CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

# APEX INSTITUTE OF TECHNOLOGY

## DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

# Introduction to Data Science (21CST-292)

**Faculty:** Dr. Jitender Kaushal (E14621)

**Lecture -**

DISCOVER . LEARN . EMPOWER

## Business Profiles in Big Data

# Introduction to Data Science: Course Objectives

## COURSE OBJECTIVES

The Course aims to:

- This course brings together several key big data problems and solutions.
- To recognize the key concepts of Extraction, Transformation and Loading
- To prepare a sample project in Hadoop Environment

# COURSE OUTCOMES

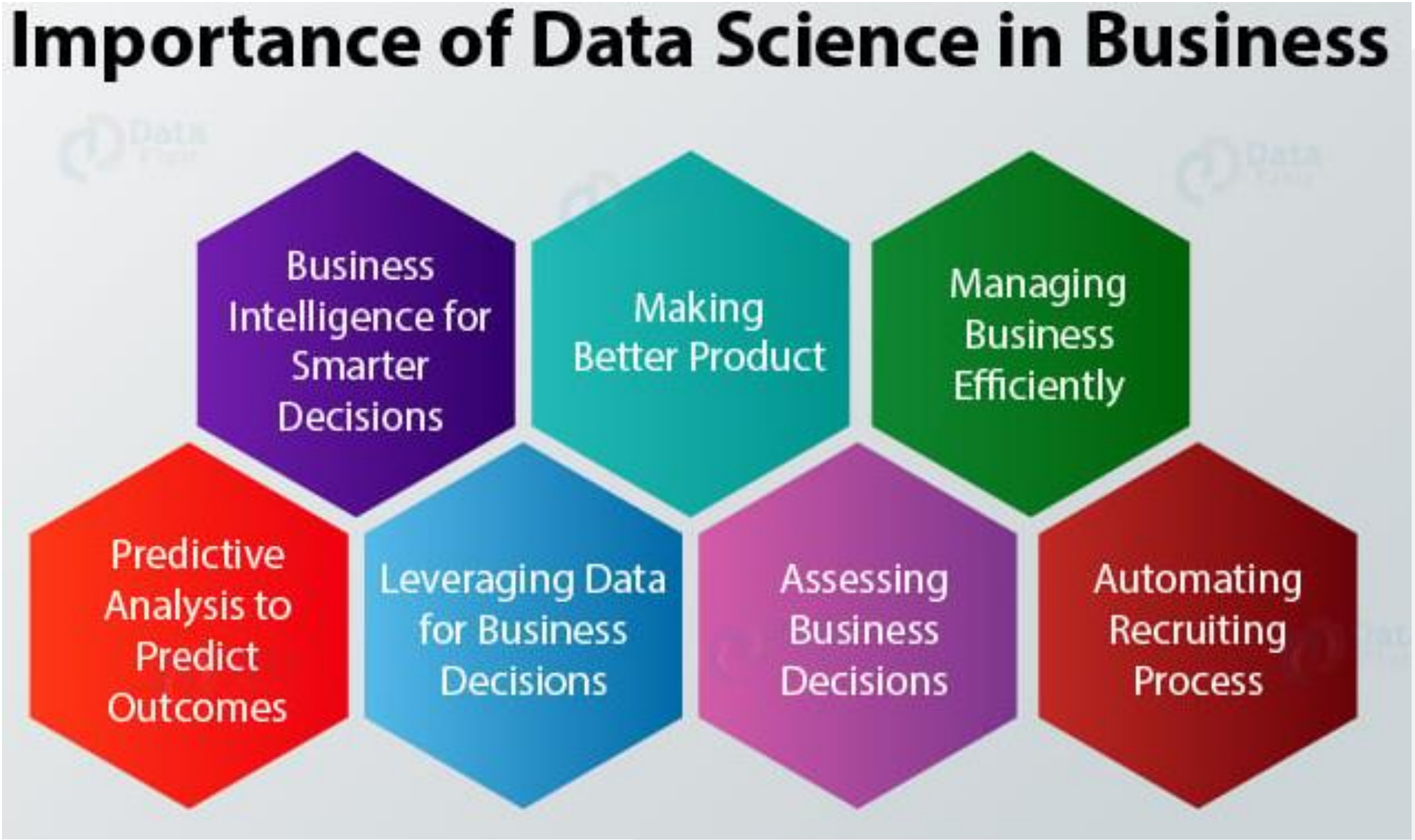
On completion of this course, the students shall be able to:-

| | |
|---|---|
| **CO2** | Describe big data processing merits in data understanding. |

