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

The Big Data processing technologies provide ways to work with large sets of structured, semi-structured, and unstructured data so that value can be derived from big data. In some of the other videos, we discussed Big Data technologies such as NoSQL databases and Data Lakes. In this video, we are going to talk about three open-source technologies and the role they play in big data analytics — ApacheHadoop, Apache Hive, and Apache Spark. Hadoop is a collection of tools that provides distributed storage and processing of big data. Hive is a data warehouse for data query and analysis built on top of Hadoop. Spark is a distributed data analytics framework designed to perform complex data analytics in real-time. Hadoop, a Java-based open-source framework, allows distributed storage and processing of large datasets across clusters of computers. In Hadoop distributed system, a node is a single computer, and a collection of nodes forms a cluster. Hadoop can scale up from a single node to any number of nodes, each offering local storage and computation. Hadoop provides a reliable, scalable, and cost-effective solution for storing data with no format requirements. Using Hadoop, you can: Incorporate emerging data formats, such as streaming audio, video, social media sentiment, and clickstream data, along with structured, semi-structured, and unstructured data not traditionally used in a data warehouse. Provide real-time, self-service access for all stakeholders. Optimize and streamline costs in your enterprise data warehouse by consolidating data across the organization and moving “cold” data, that is, data that is not in frequent use, to a Hadoop-based system. One of the four main components of Hadoop is Hadoop Distributed File System, or HDFS, which is a storage system for big data that runs on multiple commodity hardware connected through a network. HDFS provides scalable and reliable big data storage by partitioning files over multiple nodes. It splits large files across multiple computers, allowing parallel access to them. Computations can, therefore, run in parallel on each node where data is stored. It also replicates file blocks on different nodes to prevent data loss, making it fault-tolerant. Let’s understand this through an example. Consider a file that includes phone numbers for everyone in the United States; the numbers for people with last name starting with A might be stored on server 1, B on server 2, and so on. With Hadoop, pieces of this phonebook would be stored across the cluster. To reconstruct the entire phonebook, your program would need the blocks from every server

in the cluster. HDFS also replicates these smaller pieces onto two additional servers by default, ensuring availability when a server fails, In addition to higher availability, this offers multiple benefits. It allows the Hadoop cluster to break up work into smaller chunks and run those jobs on all servers in the cluster for better scalability.

Finally, you gain the benefit of data locality, which is the process of moving the computation closer to the node on which the data resides. This is critical when working with large data sets because it minimizes network congestion and increases throughput. Some of the other benefits that come from using HDFS include: Fast recovery from hardware failures, because HDFS is built to detect faults and automatically recover. Access to streaming data, because HDFS supports high data throughput rates. Accommodation of large data sets, because HDFS can scale to hundreds of nodes, or computers, in a single cluster. Portability, because HDFS is portable across multiple hardware platforms and compatible with a variety of underlying operating systems. Hive is an open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in either HDFS or other data storage systems such as Apache HBase. Hadoop is intended for long sequential scans and, because Hive is based on Hadoop, queries have very high latency—which means Hive is less appropriate for applications that need very fast response times. Also, Hive is read-based, and therefore not suitable for transaction processing that typically involves a high percentage of write operations. Hive is better suited for data warehousing tasks such as ETL, reporting, and data analysis and includes tools that enable easy access to data via SQL. This brings us to Spark, a general-purpose data processing engine designed to extract and process large volumes of data for a wide range of applications, including Interactive Analytics, Streams Processing, Machine Learning, Data Integration, and ETL. It takes advantage of in-memory processing to significantly increase the speed of computations and spilling to disk only when memory is constrained. Spark has interfaces for major programming languages, including Java, Scala, Python, R, and SQL. It can run using its standalone clustering technology as well as on top of other infrastructures such as Hadoop. And it can access data in a large variety of data sources, including HDFS and Hive, making it highly versatile. The ability to process streaming data fast and perform complex analytics in real-time is the key use case for Apache Spark.

