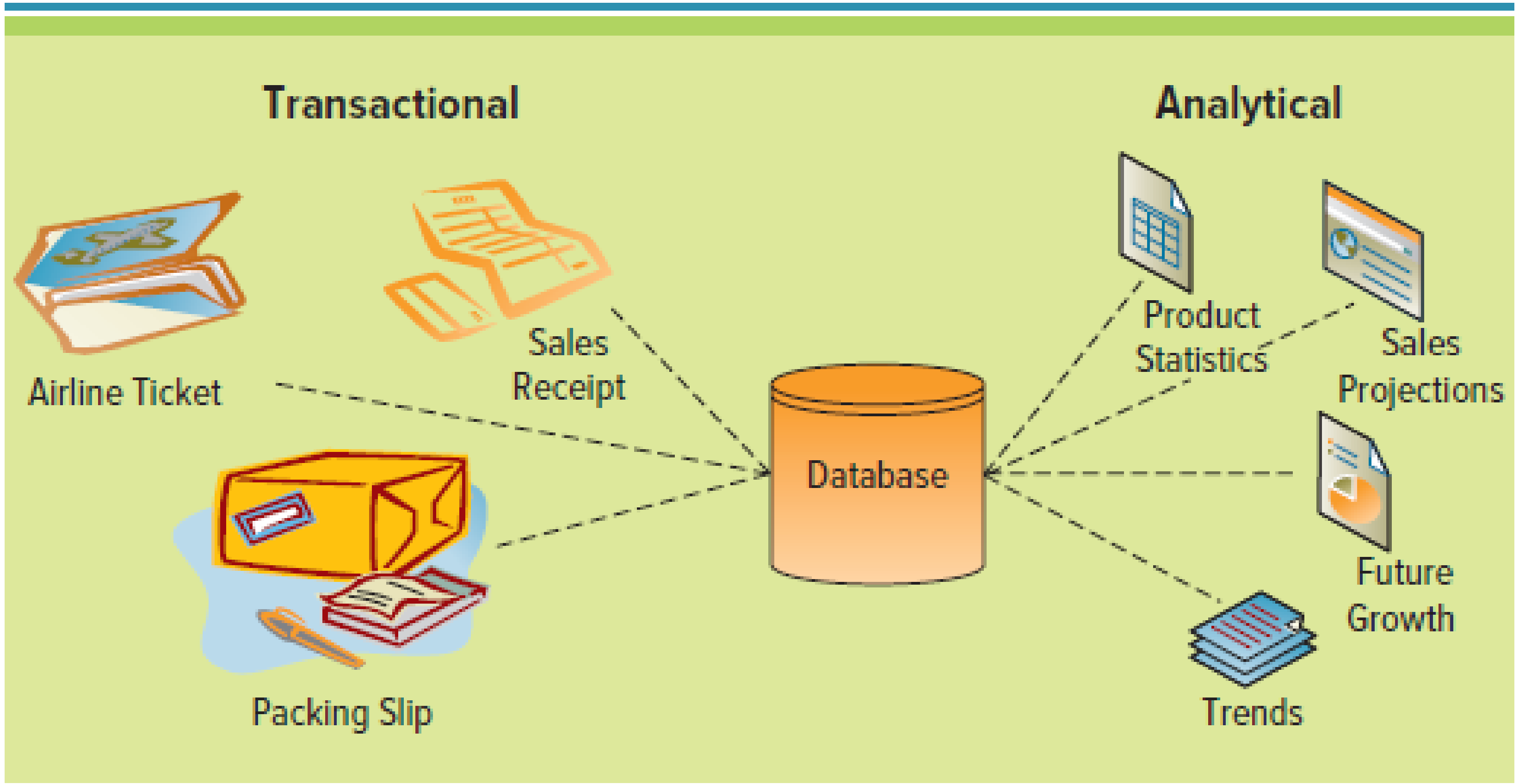
2. Organizations need to capture and store transactional data to perform operational tasks and repetitive decisions such as analyzing daily sales reports and production schedules to determine how much inventory to carry.

3. Consider Walmart, which handles more than 2 million customer transactions every hour, and Facebook, which keeps track of 800 million active users (along with their photos, friends, and web links). In addition, every time a cash register rings up a sale, a deposit or withdrawal is made from an ATM, or a receipt is given at the gas pump, the transactional data must be captured and stored.

Analytical Data Types

1. Analytical data encompasses all organizational data, and its primary purpose is to support the performance of managerial analysis tasks.
2. Analytical data is useful when making important decisions such as whether the organization should build a new manufacturing plant or hire additional sales personnel.
3. Analytical data makes it possible to do many things that previously were difficult to accomplish, such as spot business trends, prevent diseases, and fight crime. For example, credit card companies crunch through billions of transactional purchase records to identify fraudulent activity. Indicators such as charges in a foreign country or consecutive purchases of gasoline send a red flag highlighting potential fraudulent activity.

