Module-2
Data Science Topics

In the first lesson in this module, you gain insight into the impact of big data on various aspects of society, from business operations to sports, and develop an understanding of key attributes and challenges associated with big data. You will learn about the big data fundamentals, how data scientists use the cloud to handle big data, and the data mining process. Lesson two delves into machine learning and deep learning and the relationship of artificial intelligence to data science.

Learning Objectives

- Define Big Data and its distinguishing characteristics, such as velocity, volume, veracity, and value.
- Describe how Hadoop and other big data tools, combined with distributed computing power, trigger digital transformation.
- List some of the skills required to be a data scientist and analyze big data.
- Describe the five essential cloud computing characteristics
- Explain what data mining is.
- Summarize the importance of establishing goals, data selection, preprocessing, transformation, and data storage in preparation for data mining.
- Explain the difference between deep learning and machine learning.
- Describe regression and how it might be used to predict market behaviour and trend analysis.
- Describe generative AI

Big Data and Data Mining

Lesson Overview: Big Data and Data Mining

In the 'Big Data and Data Mining' lesson, delve into the world of digital transformation driven by Big Data. Explore Cloud Computing's role, foundational Big Data concepts, tools like Hadoop and Spark, and gain insights into Data Mining techniques for informed decision-making.

Asset name and type	Description
"How Big Data is Driving Digital Transformation" video	Explore the impact of Big Data on digital transformation.
"Introduction to Cloud" video	Get introduced to the fundamentals of cloud computing.
"Cloud for Data Science" video	Learn how the cloud is relevant in the field of data science.

"Foundations of Big Data" video	Build your understanding of key Big Data concepts.
"Data Scientists at New York University" video	Discover the work of data scientists at New York University.
"What is Hadoop?" video	Understand the significance of Hadoop in Big Data processing.
"Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark" video	Dive into the tools used for Big Data processing.
"Data Mining" reading	Read an excerpt on data mining explore the principles and concepts of data mining.
Practice quiz	Take a practice quiz to evaluate how well you've understood the material presented in this lesson.
Glossary	Use this glossary of terms to review the terminology presented in this lesson.
Graded quiz	Test your knowledge from this lesson by taking the graded quiz.

How Big Data is Driving Digital Transformation

1. What is Digital Transformation?

- It involves:
  - Updating existing processes.
  - Creating new operations using new technologies.
- Goal:
  - Harness benefits of digital technology.
  - Integrate digital technology into *all* areas of an organization.
  - Achieve *fundamental changes* in:
    - Operations
    - Value delivery to customers
- It is both an:
  - **Organizational change**
  - **Cultural change**

---

2. Key Driver: Data Science and Big Data

- Massive data availability enables deep analysis.
- Competitive advantage comes from insights derived through analytics.
- Industries are undergoing digital transformation due to:
  - Volume
  - Variety
  - Velocity of data

---

3. Real-World Examples of Digital Transformation

a. Netflix

- Transformed from:
  - DVD postal rental service
- To:
  - Leading global video streaming platform

- Leveraged:
    - Data analytics to understand user preferences
    - Personalized content delivery

b. Houston Rockets (NBA Team)

- **2018**: One of 4 NBA teams to install **video tracking system**.
- Collected:
    - Raw data from games using overhead cameras.
- Analysis revealed:
    - Best scoring opportunities:
        - **Two-point dunks** (inside the paint)
        - **Three-point shots**
        - Not mid-range two-point shots.
- Strategic change:
    - Increased 3-point attempts drastically.
- Result:
    - 2017–18: Most 3-point shots in NBA history.
    - Won more games than any other team that season.
- Industry impact:
    - Changed how basketball is played across the NBA.
    - Showed that **Big Data can transform an entire industry**.

c. Lufthansa

- Used **customer data** to:
    - Improve service
    - Enhance customer experience

---

4. Nature of Digital Transformation

- Not merely digitizing existing processes.
- Involves:
    - In-depth business analysis

- Process improvement
- Integration of data science into workflows

---

5. Organizational Impact

- Affects:
  - Data strategy
  - Employee roles and engagement
  - Customer interactions
  - Organizational **culture**
- **Top-down initiative**:
  - Must be led by:
    - **CEO (Chief Executive Officer)**
    - **CIO (Chief Information Officer)**
    - **CDO (Chief Data Officer)**
- Also requires:
  - Support from executives in:
    - Budgeting
    - Personnel decisions
    - Daily operations
- Involves **entire organization**:
  - Cross-functional effort.
  - Everyone must align with the transformation.

---

6. Challenges

- Requires:
  - New mindset
  - Organizational willingness to change
- Success depends on:

- Strategic vision
- Data capability
- Cultural adaptation

📌 **Summary**

Digital transformation is a profound shift in how organizations operate by integrating modern digital technologies and data science into every aspect of their functioning. Big Data is a powerful catalyst, enabling organizations to discover new efficiencies, services, and strategic models. Real-world examples like Netflix, Lufthansa, and especially the Houston Rockets show how data insights can improve performance and reshape entire industries. For successful transformation, executive leadership and organisation-wide support are essential. The process challenges existing systems but is now crucial for long-term competitiveness.

📊 **Table: What We Learnt in the Video**

Aspect	Details
Definition	Digital transformation updates and redefines operations via new tech.
Core Drivers	Data Science and Big Data
Cultural Impact	Requires a mindset and structural change across the organization
Key Technologies Used	Video tracking, data analytics, and customer data platforms
Real-World Example: NBA	The Houston Rockets used data to revolutionize gameplay strategy
Real-World Example: Netflix	Shifted from DVD rentals to the data-driven streaming platform
Real-World Example: Lufthansa	Used customer data to improve service
Key Roles Required	CEO, CIO, CDO + other execs managing resources and priorities
Transformation Nature	Not duplication, but reinvention and optimization
Organizational Requirement	Total alignment and commitment
Outcome	Improved competitiveness, innovation, and customer value delivery

Introduction to Cloud

📌 What is Cloud Computing?

- **Definition**: Cloud computing (or "the cloud") delivers on-demand computing resources over the Internet on a pay-for-use basis.
- **Resources Provided**:
  - Networks
  - Servers
  - Storage
  - Applications
  - Services
  - Data centers
- **Access**: Cloud services are accessed over the Internet instead of being stored/used locally.

💻 Examples of Cloud Computing

- Using online web applications.
- Employees accessing secure business apps online.
- Storing personal files on platforms like:
  - Google Drive
  - OneDrive
  - Dropbox

🎯 Benefits to Users

- **Cost-effective**: No need to buy full software copies; pay monthly subscriptions instead.
- **Always Updated**: Access to latest versions of software.
- **Saves Local Storage**: Apps hosted online, freeing up disk space.
- **Collaboration**: Real-time collaboration and visibility into others' edits/updates.

☁️ Essential Characteristics of Cloud Computing (5)

1. **On-demand self-service**
   - Users can access resources (compute, storage, network) via a simple interface without human interaction.
2. **Broad network access**

- Resources are accessible via standard devices like:
  - Mobile phones
  - Tablets
  - Laptops
  - Workstations

3. **Resource pooling**
   - Cloud providers use a **multitenant model**.
   - Resources are dynamically assigned and reassigned according to demand.
   - Customers don't know the physical location of the resources.
   - Achieves economies of scale, which lowers costs.
4. **Rapid elasticity**
   - Resources can scale up or down according to demand.
   - Allows provisioning and releasing of resources elastically.
5. **Measured service**
   - Pay only for what you use.
   - Usage is monitored, measured, and transparently reported.

