Content Document: Fundamentals of Big Data

Summary

This course was created for individuals who are new to the big data landscape and want to become conversant with big data terminology. It will cover foundational concepts related to the big data landscape including:

- Characteristics of big data
- Techniques used to work with big data such as artificial intelligence
- How individuals on data science teams work with big data
- How organizations from a variety of industries can use big data to enable better business decisions

Note: This course will **not** cover Databricks concepts or functionality. This is an introductory-level course focused on big data concepts.

Description

Who should take this course?:
Business leaders, managers or anyone interested in learning more about the big data landscape.

Pre-requisites:
Experience using a web browser.

Course goals:
By the end of the course, you will be able to:

- Explain foundational concepts used to define big data.
- Explain how the characteristics of big data have changed traditional organizational workflows for working with data.
- Summarize how individuals on data science teams work with big data on a daily basis to drive business outcomes.
- Articulate examples of real-world use-cases for big data in businesses across a variety of industries.

Course welcome

Welcome to Introduction to Big Data, an introductory-level course developed to help you become conversant with big data terminology.

In this course, we'll explore fundamental concepts related to the big landscape that will help you understand why the use of big data is exploding across businesses wishing to increase business value.

By the end of the course, you will be able to:

- Explain foundational concepts used to define big data.
- Explain how the characteristics of big data have changed traditional organizational workflows for working with data.
- Summarize how individuals on data science teams work with big data on a daily basis to drive business outcomes.
- Articulate examples of real-world use-cases for big data in businesses across a variety of industries.

Let's get started.

Section one - What is big data?

Lesson 1: Technology and the explosion of data generation (Video)

RISE INTRO (TEXT):
Today, our world is more interconnected than ever before and the amount of data being generated has increased dramatically.

In this video, we'll explore how technology has led to the explosion of big data.

VIDEO SCRIPT:
In the last twenty years, advances in technology have completely changed the world that we live in. Today, we have many new sources generating data that we didn't have in the past. In this video, we'll review some of these sources so that we can answer the question - where is all of this data coming from?

Take a moment to think about what you've done today.

Have you texted someone or posted something on social media?

Perhaps you did some exercise and logged your workout using a fitness tracker.

Or, you might have used a mapping application to help you get to work, school, or to run some errands.

At a minimum, you logged in to the Databricks Academy to watch this video.

And, as you did all of this, you, along with the machines and tools you used to do these things, were generating data.

When we think of human-generated data, or data coming directly from humans, we need to talk about social media.
The explosion of social media has been one of the leading factors in the increase in human-generated data. Every time we:

- Post a message
- Change our online statuses
- Upload images

  like or forward comments, and more,

we are generating data. Let's look at Facebook for example. according to Forbes, 1.5 billion people are active on Facebook every day, 510,000 comments are posted every minute, and five new Facebook profiles are created every second.

Human-generated data also consists of things we create:
Those could be spreadsheets, presentations, video files, audio files, and more.

In addition to human-generated data, exists machine-generated data, or data that is generated from machines and that doesn't rely on active human intervention. Examples of machine-generated data sources include:

- Sensors on vehicles, appliances and industrial machinery
- Cameras used for security vigilance
- Satellites
- Medical devices
- And personal tools such as smartphone apps or fitness trackers

Think of all of the machine-generated data that comes from a fitness tracker, for example. Depending on the model you have, it might generate records for your heart rate, your geographic location, calories burned and more. You aren't telling the fitness tracker to track these things - it just does it. And, if you want to look at the data later, you can probably use a smartphone application to do so.

In the big data world, you'll also often hear the term transactional data, which refers to the information generated as organizations run their businesses - when you purchase something, for example.

For every purchase you make in a physical store or online, think of the transactional data generated for each purchase - things like:

- Your customer number
- The items you purchased
- The date and time you purchased the items
- And how many of each item you purchased

All of this data - whether it's generated primarily by humans or machines, can become big data. And, all of this has been made possible through advances in technology that make it easier for all of us - humans, machines and organizations, to generate data.

RISE TRANSITION TO NEXT LESSON (TEXT):
It's important to note that although all of these sources are constantly generating data, this data isn't all necessarily considered big data. Think of yourself sending text messages throughout the day, for example. Do you send enough of them for your individual text messaging to be considered "big data" generation? Probably not.

In our next lesson, we'll explore the characteristics of big data that make it "big".

Lesson 2: What makes big data "big"? (Text + GIFs)

RISE INTRO (TEXT):
Why do we call some data "big data"? In this video, we'll review the major characteristics inherent to big data.

VIDEO SCRIPT:
Big data refers to data that is nearly impossible to process using traditional methods, because there is so much of it, coming in so quickly, and in so many different formats. It is data that cannot just be processed, for example by a single machine because it quickly outgrows that.

