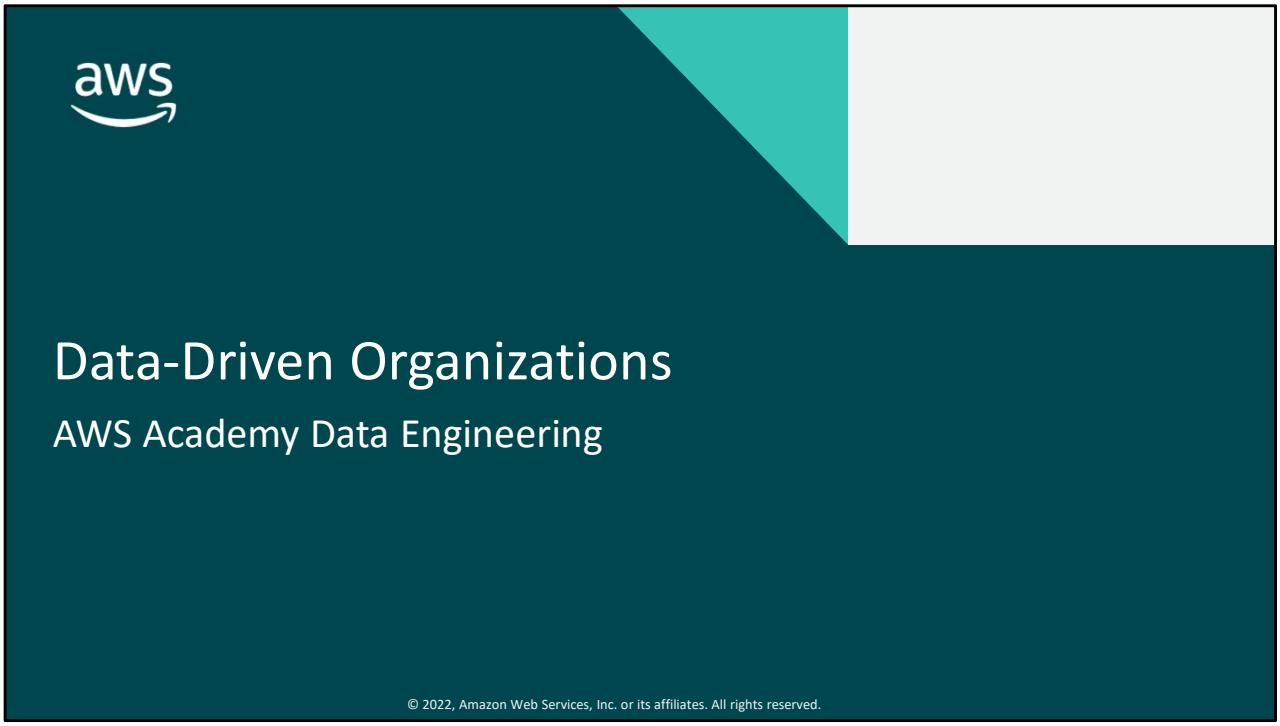
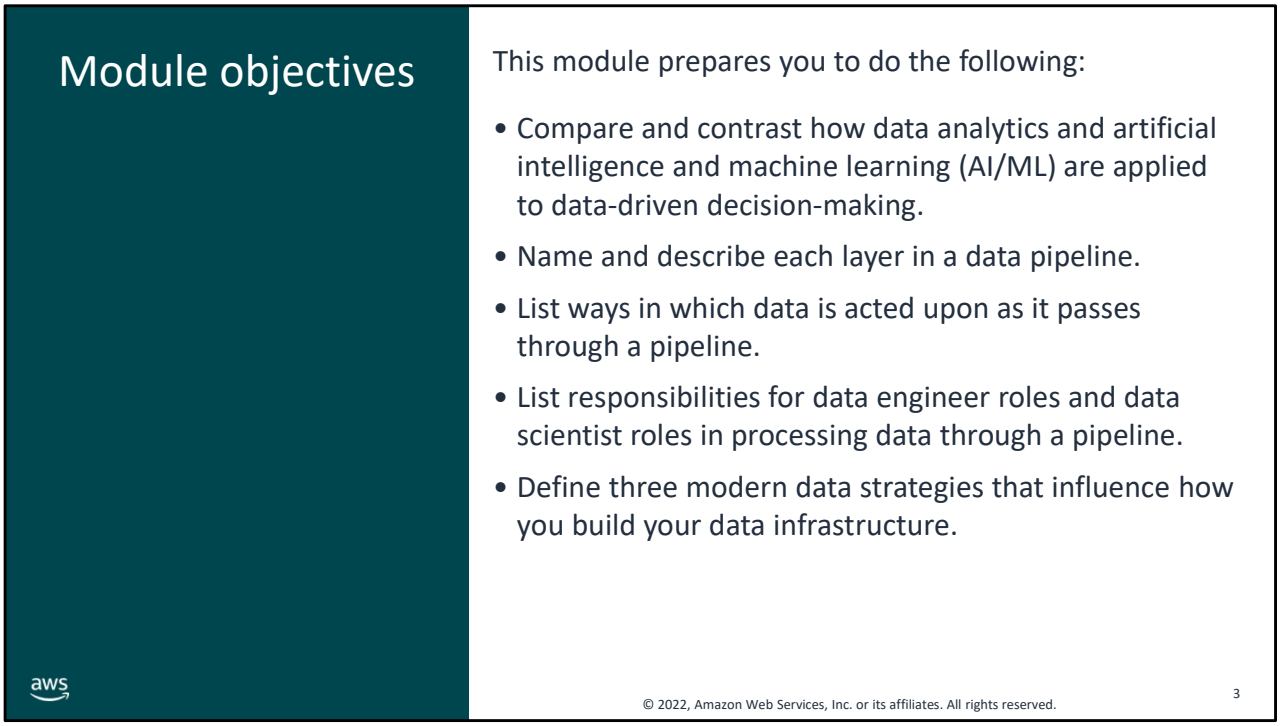


|Slide number 1
|Instructor notes
|
|Student notes
Welcome to the Data-Driven Organizations module. This module introduces how modern organizations use data and data science to make decisions.

|Slide number 2
|Instructor notes
|
|Student notes
This introduction section describes the content of this module.



|Slide number 3
|Instructor notes
|This module is intended to introduce concepts that will be revisited in following modules and iterated on in greater depth. Key points to reinforce are the need to design each pipeline to match the business problem being addressed, the iterative nature of the pipeline, and the fact that the approach often starts with "let's try something and see what we get."
|
|Student notes
The module introduces a basic vocabulary for talking about data-driven decision-making. You will learn to distinguish data analytics from artificial intelligence and machine learning (AI/ML) applications and describe how they all support data-driven decisions. You will be able to identify the layers of a data pipeline and the actions that are taken on data as it passes through the pipeline. You will be able to list the types of work that a data engineer or data scientist does when building out a data pipeline infrastructure. And finally, you will be able to talk about three strategies for building modern data infrastructures in a data-driven organization.

Module overview

Presentation sections

- Data-driven decisions
- The data pipeline – infrastructure for data-driven decisions
- The role of the data engineer in data-driven organizations
- Modern data strategies

Lab

- Accessing and Analyzing Data by Using Amazon S3

Knowledge checks

- Online knowledge check
- Sample exam question

|Slide number 4
|Instructor notes
|Each module has an introduction, content sections, and a wrap-up. The wrap-up for this module contains a sample exam question for you to review with the students.
|
|Student notes
The objectives of this module are presented across multiple sections.

You will also complete a hands-on lab that uses Amazon Simple Storage Service (Amazon S3) Select to query data that is stored in an S3 bucket.

The module wraps up with a sample exam question and an online knowledge check that covers the material presented.

|Slide number 5
|Instructor notes
|
|Student notes
This section introduces modern approaches to making data-driven decisions.

|Slide number 6
|Instructor notes
|
|Student notes
As noted in this quote, organizations are accelerating their investments in data and analytics services by a large percentage to become more data driven and digital. What does it look like to be data driven?