Cloud Deployment Models (3)

1. **Public Cloud**
   - Services offered over the Internet.
   - Infrastructure owned and managed by the cloud provider.
   - Resources are shared among multiple organizations.
2. **Private Cloud**
   - Infrastructure used exclusively by a single organization.
   - Can be on-premises or hosted by a third-party provider.
3. **Hybrid Cloud**
   - Combination of public and private clouds.

- Allows data and applications to move between the two environments seamlessly.

🧱 Cloud Service Models (3)

Based on the three layers of the computing stack:

1. **Infrastructure as a Service (IaaS)**
   - Access to physical computing resources: servers, storage, networking, data centre space.
   - Users don't manage or control the infrastructure, but can manage OS, storage, and deployed applications.
2. **Platform as a Service (PaaS)**
   - Access to tools and platforms (hardware + software) for the development and deployment of apps.
   - Developers can build and deliver applications without managing the underlying infrastructure.
3. **Software as a Service (SaaS)**
   - Centrally hosted software delivered on a subscription basis.
   - Also known as **on-demand software**.
   - End-users access applications via the Internet without needing to install or maintain them.

📚 Summary

- Cloud computing is **on-demand**, **Internet-based**, and **pay-per-use**.
- It enhances **cost-efficiency**, **scalability**, and **agility** in business operations.
- It includes:
  - **5 Essential Characteristics**: On-demand self-service, Broad network access, Resource pooling, Rapid elasticity, Measured service.
  - **3 Deployment Models**: Public, Private, Hybrid.
  - **3 Service Models**: IaaS, PaaS, SaaS.

📊 Table: What We Learnt in the Video

Category	Details
Definition	Delivery of computing resources over the Internet on a pay-as-you-go basis.
Examples	Online apps, business tools, cloud storage platforms (Google Drive, Dropbox).
User Benefits	Cost-effective, always updated apps, saves local storage, real-time

	collaboration.
Essential Characteristics	1. On-demand self-service 2. Broad network access 3. Resource pooling 4. Rapid elasticity 5. Measured service
Deployment Models	- Public: Shared over Internet - Private: Exclusive use - Hybrid: Mix of public/private
Service Models	- IaaS: Infrastructure access - PaaS: Platform for app dev - SaaS: Software on subscription

Cloud for Data Science

1. Introduction to Cloud for Data Scientists

- **Cloud** is a game changer or **godsend** for data scientists.
- Enables central storage of data and computing resources.
- Bypasses **physical limitations** of local machines or systems.
- Offers access to **advanced analytics and computing capacities** without requiring ownership of the infrastructure.

---

2. Key Capabilities of Cloud for Data Science

A. Storage and Accessibility

- Store massive datasets on **remote servers** (e.g., California, Nevada).
- Centralized data enables **easy access from anywhere**.

B. Computing Power

- Ability to run **high-performance computing (HPC)** tasks.
- Deploy complex algorithms without needing to install them locally.
- Use powerful machines that the user or the company does not own.

C. Collaborative Work

- Multiple teams across different locations (e.g., Germany, India, Ghana) can:
    - Access the **same datasets**.
    - Use the **same tools and algorithms**.
    - Collaborate in **real-time** on projects.

D. Accessibility and Device Compatibility

- Accessible from:
    - **Laptop**

    - **Tablet**
    - **Phone**
- Supports work in **any time zone** and from **any location**.

---

3. Advantages of Using Cloud

Feature	Benefit
No Local Installation	Use technologies (e.g., Apache Spark) without local setup
Always Up-to-Date	Access the latest tools and libraries
Open Source Ready	Instant access to open source tech
Remote Collaboration	Work with global teams simultaneously
Scalability	Handle large datasets and workloads efficiently

---

4. Major Cloud Providers

Company	Cloud Platform
IBM	IBM Cloud
Amazon	AWS (Amazon Web Services)
Google	Google Cloud Platform

---

5. IBM Developer Skills Network

- IBM provides **Skills Network Labs (SN Labs)** for learners.
- Tools available:
    - **Jupyter Notebooks**
    - **Spark Clusters**
- Allows learners to:
    - Create **data science projects**
    - **Develop and test solutions** in real environments.

✅ **Summary**

Cloud computing transforms how data scientists work by offering a flexible, powerful, and collaborative platform. It eliminates hardware and software limitations, promotes global teamwork, and streamlines the development

process with scalable computing and modern tools—all accessible from virtually anywhere. Leading cloud providers like IBM, Amazon, and Google offer platforms that make working on large-scale data science projects easier than ever.

📊 **Table: What We Learnt in the Video**

Topic	Key Points
Purpose of Cloud	Centralized storage and computation for data science
Benefits	Flexibility, scalability, real-time collaboration
Tools Access	Apache Spark, Jupyter Notebooks via cloud
Major Providers	IBM, AWS, Google Cloud
IBM Skills Network	Offers cloud-based tools for learners
Accessibility	Use from any device, globally available
Use Case	Enables running advanced algorithms without owning computing power

Foundations of Big Data

1. Introduction

- Every online interaction leaves behind a data trail in today's digital age.
- Activities like traveling, workouts, and entertainment generate vast data.
- The interconnected devices we use daily collect this data.
- The phenomenon is termed **Big Data**.

---

2. Definition of Big Data

- **Ernst & Young** defines Big Data as:

  "Dynamic, large and disparate volumes of data are being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value."

- No **universal definition**, but several **common elements** known as the **5 V's**:

  1. **Velocity**
  2. **Volume**
  3. **Variety**

4. **Veracity**
5. **Value**

---

3. The 5 V's of Big Data

1. Velocity (Speed of Data Generation)

- Data is generated continuously at high speed.
- Real-time/near real-time data streaming is possible.
- Technologies used: **Local and cloud-based** processing.

2. Volume (Amount of Data)

- Refers to the **scale** and **quantity** of data generated and stored.
- Drivers:
  - More data sources
  - Higher-resolution sensors
  - Scalable infrastructure
- Example: 2.5 quintillion bytes generated daily (~10 million Blu-ray DVDs).

3. Variety (Types of Data)

- Data exists in **structured** and **unstructured** forms:
  - **Structured**: Organized in databases (e.g., spreadsheets)
  - **Unstructured**: Unorganized data like tweets, images, videos, blogs, audio
- Comes from various sources: machines, people, internal/external processes.
- Drivers:
  - Mobile and wearable tech
  - Social media
  - Geo and video technologies
  - Internet of Things (IoT)

4. Veracity (Accuracy & Quality of Data)

- Concerns with data **truthfulness** and **reliability**.
- Attributes:
  - Consistency
  - Completeness

    - Integrity
    - Ambiguity
- 80% of data is **unstructured**, making analysis challenging.
- Importance of traceability and data verification.