We can summarize the major characteristics used to define big data as Volume, Velocity and Variety.

Let's start by reviewing data volume.

Volume refers to the incredible amount of data being generated every second of every day.

The International Data Corporation or IDC, forecasts that the global datasphere, or the amount of data that exists in the world will grow from 33 zettabytes in 2018 to 177 zettabytes by 2025. Just to put that into perspective, the computer I'm on right now has 256GB of storage. That's equivalent to just .000000000256 (9 zeros) zettabytes.

Another way to think about this is that if we take all the data generated in the world between the beginning of time and 2008 the same amount of data will soon be generated every single minute. (Bernard Marr, Bernard Marr & Co.)

The second characteristic that defines big data is velocity, which refers to the speed at which new data is generated and the speed at which data moves around.

Just think of social media messages going viral in seconds. We've all heard stories of videos that were posted one day and the poster is shocked when the very next day there are millions of views.

What about the speed at which credit card transactions are checked for fraudulent activities? Have you ever tried to purchase something a little out of the ordinary, only to have the transaction blocked by your credit card company? In just a matter of seconds, your credit card company received information about your purchase, was able to compare it to usual purchases you make, and decide whether or not to flag this as a fraudulent transaction.

Finally, when we look at big data, we look at data variety. There are so many different types of data that we work with today - social media posts, credit card transactions, legal contracts, biometric data, and geographic information, just to name a few.

These characteristics of big data - volume, velocity and variety, and also known as the three V's of big data. They are what set data apart as big data.

RISE TRANSITION TO NEXT LESSON (TEXT):
Now that we've defined the characteristics of big data, we'll take a closer look at data variety to explore how different types of data are structured.

Lesson 3: Types of big data (Text and GIFs)

RISE INTRO (TEXT):
In our last lesson, we reviewed the characteristics of big data. One of those characteristics was data variety. In this lesson, we'll explore data variety at a deeper level and look at how different types of data are structured.

Big data is categorized as:

Structured
Semi-structured

or
Unstructured

These categorizations are important, because data structures play a key role in how data practitioners can process and use data. We'll explore this idea as we go explore each data structure.

Structured data

The term structured data refers to any data that conforms to a certain format or schema. If we take a look at this spreadsheet for example, we see that there are clearly labeled rows and columns, and the information within those rows and columns also follows a schema.

Because structured data is clearly organized, it's generally easier to analyze. For example, if I asked you to sum the total of sales, for example, you could do it because all the data in the column for "sales" are numeric and can be summed up.

A lot of the data that organizations work with everyday can be categorized as structured data.

Some examples of structured data include:

- An address databases, or perhaps
- A product database

Unstructured data
By contrast, unstructured data. It's often referred to as "messy" data, because it isn't easily searched compared to structured data. For example, imagine that you have an hour-long video of movie reviews, and I ask you to find any reference to a comedic movie - that task is much harder to do than in the CSV file we had above.

Unstructured data is the most widespread type of data. IDC reports that almost 90% of data today is unstructured. Today, many organizations struggle with trying to make sense of unstructured data, especially when trying to use it for business insights. That's where different fields of artificial intelligence become an important part of the data analysis process.

Aside from videos, other examples of unstructured data include:

- Social media posts
- Photographs
- Emails
- Audio files, and
- Images

Semi-structured

Finally, we have semi-structured data. Semi-structured data fits in between structured and unstructured data. Semi-structured data does not reside in a formatted table, but it does have some level of organization. A good example of semi-structured data is HTML code. If you've ever right clicked in your browser and selected "inspect" or "inspect element" you've seen an example of this. Although you are not restricted to how much information you want to collect or what kind of information you want to collect, there is still a defined way to express data.

RISE OUTRO:
Congratulations! We're at the end of our first lesson. By now, you should have a good understanding of what big data is, at a high-level.

Click next to take a short quiz about this section.

Knowledge check

1. Which of the following are examples of transactional data? Choose two.
    a. **Purchase orders**
    b. Social media posts
    c. Audio files
    d. **Insurance claims**
    e. Videos files
2. Imagine that the data analyst at a movie streaming service company is given access to a a data table that contains information about the movies customers are streaming. Millions of records are added to this table every hour.

    Think about the characteristics of big data. Which characteristics does this scenario relate to? Choose two.
    a. **Volume**
    b. **Velocity**
    c. Variety
    d. Veracity
3. Which of the following statements about unstructured data are true? Choose two.
    a. It is generally easier to analyze than other types of data.
    b. **It is often referred to as "messy" data.**
    c. It fits neatly into a schema.
    d. **It is the most widespread type of data.**
    e. It is usually found in tables.

Section two - Managing big data

Lesson 1: Distributed computing (Video)

RISE INTRO (TEXT): In the last section, we reviewed fundamental characteristics about big data. Now, we'll explore how these characteristics affect the way that big data is managed.