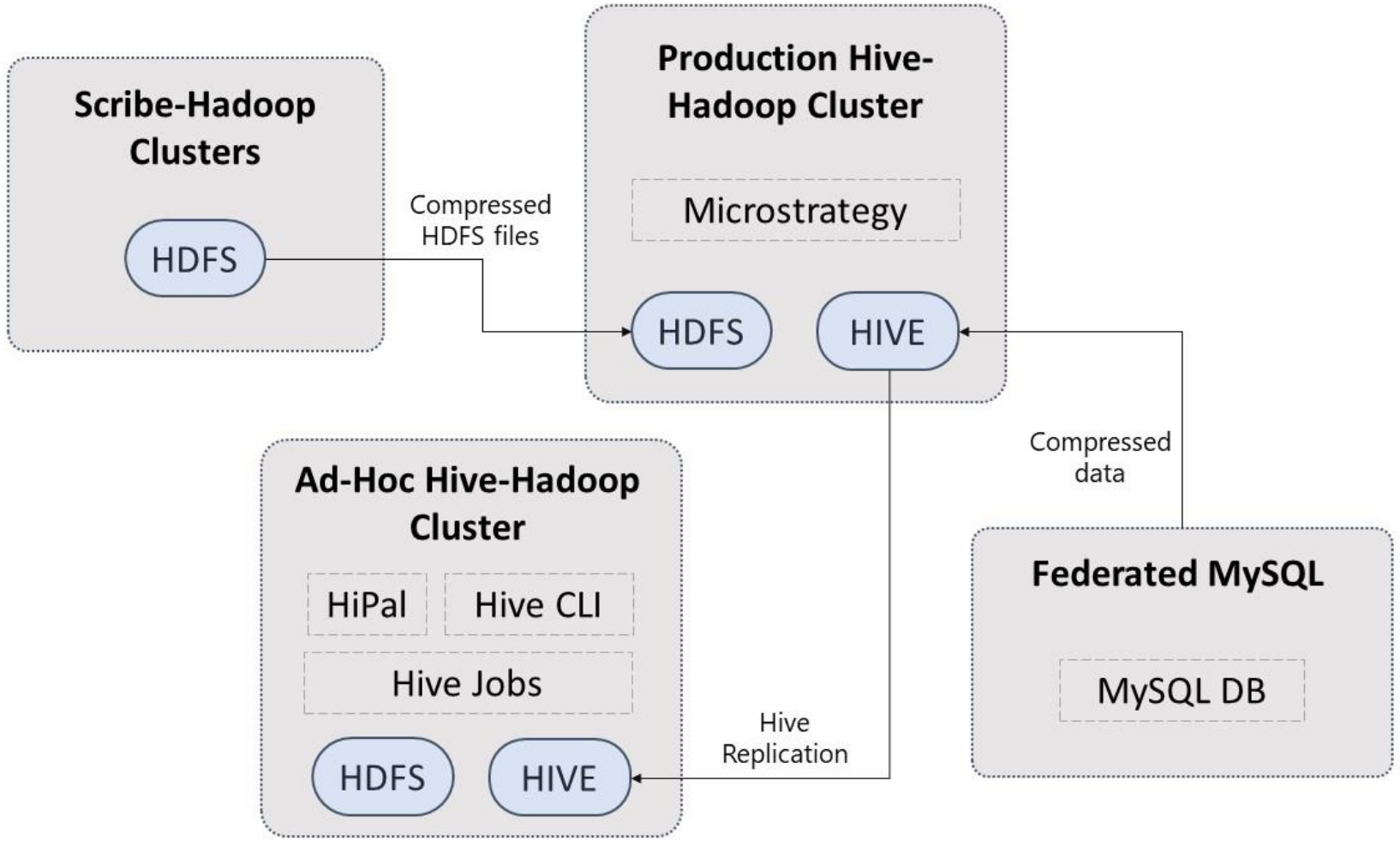
# Implementations of Data Science in Businesses

- Many popular companies are using **Data Science for easing their regular processes**.
- We will explore a use case of **Walmart** to see how it is **utilizing data to optimize its supply chain and make better decisions**.
- We will also learn the core **implementations of Data Science in businesses.**
- Businesses today have become **data-centric**. This means that the businesses of the world utilize data **to make decisions and grow** their company in the direction that the data provides.