5. Value (Usefulness of Data)

- The focus is on converting raw data into meaningful insights.
- Value isn't limited to profit:
    - Medical benefits
    - Social improvement
    - Customer satisfaction
    - Employee well-being
    - Personal achievements

---

4. Examples of the V's in Action

V's	Example
Velocity	Hours of YouTube footage uploaded every 60 seconds
Volume	7 billion+ people using digital devices generating 2.5 quintillion bytes/day
Variety	Data types: text, images, sound, video, health metrics, IoT data
Veracity	Need to make sense of 80% unstructured data
Value	Insights derived bring customer, social, and organizational benefits

---

5. Data Analysis Tools for Big Data

- Traditional tools are **inadequate** for the scale of Big Data.
- Use of **distributed computing systems** is necessary.
- Popular Tools:
    - **Apache Hadoop**
    - **Apache Spark**
- These tools enable:
    - Extracting

    - Loading
    - Analyzing
    - Processing data across multiple systems
- Provide **real-time insights** and support **scalable data solutions**.

---

6. Impact of Big Data

- Enables organizations to:
    - Better understand and connect with customers
    - Improve services
    - Drive innovation
- Everyday, personal data (from smartwatches, phones, etc.) feeds into these systems, creating a **global data journey** that may ultimately benefit the user.

Summary

Big Data refers to the massive volumes of digital information that people and devices generate. It's characterised by the five V's: **Velocity**, **Volume**, **Variety**, **Veracity**, and **Value**. The data's exponential growth and diversity demand new tools like **Apache Spark and Hadoop** for real-time processing and analysis. These insights help drive better decision-making across sectors, from business to healthcare. Whether it's a smartwatch or a social media post, every data point contributes to this global information ecosystem.

Table: What We Learned in the Video

Topic	Details
What is Big Data?	Large, dynamic, varied data created by people, tools, and machines
Key Attributes (5 V's)	Velocity, Volume, Variety, Veracity, Value
Examples of Each V	YouTube uploads, device data, unstructured content, data accuracy issues, and business insights
Tools Used	Apache Spark, Hadoop, and distributed computing platforms
Challenges	Unstructured data, traditional tools inadequate, need for new methods
Benefits of Big Data	Real-time insights, service improvements, enhanced decision-making
Everyday Impact	Wearables, smartphones, and IoT devices all contribute to Big Data

Data Science and Big Data

🎓 Programming Background Among Students

- **General Understanding**:
  - Almost everyone in the program knows how to program.
  - Backgrounds vary:
    - **MS in Computer Science**
    - **MBA students with technical experience**
    - **Students with limited exposure** (e.g., one programming course in college 4–5 years ago).
  - Common trait: **Ability to think computationally**, considered **most important**.

📊 Growth of Data Science and Business Analytics

- **Current Popularity**:
  - These fields have become **"very hot"** in the **last 4–5 years**.
- **Reasons for Popularity**:
  - Emergence of **new tools** and **approaches**.
  - **Explosion of data** which traditional techniques can't manage efficiently.
- **Corporate Awareness**:
  - Initially, only **specific companies or departments** saw the need.
  - Now, broader understanding across industries.
  - Example:
    - A **major bank**:
      - 3 years ago: **One small group** with a small cluster.
      - Now: **5–6 big data clusters**, processing **credit card data** using **advanced techniques**.

📈 Academic Interest & Trends

- **Undergraduate Course Example**:
  - Course: "Dealing with Data".
  - Enrollments:
    - Last year: **28 students**

        - This year: **140 students**

- **Parental Influence**:
    - Parents traditionally pushed students toward **finance, accounting, marketing**.
    - Now, realizing the value of **STEM** and **data science**.
    - Encouraging children to:
        - Take more STEM classes in **high school**.
        - Prepare for careers in **analytics** and **data science**.

💾 **Definition & Evolution of Big Data**

- **Different Definitions**:
    - Speaker's definition:
        - Big data = **Data too large or fast-moving** for traditional DBMS (volume + velocity).
    - Statistician's humorous definition:
        - Big data = **Anything that can't fit on a thumb drive**.
- **Historical Context**:
    - Origin traced to **Google** (Larry Page & Sergey Brin).
        - Challenge: Storing **all webpages** in the world.
        - Solution: Built their own systems (e.g., that inspired Hadoop).
- **Technological Impact**:
    - Hadoop copied Google's framework.
    - Growth of **big data clusters** and **distributed computing**.
- **Analytical Evolution**:
    - Emergence of **new statistical and analytical techniques**.
    - Increasing relevance of **deep learning** (brief mention).

🎶 **Music & Transitions**

- Musical transitions mark topic shifts (implied in transcript).

📝 **Summary**

This transcript emphasizes how **data science and business analytics** have become popular due to data growth,

new tools, and industry needs. While students vary in technical expertise, the key skill is **computational thinking**. The academic world has responded with rising enrollments, driven by parental awareness of lucrative career paths in analytics. The speaker defines **big data** as datasets too large or fast for traditional systems, tracing its roots to **Google's early challenges** and highlighting modern tools like **Hadoop** and techniques like **deep learning**.

📊 **Table: What We Learned in the Video**

Topic	Key Points
Programming Background	Most students know programming; computational thinking is crucial
Popularity of Data Science	Driven by the data explosion, new tools, and techniques
Industry Evolution	Companies like banks now heavily invest in big data infrastructure
Academic Trends	Student enrollments rising; parents support data science-focused education
Definition of Big Data	Data too large/fast for traditional systems
History of Big Data	Originated with Google; led to tools like Hadoop
Analytical Advances	Growth of statistical and deep learning techniques for large datasets

What is Hadoop?

1. Traditional Data Processing

- Traditionally, computation was data-centric:
    - Bring **data to the program**.
    - Process it on a single machine.

---

2. Innovation in Big Data: Google's Approach

- **Larry Page and Sergey Brin's breakthrough**:
    - **Slice data** into pieces.
    - **Distribute** each piece to many computers.
    - **Replicate** each piece (typically **triplicated**).
- **Send the same program** to each machine:
    - Each machine processes its piece of data independently.
    - Results are returned, **sorted**, and sent to another process.

---

3. MapReduce Paradigm

- **Two main phases**:
    - **Map (Mapper)**: Each node processes data locally.
    - **Reduce**: Results are aggregated and summarized.
- Simple concept but powerful for handling **very large datasets**.
- **Linear scalability**:
    - Add more servers → get more performance → handle more data.

---

4. Adoption and Hadoop

- **Yahoo adopted** this model.
- Hired **Doug Cutting**:
    - Developed an **open-source clone** of Google's architecture.
    - Named it **Hadoop**.
- Hadoop now part of the **big data ecosystem**:
    - **Widespread adoption**.
    - Thousands of companies involved.

---

5. Foundation of Data Science

- Core disciplines:
    - **Probability and statistics**
    - **Algebra and linear algebra**
    - **Programming**
    - **Databases**
- These elements have existed for decades.

---

6. Emergence of Machine Learning

- Shift in methodology:
    - From **hypothesis testing** → to **pattern discovery**.
- Can now handle **entire datasets**, not just samples.
- **Traditional statisticians** may find this method controversial:

    - Lack of pre-defined hypothesis.
    - Reverse flow: patterns → hypotheses.