We'll start by talking about distributed computing, which is what allows us to process big data.

SCRIPT:
You might have had a situation as a young child in school, where you and your classmates were asked to guess how many pieces of candy were in a jar. The person who guessed the correct number got to take the whole jar of candy home and eat candy for days.

Now imagine this - what if I gave you and a group of your colleagues an entire tub of candies and told you that whoever finished counting the candies first wins a prize. The only rules are the count has to reflect the exact number of candies in the jar, and, this time, you're allowed to touch the candies.

What would you do?

You might line the candies up and count them one-by-one. While this would eventually give you an accurate count, it would take a really long time.

What about weighing the candy? You could weigh one candy, weigh the entire tub of candy, and figure out how many pieces of candy there are based on the total weight. While this would give you a count relatively quickly, you can't guarantee that each individual piece of candy weighs the same, so you might not have an accurate count.

Let's look at this from a different viewpoint. Let's bring in our friends.

What if we split up the entire tub of candy between our friends? We can distribute the candies among our friends so that instead of us having to do the whole count on our own, our friends are all counting a subset of the candies at the same time, helping us get through the tub of candy much faster.

This is the idea behind distributed computing -we divide the data into smaller chunks and distribute it among different computers. We can then perform whatever our computation is, in this case a count, in parallel, on the smaller subsets of data.

An important name to know in the world of distributed computing is Apache Spark. If you haven't, Apache Spark is the defacto standard when it comes to distributed computing tools used for big data processing and analytics. Most organizations either already use Apache Spark to process their big data, or are in the process of migrating to Apache Spark to help process their big data.

If we think about Apache Spark in this example, we, the distributor of the work, would be called drivers. We are the ones distributing piles of candies to our friends and asking them to count them. We're not counting the candies ourselves, but we're keeping track of what is going on and at the end, will be given our final results. Our friends who are counting the candies could be called simple executors - they are the ones doing the work - they are counting the candies. And, the entire unit of work we're doing, counting the tub of candies, would be called a job.

Back to our candy example - once our workers count up all of the candies and have their individual counts, stage one of our job is done. To finish the job however, we need an aggregate of all of the individual counts. All we need for stage two of our job is for one of the workers to add up all of the individual counts and give us a final answer. Perfect! Now we know how many candies are in the tub. Our job is now complete.

RISE OUTRO (TEXT): Now, let's take the idea of distributed one step further and talk about the types of input data organizations have to be able to manage - batch data and streaming data.

Lesson 2: Batch vs. streaming processing (Video)

RISE INTRO (TEXT): The two types of input data that organizations must manage today are batch and streaming data. In this video, we'll revisit our candy jar example from the previous lesson and explore the characteristics of each of these.

SCRIPT:

In the big data world, there are two types of input data we can have — batch data and streaming data. These terms, batch and streaming, refer to the way that we're getting our data, and the speed at which we're getting our data.

Batch data is data that we have in storage and that we process all at once, or in a batch. Say for example that someone gives you a large jar of candies and asks you to count all of the candies in the jar. That is a simple example of a batch job. We took the candies that were already present in some form of storage (in this case, a jar) and counted them. We counted ALL of them, one time.

A real-world example of batch processing is how telecommunication companies process cellular phone usage each month to generate our monthly phone bills. To do this, they crunch batch data - the phone calls you've made, text messages you've sent, and any additional charges you've incurred through that billing cycle, to generate your bill. They use data stored in their data store and crunch all of it once a month, in a batch job.

On the other hand, we have streaming data. Streaming data is data that is being continually produced by one or more sources, and therefore must be processed incrementally as it arrives. Now, what if instead of counting candy sitting in a jar, we are asked to count candy coming towards us on a conveyor belt? As the candy reaches us, we have to count the new pieces and constantly update our overall candy count. In a streaming job, my final count is changing in real-time as more and more arrives on the conveyor belt.

A real-world example of stream processing is how heart monitors work. All day long, as you wear your heart monitor, it receives new data - dozens of thousands of data points per day as your heart beats. And, every time your heart beats, your heart monitor has new data added to its data store in real-time. If your heart monitor has a display of your average heartbeat for the day, that average must be constantly updated with the new numbers from the incoming stream of data.

Both batch and streaming data have their place when it comes to big data analytics. We need to use batch data for things like periodic reporting, and streaming data for things like fraud detection which need to be identified in real-time. Historically it's been difficult to use these different types of data in conjunction.Thanks to new advances in technology however, with new products like Delta Lake, combining batch and stream processing is possible and it leads to significant advantages in solving real-world, practical problems using data analytics.

RISE OTRO (TEXT):

At this point, we've reviewed how we process big data and have explored the types of input data we have to work with. Next, we'll discuss another topic in data management - where to store big data.