# Importance of Data Science in Business



# Relationship Between Facebook and Big Data



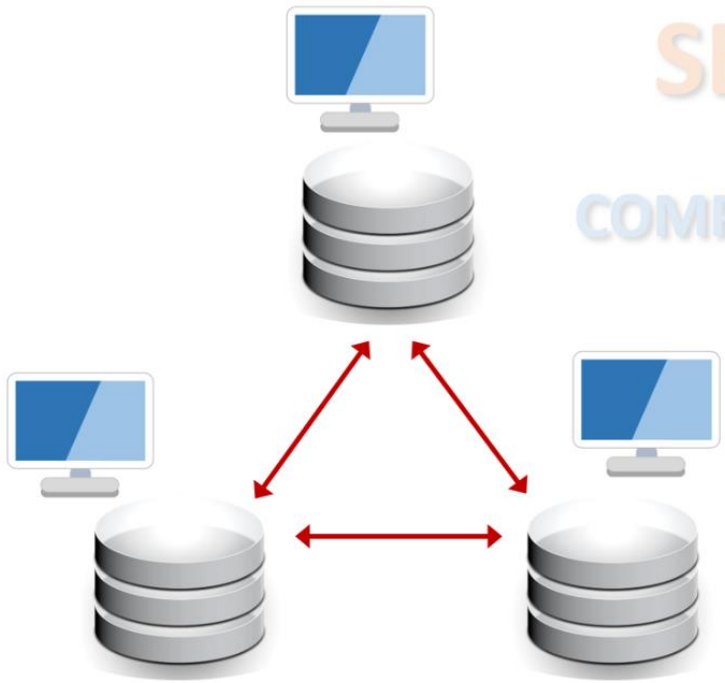

HDFS: Hadoop Distributed File System, Hive – A Petabyte Scale Data Warehouse using Hadoop, CLI (Command Line Interface)

Facebook uses Hadoop HDFS Architecture. **Facebook collects data from two sources:**

- User data is stored in the **federated MySQL layer**, and **web servers** produce event-based log data.
- Web server data is **gathered** and **sent to Scribe servers (***Scribe is a software tool developed by the USEPA's Environmental Response Team (ERT) to assist in the process of managing environmental data***)**, which run in **Hadoop clusters** (*A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment.*).
- The Hadoop Distributed File System (HDFS) **receives log data aggregated by the Scribe servers**.

- Periodically, HDFS data is **compressed before being sent to production Hive-Hadoop clusters** *(Hive allows users to read, write, and manage petabytes of data using SQL)* **for additional processing**.
- The **Production Hive-Hadoop cluster receives the Federated MySQL data, dumps, compresses and moves** it there.

[Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware.]



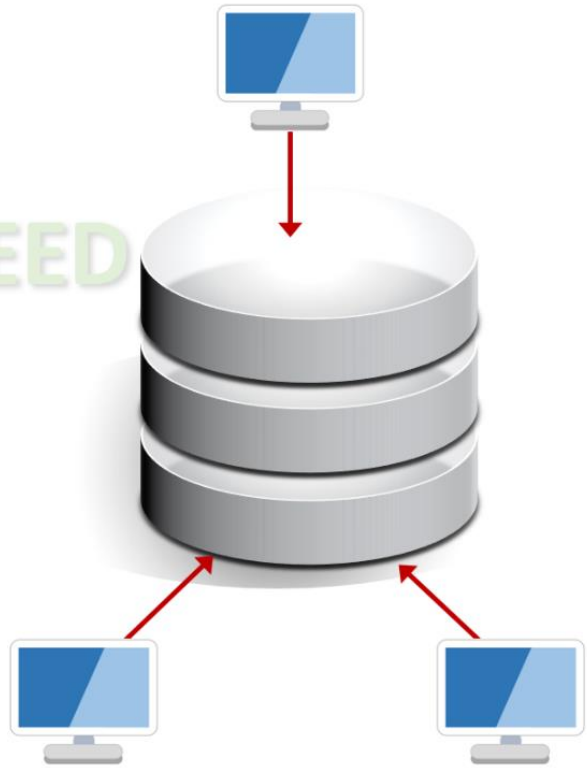
SECURITY

COMPLEXITY

SPEED

VS.