How do you decide…

- Which restaurant near me serves the best Thai food?
- Where can I find a used child's bicycle?
- Which starters should I pick for my fantasy football team?
- Which type of job should I pursue?
- What should I pay for this home?

|Slide number 7
|Instructor notes
|You might frame this as a discussion or ice breaker activity. Ask students about decisions that they make by using or not using apps. If you have a diversity of age ranges or backgrounds, you might also tap into how people made these decisions before the age of Google and Web 2.0.
|
|Student notes
When it comes to making decisions in your life, how do you decide where to eat or where to live? You probably use a mobile or web app that gives you personalized recommendations.

Fifteen to twenty years ago, you would have relied on more subjective research, such as asking a colleague or comparing information from different websites. But now, people regularly rely on apps to support large and small decisions. Those apps use data science to analyze available datasets (such as geolocation and local business information) and data that they have collected from users to find insights and suggest possible decisions to you.

How do organizations decide…

- Which of these customer transactions should be flagged as fraud?
- Which webpage design leads to the most completed sales?
- Which patients are most likely to have a relapse?
- Which type of online activity represents a security issue?
- When is the optimum time to harvest this year's crop?

|Slide number 8
|Instructor notes
|You could use this slide as more of a discussion if all or some of the audience is inexperienced in data science concepts. You might ask students for additional examples of questions that organizations might have that could be addressed with data. If there are some data scientists or data analysts in the course you might ask them to provide examples from their own work.
|
|Student notes
Like you, organizations have become more data driven over the last 20 years. Organizations have been evolving their use of data from reporting on their operational past to predicting their best options for future success. Not surprisingly, organizations have first focused their investments on critical business decisions. For example, credit card companies can predict which credit card charges are likely fraudulent, or an ecommerce site can personalize the user experience to increase sales.

One thing fueling the data-driven evolution is the explosion of richer and more abundant data from websites, mobile apps, and smart device technologies. All this

new data opens up additional areas for data-driven decision-making. For example, Internet of Things (IoT) devices can now provide continual data to help a farmer decide when to plant, water, or harvest a crop.

Fueling decisions with data science

Data analytics	AI/ML
• Is the systematic analysis of large datasets (big data) to find patterns and trends to produce actionable insights • Uses programming logic to answer questions from data • Is good for structured data with a limited number of variables	• Is a set of mathematical models that are used to make predictions from data at a scale that is difficult or impossible for humans • Uses examples from large amounts of data to learn about the data and answer questions • Is good for unstructured data and where the variables are complex

aws

|Slide number 9
|Instructor notes
|This is the initial introduction of these topics to start setting context and differentiating concepts that will be retuned to. The main point of distinction that will be carried through the course is that analytics uses rule-based (if/then) logic for analysis, whereas ML learns by comparing many examples.
|
|Student notes
The data science behind data-driven decisions falls into two main categories: data analytics, and artificial intelligence (AI) or its subfield, machine learning (ML). The primary distinction between the two is how data scientists use them to arrive at their results.

Data analytics is a broad term that describes the systematic analysis of large datasets (also known as big data) to derive insights. For example, you might use data analytics to answer the question, "Which factors appear to be the most important to the selection of a restaurant?" The data scientist would hypothesize about the data variables that will help to answer this question and then use statistical methods and programming techniques to derive results. Data analytics might also include

predictions, but with traditional methods, data scientists create rule-based logic to make predictions. Analysts might use a variety of business intelligence (BI) tools to analyze collected data. This might include simple tools, such as spreadsheets, or more complex data and visualization tools.

AI and ML take the process of prediction to a level that hadn't previously been possible. With ML models, a data scientist can automate an application's ability to actually learn and improve the quality of its predictions. For example, an ML model could use data about what diners actually choose to do after receiving a recommendation to refine future recommendations. The ML model learns by getting more and more examples. ML models might also be able to find relationships among the data that would be much more difficult for a human observer to hypothesize. The more high quality data that you can provide, the better the predictions that ML can produce. Better predictions provide a more accurate basis for decision-making.

Example: Identify pictures of dogs

Data analytics approach	AI/ML approach
Based on a set of defined features, which of these pictures are dogs?	Which of these images have the same features as the pictures that I have seen that were labeled as dogs?




|Slide number 10
|Instructor notes
|Depending on the audience and timing, you could extend this conversation by asking students to list the traits and the if/then logic that they would use to capture the first three pictures as dogs and the last one as a cat. You might also discuss the miscellaneous and confusing information in the images that makes the task challenging (for example, two dogs at a bit of a distance, one looking up into the camera as a puppy, and one buried among stuffed animals).
|
|Student notes
Here's a simplified example to distinguish a data analytics approach from an AI/ML approach. This example demonstrates the type of complexity where an ML solution might produce better results than data analytics.

If you had a large dataset of images and you wanted to create a website that highlighted only dog images, how could you apply data science to automatically identify which pictures are dogs?

Using traditional data analytics, you would need to create logic to identify the

features that make a dog a dog. Take a minute to think of all the different characteristics that two dogs might have while still being recognized by humans as dogs. What logic would you use to avoid the cat fitting into your model?

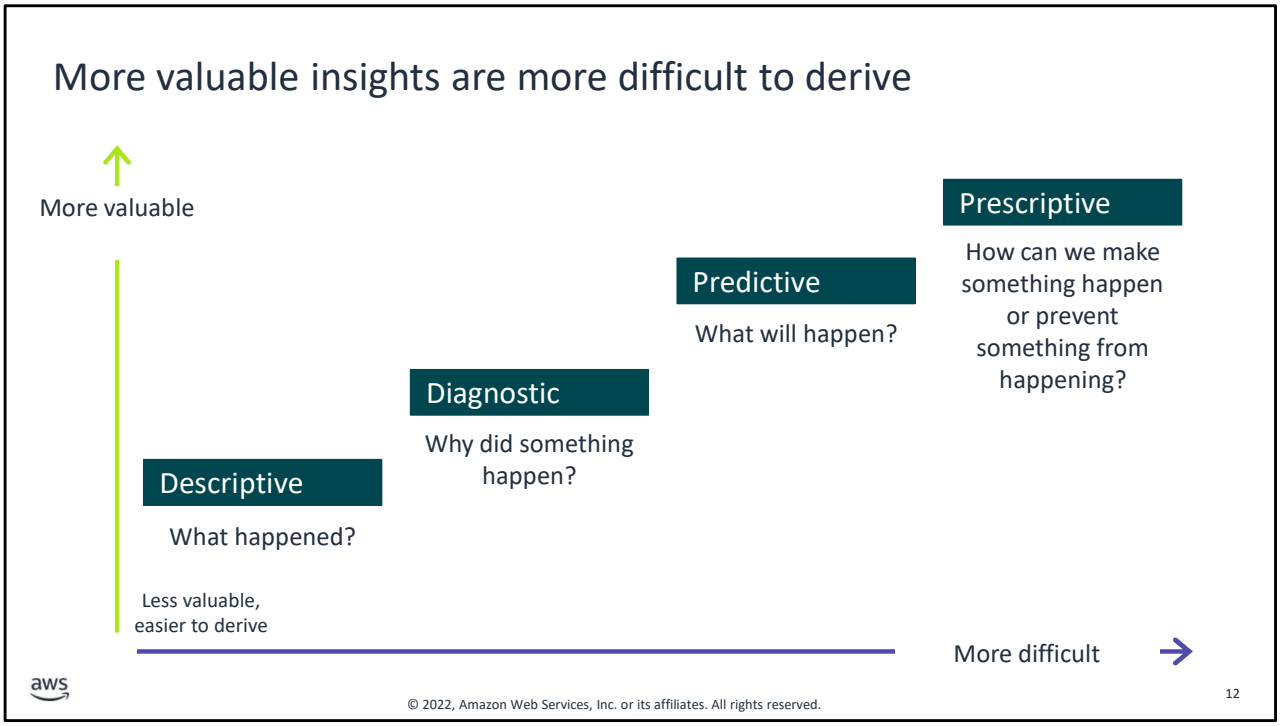
With ML, you would train your model by showing it a lot of pictures that are labeled as dogs and letting it *learn* to identify new, unlabeled pictures as dogs.