Lesson 3: Data storage systems (Text + GIF)

RISE INTRO (TEXT): Once we have collected and processed batch and streaming data, we need somewhere to put it. In this lesson, we'll review the most popular types of big data storage systems and review the benefits and drawbacks of each. Specifically, we'll explore data warehouses, data lakes, and cloud data platforms.

VIDEO SCRIPT:

As you can imagine, storing big data requires a lot of space. It is no longer the case that an organization can store all of their data on a single computer or server.

Today, most organizations are storing their big data in one or a combination of the following storage systems:

- Data warehouses
- Data lakes
- And unified data platforms

Data warehouses

We'll start with looking at data warehouses.

Data warehouse technology emerged in the 80's and provides a centralized repository for storing all of an organization's data. They can be on-premises or in the cloud.

Benefits of data warehouses include that:

- They've been around for decades, work well for structured data and are reliable.
- Since they generally only take structured data, data is typically clean and easy to query.

Challenges with data warehouse include that:

- They can be hard and expensive to scale (if you need more space, for example)
- You lose a lot of valuable potential by not taking advantage of unstructured data
- Plus, you often have to deal with vendor lock-in, when your data is stored in a system that does not belong to you.
- And, data warehouses are very expensive to build, license and maintain, especially for large data volumes, despite the availability of cloud storage.

Data lakes

Unlike data warehouses, data lakes store data in its raw format. Data lakes can store unstructured as well as structured data, and are known to be more horizontally scalable (in other words, it's easy to keep adding more data in).

Benefits of using data lakes include:

- The ability to hold different types of data
- They're easier to scale since they are usually cloud-based
- They allow organizations to capture and keep all of their data even (which is great, if you're not quite sure what to do with it), since cloud storage is so cheap
- Plus, they allow you to separate data storage from data compute, leading to cost savings. In other words, you pay only to store your data, until you need to do something with it.

Some issues with data lakes include that:

- Individuals unfamiliar with working with raw data can experience a bit of a learning curve or difficulties navigating a data lake
- Due to larger volumes of data and occasional lack of structure, query speeds can be impacted in traditional data lakes

Unified data platforms

Finally, a data storage system that is quickly gaining popularity today is the unified data platform. These provide all of the benefits of data lakes, with the addition of some data warehousing capabilities, all wrapped up in a platform that your data teams can work in together.

Benefits of using unified data platforms include:

- They give you a single source of truth for your data and can guarantee data validity
- You don't have to worry about maintaining your own physical infrastructure - a third party service does that for you - like Amazon Web Services or Microsoft Azure
- They are easily scalable - both in terms of storage and compute
- They offer you a collaborative space where members of your data team, with various levels of programming ability can work together

Some issues with unified data platforms include:

- The ability to scale instantly can become very costly if you don't pay attention to how much compute your organization is using
- Vendor lock-in can become a problem, if the platform provider requires you to store your data with them

RISE OUTRO (TEXT): Congratulations! We're at the end of our second lesson. By now, you should have a good understanding of what big data is, at a high-level, and how it is managed.

Click next to take a short quiz about this section.

Knowledge check

1. In distributed computing using Apache Spark, the individual computers tasked with processing stages of a job are called:
    a. Workers
    b. Drivers
    c. Processors
    d. Executors
2. Which of the following scenarios below is an example of stream processing? Choose two.
    a. Financial transactions processed once a day by a bank
    b. Cell phone bills generated each month by a phone company
    c. Heart rate collected continuously by a fitness tracker
    d. Geographic coordinates updated by a smart car sensor as it is driven
3. Which of the following statements describes how unified cloud data platforms differ from data lakes?
    a. They are easy to scale since they are usually cloud-based
    b. They provide a collaborative space where your teams can work together
    c. They have the ability to hold structured, semi-structured and unstructured data
    d. They allow you to separate data storage from data compute

Section three - Extracting insights from big data

Lesson 1: Techniques for working with big data (Combo - text / GIFs + video)

RISE INTRO (TEXT):

At this point, we've covered the big data landscape and concepts related to how organizations manage big data. Now, we'll explore techniques that organizations use to work with big data. We'll start by looking at data science, artificial intelligence and machine learning.

SCRIPT:
Data science, artificial intelligence and machine learning are techniques used by organizations to extract insights from big data. While they are related, they each mean something different.

Data science

Data science is very popular in businesses today as a way to extract insights from big data to help inform business decisions. It is a field that combines tools and workflows from disciplines like math and statistics, computer science and business, to process, manage and analyze massive amounts of data. Some artificial and machine learning techniques fall under the umbrella of data science, as they are also techniques that are used to help extract insights from data.

Artificial intelligence

Machines enabled with artificial intelligence can be taught to understand, learn and make decisions that mimic the decisions that a human would make.

Say for example that you own a company and want to know if customers are buying your product. Using algorithms, a computer can process different scenarios and make decisions.