LOGISTICS



***federated***

**(no data exchange – analysis parameters optmized across the network)**

***centralized***

**(data combined into single database)**

CU CHANDIGARH UNIVERSITY

hadoop.apache.org

Apache Hadoop | Download | Documentation | Community | Development | Help | Apache Software Foundation

# Apache Hadoop

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Learn more » | Download » | Getting started »

## Latest news

## Modules

## Related projects

**Facebook uses two different clusters for data analysis:**

- The **Production Hive-Hadoop cluster** is used **to process tasks with severe deadlines**.
- The **Ad-hoc Hive-Hadoop cluster** does **lower-priority jobs and ad hoc analysis activities**.

  The Ad hoc cluster **receives data replication from the Production cluster**.

The **data analysis outcomes** are **saved to the Hadoop Hive cluster** or, **for Facebook users**, **to the MySQL tier**. A **graphical user interface (HiPal)** or a **command-line interface (Hive)** is used to specify **ad hoc analysis queries (Hive CLI)**. Facebook uses a **Python framework to execute (Database) and schedule periodic batch jobs** in the Production cluster.

Organizations are **leveraging** **Big Data analytics** to engage with customers by understanding their behaviour more precisely. **Some of the key takeaways from the article are –**

1. Facebook uses all the available **user information** with its **deep neural network models** and finds the **target audience for a particular advertisement**. This helps to serve the users' advertisements more insightfully.
2. Due to this, Facebook has emerged as one of the **toughest competitors for Google Search Engine and Youtube** in the digital marketing race.
3. With the humungous amount of user **data present with Facebook and continuously increasing**, there are still a lot of other use cases we might see in the future from Facebook.

# Data Science Case Study
# Walmart – Leveraging Data to Make Business Better

# Data Science Case Study
# Walmart – Leveraging Data to Make Business Better

- Walmart is the world's largest retailer. It is one of the many major industries that is leveraging Big Data to make the business more efficient. Walmart handles a plethora of customer data. A staggering amount of about 2.5 petabytes of data is collected from the customers every hour.

- This data is unstructured that is utilized through Hadoop and NoSQL. It tracks and monitors various factors that might affect the sales at Walmart stores.

Some of the ways in which Walmart is using data science are:

- Walmart is using data science to make **store checkouts more efficient**. There are certain times of the day where the checkouts can become crowdy. This makes it difficult for Walmart employees to manage customers during rush hours. However, with the help of predictive analytics, Walmart can analyze data and determine the best form of checkout for each store, that is, self-checkout and facilitated checkout.
- Walmart is using real-time analytics to **analyze the purchasing patterns** of the customers. This allows them to stock up on products that are in demand and also the products which will be in future demand based on several factors.

- Walmart is **managing supply chain and logistics** with the help of data science. It manages its inventory and analyzes the rate of its depletion, thereby taking the necessary steps to mitigate it through efficient logistics. Walmart also analyzes the transportation lanes for the company's trucks to follow. It specifies an optimized route using data science, thereby reducing the cost and time.

- Walmart is personalizing the shopping experience by **analyzing the preferences and behaviour** of the customers. Using data science, it tracks the purchasing patterns of the customers and recommends them further products and discounts to improve their shopping experience.