Given the vast number of dog images that are currently available for analysis, this might be a good candidate for ML.

Business example: Customer relationship management

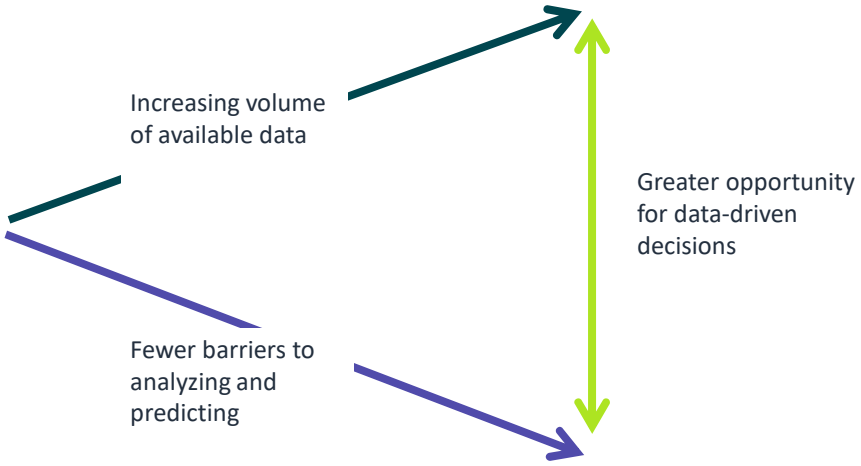
Data analytics approach	AI/ML approach
• A retail business analyzes total revenue per customer and segments customers into categories based on spending. • The segmentation might be used to give a higher level of customer service to customers who spend more.	• A retail business uses AI/ML to analyze customer churn (why and how often customers come and go). • AI/ML might uncover factors that influence churn so that the business can make changes to better retain customers.

|Slide number 11
|Instructor notes
|
|Student notes
For many business scenarios, your organization might use both data analytics and an AI/ML approach as part of a unified decision-making strategy. The slide highlights a simple example.



|Slide number 12
|Instructor notes
|
|Student notes
You can categorize analytics into these four increasingly valuable—but increasingly difficult to derive,—types: descriptive, diagnostic, predictive, and prescriptive. The volume of data needed, amount of compute and storage required, and complexity of the data science all increase as you move from descriptive to predictive and prescriptive decision support.

More data + fewer barriers = more data-driven decisions



|Slide number 13
|Instructor notes
|
|Student notes

The explosion of richer data sources has coincided with the availability of much more powerful compute at lower costs than what had been available. In addition, pay-as-you-go cloud-based infrastructure increases resource flexibility. Evolving cloud service offerings have also lowered the technical complexity required to derive higher-value insights. This lowers the overall cost and complexity of using big data analytics and AI/ML in your decision-making. The easier it is to make predictions, the more an organization can extend data-driven decisions into additional scenarios and drive incremental improvements in the quality of their analyses and predictions. The more data that is available, the greater the opportunity is to uncover new insights or increase the accuracy of predictions.

More data science in daily life

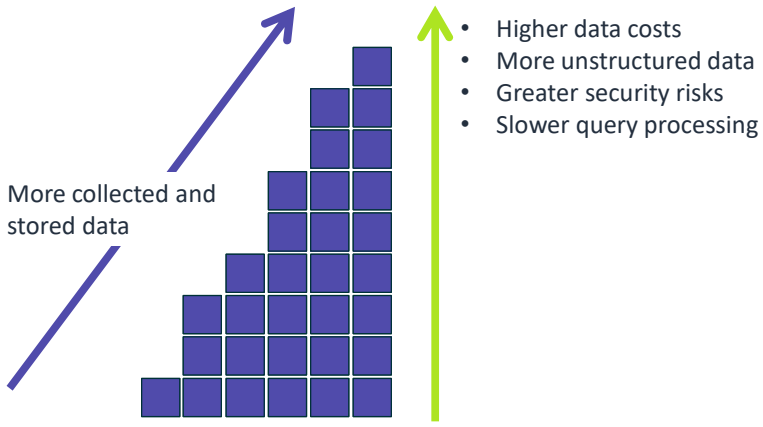
- You are shopping for shoes online and start to see a lot of shoe-related advertisements.
- You watch a movie or listen to a song on a streaming platform and begin to get recommendations for movies or music you might also like.
- You order pizza online and are kept up to date about each step in the preparation and delivery process.
- You use your credit card outside of your usual geographic area and get additional fraud alerts from your bank.
- Your navigation app on your phone alerts you to a traffic jam.

|Slide number 14
|Instructor notes
|If time permits, this is a good opportunity to engage students in thinking about their online experiences in light of AI/ML and how it used. Depending on your audience, you could engage in a discussion of how students perceive these types of applications. What makes them feel helpful rather than intrusive? What is the benefit to the business in providing these benefits to the user?
|
|Student notes
The increasing use of data and the application of AI/ML approaches is recognizable when you interact with online systems. The slide lists a few examples, and you can probably think of other ways in which your online experience is personalized. Data about you is used to predict what you might like or to prevent something bad from happening.

The navigation system on your phone is a great example of large amounts of how geospatial data are brought together to personalize your navigation experience. Geospatial data is data about objects, events, or phenomena that have a location on the surface of the earth. The location may be static in the short-term (e.g., the

location of a road, an earthquake event, children living in poverty), or dynamic (e.g., a moving vehicle or pedestrian, the spread of an infectious disease). Geospatial data combines location information (usually coordinates on the earth), attribute information (the characteristics of the object, event, or phenomena concerned), and often also temporal information (the time or life span at which the location and attributes exist). (source: https://www.sciencedirect.com/topics/computer-science/geospatial-data).

More data doesn't necessarily equal more value



|Slide number 15
|Instructor notes
|
|Student notes

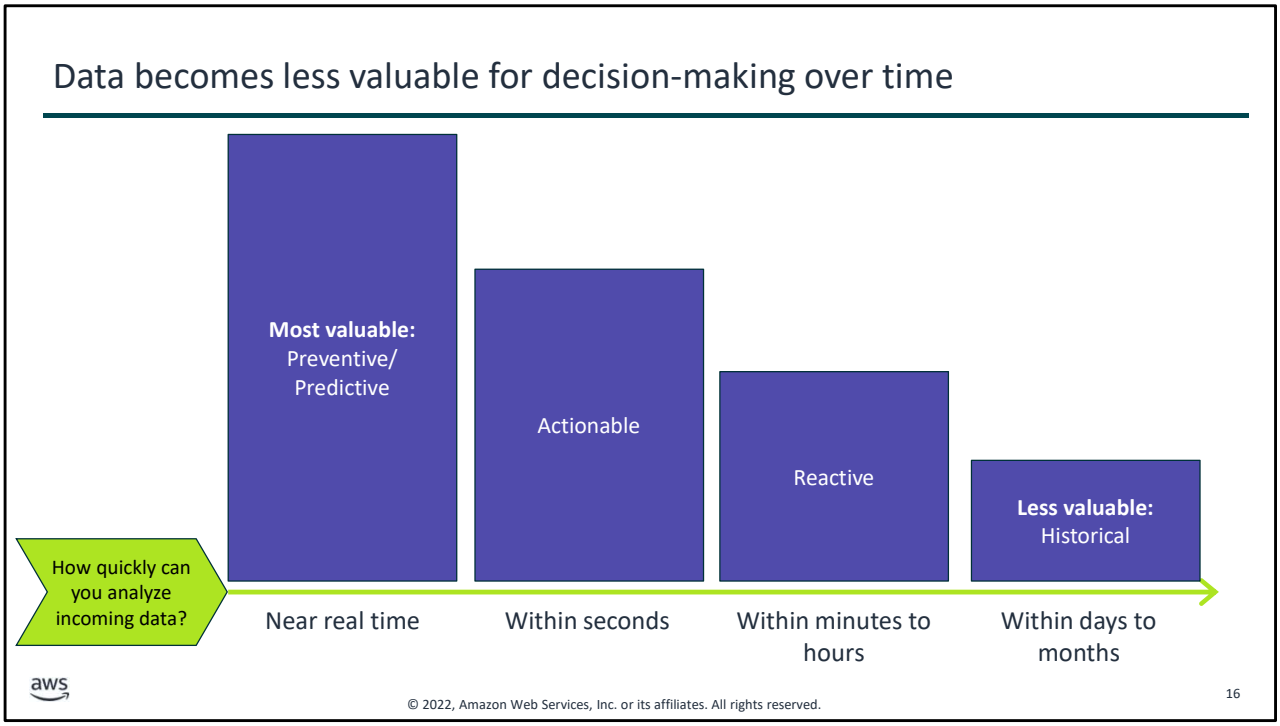
All that data presents challenges as well as opportunities. Organizations must decide how to store the data and account for both traditional structured data (rows and columns) as well as unstructured data, such as images, documents, handwritten notes, and geospatial information.