For example - are customers buying my product? Yes!
Okay, is customer churn low? Yes!
Great! We can operate business as usual.

What about - customers are buying my product but customer churn is not low? Well, then you're going to want to make some changes.

And, of course, if customers are not buying your products, you'll also want to make some changes.

This is a high-level example, but this is what organizations are trying to do, at a much deeper and more complicated level, to understand the nuances of their business.

Machine learning

What is machine learning?

Machine learning is a subset of artificial intelligence. It refers to a process by which humans program computers to learn on their own. They learn something from data, they evaluate how well they're learning, and they can adjust to improve their learning as new data becomes available. This process helps them improve their learning over time, allowing them to answer more complex questions to solve more complex problems. Our last example was not machine learning - if we implemented that, it would be a very short uncomplicated program, but also not all that useful in the long term.

How does machine learning work?

Say that someone creates an algorithm to help train a computer to recognize a specific image - a koala.

The algorithm is fed into the computer with hundreds, thousands, or millions of pictures - some of these showing koalas so that it can learn what a koala looks like, and others not showing koalas so that it can learn what a koala does not look like. The more pictures that are fed into the computer, the more it learns. Over time, it can more easily and quickly identify a koala over other images. And, while humans might recognize koalas by their fluffy ears or large oval-shaped noses, a computer will detect things that we cannot - things like distances between the koala's eyes, or patterns in the lines in it's fur.

Let's go back to our customer analogy. With machine learning, we could train a computer to look for what a churning customer looks like by feeding it customer data and writing algorithms that tell it to look for signs like a decrease in product usage, increase in customer service calls and low rating on customer surveys. Over time, the computer will pick up on other patterns as well, and can help us identify a churning customer before they leave us.

Machine learning is incredibly powerful because it helps to scale programming. Instead of humans processing all of this data and making decisions manually, computers do it with human oversight and find patterns that we cannot.

RISE OUTRO (TEXT): While this was not an exhaustive list, these are some of the most popular techniques for working with big data. In the next lesson, we'll take a deeper look at the data science workflow, a cyclical process for working with big data.

Lesson 2: The data science workflow (Video)

RISE INTRO (TEXT): The data science workflow is a series of steps that data practitioners follow when working with big data. In this video, we'll explore each step of the data science workflow.

SCRIPT:

The data science workflow is a series of steps that data practitioners follow to work with big data. It starts with identifying business problems and ends with delivering business value. An important idea to note here is that the data science workflow is cyclical in nature.

The steps in the data science workflow include:

Identifying business needs
Ingesting data
Preparing data
Analyzing data
Sharing data insights

Identifying business needs

Identifying business needs is the first step we'll discuss in the data science workflow. It's all about establishing the questions that managers or business leaders want answered. Questions like - should we make changes to our product? Which of our customers are at a greatest risk for churn and why? Or, can we save money by changing the way we're pricing our products? In this phase of the data science workflow, business leaders identify a set of questions or business goals for data practitioners to work towards.

Data ingestion

In the data ingestion phase, an organization is taking data in. This data could be real-time data which comes in streams, such as customer transactions that get added to your data store every time a customer purchases something, or the continuous data from a heart rate monitor or fitness tracker. Other times, data is ingested in batches, such as loading customer records into your data store that exist in spreadsheets somewhere else (on a local drive, perhaps). In the data ingestion phase, data is in its raw, and often messy, state.

Data cleansing / preparation

After your organization ingests raw data, it needs to be prepared for use through a data cleaning / preparation process referred to as data munging. During data munging, raw data is cleansed, aggregated, and/or augmented with the intent to serve the needs of the team members who need to use it for something like machine learning or business analytics purposes. Munging data can mean anything like cleaning up, extracting, standardizing, joining, consolidating or filtering data. The point is to prepare data enough so that the people using it to train machine learning models or generate business insights don't have to fix their input data.

Data analysis

During data analysis, your data teams are exploring this prepared data to find data insights. Oftentimes, this is where machine learning comes into play. Although machine learning isn't the only type of data analysis that can be applied to your data, it is becoming more and more popular today, especially in conjunction with big data. Aside from machine learning, individuals can also query data or use more traditional data science methods to produce insights.

Sharing insights

Finally, once business insights are discovered, they are shared with you or other stakeholders who use them to make business decisions. This could be through interactive dashboards, emails, presentations, or more.

RISE INTRO (TEXT): The data science workflow lays out the steps that data practitioners take to work with data. Depending on the stage of the workflow, a data practitioner needs specific skill sets to do their jobs effectively. In the next lesson, we'll review the individual roles on data science teams.

Lesson 3: Roles on a data science team (Text + GIFs)

RISE INTRO (TEXT): Data science teams usually include several individuals that have different skill sets to work with big data. And, while no two data teams look the same, the overall mission of a data team is to follow the steps in the data science workflow to help organizations make more informed business decisions.