**Data Mining**

# 1. Establishing Data Mining Goals

The first step in data mining requires you to set up goals for the exercise. Obviously, you must identify the key questions that need to be answered. However, going beyond identifying the key questions are the concerns about the costs and benefits of the exercise. Furthermore, you must determine, in advance, the expected level of accuracy and usefulness of the results obtained from data mining. If money were no object, you could throw as many funds as necessary to get the answers required. However, the cost-benefit trade-off is always instrumental in determining the goals and scope of the data mining exercise. The level of accuracy expected from the results also influences the costs. High levels of accuracy from data mining would cost more and vice versa. Furthermore, beyond a certain level of accuracy, you do not gain much from the exercise, given the diminishing returns. Thus, the cost-benefit trade-offs for the desired level of accuracy are important considerations for data mining goals.

## **2. Selecting Data**

The output of a data-mining exercise largely depends upon the quality of data being used. At times, data are readily available for further processing. For instance, retailers often possess large databases of customer purchases and demographics. On the other hand, data may not be readily available for data mining. In such cases, you must identify other sources of data or even plan new data collection initiatives, including surveys. The type of data, its size, and frequency of collection have a direct bearing on the cost of data mining exercise. Therefore, identifying the right kind of data needed for data mining that could answer the questions at reasonable costs is critical.

## **3. Preprocessing Data**

Preprocessing data is an important step in data mining. Often raw data are messy, containing erroneous or irrelevant data. In addition, even with relevant data, information is sometimes missing. In the preprocessing stage, you identify the irrelevant attributes of data and expunge such attributes from further consideration. At the same time, identifying the erroneous aspects of the data set and flagging them as such is necessary. For instance, human error might lead to inadvertent merging or incorrect parsing of information between columns. Data should be subject to checks to ensure integrity. Lastly, you must develop a formal method of dealing with missing data and determine whether the data are missing randomly or systematically.

If the data were missing randomly, a simple set of solutions would suffice. However, when data are missing in a systematic way, you must determine the impact of missing data on the results. For instance, a particular subset of individuals in a large data set may have refused to disclose their income. Findings relying on an individual's income as input would exclude details of those individuals whose income was not reported. This would lead to systematic biases in the analysis. Therefore, you must consider in advance if observations or variables containing missing data be excluded from the entire analysis or parts of it.

## **4. Transforming Data**

After the relevant attributes of data have been retained, the next step is to determine the appropriate format in which data must be stored. An important consideration in data mining is to reduce the number of attributes needed to explain the phenomena. This may require transforming data Data reduction algorithms, such as Principal Component Analysis (demonstrated and explained later in the chapter), can reduce the number of attributes without a significant loss in information. In addition, variables may need to be transformed to help explain the phenomenon being studied. For instance, an individual's income may be recorded in the data set as wage income; income from other sources, such as rental properties; support payments from the government, and the like. Aggregating income from all sources will develop a representative indicator for the individual income.

Often you need to transform variables from one type to another. It may be prudent to transform the continuous variable for income into a categorical variable where each record in the database is identified as low, medium, and high-income individual. This could help capture the non-linearities in the underlying behaviors.

## **5. Storing Data**

The transformed data must be stored in a format that makes it conducive for data mining. The data must be stored in a format that gives unrestricted and immediate read/write privileges to the data scientist. During data mining, new variables are created, which are written back to the original database, which is why the data storage scheme should facilitate efficiently reading from and writing to the database. It is also important to store data on servers or storage media that keeps the data secure and also prevents the data mining algorithm from unnecessarily searching for pieces of data scattered on different servers or storage media. Data safety and privacy should be a prime concern for storing data.

## **6. Mining Data**

After data is appropriately processed, transformed, and stored, it is subject to data mining. This step covers data analysis methods, including parametric and non-parametric methods, and machine-learning algorithms. A good starting point for data mining is data visualization. Multidimensional views of the data using the advanced graphing capabilities of data mining software are very helpful in developing a preliminary understanding of the trends hidden in the data set.

Later sections in this chapter detail data mining algorithms and methods.

## **7. Evaluating Mining Results**

After results have been extracted from data mining, you do a formal evaluation of the results. Formal evaluation could include testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data. This is known as an "in-sample forecast". In addition, the results are shared with the key stakeholders for feedback, which is then incorporated in the later iterations of data mining to improve the process.

Data mining and evaluating the results becomes an iterative process such that the analysts use better and improved algorithms to improve the quality of results generated in light of the feedback received from the key stakeholders.