If an organization isn't thoughtful about its approach, it might spend a lot of money to pay for the equivalent of a pile of unmarked boxes in the back of a garage.

Organizations also need to manage the security of all that data they're collecting and storing. This is especially critical if they are collecting personal data.

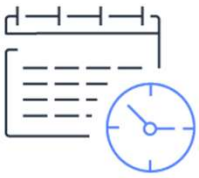
Additionally, an organization must think about scaling their solutions to handle the volume and type of data to be processed. Data that hasn't been properly prepared, or isn't held in an appropriate type of storage, won't yield the best results.

The trade-off of collecting more data in hopes of fueling future decisions or providing greater accuracy has to be weighed against the cost and speed impacts of managing and processing that data.



|Slide number 16
|Instructor notes
|
|Student notes
Another consideration to decide how much and what type of data to store is how quickly you can act upon the data. As the slide highlights, the business can derive the most value from data when it is fresh and can be used immediately (for example, to prevent a credit card transaction) or within a few seconds (for example, to recommend a secondary purchase before you checkout). As data ages, it loses its ability to inform proactive decisions and becomes an asset to take corrective action after the fact. Using data for reactive and historical analysis might still provide value to the organization, but when deciding what to keep and how long to keep it, an organization should think about whether they have the capacity to process the data within the timeframe that maximizes its value.

The trade-offs of data-driven decisions



Cost	Speed	Accuracy
• How much should you invest to go faster or predict more accurately? • How much incremental improvement justifies additional cost?	• How quickly do you need an answer? • Can you sacrifice accuracy for speed?	• How accurate does the prediction need to be? • Does waiting for a better answer outweigh answering more quickly?

aws

|Slide number 17
|Instructor notes
|
|Student notes

The trade-offs you need to consider to decide how much data to collect are reflected across your infrastructure choices. Many options are available to build your infrastructure, so it's important to choose the components that provide the best value for the decisions that those components must support. Balancing cost, speed, and accuracy is at the core of building the data-driven infrastructure that an organization needs to support its data-driven decisions.

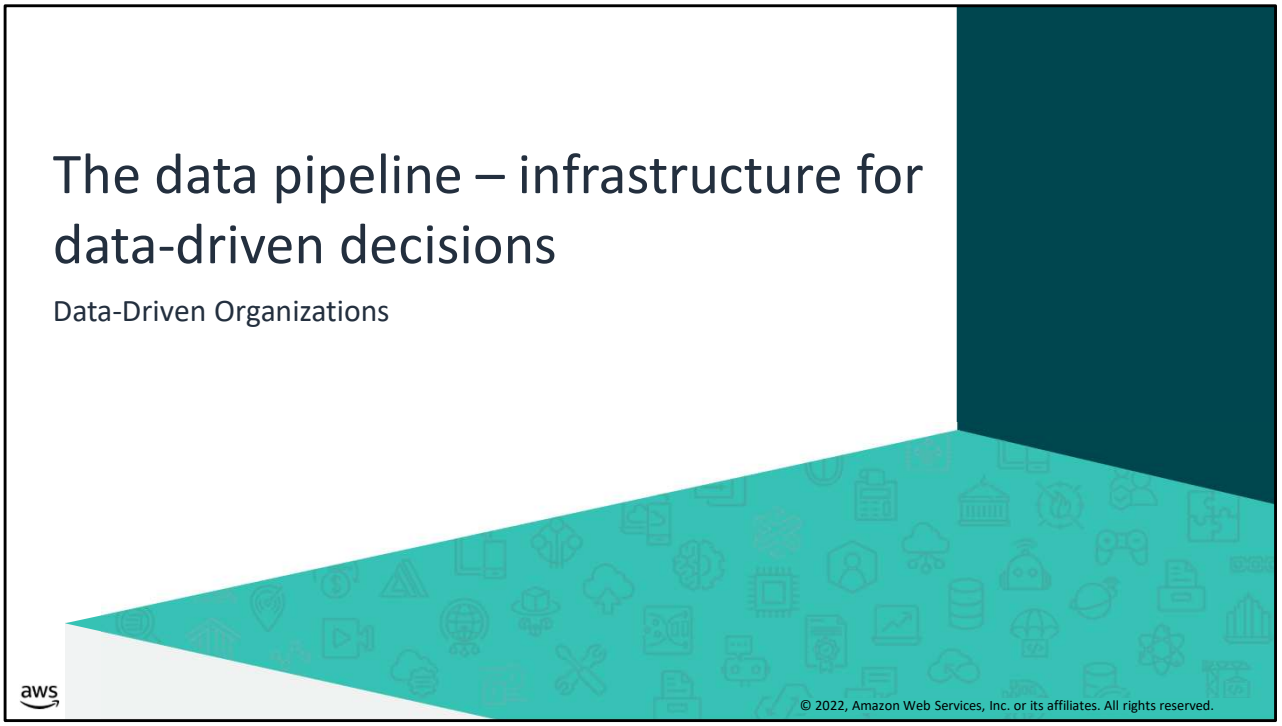
Key takeaways: Data-driven decisions

- Data-driven organizations use data science to make informed decisions.
- Data analytics relies on programming logic and tools to arrive at predictions from data. This approach is good for structured data with a limited number of variables.
- AI/ML can learn to make predictions from examples in data. This approach is good for unstructured data and where the variables are complex.
- A large increase of available data, coupled with a decrease in the cost of relevant technology, has increased the opportunity for data-driven decisions.

|Slide number 18
|Instructor notes
|
|Student notes

Here are a few key points to summarize this section:

- Data-driven organizations use data science to inform decisions. This includes data analytics and AI, along with its subfield of ML.
- The main distinction between data analytics and AI/ML is the way that the analysis is performed. Data analytics relies on programming logic while AI/ML applications can learn from examples in data to make predictions. This makes AI/ML good for unstructured data where the variables are complex.
- The rapidly increasing availability of data, coupled with decreasing costs of supporting technology, increases the opportunity for organizations to make data-driven decisions.



|Slide number 19
|Instructor notes
|
|Student notes
A data pipeline provides the infrastructure to turn data into insights and make data-driven decisions. This section introduces the basic layers of a data pipeline.

A data pipeline in its simplest terms

Collect data

Store and process data

Build something useful with data

|Slide number 20

|Instructor notes

|Modules 3 and 4 will iterate further on this topic. Those modules will apply more technical details about the types of data and data processing, and match that information to the AWS services and reference architectures that will be used. The goal of this section is to introduce the high-level vocabulary and provide a simplified analogy by using plastic building blocks. It might be helpful to return to this analogy as the course becomes more complex. Modules 6–11 will go deeper into each individual layer and the services and considerations that impact design of that layer.
|

|Student notes

The data pipeline is the infrastructure that you build to support data-driven decision-making. At the most basic level, any data pipeline infrastructure must be able to bring data in, store it, and provide the means to work with the data to derive insights. As the data engineer or data scientist, it's your job to build the pipeline to figure out what's appropriate and how you're going to do it. Often this will include exploring and experimenting to get to know the data and the best way to derive value from it.