---

7. Applications in Social Media

- Massive social data sets made traditional statistics hard to apply.
- Machine learning became the **only viable solution** for:
    - Pattern recognition
    - Trend prediction

---

8. Decision Sciences: A New Integration

- Combines:
    - **Computer science**
    - **Mathematics**
    - **Probability & statistics**
- Term: **Decision Sciences**
- NYU Stern's advantage:
    - Strong departments in:
        - **Statistics** (Phd-level faculty)
        - **Operations management**
        - **Information systems**
    - Enabled early adoption of the data science wave.

---

9. Evolution of Data Science as a Field

- The rise of the term **"data science"**:
    - Barely known 5 years ago.
    - Rapid spike in interest/searches.
- Mirror trend of **"big data"** from 7–8 years ago.

---

10. Neural Networks and Deep Learning

- Neural networks have existed for 20–30 years.

- Previously limited in capability.

- Breakthrough in **multi-layer neural networks**:
  - Researchers at **University of Toronto**.

- Rise of **deep learning** in the last **3 years**.
  - Now used by major companies: **Google, Facebook**, etc.

📌 **Summary**

This transcript covers the evolution of **data processing**, the rise of **big data infrastructure** pioneered by **Google** (through **MapReduce**), and its open-source parallel, **Hadoop**. It explains the convergence of classical disciplines like **statistics, algebra, databases**, and **programming** under the umbrella of **data science**. It also describes how machine learning has revolutionised data analysis, especially in handling **massive social media datasets**, and has led to the development of **Decision Sciences**. Finally, it discusses how **deep learning**—especially neural networks—has surged due to computational advances and research breakthroughs.

📝 **Table: What We Learnt in the Video**

Topic	Details
Traditional Processing	Data brought to the program on a single computer
Google's Innovation	Distributed processing using MapReduce
MapReduce	Map (process pieces) + Reduce (aggregate results)
Hadoop	Open-source implementation of Google's system
Big Data Benefits	Linear scalability, handles vast datasets
Foundations of Data Science	Probability, statistics, linear algebra, programming, databases
Machine Learning	Focuses on pattern recognition over hypothesis testing
Social Media Data	Too large for traditional methods → machine learning necessary
Decision Sciences	Integration of CS, math, stats—adopted early by NYU Stern
Rise of Data Science	Term gained popularity rapidly in last 5 years
Deep Learning	Multi-layer neural networks, Toronto breakthrough, widespread corporate use

Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark

1. Introduction to Big Data Technologies

- Big Data involves processing large structured, semi-structured, and unstructured data to extract value.
- Technologies discussed:
  - **NoSQL Databases** and **Data Lakes** (mentioned in other videos)
  - **Apache Hadoop, Apache Hive, and Apache Spark** (covered in this video)

---

2. Apache Hadoop

- **Definition**: Java-based open-source framework for distributed storage and processing of large datasets.
- **Cluster Architecture**:
  - **Node**: A single computer
  - **Cluster**: A group of nodes
  - Scales from one node to many, each with local storage and processing power.
- **Key Features**:
  - Handles structured, semi-structured, and unstructured data (e.g., streaming audio, video, social media, clickstream).
  - Offers real-time self-service access for stakeholders.
  - Cost-effective by moving "cold" data from expensive systems to Hadoop.
- **Use Cases**:
  - Data consolidation across organizations.
  - Optimization of enterprise data warehouse costs.

---

3. Hadoop Distributed File System (HDFS)

- **Role**: Core component of Hadoop for data storage.
- **Functionality**:
  - Splits large files into blocks stored across multiple nodes.
  - Allows **parallel processing** of data.
  - **Replication**: Each file block is replicated on two additional nodes by default.
- **Example**:

    - A U.S. phonebook: Names starting with "A" on server 1, "B" on server 2, etc.
    - All servers needed to reconstruct the full phonebook.
- **Benefits**:
    - **Fault Tolerance**: Quick recovery from hardware failure.
    - **High Throughput**: Supports access to streaming data.
    - **Scalability**: Handles hundreds of nodes.
    - **Portability**: Compatible across various platforms and OS.
    - **Data Locality**: Brings computation close to data to reduce network load and increase speed.

---

4. Apache Hive

- **Definition**: Open-source data warehouse software built on Hadoop.
- **Data Management**:
    - Manages large datasets in **HDFS** or **Apache HBase**.
- **Characteristics**:
    - Designed for **long sequential scans**.
    - High query **latency**—not suitable for real-time applications.
    - **Read-based**, not ideal for **transaction processing** (write-heavy).
- **Best Uses**:
    - ETL processes (Extract, Transform, Load)
    - Reporting
    - Data analysis
- **Tools**:
    - SQL-like interface for easy data access and querying.

---

5. Apache Spark

- **Definition**: General-purpose distributed data processing engine.
- **Applications**:
    - Interactive analytics

    - Stream processing
    - Machine learning
    - Data integration and ETL
- **Performance**:
    - **In-memory processing**: Fast computation
    - Falls back to disk when memory is constrained
- **Language Support**:
    - Java, Scala, Python, R, and SQL
- **Deployment**:
    - Standalone or on top of Hadoop and other infrastructures
- **Data Access**:
    - Compatible with HDFS, Hive, and other sources
- **Key Benefit**:
    - Handles **real-time streaming data** and **complex analytics** efficiently

---

Summary
This video introduces three essential open-source technologies in the Big Data ecosystem: Apache Hadoop, Hive, and Spark. Hadoop provides a distributed framework for storing and processing large datasets, with HDFS at its core for scalable and fault-tolerant storage. Hive functions as a data warehouse on top of Hadoop, suitable for analytical workloads but not for fast transactions. Spark is a powerful, in-memory processing engine that supports real-time data processing and a variety of analytics tasks. Together, these technologies enable organisations to manage and analyse vast and diverse datasets efficiently.

Table: What We Learnt in the Video

Topic/Section	What We Learnt
Big Data Technologies	Key technologies include Hadoop, Hive, and Spark; support structured & unstructured data
Apache Hadoop	Java-based, scalable, distributed data processing and storage framework
Hadoop Cluster & Nodes	A node is one computer; a cluster is a group of nodes that work together
HDFS	Splits and replicates files across nodes for parallelism, fault tolerance, and scalability
Data Locality	Computation occurs where data resides to boost performance

Apache Hive	Data warehouse on Hadoop; supports SQL-like queries; best for ETL and analysis
Hive Limitations	High latency, not suitable for transaction processing
Apache Spark	Fast, general-purpose processing engine; supports real-time analytics and machine learning
Spark Features	In-memory processing, multiple language support, and integration with Hadoop and Hive

Lesson Summary: Big Data and Data Mining

Introduction to Big Data and Its Societal Impact

- **Big Data's Influence**:
    - Transforms business operations, sports, and daily life.
    - Requires organizations to fundamentally change their approach to business.
- **Need for Innovation**:
    - Vast amounts of data created by people, tools, and machines demand:
        - New technologies
        - Innovative, scalable solutions

Big Data Fundamentals

- **Key Attributes of Big Data (The 5 V's)**:
    1. **Value**:
        - Data must provide meaningful insights and business value.
    2. **Volume**:
        - Refers to the scale of data being generated and stored.
        - Driven by:
            - Increased number of data sources
            - Scalable infrastructure
    3. **Velocity**:
        - Speed at which data is generated, collected, and processed.

        - Comes from continuous and rapid data sources.