**Lesson Summary: Big Data and Data Mining**

Welcome to the big data and data mining lesson summary video. In this lesson, you gained insights into the impact of big data on various aspects of society, from business operations to sports. And developed an understanding of key attributes and challenges associated with big data. In this video, we will recap fundamentals of big data and how big data drives digital transformation. How data scientists leverage the essential characteristics of the Cloud to gain insights from big data, the data mining process, and common tools used to process big data. The availability of vast amounts of data, resulting in what we now call big data, is driving transformation in business and industry and consequently, how we live our daily lives. Organizations realize that we require fundamental changes to their approach to business, impacting every aspect of the organization. The availability of so many disparate amounts of data created by people, tools, and machines requires new, innovative, and scalable technology. Big data drives us to derive real-time business insights relating to consumers risk, profit, performance productivity management, and ultimately enhancing business values. Not everyone agrees on the definition of big data, but people generally agree on the five characteristics of this data value, volume, velocity, variety, and veracity. People expect investing time in studying big data will create value. Volume refers to the scale of the data, drivers of volume include increasing collectible data sources and scalable infrastructure. Velocity indicates ever-increasing sources of nonstop processes that generate data quickly. Variety reflects that related data comes from different sources, both structured and unstructured. Veracity refers to the quality and origin of data and that it accurately conforms to facts.

The development of cloud and cloud technologies enables us to work with big data. The cloud refers to the delivery of on-demand computing resources on a pay-for-use basis. Cloud computing has five essential characteristics, on-demand, network access, resource pooling, elasticity, and measured service. On-demand means access to processing, power, storage, and network that you need. These computing resources can be accessed via a network with Internet access. Resource pooling allows providers to service multiple consumers with the resources dynamically assigned according to demand, making cloud computing cost-efficient. Elasticity means that you can access resources as you need them and automatically scale back when you don't. With measured service, you only pay for what you use or reserve as you go. You also gain an understanding of how cloud computing addresses challenges related to scalability, collaboration, and accessibility. And software maintenance, making it a valuable resource for data analysis and other computational tasks. The Cloud gives you instant access to technologies without needing to install or configure them and provides updated versions of these tools as they get released. Popular open-source tools to compute using big data include Apache Hadoop,

Apache Hive, and Apache Spark. Hadoop provides distributed storage and processing tools across clusters of computers. Hive is a data warehouse for data query and analysis built on top of Hadoop. Hive allows you to read, write, and manage large datasets directly in the Hadoop File system or HDFS or Apache HBase. Spark provides a general-purpose data processing engine designed to extract and process large volumes of data for a wide range of applications. Big data requires a process called data mining to make use of. This six-step process includes goal setting, selecting data sources, preprocessing, transforming, mining, and evaluation. In the first step, goal setting, you identify key questions you want to answer, concerns about cost and benefits should inform this step. Once you identify the questions, select the data by identifying sources or planning data collection initiatives.

In the next step, preprocessing, you identify irrelevant attributes of data and enormous aspects of the data by flagging them as necessary. After preprocessing, you transform the data by determining the appropriate format to store the data. Now you get to mine the data, which includes determining analysis methods and the machine learning algorithms you will use to process the data. Once the data has been mined, you finally must evaluate your outcomes. By testing the predictive capabilities of the models on the observed data to find the effectiveness and efficiency of your algorithms. In addition, you share your results with stakeholders. This entire process should be conducted iteratively as your results from this iteration will inform further data mining efforts. In summary, big data characteristics that data scientists agree on, even though they might not agree on the exact definition, include value, volume, velocity, and veracity. Data with these qualities is driving transformation across industries and in our daily lives. In large part, cloud technologies enable us to handle big data because they provide ubiquitous access to computational power and storage capacity. Open-source cloud tools such as Hadoop, Hive, and Spark leverage these advantages, allowing us to effectively and efficiently mine big data.