Work backwards to design your infrastructure

2. What data do you need to support this?

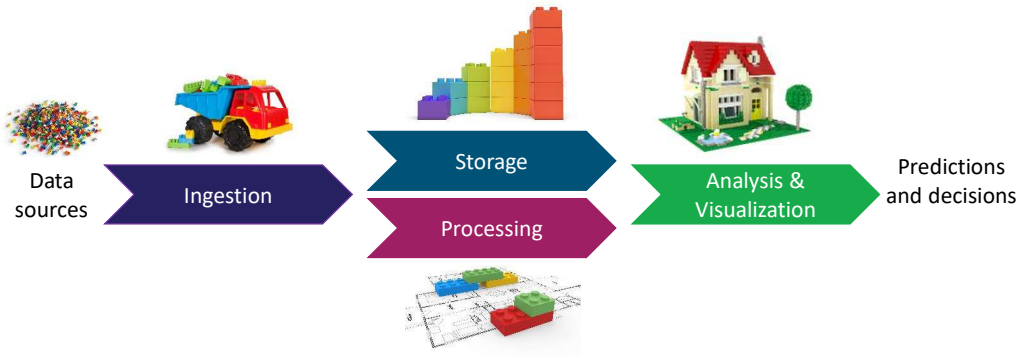
1. What decision are you trying to make?

Weigh the trade-offs of cost, speed, and accuracy.

|Slide number 21
|Instructor notes
|
|Student notes

The number of variations and options to create a data pipeline are broad. The key to designing an effective decision-making infrastructure is to start with the business problem to be solved or decision to be made. Then, build the pipeline that best suits that use case. The cloud makes it easier to start up different infrastructure to serve different use cases. An organization doesn't need to invest in an infrastructure that tries to serve all business needs with a single database, server, or technology stack. For each use case, start with the end in mind, and build the data pipeline that supports it.

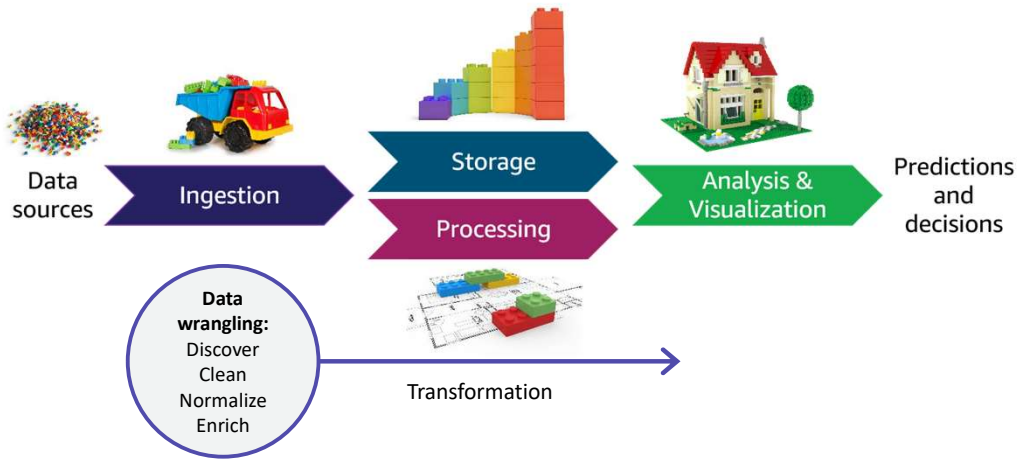
Layers of the pipeline infrastructure



|Slide number 22
|Instructor notes
|
|Student notes

The infrastructure layers that you need to build include methods to ingest and store data from the data sources that you have identified. You need to make the stored data accessible for decision-making processes. You must also create the infrastructure to process, analyze, and visualize data by using tools that are appropriate to the use case. To build an appropriate infrastructure, you will need to understand the nature of the data and the intended type of processing and analysis to be performed.

Actions taken with data in the pipeline



|Slide number 23

|Instructor notes

|This slide includes an animation that runs automatically when you present the slide. It moves the data wrangling circle across the pipeline illustrating that data wrangling happens in multiple stages across the pipeline.

|

|Student notes

Looking at the pipeline from the perspective of the data that travels through it, the pipeline must support the ability to discover details about the data and how it fits (or doesn't fit) your intended outcomes. You might start with a hypothesis and use BI tools or other mechanisms to discover more information about the data, and then experiment and see where it takes you.

Data might need to be cleaned or normalized. For example, you might merge multiple data sources that share some similar data and you need to rectify inconsistencies, or you might have a data source that hasn't been well-maintained or validated at its source. You need to prepare the data before it's viable for analysis.

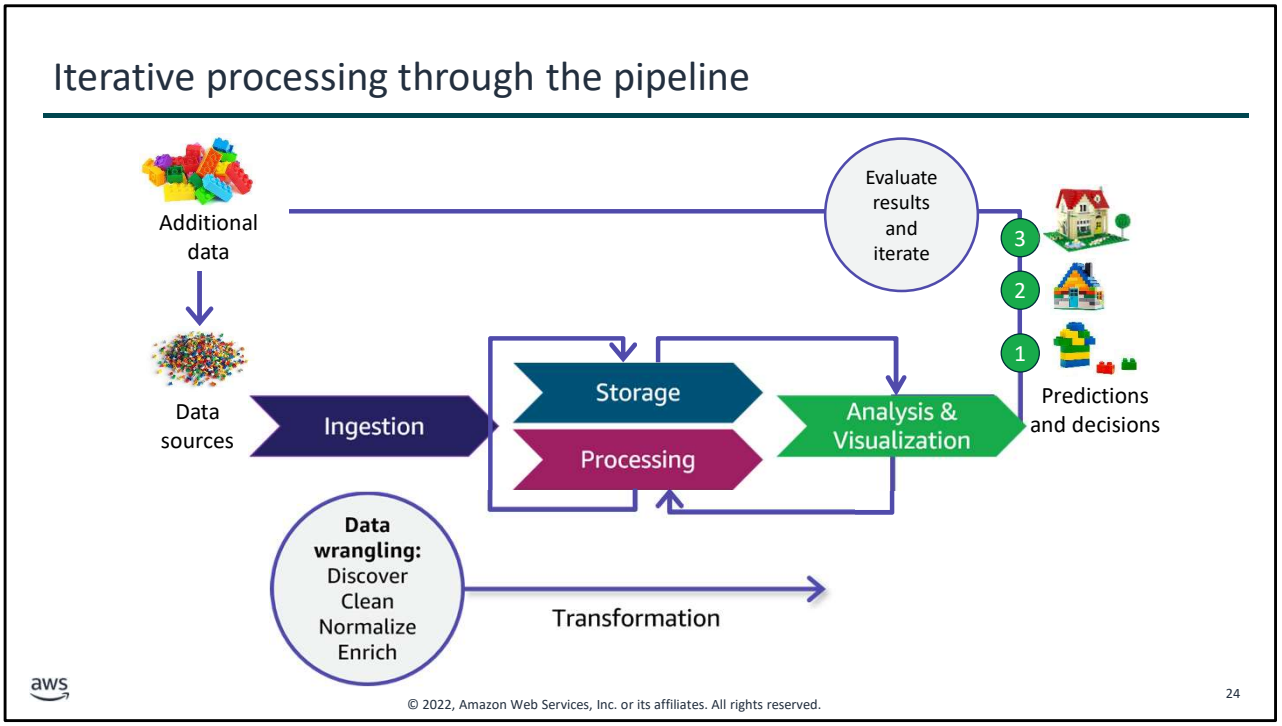
Data will almost always be transformed as it moves through the pipeline. This might

include modifying the format to support a specific analysis tool or replacing values (for example, zeroes in place of nulls). You might also augment the dataset by filling in gaps or enriching it with additional information. For example, you might calculate and save additional data values to be used during analysis. You might also add metadata to categorize and catalog data. *Data wrangling* is a term that is used to describe the ways in which data is manipulated and transformed from its raw state into more meaningful states that downstream processes or users can use.

Within each layer, or as part of transitioning between pipeline layers, you might perform these tasks on data to prepare it for the desired analysis or predictions. Some task types can happen in multiple places across the pipeline, and some tasks can be performed together as part of a single process. For example, in traditional data science architectures, the extract, transform, and load (ETL) process takes data from one source, performs some type of transformation on it, and then loads it into another location where the analysis is done.