4. **Variety**:
    - Refers to different data types and sources.
    - Includes structured (e.g., databases) and unstructured data (e.g., videos, social media).
5. **Veracity**:
    - Data quality and accuracy.
    - Ensures data conforms to reality and is reliable.

Cloud Computing and Its Role in Big Data

- **Definition**:
    - On-demand access to computing resources via the internet, on a pay-as-you-go basis.
- **Five Essential Characteristics**:
    - **On-demand self-service**:
        - Users can access resources as needed without human interaction with providers.
    - **Broad network access**:
        - Resources accessible over the internet from various devices.
    - **Resource pooling**:
        - Resources dynamically allocated to multiple users.
        - Enhances cost-efficiency.
    - **Elasticity**:
        - Resources can scale up/down based on demand.
    - **Measured service**:
        - Users only pay for what they use or reserve.
- **Benefits for Big Data**:
    - Scalability
    - Collaboration
    - Accessibility

- Simplified software maintenance
- Instant access to updated tools

Open Source Tools for Big Data Processing

- **Apache Hadoop**:
  - Framework for distributed storage and processing.
  - Works across clusters of computers.
- **Apache Hive**:
  - Data warehouse infrastructure built on Hadoop.
  - Used for querying and managing large datasets.
  - Interfaces with Hadoop Distributed File System (HDFS) and Apache HBase.
- **Apache Spark**:
  - General-purpose data processing engine.
  - Efficiently handles large-scale data analytics tasks.

Data Mining Process

- **Six-Step Process**:
  1. **Goal Setting**:
     - Define key questions and objectives.
     - Consider cost-benefit implications.
  2. **Selecting Data Sources**:
     - Identify existing data sources or plan data collection.
  3. **Preprocessing**:
     - Clean data by removing irrelevant attributes and correcting errors.
  4. **Transforming**:
     - Format and structure data appropriately for analysis.
  5. **Mining**:
     - Apply machine learning algorithms and analysis methods.

6. **Evaluation**:
    - Test predictive models on real data.
    - Evaluate effectiveness and efficiency.
    - Share results with stakeholders.
    - Iterative process – results guide future mining efforts.

Recap and Key Insights

- Despite varying definitions, the **5 V's** are universally recognized.
- Big data is **driving change across industries**.
- Cloud computing is **critical to big data success**, offering:
    - Scalability
    - Efficiency
    - Up-to-date tools
- Tools like Hadoop, Hive, and Spark are **essential for processing** and **analyzing big data**.
- The data mining process enables organizations to **extract valuable insights** effectively.

Summary

This video provides a comprehensive overview of big data and its transformative impact on society, emphasizing its defining characteristics: value, volume, velocity, variety, and veracity. It highlights how cloud computing facilitates big data storage, processing, and analysis through scalable and cost-effective resources. Open-source tools such as Apache Hadoop, Hive, and Spark play a central role in managing large datasets. The data mining process is detailed in six iterative steps, from setting goals to evaluating outcomes, enabling organizations to derive actionable insights. The video illustrates how big data and cloud technologies are central to digital transformation and informed decision-making.

Table: What We Learnt in the Video

Topic/Section	What We Learnt
Impact of Big Data	Transforms business, industry, and daily life; necessitates innovation
5 V's of Big Data	Value, Volume, Velocity, Variety, Veracity – core characteristics of big data
Cloud Computing	Enables scalable, accessible, and cost-effective data handling
Cloud Characteristics	On-demand, network access, resource pooling, elasticity, measured service

Benefits of Cloud for Big Data	Scalability, simplified maintenance, accessibility, collaborative potential
Hadoop	Framework for distributed storage and processing of big data
Hive	Data warehouse tool for managing and querying data on Hadoop
Spark	High-performance processing engine for big data analytics
Data Mining Process	Six steps: Goal setting, data selection, preprocessing, transforming, mining, evaluation
Role of Evaluation	Validates models and informs stakeholders and future data mining iterations.

Deep Learning and Machine Learning

Lesson Overview: Deep Learning and Machine Learning

In this lesson, "Deep Learning and Machine Learning," you'll dive into the exciting concepts of artificial intelligence and data science. Throughout this module, you will explore various machine and deep learning aspects, gaining valuable insights and skills.

Asset name and type	Description
“Artificial Intelligence and Data Science” video	Get introduced to the captivating field of artificial intelligence and its role in data science.
“Generative AI and Data Science” video	Discover the exciting realm of generative artificial intelligence and its applications in data science.
“Neural Networks and Deep Learning” video	Explore the fundamentals of neural networks and delve into the depths of deep learning.
“Applications of Machine Learning” video	Uncover the real-world applications of machine learning and its impact on various industries.
“Regression” reading	Learn about regression analysis, a fundamental statistical technique used in machine learning.
Practice quiz	Test your understanding of the previous reading.
“Exploring Data using IBM Cloud Gallery” lab	Engage in hands-on data exploration using the IBM Cloud Gallery, gaining practical experience in data analysis.
“Lesson Summary” video	Sum up your learning from this module with a concise lesson summary.
Practice quiz	Take a practice quiz to evaluate how well you’ve understood the material presented in this lesson.

Glossary	Use this glossary of terms to review the terminology presented in this lesson.
Graded quiz	Test your knowledge from this lesson by taking the graded quiz.

Artificial Intelligence and Data Science

🔍 Understanding Big Data

- **Definition**:
  - Big Data refers to data sets that are:
    - Massive in *volume*
    - Rapidly generated (*velocity*)
    - Highly varied in type (*variety*)
    - Often uncertain in quality (*veracity*)
    - Contain potential *value*
  - These data sets are too complex for traditional analysis tools like relational databases.
- **Technological Response**:
  - Emergence of:
    - Distributed computing
    - Advanced analytics tools
  - Organizations can now extract new insights from vast, complex data.

🔍 The 5 V's of Big Data

V	Description
Velocity	The speed at which data is generated and processed
Volume	The massive amount of data produced
Variety	Different types of data (structured, unstructured, etc.)
Veracity	Accuracy and trustworthiness of the data
Value	Potential insights and benefits derived from analyzing the data

---

🔎 Data Mining

- **Definition**:
    1. Automatic search and analysis of large data sets to find patterns and insights.
- **Steps in Data Mining**:
    1. **Preprocessing**:
        - Clean and prepare data
    2. **Transformation**:
        - Convert data into suitable format
    3. **Mining**:
        - Use tools such as:
            - Visualization
            - Machine learning models
            - Statistical analysis

🤖 Machine Learning (ML)

- **Definition**:
    - A subset of AI where algorithms learn from data without being explicitly programmed.
- **Key Features**:
    - Learns from examples
    - Adapts based on input
    - Makes predictions and decisions autonomously
    - Not rules-based
- **Applications**:
    - Predictive analytics
    - Natural language processing
    - Recommendation systems

🧠 Deep Learning (DL)

- **Definition**:

- A specialized subset of ML using layered neural networks to simulate human decision-making.