# Glossary: Big Data and Data Mining

# Artificial Intelligence and Data Science

In data science, there are many terms that are used interchangeably, so let's explore the most common ones. The term big data refers to data sets that are so massive, so quickly built, and so varied that they defy traditional analysis methods such as you might perform with a relational database. The concurrent development of enormous compute power in distributed networks and new tools and techniques fordata analysis means that organizations now have the power to analyze these vast data sets. A new knowledge and insights are becoming available to everyone. Big data is often described in terms of five V's; velocity, volume, variety, veracity, and value. Data mining is the process of automatically searching and analyzing data, discovering previously unrevealed patterns. It involves preprocessing the data to prepare it and transforming it into an appropriate format. Once this is done, insights and patterns are mined and extracted using various tools and techniques ranging from simple data visualization tools to machine learning and statistical models. Machine learning is a subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it is learned without being explicitly programmed. Machine learning algorithms are trained with large sets of data and they learn from examples. They do not follow rules-based algorithms. Machine learning is what enables machines to solve problems on their own and make accurate predictions using the provided data. Deep learning is a specialized subset of machine learning that uses layered neural networks to simulate human decision-making. Deep learning algorithms can label and categorize information and identify patterns. Itis what enables AI systems to continuously learn on the job and improve the quality and accuracy of results by determining whether decisions were correct. Artificial neural networks, often referred to simply as neural networks, take inspiration from biological neural networks, although they work quite a bit differently. A neural network in AI is a collection of small computing units called neurons that take incoming data and learn to make decisions over time. Neural networks are often layer-deep and are the reason deep learning algorithms become more efficient as the data sets increase in volume, as opposed to other machine learning algorithms that may plateau as data increases. Now that you have a broad understanding of the differences between some key AI concepts, there is one more differentiation that is important to understand that between Artificial Intelligence and Data Science. Data Science is the process and method for extracting knowledge and insights from large volumes of disparate data. It's an interdisciplinary field involving mathematics, statistical analysis, data visualization, machine learning, and more. It's what makes it possible for us to appropriate information, see patterns, find meaning from large volumes of data and use it to make decisions that drive business. Data Science can use many of the AI techniques to derive insight from data. For example, it could use machine learning algorithms and even deep learning models to extract meaning and draw inferences from data. There is some interaction between AI and Data Science, but one is not a subset of the other. Rather, Data Science is a broad term that encompasses the entire data processing methodology while AI includes everything that allows computers to learn how to solve problems and make intelligent decisions. Both AI and Data Science can involve the use of big data. That is, significantly large volumes of data.

# Generative AI and Data Science

Welcome to Generative AI and Data Science. After watching this video, you will be able to: Describe generative AI and Explain how data scientists use generative AI in data science. 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. How does generative AI operate, though? Deep learning models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are at the foundation of this technique. These models create new instances that replicate the underlying distribution of the original data by learning patterns from enormous volumes of data. Generative AI has found diverse applications across various industries. Let’s look at some fascinating examples! Natural language processing like OpenAI’s GPT-3 can generate human-like text, revolutionizing content creation, and chatbots. In healthcare, Generative AI can synthesize medical images, aiding in the training of medical professionals. Generative AI can create unique and visually stunning artworks, generating endless creative visual compositions. Game developers use Generative AI to generate realistic environments, characters, and game levels. Generative AI assists in fashion by designing new styles and creating personalized shopping recommendations. Now, let's discuss how data scientists use Generative AI. Training and testing a model takes lots of data. Sometimes, the data you want to study doesn’t have enough observations to build a model. Interest in synthetic data as a tool for analysis and model creation has increased due to recent developments in generative AI. Data scientists can augment their data sets using generative AI to create synthetic data. It creates this data with similar properties as the real data, such as its distribution, clustering, and many other factors the AI learned about the real data set. Data scientists can then use the synthetic data along with the real data for model training and testing. Data scientists, researchers, and analysts frequently find themselves confined by time when examining data patterns and insights. Due to this restriction, they can only conceive, develop, and evaluate a small number of hypotheses, leaving many other possibilities untested. With generative AI, data scientists can leverage generative AI to generate and test software code for constructing analytical models. Coding automation has the potential to revolutionize the field of analytics, allowing the data scientist to focus on higher-level tasks such as identifying and clarifying the problem the models intend to solve and evaluating hypotheses from a wider range of data sources. Generative AI can generate accurate business insights and comprehensive reports, making it possible to update these insights as the data evolves. Furthermore, it can autonomously explore data to uncover hidden patterns and insights that might go unnoticed during manual analysis and enhance decision-making. For instance, IBM’s Cognos Analytics, which provides AI-powered automation, enables you to unlock the full potential of your data with the help of its natural language AI assistant. You describe the question you are looking to answer or the hypothesis you want to test, and it helps generate insights you need.