As you learn about different scenarios, you will see these tasks performed at different points in the pipeline, using different services and methods dependent on the uses case.

The tools that you use and the way you perform these tasks will differ based on characteristics of your data, the business problem that you're addressing, and the methods or models that you're applying.

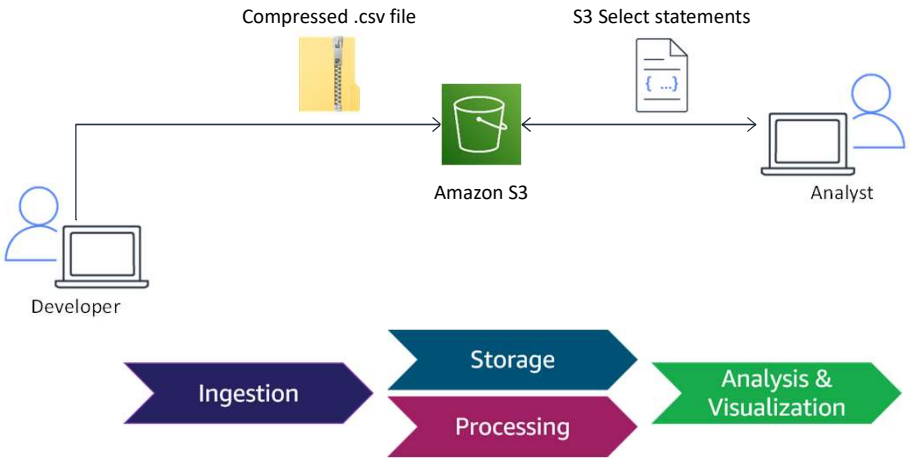


|Slide number 24
|Instructor notes
|
|Student notes

Another key characteristic of deriving insights by using your data pipeline is that the process will almost always be iterative. You have a hypothesis about what you expect to find in the data, and you need to experiment and see where it takes you. You might develop your hypothesis by using BI tools to do initial discovery and analysis of data that has already been collected. You might iterate within a pipeline segment, or you might iterate across the entire pipeline. For example, in this illustration, the initial iteration (number 1) yielded a result that wasn't as defined as was desired. Therefore, the data scientist refined the model and reprocessed the data to get a better result (number 2). After reviewing those results, they determined that additional data could improve the detail available in their result, so an additional data source was tapped and ingested through the pipeline to produce the desired result (number 3).

A pipeline often has iterations of storage and processing. For example, after the external data is ingested into pipeline storage, iterative processing transforms the data into different levels of refinement for different needs.

Example: A simple pipeline



|Slide number 25
|Instructor notes
|
|Student notes
This example is similar to the pipeline that you will work with in the lab titled Accessing and Analyzing Data by Using Amazon S3.

The dataset to be ingested into the pipeline is a .csv file that contains structured data. A developer uses a utility to compress the file, then uploads the file to an S3 bucket for storage. The data science team who will analyze the data can use S3 Select statements to directly query the data in Amazon S3.

Key takeaways: The data pipeline – infrastructure for data-driven decisions

- A data pipeline provides the infrastructure for data-driven decision-making.
- When designing a pipeline, start with the business problem to be solved and work backwards to the data.
- The pipeline includes layers to ingest, store, process, and analyze and visualize data passing through it.
- Data wrangling refers to how data is acted upon as it passes through the pipeline. Tasks include discovery, cleaning, normalization, transformation, and augmentation.
- Data is processed iteratively to evaluate and improve upon results.

aws

|Slide number 26
|Instructor notes
|
|Student notes

Here are a few key points to summarize this section:

- The data pipeline provides the infrastructure for data-driven decisions and includes layers to ingest, store, process, and analyze and visualize data.
- Data wrangling refers to the work to transform data to prepare it for analysis. Processing is typically iterative. The business problem should drive the pipeline design.

|Slide number 27
|Instructor notes
|
|Student notes
This section introduces the responsibilities of the data engineer and data scientist roles to plan and manage a data infrastructure.

Common data pipeline questions

Data engineer

- Does the organization have the data that addresses the need? Where is the data stored and in what format?
- Will I need to combine data from multiple sources?
- What is the type and quality of data? What will be the source of truth?
- What are the security requirements for this data? Who needs access to it and in what state?
- What types of mechanisms are needed to transfer the data from its locations into the pipeline?
- How much data is there, and how frequently is it updated or processed?
- How important is speed when data is requested?

- What can the data tell me?
- How will I evaluate the results?
- What kind of visualization do I need?
- Which formats and tools are the analysts familiar with?
- Do I need a big data framework?
- Which type of AI/ML models fit?
- What is the simplest way to implement AI/ML?

Data scientist

aws

|Slide number 28

|Instructor notes

|This slide is a baseline setup for activities that the student will perform throughout the course. You might withhold the questions and instead treat this as a precursor to the data module that follows. The slide is animated to support withholding the questions and treating this as a discussion.

|Ask students to think of the type of questions that they believe will need to be answered by the data engineer based on what they have learned so far.

|

|Student notes

To build the correct pipeline, the person responsible for building the pipeline needs to ask a lot of questions about the nature of the data and the intent for its processing. You can classify the questions broadly into data engineering questions, which are mostly focused on getting the data into the pipeline, and data science questions, which are mostly about what you are trying to get out of the data.

In the remainder of this course, you will learn how the answers to these types of data engineering questions lead you to select components and build pipelines that provide the appropriate infrastructure for data-driven decisions.

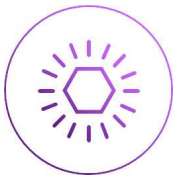
Key takeaways: The role of the data engineer in data-driven organizations

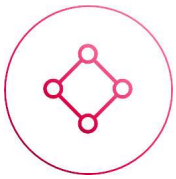
- Both data scientist roles and data engineering roles work with the data pipeline, and tasks might be performed by someone in either role.
- Data engineering is primarily focused on the infrastructure that the data passes through while the data scientist works with the data in the pipeline.
- To build the right pipeline, you need to ask questions about the desired outcomes and the data and then iterate on your answers as you learn more.

|Slide number 29
|Instructor notes
|
|Student notes
Here are a few key points to summarize this section:
Data scientists and data engineers both work with the data pipeline, and depending on your organization some tasks might be performed by either role. But generally speaking, the data engineer is focused on the infrastructure the data passes through and the data scientist is focused on analyzing the data in the pipeline. Both roles need to ask questions about how they data meets their needs, and iterate on their answers and design as they learn more.

|Slide number 30
|Instructor notes
|
|Student notes
This section introduces strategies that an organization can use to make the most of its data infrastructure.

Three-pronged strategy to build data infrastructure





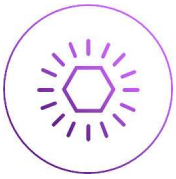
Modernize | Unify | Innovate

|Slide number 31
|Instructor notes
|
|Student notes
If you go to work for an organization that aspires to use its data as a strategic asset, you will likely spend time helping the organization to adapt more traditional infrastructures toward those that are designed to better support data analytics and AI/ML. Amazon Web Services (AWS) advises its customers to approach this type of data infrastructure along three strategies: modernize, unify, and innovate.

Modernize

Increase agility and reduce undifferentiated lifting

- Move from on premises to cloud-based services.
- Migrate to purpose-built tools and data stores.
- Build loosely coupled pipelines.



|Slide number 32
|Instructor notes
|
|Student notes

Modernizing the infrastructure gives an organization the agility it needs to respond quickly to changes in the business landscape and incorporate new features. The practices listed here are actually applicable to any modern application architecture, but in this course, you will look at them through the lens of data analytics and AI/ML use cases.

AWS use the term undifferentiated heavy lifting to refer to the IT tasks that don't really help a business focus on their core value to customers, but take a lot of their time and resources. For example, managing data centers and hardware, or patching operating systems and databases. By migrating applications and databases to the cloud, an organization can reduce its focus on administrative activities and instead focus development resources on business functionality. The cloud also makes it easier to start up different infrastructure to serve different use cases. An organization can choose the best set of tools and resources for each business problem. They can also experiment with new environment types or services without committing to

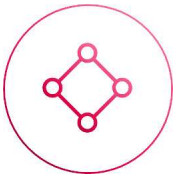
resources long term.