- **Capabilities**:
  - Labelling, categorizing, and pattern identification
  - Continuous learning and improvement
  - Enhances accuracy over time

🧬 **Neural Networks**

- **Inspiration**:
  - Modelled loosely on biological neural networks
- **Structure**:
  - Made up of "neurons" (computing units)
  - Often organized in multiple layers (*deep layers*)
- **Key Points**:
  - Adapt and improve with larger data sets
  - Unlike traditional ML, they scale better with more data

💉 **Artificial Intelligence vs. Data Science**

Aspect	**Artificial Intelligence (AI)**	**Data Science**
Definition	Systems that mimic intelligent human behaviour	Methods to extract insights from complex data
Scope	Focused on decision-making and automation	Broad interdisciplinary field
Techniques Used	Machine Learning, Deep Learning	ML, Stats, Data Viz, Programming
Relationship	AI is not a subset of Data Science	DS may use AI techniques
Commonality	Both can use Big Data	Both deal with complex data-driven problems

- **Conclusion**:
  - Data Science = Full data processing journey

- AI = Techniques for learning, problem-solving, and intelligent decisions

---

Summary
This video explains fundamental concepts in data science and artificial intelligence, focusing on differentiating between Big Data, Data Mining, Machine Learning, Deep Learning, Neural Networks, AI, and Data Science. It outlines how big data is characterized by the five V's and highlights how modern technologies have enabled powerful analytics. Machine Learning and Deep Learning are presented as vital AI tools, while Data Science is shown as a broader interdisciplinary field. The video emphasises the complementary yet distinct roles of AI and Data Science in leveraging big data for decision-making.

Table: What We Learnt in the Video

Topic/Section	What We Learnt
Big Data	Massive, fast, varied data sets needing advanced analysis tools
The 5 V's of Big Data	Velocity, Volume, Variety, Veracity, and Value define big data characteristics
Data Mining	Process of discovering patterns via preprocessing, transformation, and analysis
Machine Learning	Algorithms learn from data, enabling intelligent decisions without manual coding
Deep Learning	Subset of ML using neural networks for high-level pattern recognition
Neural Networks	Computing units mimicking neurons, enabling scalable learning
AI vs. Data Science	AI solves problems intelligently; Data Science extracts insights from data

Generative AI and Data Science

🧠 **Introduction to Generative AI**

- **Definition**:
    - A **subset of artificial intelligence** is focused on **producing new data** rather than analyzing existing data.
    - Capable of creating **content** such as:
        - Images 🖼️
        - Music 🎵
        - Language 🗣️
        - Code 💻
- **Purpose**:

- Mimics human-created content by learning from large datasets.

🛠️ **How Generative AI Works**

- **Foundational Models**:
  - **Generative Adversarial Networks (GANS)**:
    - Two neural networks compete: Generator vs. Discriminator.
    - The generator tries to create realistic data; the discriminator attempts to distinguish real from fake.
  - **Variational Autoencoders (VAEs)**:
    - Encodes input data into a compressed representation.
    - Decodes to generate new data similar to the original.
- **Process**:
  - Learns data **patterns and distributions** from large datasets.
  - Generates new instances that resemble the **original data**.

🧩 **Applications of Generative AI Across Industries**

📝 **Natural Language Processing**

- **Example**: GPT-3 (by Openai)
- **Uses**:
  - Content creation
  - Chatbots
  - Generating human-like text

🏥 **Healthcare**

- **Use**:
  - **Synthetic medical images** for training and simulation.
  - Assists medical professionals in learning and diagnosing.

🎨 **Art & Creativity**

- **Use**:
  - Creating **unique and stunning artworks**
  - Endless generation of creative visual compositions

🎮 **Gaming Industry**

- **Use**:
    - **Realistic game environments**
    - Designing characters and levels automatically

👗 **Fashion Industry**

- **Use**:
    - Designing new styles
    - Personalised shopping recommendations

📊 **Use of Generative AI in Data Science**

🧪 **Data Augmentation with Synthetic Data**

- **Problem**:
    - Not enough real-world data to build effective models.
- **Solution**:
    - Generative AI **creates synthetic data** with similar properties:
        - Distribution
        - Clustering
        - Other statistical traits
- **Outcome**:
    - Combines real and synthetic data for **better model training and testing**.

🧠 **Hypothesis Generation & Code Automation**

- **Challenge**:
    - Data scientists face **time constraints** which limit:
        - Number of hypotheses tested
        - Speed of analysis
- **Solution**:
    - Generative AI automates code generation for:
        - Analytical models
        - Hypothesis testing
    - Frees data scientists to focus on:

        - Defining problems
        - Evaluating a broader range of hypotheses

📈 **Business Insights & Reporting**

- **Use**:
    - Generates accurate reports and **dynamic insights**
    - Updates insights **as data changes**
    - Identifies **hidden patterns** missed in manual analysis
- **Tool Example**:
    - **IBM Cognos Analytics**:
        - AI assistant helps users frame questions or hypotheses
        - Automatically generates insights

📝 **Summary**

This video introduces **Generative AI** as a powerful subset of artificial intelligence focused on creating new data instead of merely analysing existing data. It operates through deep learning models such as GANs and VAEs that learn from large datasets to generate new content. Generative AI is widely used in various industries, from text generation and healthcare imaging to art, gaming, and fashion. In data science, it helps solve data scarcity through synthetic data, accelerates model development through code generation, and enhances analytics by automating insights and uncovering hidden patterns. These capabilities make it an essential tool for data-driven decision-making.

📋 **Table: What We Learnt in the Video**

Topic/Section	What We Learnt
Generative AI Definition	Focuses on creating new data mimicking human-made content
Core Technologies	GANs and VAEs drive the creation of new, realistic data
NLP (e.g., GPT-3)	Used to generate human-like text, transforming chatbots and content writing
Healthcare	Generates synthetic medical images for training
Art	Can autonomously create visually compelling artworks
Gaming	Used to build environments, characters, and levels
Fashion	Helps design styles and personalize shopping
Synthetic Data in Data Science	Supplements real data, improves model training

Coding Automation	Accelerates code generation and hypothesis testing
Business Analytics	Enables real-time insight generation and hidden pattern detection
IBM Cognos Analytics	An AI assistant that interprets user questions to deliver useful data insights

Neural Networks and Deep Learning

🧠 Introduction to Neural Networks

- **Definition**:
  - An attempt by computer science to mimic the brain's neural structure using computer programs.
  - Neural networks replicate the way biological neurons and synapses work to process information.
- **Structure of Neural Networks**:
  - Inputs → Processing nodes (neurons) → Transformations & Aggregation → Further layers → Output.
  - Originally used for tasks like **recognizing handwritten digits**.
- **Training Process**:
  - Input data is repeatedly passed through the network.
  - Weights are adjusted through iterations to match outputs with expected results (convergence).

📉 Historical Challenges of Neural Networks

- **Computational Limitations**:
  - Early neural networks were computationally intensive.
  - Despite working for small-scale problems, they fell out of favor due to hardware constraints.
- **Teaching Timeline**:
  - The speaker stopped teaching neural networks approximately 15 years ago due to their impracticality at the time.

🚀 Emergence of Deep Learning

- **Revival**:
  - The term *deep learning* started gaining popularity around 4–5 years ago (from the speaker's perspective).

- **Definition & Difference**:
    - Deep learning = Neural networks with **many layers** (hence “deep”) and immense computing power.
    - “Neural networks on steroids.”
- **Technological Shift**:
    - Advances in **hardware** and **software** enabled training of deep neural networks on large datasets.
    - Now capable of performing sophisticated tasks previously unachievable.