In this video, you learned that: Generative AI, a subset of artificial intelligence, focuses on producing new data rather than analyzing existing data. Generative AI augments data science efforts, enabling deeper insights, addressing data limitations, and improving the overall quality of data-driven outcomes.

# Neural Networks and Deep Learning

It's, I guess, Computer Sciences attempt to mimic real, the neurons, in how our brain actually functions. So 20-23 years ago, a neural network would have some inputs that would come in. They would be fed into different processing nodes that would then do some transformation on them and aggregate them or something, and then maybe go to another level of nodes. And finally there would some output would come out, and I can remember training a neural network to recognize digits, handwritten digits and stuff. So a neural network is trying to use a computer, a computer program that will mimic how neurons, how our brains use neurons to process things, neurons, and synapses, and building these complex networks that can be trained. So this neural network starts out with some inputs and some outputs, and you keep feeding these inputs in to try to see what kinds of transformations will get to these outputs. And you keep doing this over, and over, and over again in a way that this network should converge. So this input, the transformations will eventually get these outputs. The problem with neural networks was that even though the theory was there they did work on small problems like recognizing handwritten digits and things like that. They were computationally very intensive and so they went out of favor and I stopped teaching them probably 15 years ago. And then all of a sudden we started hearing about deep learning, heard the term deep learning. This is another term, when did you first hear it? Four years ago, five years ago? And so, I finally said, what the hell is deep learning? It's really doing all this great stuff, what is it? And I Google, I was like, this is neural networks on steroids. What they did was they just had multiple layers of neural networks, and they used lots, lots, and lots of computing power to solve them. Just before this interview, I had a young faculty member in the marketing department whose research is partially based on deep learning. And so she needs a computer that has a Graphics Processing Unit in it because it takes an enormous amount of matrix and linear algebra calculations to actually do all of the mathematics that you need in neural networks. But they've been they are now quite capable. We now have neural networks and deep learning that can recognize speech, recognize people, you got there, getting your face recognized. I guarantee that NSA has a lot of work going on in neural networks. At the university right now, as director of research computing, I have some small set of machines down at our south data center, and I went in there last week and there were just piles, and piles, and piles of cardboard boxes all from Dell with a GPU on the side. Well, the GPU is a Graphics Processing Unit. There's only one application in this University that needs two hundred servers each with Graphics Processing Units in it, and each Graphics Processing Unit has the equivalent of 600 cores of processing. So this is tens of thousands of processing cores that are for deep learning, I guarantee. Some of the first ones are speech recognition, who teaches the deep learning class at NYU and is also the head data scientist at Facebook comes into class with a notebook, and it's a pretty thick notebook. It looks a little odd because it's like this and it's that thick because it has a couple of Graphics Processing Units in it, and then he will ask the class to start to speak to this thing. And it will train while he's in class, he will train a neural network to recognize speech. So recognizing speech, recognizing people, images, classifying images, almost all of the traditional tasks that neural nets used to work on in little tiny things. Now, they can do really, really, really large things. It will learn on its own, the difference between a cat and a dog, and different kinds of objects, it doesn't have to be taught. It doesn't, it just learns that's why they call it deep learning, and if you hear, he plays this, if you hear how it recognizes speech and generates speech.

It sounds like a baby who learning to talk. You can just, you're like really do about all of a sudden this stupid machine is talking to you and learned how to talk. That's cool. I need to learn some linear algebra, a lot of a lot of this stuff is based on matrix and linear algebra. So you need to know how to use linear algebra to do transformations. Now, on the other hand, there are lots of packages out there that will do deep learning and they'll do all the linear algebra for you, but you should have some idea of what is happening underneath. Deep learning, particularly needs really high-powered computational power. So it's not something that you're going to go out and do on your notebook for it. You could play with it. But if you really want to do it, seriously, you have to have some special computational resources.