Transactional and Analytical Data Types

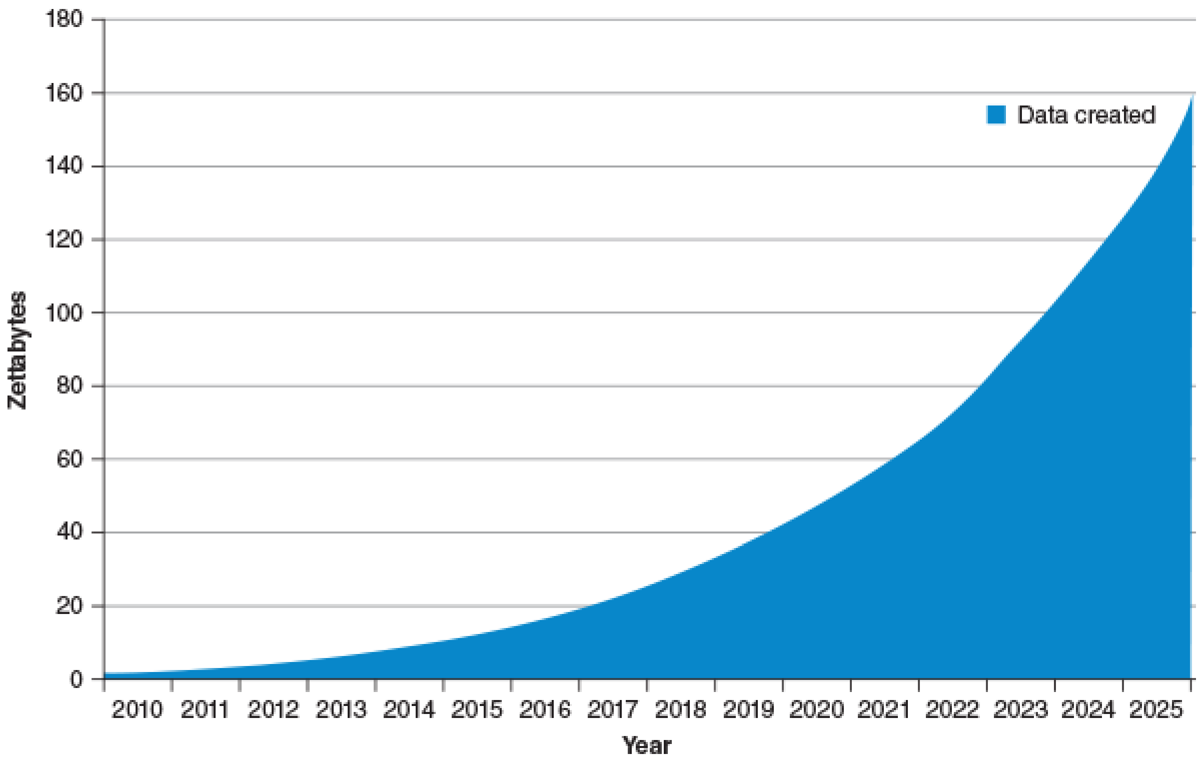


Five Characteristics Of Big Data

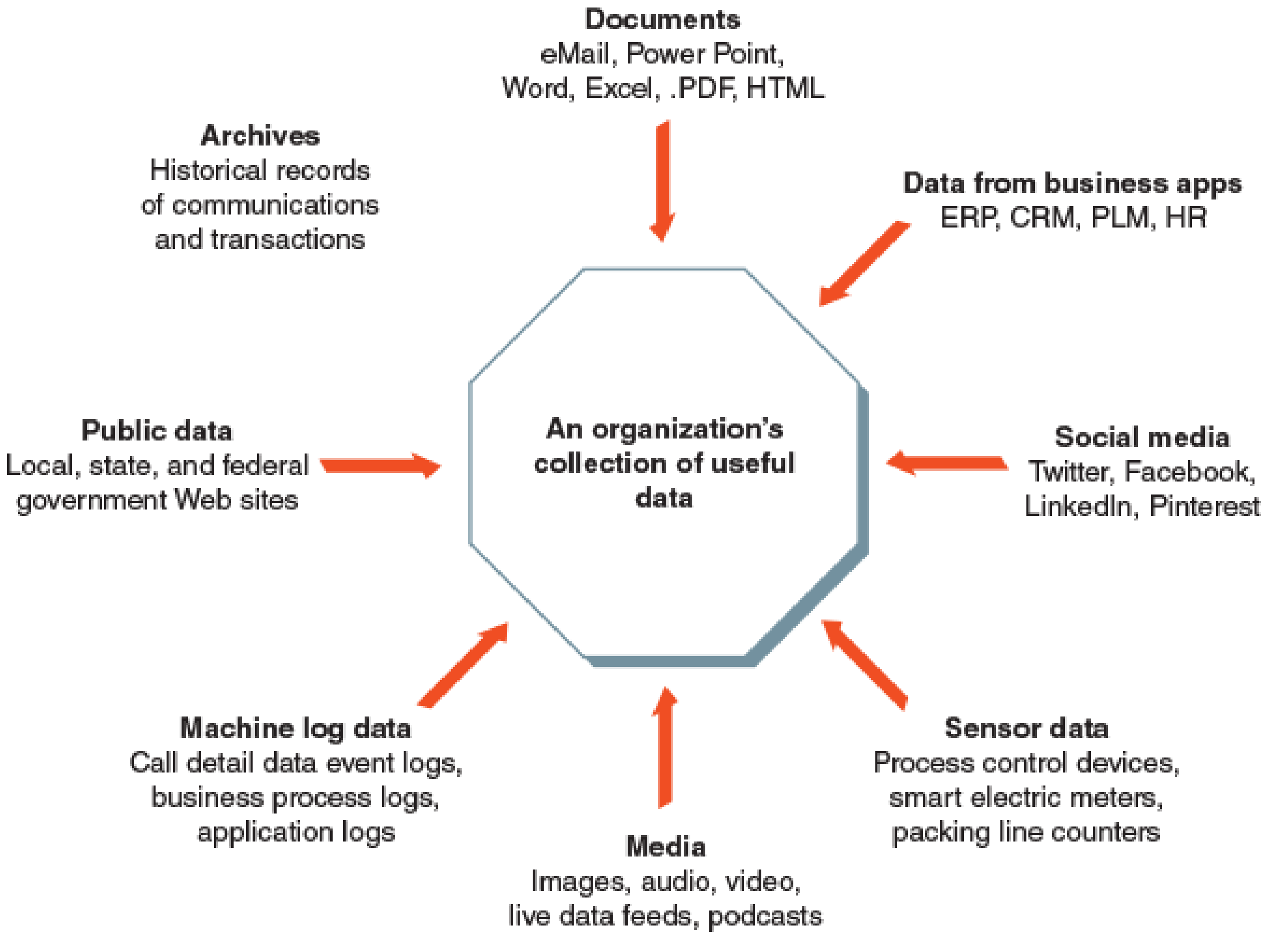
1. Volume refers to Zettabyte (one zettabyte equals one trillion gigabytes).
2. Velocity refers to the rate at which new data is being generated.
3. Value refers to the worth of the data in decision making.
4. Variety refers to Data today comes in a variety of formats. Some of the data is what computer scientists call structured data making.
5. Veracity refers to a measure of the quality of the data

Data Volume

Figure 6.1 Increase in Size of the Global Datasphere



Sources of Big Data



Documents
eMail, Power Point,
Word, Excel, .PDF, HTML

Archives
Historical records
of communications
and transactions

Data from business apps
ERP, CRM, PLM, HR

Public data
Local, state, and federal
government Web sites

An organization's collection of useful data

Social media
Twitter, Facebook,
LinkedIn, Pinterest

Machine log data
Call detail data event logs,
business process logs,
application logs

Sensor data
Process control devices,
smart electric meters,
packing line counters

Media
Images, audio, video,
live data feeds, podcasts

Big Data Uses

Retail organizations monitor social networks such as Facebook, Google, LinkedIn, Twitter, and Yahoo to engage brand advocates\

Identify brand adversaries (and attempt to reverse their negative opinions), and even enable passionate customers to sell their products.

Advertising and marketing agencies track comments on social media to understand consumers' responsiveness to ads, campaigns, and promotions.

Hospitals analyze medical data and patient records to try to identify patients likely to need readmission within a few months of discharge, with the goal of engaging with those patients to prevent another expensive hospital stay.

Consumer product companies monitor social networks to gain insight into customer behavior, likes and dislikes, and product perception to identify necessary changes to their products, services, and advertising.

Financial services organizations use data from customer interactions to identify customers who are likely to be attracted to increasingly targeted and sophisticated offers.