Another advantage of using cloud services is the ability to use purpose-built data stores rather than trying to use a single type of database to serve all needs. Choose a data store that matches the data structure and scalability that are required for each use case. Together with purpose-built data stores, use loosely coupled components so that you can use services that are designed for a targeted purpose and modify or scale them independently from each other. This greatly increases the flexibility of the infrastructure.

Unify

Create a single source of truth

- Break down data silos.
- Democratize access.
- Equip users with tools to visualize their own insights.
- Use a data lake, and run queries directly on data.
- Support simplified governance and movement between the data lake and purpose-built stores.



|Slide number 33
|Instructor notes
|
|Student notes

One of the most important and most challenging strategies for successful data-driven organizations is to unify data to create a single source of truth. This means moving from siloed databases that one group of users access to a data lake, which brings all the data together and democratizes access. Control of data moves away from a centralized team such as IT, where users must request data. With modern data strategies, users can directly access the data that is relevant to them and build their own visualizations. Users can run queries directly against the data in the lake, but your infrastructure should also support movement between the data lake and purpose-built data stores that are needed for specific use cases.

Moving to this type of data structure often requires significant changes to the cultural aspects of who owns the data and the governance needed to maintain the single source of truth.

Innovate

Use AI and ML to discover new insights faster

- Move from reactive to proactive decision-making.
- Incorporate AI/ML into decision-making, and tap into new insights in vast amounts of unstructured data.
- Take advantage of cloud services with AI/ML features that democratize who can use ML.

|Slide number 34
|Instructor notes
|
|Student notes

As noted in an earlier section, the exploding availability of data and the lower costs of applying AI make it possible to incorporate AI and ML into more decision-making. Organizations that take advantage of this can drive value in new ways and adapt to changing business needs more proactively. But it takes conscious effort to reframe business problems as prediction problems and experiment with ML models.

One trend in AWS service features is an increasing ability to apply an AI/ML type of functionality, without as much overhead or required experience as more traditional ML infrastructures. For example, Amazon SageMaker Studio provides fully managed ML infrastructure and a no-code visual interface for business analysts. Amazon Redshift is a data warehouse service, and one of its features, Redshift ML, helps users to create, train, and deploy SageMaker models by using familiar SQL syntax.

|Slide number 35
|Instructor notes
|
|Student notes
Organizations that can put their data to work to make better, more informed decisions, can respond faster to the unexpected and uncover completely new opportunities. To be successful, organizations need to treat data as a strategic asset and build an infrastructure that allows them to use that strategic asset for big and small decisions all across the organization.

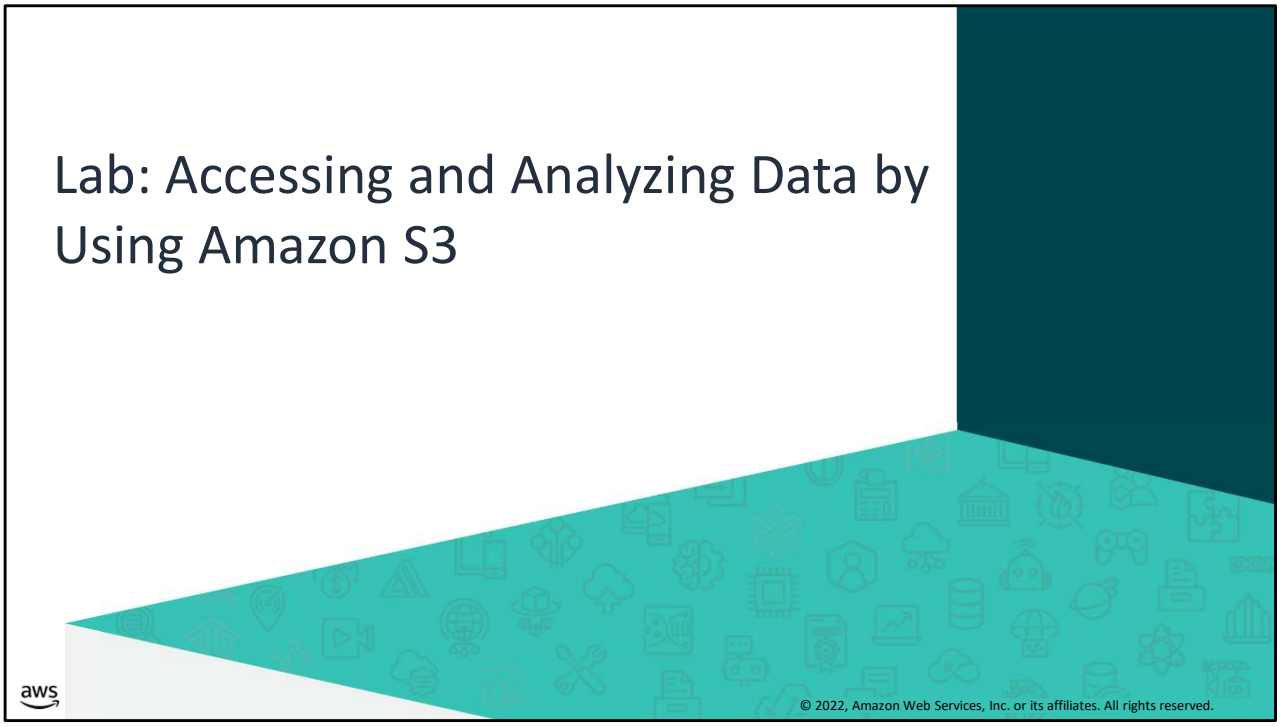
Key takeaways: Modern data strategies

- Organizations that want to become data driven should modernize, unify, and innovate with their data infrastructures.
- Modernizing is about moving to cloud-based infrastructures and purpose-built services to reduce administrative and operational effort.
- Unifying is about creating a single source of truth for data and making the data available across the organization.
- Innovating is about looking for new ways to find value in the data—specifically, applying AI and ML to find new insights.

|Slide number 36
|Instructor notes
|
|Student notes

Here are a few key points to summarize this section:

- Organizations that want to become data driven should modernize, unify, and innovate. This means moving to cloud-based, purpose-built services and reducing operational overhead. It also means creating a single source of truth for data and making data available across the organization.
- Organizations should innovate by finding new value in their data with the use of AI and ML.



|Slide number 37
|Instructor notes
|Lab details are provided in the Vocareum environment in the readme file associated with the lab.
|
|Student notes
You will now complete a lab. The next slide summarizes what you will do in the lab, and you will find the detailed instructions in the lab environment.

Lab introduction: Accessing and Analyzing Data by Using Amazon S3

- In this lab, you will use a simple pipeline to upload compressed files to Amazon Simple Storage Service (Amazon S3) and use S3 Select to run queries directly on the files.
- You will work with AWS Identity and Access Management (IAM) to set access permissions and AWS CloudFormation to automate creation of your infrastructure.
- Open your lab environment to start the lab and find additional details about the tasks that you will perform during this lab.

|Slide number 38
|Instructor notes
|
|Student notes
Access the lab environment through your online course to get additional details and complete the lab.

Debrief: Accessing and Analyzing Data by Using Amazon S3

- Now that you have used S3 Select in this lab, how do you think that you could use it on the job?
- What use cases would work for S3 Select?
- How might using S3 Select impact cost?
- What blockers exist for using this type of workflow with a team? What knowledge and skills would team members need to use this workflow?

|Slide number 39
|Instructor notes
|Q1- **Example strong response:** With S3 Select, you can use SQL statements to filter the contents of an S3 object and retrieve the subset of data that you need.
|Q2 - **Example strong response:** Using S3 Select could be helpful any time that you need to quickly query an object that is stored in an S3 bucket and is formatted as CSV, JSON, or Apache Parquet.
|Q3 - **Example strong response:** By using S3 Select to filter data that is stored in AmazonS3, you can reduce the amount of data that S3 transfers. This reduces the cost and latency to retrieve the data.
|Q4 - **Example strong response:** Teammates would need to have access to the data in the applicable S3 buckets. They would need to understand where the data is stored in folders and why the data is stored in a certain taxonomy. Teammates would also need to understand the different file types that S3 Select supports and how to use SQL to query the data. They would also need to understand the requirements and limits that are associated with using S3 Select.
|
|Student notes
Your instructor might review these questions with you, or you might review them on

your own. Use this opportunity to extend your thinking about the tasks that you performed during the lab.

|Slide number 40
|Instructor notes
|
|Student notes
This section summarizes what you have learned and brings the Data-Driven Organizations module to a close.