# Applications of Machine Learning

Everybody now deals with machine learning. But recommender systems are certainly one of the major applications. Classifications, cluster analysis, trying to find some of the marketing questions from 20 years ago, market basket analysis, and what goods tend to be bought together. That was computationally a very difficult problem, I mean we're now doing that all the time with machine learning. So predictive analytics is another area of machine learning. We're using new techniques to predict things that statisticians don't particularly like. Decision trees, Bayesian Analysis, naive Bayes, and lots of different techniques. The nice thing about them is that in packages like R now, you really have to understand how these techniques can be used and you don't have to know exactly how to do them but you have to understand what their meanings are. Precision versus recall and the problems of over-sampling and over-fitting so you can, someone who knows a little about data science can apply these techniques but they really need to know, maybe not the details of the technique as much as what the trade-offs are. So, some applications of machine learning in fintech are probably a - couple of different things I could talk about there. One of them is recommendations. Right, so, when you use Netflix, or you use Facebook, or a lot of different software services, the recommendations are served to you. Meaning, "Hey, you're a user, you've watched this show, so maybe you'd like to see this other show." Right, or, you follow this person, so maybe you should follow this other person. It's actually kind of the same thing in fintech, right. Because you've looked at - if you're an investment professional, right, and because you've looked at this investment idea, it might be really cool for you to look at this other investment idea, which is kind of similar. Right, it's a similar kind of asset, it's a similar kind of company. Or it's a similar kind of technique for doing the investment. So, We can apply recommendations using machine learning throughout a lot of different parts of fintech. Another one that people talk about, and is important, especially on retail, in the retail aspects of banking and finance is fraud detection. Trying to determine whether a charge that comes from a credit card is fraudulent or not, in real time, is a machine learning problem. Right, you have to learn from all of the transactions that have happened previously and build a model, and when the charge comes through you have to compute all this stuff and say, "Yeah we think that's ok," or "hmm, that's not so good. Let's route it to, you know, our fraud people to check."

# Lesson Summary: Deep Learning and Machine Learning

Welcome to the Deep Learning and Machine Learning lesson summary. In this video, we’ll review what you learned about Deep Learning and Machine Learning from the videos and reading in this lesson. We will recap: Many of the terms commonly used in artificial intelligence and how to differentiate them, How data scientists use Artificial intelligence OR AI, The relationship between machine learning and data science, AND The regression model used to show relationships between things. Artificial intelligence (AI) has boomed and become accessible to almost everyone. Data scientists use AI regularly when analyzing data. Let’s discuss some of the terms used in this lesson related to AI, such as machine learning, deep learning, neural networks, and generative AI. AI is the branch of computer science that includes the development of systems that can mimic many of the tasks associated with human intelligence. Machine learning is a subset of AI that uses computer algorithms to learn about data and make predictions with it without needing to explicitly program the analysis methods into the system. Deep learning is a subset of machine learning that uses layered neural networks to simulate human decision-making. A neural network is a collection of small computing units, called neurons, that take incoming data and learn to make decisions over time, such as the difference between a dog and a cat. Deep learning algorithms become more efficient as the amount of data increases in volume, unlike other machine learning algorithms, which tend to plateau. Another subset of AI, generative AI, focuses on producing new data rather than just analyzing existing data. It allows machines to create content, including images, music, languages, and computer code, mimicking human creations. Generative AI also can make data sets with similar traits to a raw data set. Sometimes, data scientists, when they do not have enough data, can create synthetic data and use it to train and test their models. As a data scientist, you will use machine learning algorithms to derive insights from data. You will frequently apply machine learning algorithms for predictive analytics or make recommendations. For example, you may also use these algorithms for fraud detection to identify fraudulent credit card purchases based on previous purchasing habits. Machine learning algorithms rely heavily on statistical technique called regression. Regression identifies the strength and amount of the correlation between one or more inputs and an output. For instance, how much does the price of a house increase based on its square footage and number of bedrooms, and how confident can you be of this relationship?

In summary, Generative AI produces new data like a raw data set. deep learning is a subset of machine learning, and machine learning is a subset of artificial intelligence. Deep learning utilizes neural networks to teach itself patterns in inputs and outputs. Using big data, data scientists use all of these areas of AI to make predictions.