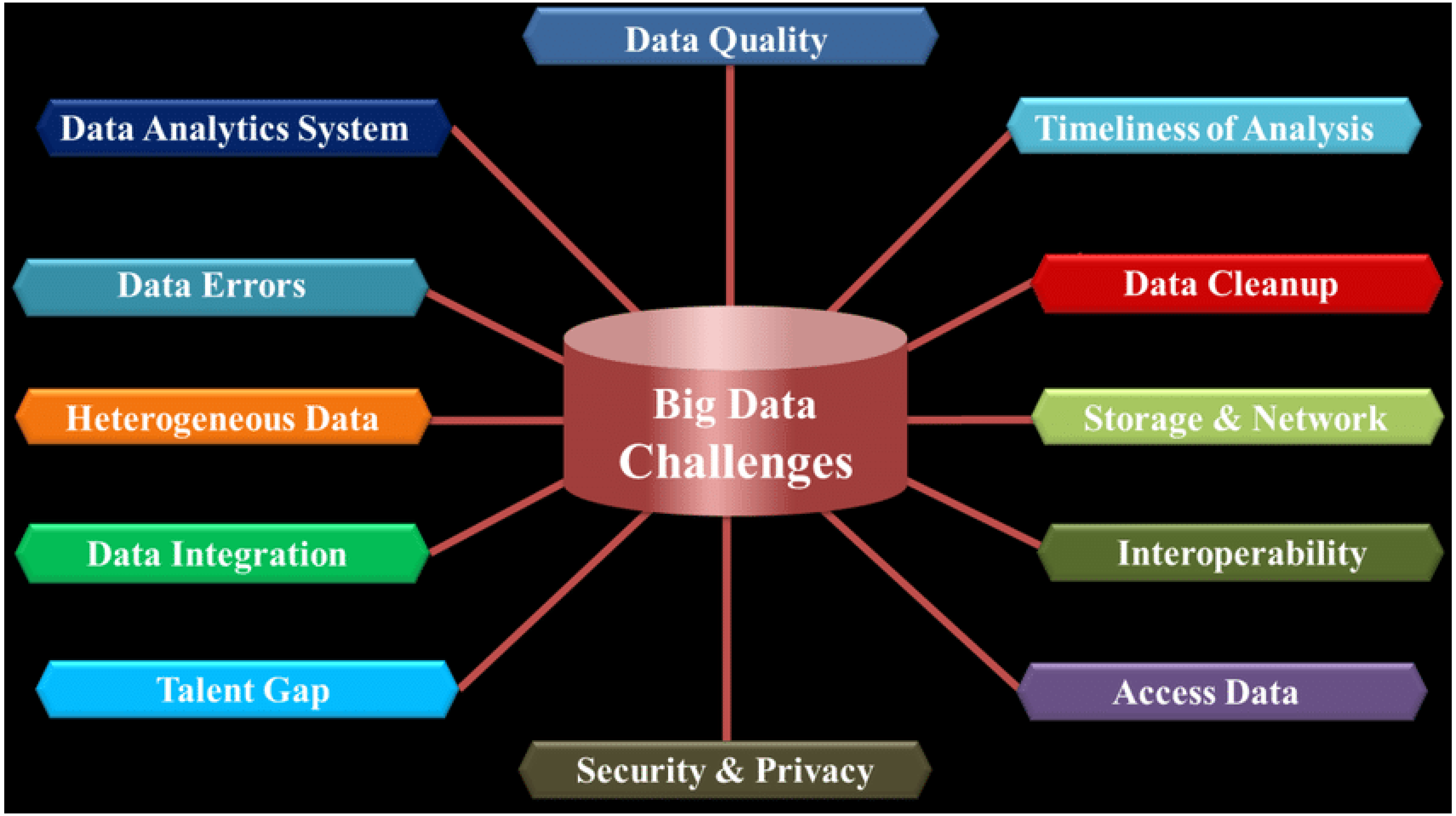
Manufacturers analyze minute vibration data from their equipment, which changes slightly as it wears down, to predict the optimal time to perform maintenance or replace the equipment to avoid expensive repairs or potentially catastrophic failure.

Five key challenges associated with big data

1. How to choose what subset of data to store.
2. Where and how to store the data?
3. How to find those nuggets of data that are relevant to the decision making at hand.
4. How to derive value from the relevant data.
5. How to identify which data needs to be protected from unauthorized access.

Case Study: Big Data Challenges

.



Source: https://www.researchgate.net/figure/Big-Data-challenges-in-Smart-city_fig1_341756095

Group Discussion

You need to work in a group of 4 students and discuss the following:

1. List five big data industries.
2. Which are the top 5 sources of big data?
3. What are the fastest growing sources of big data?
4. Why big data is important?

Data Warehouse
Data Mart and data Lake

1. A **data warehouse** is a large database that holds business information from many sources in the enterprise, covering all aspects of the company's processes, products, and customers.