Data engineers
Data engineers develop, construct, test and maintain data pipelines, which are mechanisms that allow data to move between systems or people. If we think back to the data science workflow, we talked about data ingestion and that once data is ingested, it needs to be prepared for use for machine learning and business analytics. This is where a data pipeline fits in - taking data from its raw data source, and moving it along that pipeline to where it can be used at different stages of a machine learning or data analytics project.

To perform their duties, data engineers use a set of tools to build and maintain these pipelines including programming languages like Python and Scala, different data storage solutions, and data processing engines like Apache Spark.

Data scientists
Data scientists take the data prepared by data engineers and use a variety of methods to extract insights. Data scientists usually have a strong background in disciplines like math, statistics, and computer science. They are often tasked with building machine learning models, testing those models, and keeping track of their machine learning experiments.

To perform their duties, data scientists use tools like the programming languages Python, R and SQL, machine learning libraries, and notebook interfaces like Jupyter.

Data analysts
Data analysts also take data prepared by data engineers to extract insights. Typically, a data analyst will also present data in the form of graphs, charts and dashboards to stakeholders to help them make business decisions. Data analysts can also take advantage of the work of machine learning engineers to help derive insights from data. They are typically well-versed in data visualization tools and business intelligence concepts and can be in charge of interpreting data insights and effectively communicating their findings with stakeholders.

To perform their duties, data analysts often use the SQL programming language and visualization tools like Tableau, PowerBI, Looker and others.

Platform administrators
Platform administrators can also be called devops engineers, infrastructure engineers and cloud engineers. They are responsible for managing and supporting big data infrastructure. This could include setup of big data infrastructure, updating and maintenance, performing health checks, etc. The platform administrator typically implements best practices for managing data and is interested in monitoring how team members are using the platform. This involves setting up alerts, monitoring, alerts and planning and evaluating new tools and technologies around the big data analytics platform infrastructure. Additionally, platform administrators provide governance to development teams around change, configuration and upgrades.

To perform their duties, platform administrators often use tools like the various infrastructure and monitoring services the major cloud providers offer, to help them keep data secure and scale and manage their infrastructure.

RISE INTRO (TEXT): Although not included in the video above, it's important to note that in many small organizations, data science teams can often consist of one individual trying to do all of this on their own or with little help. While it certainly takes a talented individual to be able to do all of this work, this type of set-up is not scalable. Over time, organizations would benefit from having multiple individuals on a data science team that can work together and tackle these tasks.

Now, we're at the end of our third lesson. Congratulations! By now, you should have a good understanding of what big data is, how it is managed, and how data science teams work with big data.

Click next to take a short quiz about this section.

Knowledge check

1. The image below shows three circles, each one within another. If you had to use this diagram to illustrate the relationship between artificial intelligence, machine learning and data science, how would you label each circle?
   a. Artificial intelligence → A
   b. Machine learning → B
   c. Data science → C
2. Each description below describes a stage in the data science workflow. Match each description to the role it describes.
   a. Identifying business needs
   b. Ingesting data
   c. Preparing data

    d. Analyzing data
    e. Sharing data insights
3. Each description below describes the duties of an individual on a data science team. Match each description to the role it describes.
    a. Platform administrator: responsible for managing and supporting big data infrastructure
    b. Data engineer: develops, constructs, tests and maintains data pipelines
    c. Data analyst:presents data in the form of graphs, charts and dashboards to stakeholders to help them make business decisions
    d. Data scientist: builds and tests machine learning models and keeps track of machine learning experiments

Section four - Big data for business intelligence

Lesson 1: Big data for business decision-making (Video)

RISE INTRO (TEXT): At this point, we've reviewed the big data landscape, how big data is managed, and how organizations work to extract insights from data. In this lesson, we'll answer the question "Why does all of this matter"? Let's get started.

SCRIPT:

Today, many organizations are including big data analytics as part of their strategic priorities.

But - why? How can big data help? We know that we can run analytics on big data, but what are some of the questions that we can answer? Before we talk about that, let me ask you a few questions.
Imagine that you and I work at a bank, and we accidentally transferred 5,000 into the wrong customer's account. All we know about this customer is that he is:
Male
20 years old
Single
Resides in New York

What do you think our customer would do with the $5,000?
A. Give the money back
B. Take the money and run

Take a second to think about this. Do you have an answer? Why did you come to your answer? Do you need more information?

What if now I told you that this error has happened twice in the past, to the same individual, and each time, they've given the money back?

How would knowing this change our assessment of what would happen to the money?