Module summary

This module prepared you to do the following:

- Compare and contrast how data analytics and AI/ML are applied to data-driven decision-making.
- Name and describe each layer in a data pipeline.
- List ways in which data is acted upon as it passes through a pipeline.
- List responsibilities for data engineer roles and data scientist roles in processing data through a pipeline.
- Define three modern data strategies that influence how you build your data infrastructure.

|Slide number 41
|Instructor notes
|This is a good opportunity to use an online group or discussion board to ask students to reflect on what they have learned. You might ask the students to recall a point from the module that aligns to one of the listed objectives. This provides a good segue to the knowledge check and sample exam question.
|
|Student notes
This module introduced data-driven decision-making and how data analytics and AI/ML can support decision-making. You learned that both data analytics and AI/ML can help with predictions, but AI/ML does it by learning about the data rather than by using programmed rules. You learned about the four main layers in a data pipeline: ingestion, storage, processing, and analysis and visualization. You were also introduced to the type of actions that are taken on data as it passes through the pipeline; for example, cleaning and transformation. You heard about the types of concerns that a data engineer or data scientist might have when building out and using a data pipeline. And finally, you learned about the modernize, unify, and innovate approach for building modern data infrastructures.

Module knowledge check

- The knowledge check is delivered online within your course.
- The knowledge check includes 10 questions that are based on material from the slides and slide notes.
- You can retake the knowledge check as many times as you would like.

|Slide number 42
|Instructor notes
|
|Student notes
Use your online course to access the knowledge check for this module.

Sample exam question

A traditional retailer with a relatively small online presence recently decided that they need to expand their online presence to increase sales. After a major marketing push and investing in IT infrastructure to support increased online transactions, quarterly results show that online sales results are far below expectations. The business wants to find ways to improve the poor online sales performance.

Which action represents an appropriate first step to address the request in a data-driven way?

Identify the key words and phrases before continuing.

The following are the key words and phrases:

- **Improve online sales**
- Appropriate **first step**
- **Data-driven** way

|Slide number 43

|Instructor notes: The key words section is animated to be revealed on click.

|

|Student notes

The question implies the need for some type of analytics solution, based on the desired outcome of improved sales and the desire to address the request in a data-driven way. The other key consideration is that the question asks for the first step.

Sample exam question: Response choices

A traditional retailer with a relatively small online presence recently decided that they need to expand their online presence to increase sales. After a major marketing push and investing in IT infrastructure to support increased online transactions, quarterly results show that online sales results are far below expectations. The business wants to find ways to **improve the poor online sales** performance.

Which action represents an appropriate **first step** to address the request in a **data-driven way**?

Choice	Response
A	Develop an ML model that analyzes customer comments to determine where customers are unsatisfied.
B	Use A/B testing to provide some customers with a new UI that is designed to simplify the purchasing process. Then, compare sales results from the old and new UIs.
C	Use a business intelligence (BI) tool to analyze sales data for the previous two quarters. Then, generate a hypothesis as to why the online sales are not in line with expectations.
D	Analyze the total revenue per customer, and create customer segments that receive different types of offers based on their segment.

|Slide number 44
|Instructor notes
|
|Student notes
Use the key words that you identified on the previous slide, and review each of the possible responses to determine which one best addresses the question.

Sample exam question: Answer

The correct answer is C.

Choice	Response
C	Use a business intelligence (BI) tool to analyze sales data for the previous two quarters. Then, generate a hypothesis as to why the online sales are not in line with expectations.

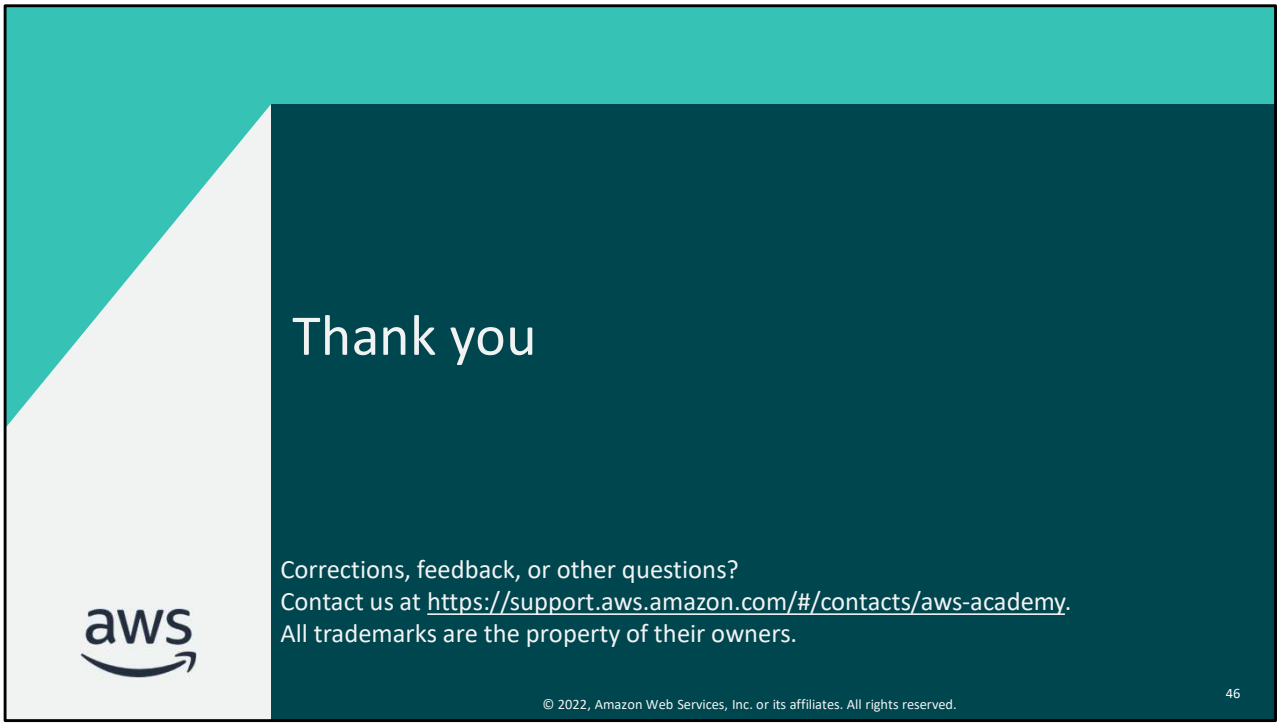
|Slide number 45
|Instructor notes
|
|Student notes

Choice A (Develop an ML model that analyzes customer comments to determine where customers are unsatisfied) might be appropriate if the retailer believes that the cause of poor sales is related to customer satisfaction with their products and services.

Choice B (Use A/B testing to provide some customers with a new UI that is designed to simplify the purchasing process, and then compare sales results from the old and new UIs) might be appropriate if the retailer believes that sales are low because customers are having difficulty making purchases through the older UI.

Choice D (Analyze the total revenue per customer, and create customer segments that receive different types of offers based on their segment) might be appropriate if the retailer believes that low sales can be attributed to the way that different customer segments shop on their website.

Each of these answers (A, B, and D) represents ways that the organization might gather and analyze additional data depending on their hypothesis. But the first step should be to work backwards from the business problem (poor sales) to form a hypothesis and then build a data pipeline to collect and process data iteratively. So choice C is the correct answer.



|Slide number 46
|Instructor notes
|
|Student notes
That concludes this module. The Content Resources page of your course includes links to additional resources that are related to this module.