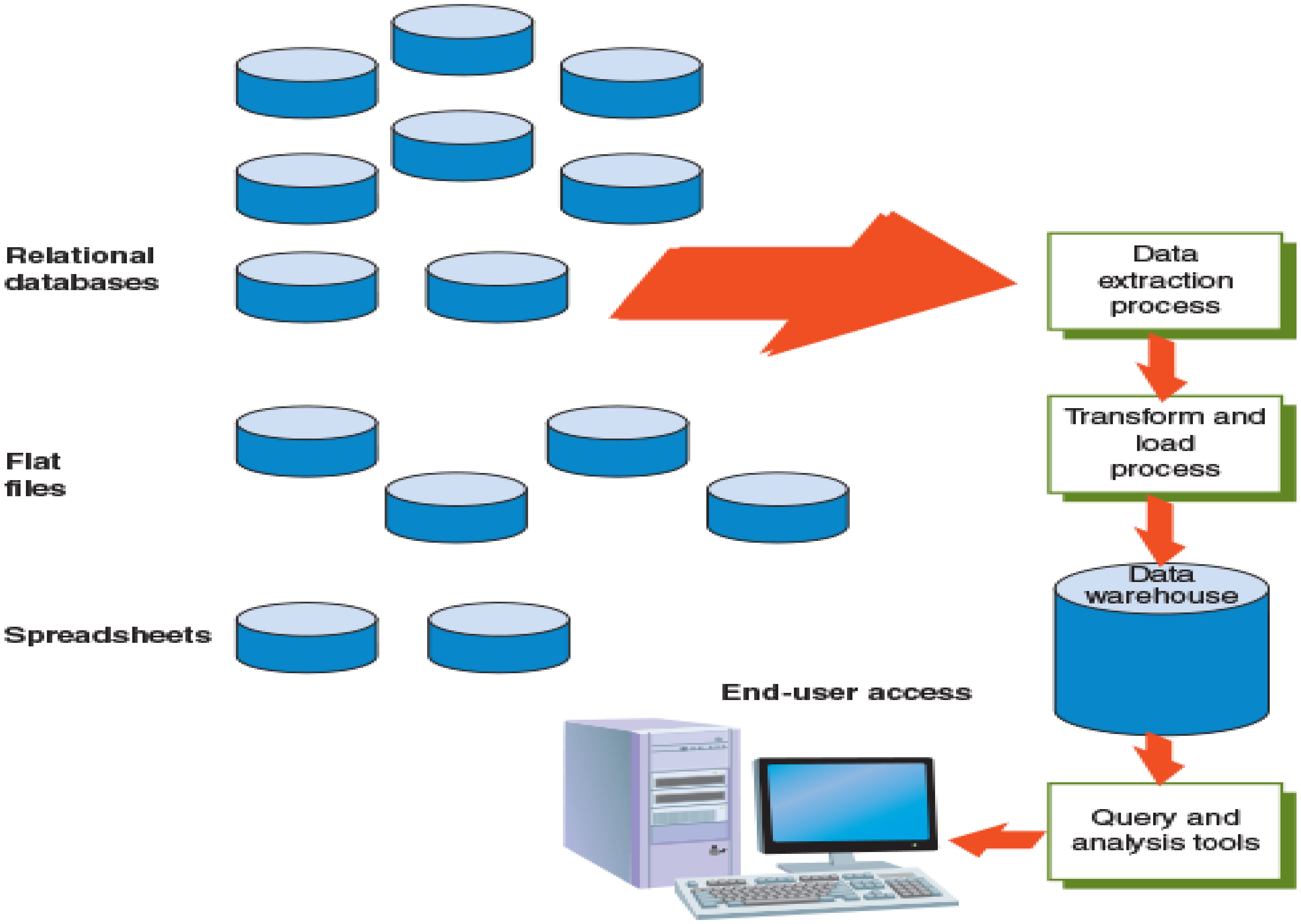
2. A **data mart** is a subset of a data warehouse. Data marts bring the data warehouse concept—lots of data from many sources—to small- and medium-sized businesses and to departments within larger companies.

3. A **data lake** takes a "store everything" approach to big data, saving all the data in its raw and unaltered form. The raw data residing in a data lake is available when users decide just how they want to use the data to glean new insights.

Data Warehouse

1. This data is used by people across the organization to support various processes and decision making.
2. The data in a data warehouse is historical data often going back 5 years or more.
3. The data can be analyzed in many ways. For example, data warehouses allow users to “drill down” to get greater detail or “roll up” to generate aggregate or summary reports.
4. The primary purpose is to relate information in innovative ways and help managers and executives make better decisions.

Data Warehouse



Data Mart

1. Data marts bring the data warehouse concept—lots of data from many sources—to small- and medium-sized businesses and to departments within larger companies.
2. Rather than store all enterprise data in one monolithic database, data marts contain a subset of the data for a single aspect of a company's business—for example, finance, inventory, or personnel.

Data Lakes

1. The raw data residing in a data lake is available when users decide just how they want to use the data to glean new insights.
2. Only when the data is accessed for a specific analysis is it extracted from the data lake, classified, organized, edited, or transformed. Thus, a data lake serves as the definitive source of data in its original, unaltered form.
3. Its contents can include business transactions, clickstream data, sensor data, server logs, social media, videos, and more.

Group Discussion

In the same group, discuss when an organization would use a data warehouse and a data lake?