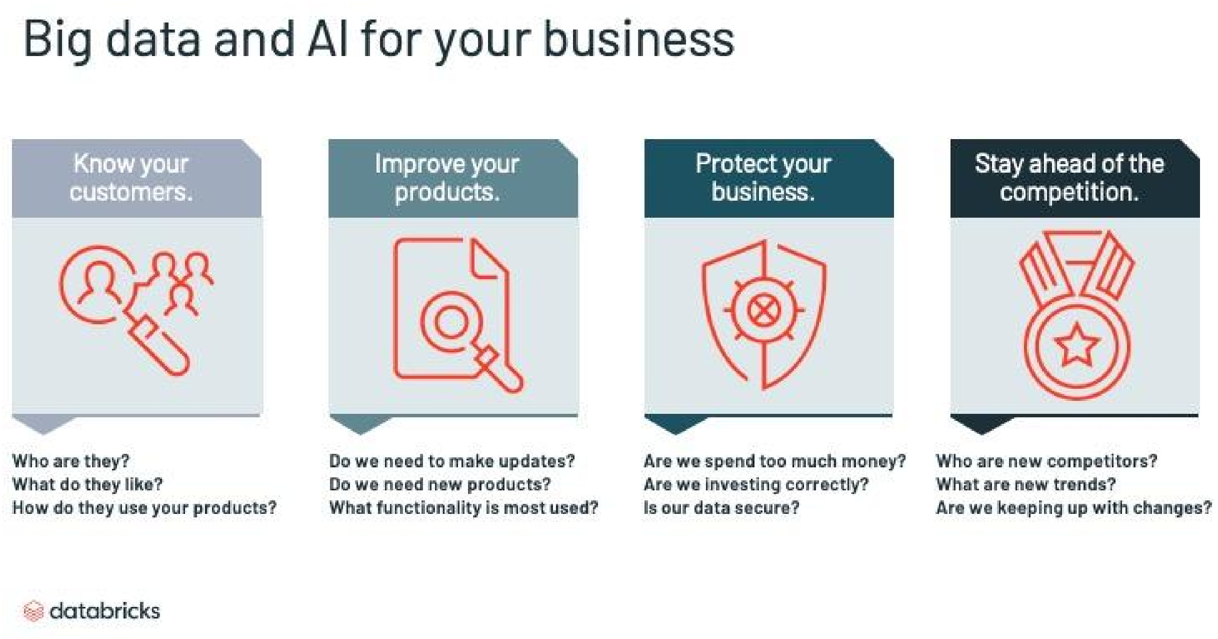
At this point, we're feeling a lot better about the money. After all, our customer has proven to be a pretty honest guy who isn't going to keep money that isn't his.

But, without this additional piece of information, there's really no way to know, or even make an educated guess about what our customer would do.

Now... think about this at a much more complex and MASSIVE scale. When we think unlocking the full potential of our big data, this is what's so powerful - imagine if we could use data to understand our customer's next moves. Imagine if we could go into a business decision having a fairly certain idea of what the outcome would be. Big data analytics makes this possible.

Think about your own line of work - what are some questions that would help you streamline business value? Maybe you're asking yourself:

- Who are my customers? What do they like? How do they use our products?
- What changes should we make to our products? Do we need to make changes? What do people like the most about our products?
- Maybe you're most concerned about protecting your business. Are you spending too much money? Are you investing money correctly?
- Or how do you stay ahead of the competition? Who are new competitors? What are new trends? Are we keeping up with changes?



Big data and AI for your business

Know your customers.

Who are they?
What do they like?
How do they use your products?

Improve your products.

Do we need to make updates?
Do we need new products?
What functionality is most used?

Protect your business.

Are we spend too much money?
Are we investing correctly?
Is our data secure?

Stay ahead of the competition.

Who are new competitors?
What are new trends?
Are we keeping up with changes?

databricks

Regardless of your industry, big data and analytics can completely change your organization. It enables you to find answers to questions you want to know and also allows you to find patterns to answer questions you didn't even know to ask.

OUTRO: In this lesson, we explored, at a high-level, why big data matters to businesses. In the next and final lesson, we'll take a closer look at how different industries are using big data to help streamline their businesses.

Lesson 2: Big data use-cases in different industries (Text)

RISE INTRO (TEXT): Thousands of organizations around the world are applying advanced analytics to big data to enrich and accelerate business outcomes. In this section, we'll review some of these examples, by industry.

Please feel free to review the ones of most interest to you.

Health and life sciences

- Who: Large integrated healthcare systems, major pharmaceutical companies, diagnostic labs
- Goal: Apply advanced analytics to their large volumes of clinical and research data to accelerate R&D and improve patient outcomes
- Examples:
  - Precision medicine: Analyzing clinical and genomic datasets in order to prescribe

    - targeted treatments specific to an individual's biology.
    - Disease prediction: Using real world evidence and public datasets to identify biomarkers that have high probability of driving the onset of disease.
    - Claims analysis: Applying machine learning to large volumes of claims to determine preventative measures to improve patient health and identify fraud patterns.