💻 **Computational Requirements for Deep Learning**

- **Hardware Needs**:
    - Deep learning requires **GPUs (Graphics Processing Units)**.
    - GPUs perform large-scale **matrix and linear algebra operations** critical for training.
- **University Use Case**:
    - At the university’s South Data Center:
        - 200+ servers equipped with GPUs (each with ~600 cores).
        - Indicates high demand for deep learning infrastructure.
- **Faculty Application**:
    - A marketing professor uses deep learning in research.
    - Needs a GPU-enabled computer for heavy mathematical computations.

🎓 **Real-World Teaching Example**

- **Instructor from NYU / Facebook**:
    - Bring a thick notebook to class with embedded GPUs.
    - Trains a speech recognition model live in class using real-time student input.
    - Demonstrates how quickly deep learning models can learn and respond.

🔍 **Capabilities of Deep Learning**

- **Applications**:
    - Speech recognition

    - Facial recognition
    - Image classification
    - Object detection
    - Self-learning capabilities (e.g., distinguishing cats vs. dogs without explicit instruction)
- **Learning Behavior**:
    - Models behave like babies learning to talk—initial awkwardness followed by improved fluency.
    - Can train themselves without being manually programmed for each rule.

🧮 **Importance of Linear Algebra**

- **Mathematical Foundation**:
    - Deep learning heavily relies on **matrix operations** and **linear transformations**.
    - Understanding linear algebra is important for grasping what's happening under the hood.
- **Tools**:
    - Many packages (e.g., TensorFlow, PyTorch) handle the math, but **basic understanding** is still recommended.

🖥️ **Practical Considerations**

- **Not for Basic Laptops**:
    - Deep learning is **computationally heavy** and typically requires powerful machines.
    - You can experiment on a notebook, but serious work needs **dedicated computing resources**.

Summary

This video offers a historical and conceptual overview of neural networks and the evolution into deep learning. It explains how early neural networks aimed to simulate the brain's processing but were limited by computational resources. "Deep learning" refers to stacking multiple neural network layers and harnessing massive computational power, especially GPUS, to solve complex tasks such as speech and image recognition. Real-world applications and teaching examples demonstrate how accessible and powerful deep learning has become. Despite the availability of libraries that simplify implementation, foundational knowledge in linear algebra is still valuable.

Table: What We Learnt in the Video

Topic/Section	What We Learnt
Neural Network Basics	Simulate the brain's neurons; involve inputs, processing nodes, and outputs.
Early Limitations	Computationally intensive, fell out of favor.
Deep Learning Definition	Neural networks with multiple layers and high computational demands.
Importance of GPUs	Essential for deep learning; enable large matrix and algebra operations.
Real-World Applications	Used in speech recognition, image classification, and object recognition.
University Infrastructure	Example of large-scale GPU setup for deep learning research.
Teaching Example (NYU/Facebook)	Live speech recognition training in class using GPU-powered notebook.
Role of Linear Algebra	Critical for understanding deep learning; powers the underlying math.
Software Tools	Packages handle computations, but basic knowledge of math is still important.
Practical Needs	Serious deep learning requires specialized hardware, not just a regular notebook.

Applications of Machine Learning

🤖 **Machine Learning Overview**

- **Ubiquity**: Machine learning is now a standard part of many industries.
- **Major Applications**:
    - **Recommender Systems**
    - **Classification**
    - **Clustering**
    - **Predictive Analytics**
    - **Fraud Detection**

🛒 **Market Basket Analysis (Past → Present)**

- **Definition**: Identifying items frequently bought together (e.g., beer and diapers).
- **History**:
    - 20 years ago: Computationally difficult.
    - Today: Easily handled with modern machine learning techniques.

📈 **Predictive Analytics**

- **Usage**: Forecasting trends, behaviors, or outcomes using historical data.
- **Techniques Involved**:
    - Decision Trees
    - Bayesian Analysis
    - Naive Bayes
- **Statistical Perspective**:
    - These techniques are not always favored by traditional statisticians.
    - Still, they provide valuable predictive capabilities.

📊 **Practical Understanding in Tools like R**

- **User Role**:
    - Users need to understand **how** to use techniques, not necessarily **how they work in detail**.
    - Importance is placed on interpreting results and understanding **trade-offs**.
- **Key Concepts to Know**:
    - **Precision vs. Recall**:
        - Precision: Correct positive predictions / total predicted positives.
        - Recall: Correct positive predictions / total actual positives.
    - **Overfitting**: Model performs well on training data but poorly on new data.
    - **Oversampling**: Can distort data balance; needs careful handling.

💼 **Applications in Fintech**

1. 🧠 **Recommender Systems in Finance**

- **Concept**: Similar to Netflix or Facebook recommendations.
- **Application in Fintech**:
    - Investment suggestions based on user behavior.
    - Helps financial advisors or retail investors discover related assets or strategies.
    - Based on **similar assets**, **companies**, or **investment approaches**.

2. 🦹 Fraud Detection

- **Objective**: Detect fraudulent transactions in real-time.
- **Methodology**:
    - Learn from historical transaction data.
    - Build models to assess new transactions.
    - Flag or approve charges based on model outputs.
- **Execution**:
    - Real-time analysis.
    - Escalation to fraud teams if the transaction appears suspicious.

Summary

The video discusses the growing role of machine learning across sectors, focusing on its use in recommender systems, predictive analytics, and fraud detection, especially within fintech. Traditional challenges like market basket analysis have become manageable with modern techniques. Tools like R make machine learning accessible, even to those without deep technical knowledge, although understanding trade-offs like precision vs. recall or overfitting remains essential. In fintech, machine learning powers recommendation engines for investment advice and real-time fraud detection systems, showing its vital and widespread impact.

Table: What We Learnt in the Video

Topic/Section	What We Learnt
Machine Learning Applications	Used widely today for classification, prediction, clustering, and recommendations.
Market Basket Analysis	Once complex, now efficiently handled via ML.
Predictive Analytics	Uses models like decision trees and Bayesian analysis to forecast outcomes.
Tools like R	Users don't need full technical knowledge but must understand trade-offs.

Key ML Concepts	Precision, recall, overfitting, and oversampling are critical to model success.
Recommender Systems in Fintech	Suggests related investments based on user behaviour, like media platforms do.
Fraud Detection	Uses real-time ML to flag or approve credit card transactions.

Regression

*From the book **Getting Started with Data Science** by Murtaza Haider*

📏 The Concept of Regression Toward the Mean

- **Main Idea**: Tall parents tend to have tall children, but not necessarily taller than themselves.
- **Implication**: If height increased each generation based on parents' height, future generations would become abnormally tall.
- **Historical Reference**:
    - Sir Francis Galton (1886) studied this phenomenon.
    - Developed the statistical concept we now call **regression**.
    - Found that offspring of tall parents tend to "regress" toward the mean (average) height.