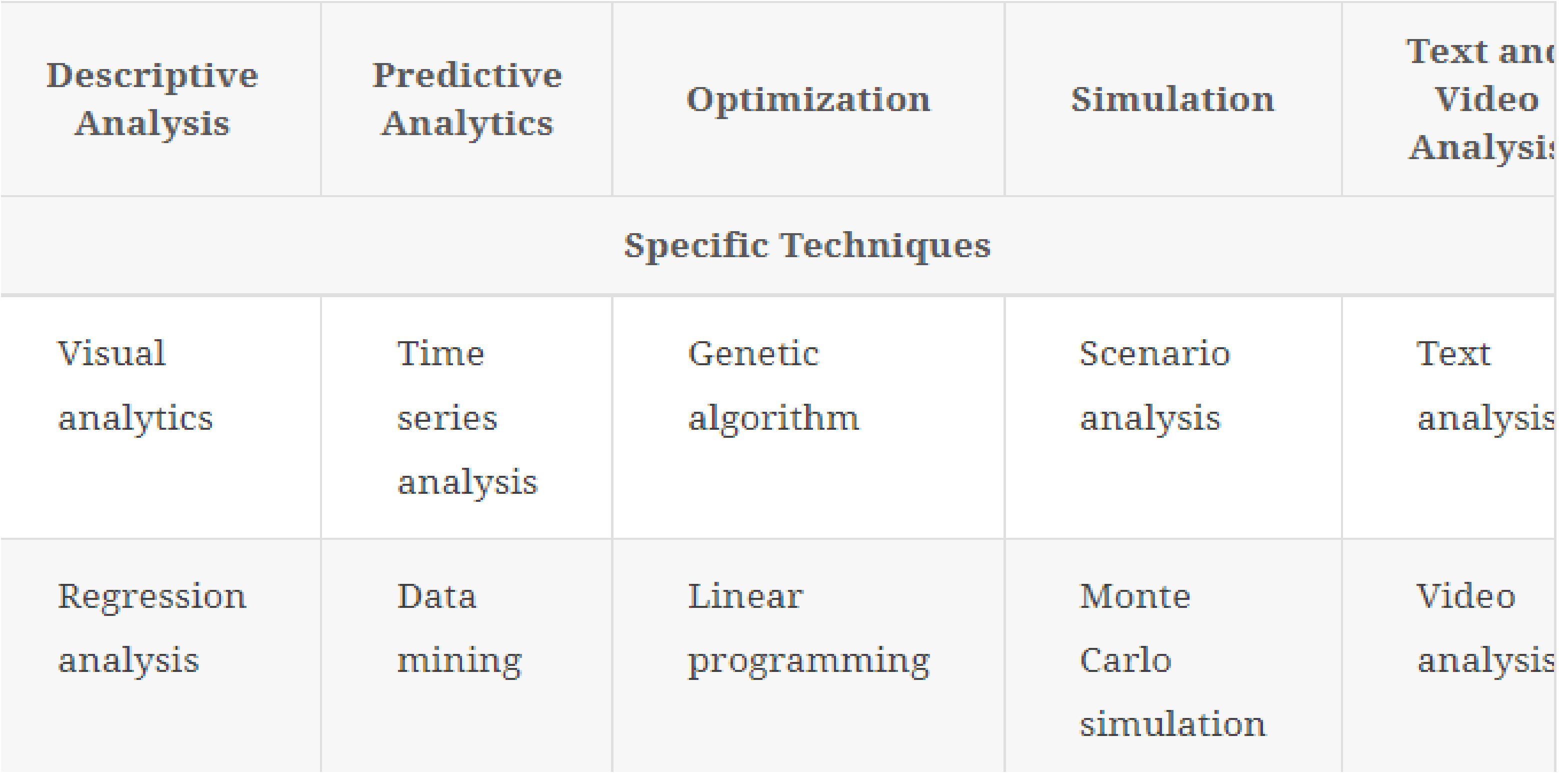
Three key organizational components

1. A solid data management program, including data governance.
2. Creative data scientists.
3. The success of a BI and analytics program, the management team within an organization must have a strong commitment to data-driven decision making.

Five Categories Of Business Intelligence

1. Descriptive analysis.
2. Predictive analysis.
3. Optimization
4. Simulation.
5. Text and video analysis.

General Categories of BI/Analytic Techniques



Descriptive Analysis	Predictive Analytics	Optimization	Simulation	Text and Video Analysis
Specific Techniques				
Visual analytics	Time series analysis	Genetic algorithm	Scenario analysis	Text analysis
Regression analysis	Data mining	Linear programming	Monte Carlo simulation	Video analysis

Descriptive Analysis

1. Descriptive analysis is a preliminary data processing stage used to identify patterns in the data and answer questions about who, what, where, when, and to what extent.
2. It is used to provide information about what happened and why. You might see, for example, an increase in a stock price following a series of positive tweets on Twitter by popular market analysts.
3. There are many descriptive analysis techniques. We will cover two: visual analytics and regression analysis.