Financial services

- Who: Retail and commercial banks, hedge funds, fintech innovators and more
- Goal: Apply advanced analytics to large volumes of customer and transaction data to reduce risk, boost returns and improve customer satisfaction
- Examples:
    - Investment decisions: Maximizing returns with AI-powered insights based on billions of market signals and alternative data sources
    - Personalized banking: Delivering the right financial products and guidance to customers with real-time customer insights and predictive analytics
    - Fraud prevention: Detecting and preventing fraudulent activities (e.g. money laundering, credit card fraud) by leveraging machine learning to predict anomalies in real-time.

Media and entertainment

- Who: Major publishers, streamers, gaming companies and more
- Goal: Apply advanced analytics to large volumes of audience and content data to deepen audience engagement, reduce churn and optimize advertising revenues
- Examples:
    - Content personalization: Driving 1:1 experiences to drive engagement and customer satisfaction
    - Sentiment analytics: Understanding how content is resonating in social channels and using data to find the next most popular article, show or game
    - Churn management: Determining which customers are likely to churn to drive personalization and prevent them from churning

Advertising and marketing

- Who: Global agencies, small ad tech companies and more
- Goal: Apply advanced analytics to large volumes of consumer, clickstream and ad data to improve return on ad spend, inventory management and audience segmentation efforts
- Examples:
    - Return on ad spend: Understanding where marketing dollars are going and what are the most effective campaign and channel strategies.
    - Customer journeys: Finding ways to improve customer journeys and maximize revenue outcomes using customer-use data
    - Recommendations: Leveraging real-time interactions to optimize marketing, sales and service channels with recommended products and services.

Telecom

- Who: Global communication service providers, network and equipment providers and more
- Goal: Apply advanced analytics to large volumes of customer and network data to improve network services and performance while reducing customer churn
- Examples:
    - Network performance: Understanding network chokepoints and automate load balancing in real time
    - Upselling services: Maximizing customer revenue by using customer usage data to drive cross and upsell services and products
    - Fraud prevention: Analyzing SIM cards and other data sources to minimize fraudulent transactions

Retail

- Who: Traditional brick and mortar companies, ecommerce companies
- Goal: Apply advanced analytics to large volumes of customer, product and supply chain data to better attract customers, increase basket size and reduce costs
- Examples:
    - Targeted recommendations: Using machine learning to mine clickstream, purchase and customer data to provide personalized recommendations
    - Demand forecasting: Predicting real-time demand and returns at a granular level using new and non-traditional data sources to optimize inventory
    - Optimized pricing: Improving campaign conversion and return-on-ad-spend by using big data to serve the right ad, at the right time, to the right person

Oil, gas and energy

- Who: Oil upstream and downstream organizations utility companies and more
- Goal: Apply advanced analytics to large volumes of sensor, supply chain, and customer data to improve exploration, reduce machinery downtime and optimize sales and supply chain operations
- Examples:
    - Smart grids: Analyzing e-sensor data from IoT devices to detect energy consumption patterns, predict future usage and optimize production, storage and distribution.
    - Predictive maintenance: Avoiding production failures by analyzing real-time machine data, maintenance schedules and other historical data to predict equipment maintenance.
    - Improved well production: Analyzing geospatial data to determine optimal well placement and real-time insights to improve drill and well efficiency

Automotive

- Who: Automotive and automotive parts manufacturers

- Goal: Apply advanced analytics to their large volumes of driver, vehicle, supply chain and IoT data to improve manufacturing efficiency and create better and safer autonomous driver experiences
- Examples:
    - Predictive maintenance: Improving vehicle safety to help drivers protect their investments with predictive maintenance recommendations
    - Supply chain optimization: Using big data analytics and ML to identify demand levels at a more granular level and predict the right amount of inventory to produce and distribute to the right customers
    - Autonomous applications: Processing large volumes of driver video and sensor data to build AI applications for self-driving car features

RISE OUTRO (TEXT): Harnessing the power of big data through data science workflows is relevant to every organization, regardless of size, geographic location, or overall mission. In order to successfully take advantage of big data, many organizations are adopting a unified data analytics approach to big data. If you'd like to learn more about the details behind a unified data analytics approach to big data, please visit the Databricks Academy and look for our Introduction to Unified Data Analytics course.

Summary and next steps

Congratulations!

You completed the Introduction to Big Data course.

By now, you should be able to:

- Explain foundational concepts used to define big data.
- Describe how organizations manage big data.
- Summarize how individuals on data science teams work with big data on a daily basis
- Describe real-world use cases for big data in a variety of industries

Next steps:

- Evaluate this course by taking a brief, three question survey. Your feedback is valuable and helps us improve our courses.
- Continue your learning journey by visiting the Databricks Academy.