# 1. Business Intelligence for Making Smarter Decisions

- **Traditional Business Intelligence was more descriptive and static in nature**. However, with the addition of data science, it has transformed itself to become a **more dynamic field**. Data Science has **rendered Business Intelligence to incorporate a wide range of business operations**.
- **With the massive increase in the volume of data, businesses need data scientists to analyze and derive meaningful insights from the data.**
- The meaningful insights will help the data science companies **to analyze information at a large scale and gain necessary decision-making strategies.** The process of decision-making involves **the evaluation and assessment** of various factors involved in it. Decision Making is a four-step process:

# Decision Making Process

# 2. Making Better Products

- The process involves the **analysis of customer reviews to find the best fit for the products**. This analysis is carried out with the advanced analytical tools of Data Science.
- Furthermore, **industries utilize the current market trends to devise a product for the masses**. These market trends provide businesses with clues about the **current need for the product**. Businesses evolve with **innovation**.
- **For example** – Airbnb uses data science to improve its services. The data generated by the customers is processed and analyzed. It is then used by Airbnb **to address the requirements and offers premier facilities to its customers.**

# 3. Managing Businesses Efficiently

- With Data Science, **businesses can manage themselves more efficiently**. Both large-scale businesses and small startups can **benefit from data science in order to grow further.**
- Data Scientists help **to analyze the health of the businesses**. With data science, companies can **predict the success rate of their strategies**. **Data Scientists are responsible for turning raw data into cooked data**.
- Based on this, the business can take important measures **to quantify and evaluate its performance and take appropriate management steps**. It can also help the managers **to analyze and determine the potential candidates for the business**.
- **For example –** Data Science can be used **to monitor the performance of employees**. Using this, **managers can analyze the contributions made by the employees and determine when they should be promoted, manage their perks**, etc.

# 4. Predictive Analytics to Predict Outcomes

- **Predictive analytics is the most important part of businesses**. With the advent of advanced predictive tools and technologies, companies have expanded their **capability to deal with diverse forms of data.**
- In formal terms, predictive analytics is the **statistical analysis of data** that involves several machine learning algorithms **for predicting the future outcome** using **historical data**. There are several predictive analytics tools like SAS, IBM SPSS, SAP, etc.
- Predictive Analytics has its **own specific implementation based on the type of industry**. However, regardless of that, **it shares a common role in predicting future events**.

# 5. Leveraging Data for Business Decisions

- **In the past**, many businesses would take **poor decisions due to the lack of surveys or sole reliance on 'gut feelings'**. It would **result in some disastrous decisions leading to losses of millions**.

- However, with the presence of a plethora of data and necessary data tools, it is now possible for the data industries to make **calculated data-driven decisions.**

- Furthermore, **business decisions can be made with the help of powerful tools that can not only process data faster but also provide accurate results**.

# 6. Assessing Business Decisions

- **After making decisions through the forecast of future occurrences, it is a requirement for the companies to assess them**. This is possible through several **hypothesis testing tools**.

- After implementing the decisions, businesses should understand how these **decisions affect their performance and growth**. If the decision leads to any negative factor, then they should analyze it and eliminate the problem that is slowing down their performance.

- Furthermore, in order to assess future growth through the present course of actions, businesses can make profits considerably with the help of data science.

# 7. Automating Recruitment Processes

- **Data Science has played a key role in bringing automation to several industries.** It has taken away mundane and repetitive jobs. One such job is that of resume screening. Every day, **companies have to deal with hordes of applicants' resumes**.

- Data science technologies like image recognition are able to convert the visual information from the resume into a digital format. It then processes the data using various **analytical algorithms like clustering and classification** to churn out the right candidate for the job.

- Furthermore, businesses study the right trends and analyze potential applicants for the job. This allows them to reach out to candidates and have an in-depth insight into the job-seeker market.

# Summary

- In the end, we understand how data science plays an important role in businesses. We realized how data science is being used for business intelligence, for improving products, for increasing the management capabilities of companies and for predictive analytics.

- We also went through a use case of Walmart and how they utilize the data science to increase their efficiency.

# FAQs

- Explain the importance of data science in businesses.
- Discuss the different factors used for setting up large-scale business profiles.
- Give examples of the use of big data in large-scale industries.

# References


- RamezElmasri and Shamkant B. Navathe, “Fundamentals of Database System”, The Benjamin / Cummings Publishing Co.
- https://www.geeksforgeeks.org/
- Sinan Ozdemir, “Principles of data science”, Packt> Publications, 2016..

# THANK YOU