Visual Analysis

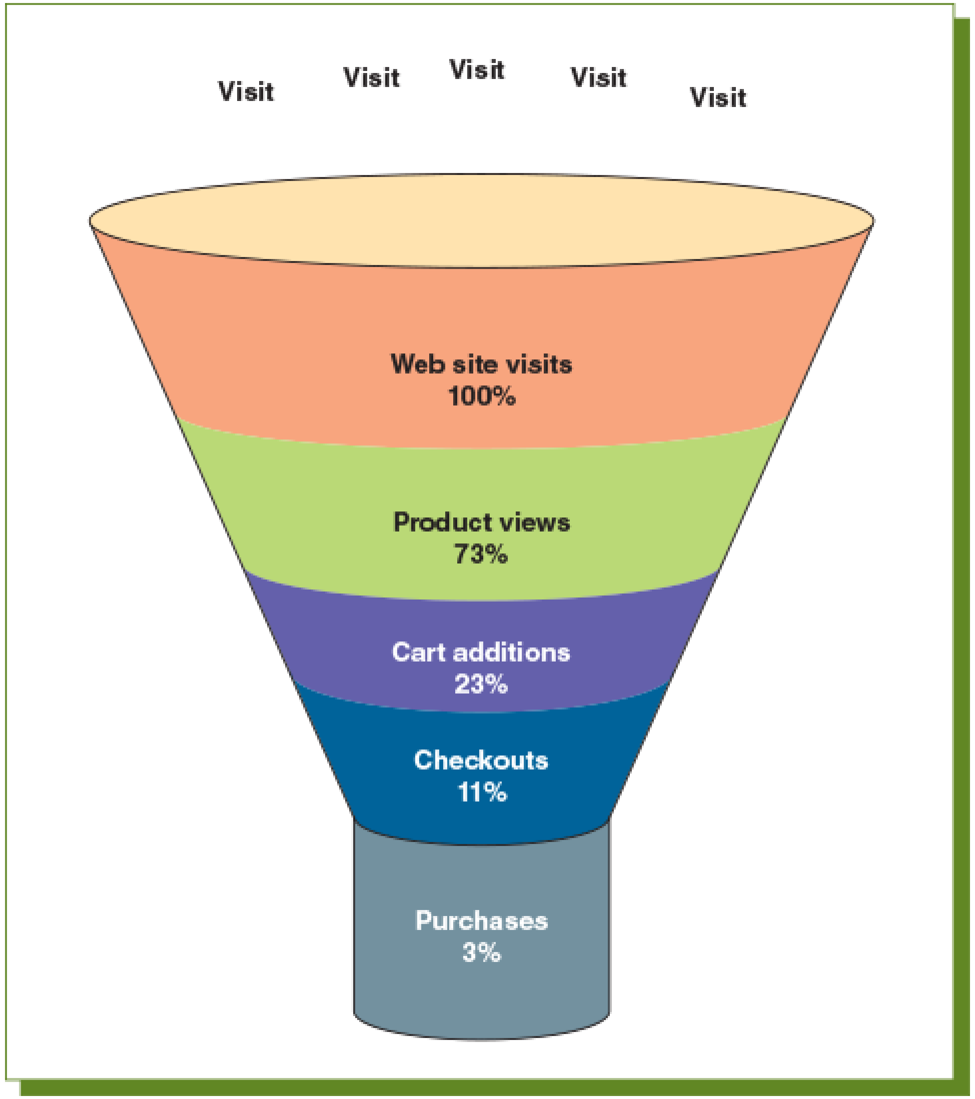
1. Visual analytics is the presentation of data in a pictorial or graphical format.
2. The human brain works such that most people are better able to see significant trends, patterns, and relationships in data that is presented in a graphical format rather than in tabular reports and spreadsheets.
3. As a result, decision makers welcome data visualization software that presents analytical results visually. In addition, representing data in visual form is a recognized technique to bring immediate impact to dull and boring numbers.
4. A wide array of tools and techniques are available for creating visual representations that can immediately reveal otherwise difficult-to-perceive patterns or relationships in the underlying data.

A Conversion Funnel

1. A conversion funnel is a graphical representation that summarizes the steps a consumer takes in making the decision to buy your product and become a customer.

2. It provides a visual representation of the conversion data between each step and enables decision makers to see what steps are causing customers confusion or trouble.