📊 Introduction to Regression Models

- **Regression Models**:
    - Core tool of statistical analysis.
    - Ubiquitous across disciplines: economics, medicine, social sciences, business, etc.
- **Uses**:
    - **Medical Science**:
        - Developing effective drugs
        - Optimizing surgeries and hospital operations
    - **Business Analysis**:
        - Understanding consumer behavior
        - Measuring firm productivity

        - Comparing public vs private sector efficiency

🧠 A Personal Story: The Department of Obvious Conclusions

- **Author's Thesis**:
    - Topic: *Hedonic price models* in real estate.
    - Data: 500,000 residential property transactions over 3 years.
- **Initial Finding**:
    - "Larger homes sell for more than smaller homes."
    - Dismissed by author's wife as obvious.
- **Deeper Insights**:
    - The value lies in **quantifying** these relationships:
        - **Washrooms vs Bedrooms**:
            - One additional washroom adds *more* value than an additional bedroom.
        - **Proximity to Infrastructure**:
            - Closer to subway → higher prices.
            - Closer to freeways → lower prices.
        - **Shopping Centers (Nonlinear Impact)**:
            - <2.5 km → *lower* prices (possibly due to congestion/noise).
            - 2.5–5 km → *higher* prices.
        - **Distance from Downtown**:
            - Further from downtown = lower home values.
- **Conclusion**:
    - The study went beyond the obvious by measuring *how much* each factor influences price.

📈 Why Use Regression? (Regression Analysis Applications)

- **Typical Questions Regression Can Answer**:
    - How much does an extra bedroom increase a home’s price?

    - How does lot size affect pricing?
    - Do homes with certain exteriors (brick vs stone) sell differently?
    - What is the price premium for a finished basement?
    - How do high-voltage power lines affect home prices?
- **Core Function**:
    - Helps isolate individual variables' effects on a dependent variable (e.g., housing price), holding all else constant.

📝 Summary

This chapter introduces regression models through an accessible analogy about why children of tall parents are not always taller. It presents regression as a key statistical method first studied by Galton and now fundamental in research and industry. The author personalizes the lesson with a story of his Master's thesis, explaining how regression helped quantify the influence of home features and location on housing prices. The value of regression lies in measuring *how much* each variable contributes, going beyond obvious conclusions to deliver actionable, data-driven insights.

📚 Table: What We Learnt in the Chapter

Topic/Section	What We Learnt
Regression to the Mean	Tall parents have tall kids, but usually not taller than themselves
Historical Origin	Galton’s 1886 study led to the development of regression models
Regression Models in Research	Widely used in medicine, business, and social science
Author’s Thesis Story	Real estate prices were studied using hedonic regression over 500,000 transactions
Key Real Estate Insights	Washrooms add more value than bedrooms; subways increase value, freeways reduce it
Nonlinear Shopping Centre Effect	Very close proximity lowers value; moderate proximity increases it
Distance from Downtown	Greater distance = lower value
Value of Regression	Not just correlation but *quantification* of relationships
Sample Regression Questions	E.g., effect of extra bedroom, lot size, basement, power lines on house prices

Lesson Summary: Deep Learning and Machine Learning

🎈 Introduction to the Lesson

- **Purpose of the Video:**
    - Review key concepts from the lesson on Deep Learning and Machine Learning.
    - Recap essential AI-related terms and how data scientists apply them.
    - Explain relationships between AI, Machine Learning, and Data Science.
    - Introduce regression models used for analyzing data relationships.

🤖 Understanding Artificial Intelligence (AI)

- **Definition:**
    - A branch of computer science that develops systems capable of mimicking tasks typically associated with human intelligence.
- **Applications:**
    - Used in analyzing large datasets.
    - Powers tasks like image recognition, decision-making, and language processing.

🧠 Machine Learning (ML)

- **Definition:**
    - A **subset of AI**.
    - Uses algorithms to learn from data and make predictions.
    - Does not require explicit programming of the decision logic.
- **Applications:**
    - Predictive analytics.
    - Recommendation systems.
    - Fraud detection (e.g., identifying suspicious credit card purchases).

🧬 Deep Learning

- **Definition:**

- A **subset of machine learning**.
- Employs **layered neural networks** to simulate human decision-making.

- **Neural Networks:**
  - Consist of computing units called **neurons**.
  - Learn to identify patterns from input data.
  - Example: Distinguishing between a dog and a cat.
- **Advantages:**
  - **Improves with larger volumes of data** (unlike some ML models that plateau in performance).

🧠 **Neural Networks Explained**

- **Structure:**
  - Composed of layers of artificial neurons.
  - Each neuron processes inputs and passes the result forward.
- **Function:**
  - Learns to make decisions through exposure to large datasets.

🎨 **Generative AI**

- **Definition:**
  - A **subset of AI** that focuses on **creating new content**, rather than analyzing existing data.
- **Capabilities:**
  - Generates images, music, language, and computer code.
  - Mimics human creativity.
- **Synthetic Data Generation:**
  - Used when there's insufficient real-world data.
  - Allows training and testing of models using artificially created datasets.

📈 **Regression Models in Machine Learning**

- **Purpose:**
    - Measure the relationship between input variables and outputs.
- **Usage:**
    - Used heavily in ML for:
        - Predictive modeling.
        - Statistical analysis.
- **Example:**
    - Predicting house prices based on size and number of bedrooms.
    - Helps determine **how strongly** these features influence price.

👷 AI in Data Science

- **Role of Data Scientists:**
    - Use AI tools, particularly ML and deep learning, to extract insights.
    - Apply models for tasks like prediction, classification, recommendation, and fraud detection.
- **Importance of Data:**
    - Larger and higher-quality datasets enhance model accuracy, especially for deep learning.

✨ Summary

This video summarizes the essential concepts of Artificial Intelligence, Machine Learning, and Deep Learning. It explains how data scientists utilize these technologies to extract insights, make predictions, and solve real-world problems. Key terms such as neural networks and generative AI are introduced, highlighting how AI mimics human intelligence and creates new content. Regression models are emphasized as foundational tools for analyzing relationships between variables. The lesson underlines the growing accessibility of AI and its increasing integration into data science workflows.

📊 Table: What We Learnt in the Video

Topic/Section	What We Learnt
Artificial Intelligence (AI)	The broad field that simulates human intelligence in machines.
Machine Learning (ML)	A subset of AI that learns from data to make predictions.
Deep Learning	The subset of ML uses neural networks; it improves with large datasets.

Neural Networks	Composed of neurons, learn to recognize patterns in data.
Generative AI	AI that creates new content (e.g., images, code); applicable for synthetic data.
Regression	A statistical method used in ML to measure input-output relationships.
Role of Data Scientists	Use AI tools to derive insights, predict trends, and identify anomalies.

Summary: Deep Learning and Machine Learning

Congratulations! You have completed this lesson. At this point in the course, you know:

- Big Data has five characteristics: velocity, volume, variety, veracity, and value.
- The five cloud computing characteristics are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.
- Data mining has a six-step process: goal setting, selecting data sources, preprocessing, transforming, mining, and evaluation.
- The availability of so many disparate amounts of data created by people, tools, and machines requires new, innovative, and scalable technology to drive transformation.
- Deep learning utilizes neural networks to teach itself patterns in inputs and outputs. Machine learning is a subset of AI that uses computer algorithms to learn about data and make predictions without explicitly programming the analysis methods into the system.
- Regression identifies the strength and amount of the correlation between one or more inputs and an output.
- Skills involved in processing Big Data include the application of statistics, machine learning models, and some computer programming.
- Generative AI, a subset of artificial intelligence, focuses on producing new data rather than just analyzing existing data. It allows machines to create content, including images, music, language, computer code, and more, mimicking creations by people.