A Conversion Funnel



Regression Analysis

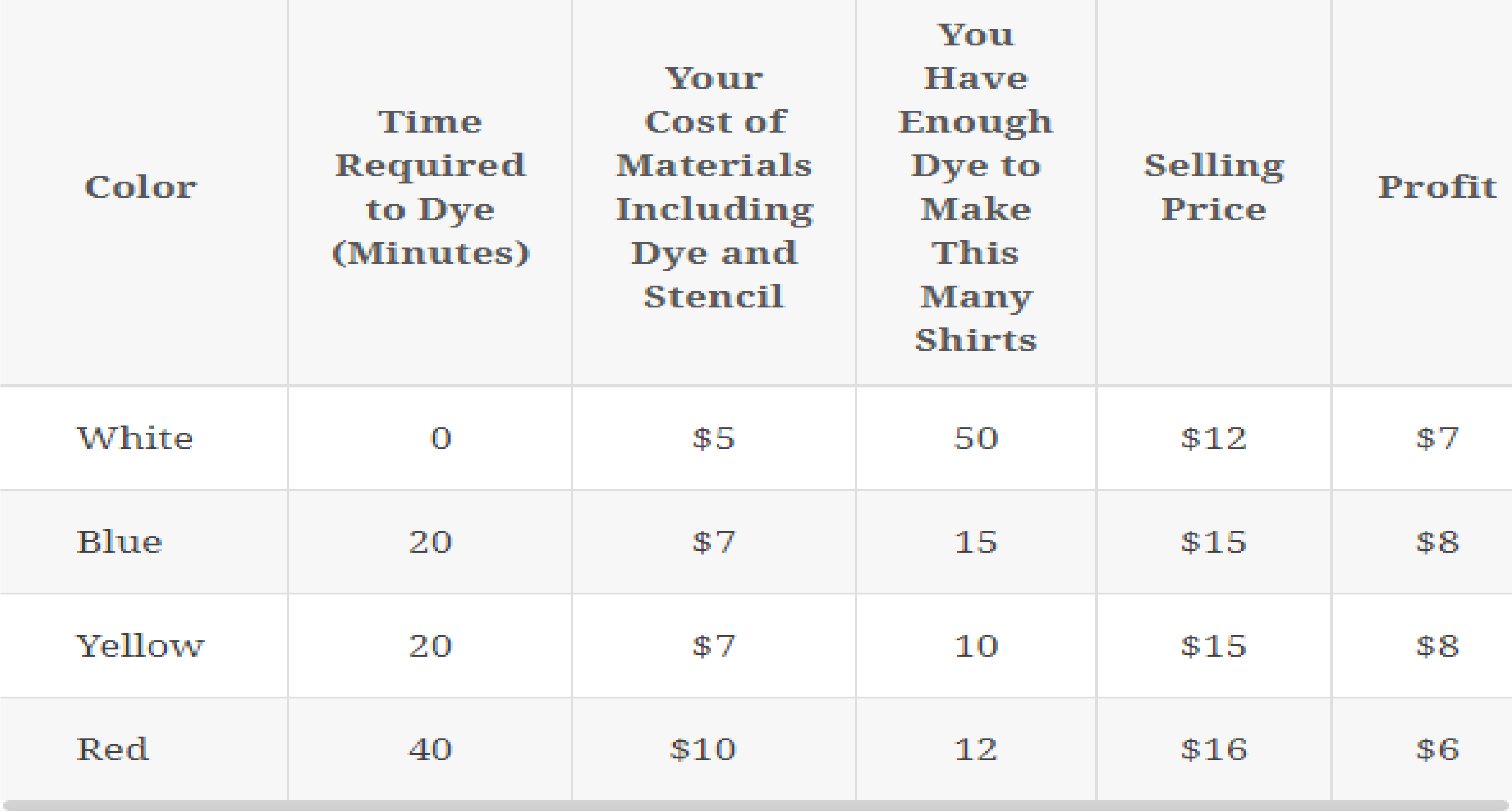
1. Regression analysis involves determining the relationship between a dependent variable (y) and one or more independent variables.
2. It is a proven method for determining which variables have an impact on the dependent variable

Case Study

Use one of the BI/analytics techniques to find the optimal solution to this problem. You make custom T-shirts with inspirational sayings on them. You just found out about a community flea market sale that is starting tomorrow. You have just 8 hours to prepare product for this sale. You start with a plain white T-shirt. This is your most popular color. But you can dye the white T-shirt blue, yellow, or red—but only one shirt at a time. Your current inventory is 50 white T-shirts and you have enough dye to make **12 red shirts, 10 yellow shirts, and 15 blue shirts.**

You take the various color shirts to the sale and then stencil on an inspirational saying—whatever the customer wants up to **35 characters**. Based on experience, you know that at a sale like this, you will be able to sell all **50 shirts**. Use the data in the table on the next page to determine how many shirts of each color you should bring to the sale to maximize your profits

Case Study



Color	Time Required to Dye (Minutes)	Your Cost of Materials Including Dye and Stencil	You Have Enough Dye to Make This Many Shirts	Selling Price	Profit
White	0	$5	50	$12	$7
Blue	20	$7	15	$15	$8
Yellow	20	$7	10	$15	$8
Red	40	$10	12	$16	$6

Quiz

1. _________ is a measure of the quality of big data.

2. The fact that big data comes in many formats and may be structured or unstructured is an indicator of its __________

3. An enterprise system used for the analysis and reporting of structured and semi-structured data from multiple sources. Answer: Data Warehouse/ Data lake?
4. It is a centralized repository designed to store, process, and secure large amounts of structured, semi structured and unstructured data. Answer: Data Warehouse/ Data lake?

Any Questions?