**Data Consolidation Analytics.**

**Introduction.**

In the previous module, you learned that unlocking the value of data is central to digital transformation.

Businesses that leverage data effectively can improve efficiency and productivity, deliver fresh, personalized customer experiences, and create new business value.

In particular, I discussed the different types of data that businesses can access, and how you can combine them to generate insights and take intelligent action.

The way that data is collected, stored, and managed is foundational to what you can do with it.

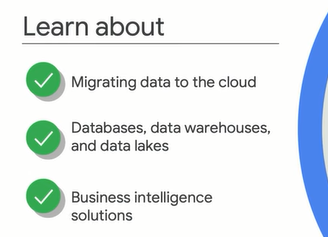
n this module, I'll start by considering where data is now and the benefits of migrating your data to the Cloud.

Then I'll define key terms related to data storage, including database, data warehouse, and data lake.

For each term, I'll cover some use cases in applicable Google Cloud Solutions.

Finally, I'll close the module by exploring business intelligence solutions like Looker, which enable businesses to gain insight into their data.

Let's get started.



**Migrating your data to the cloud.**

Think about your company or customer data.

Where is it currently stored?

Some of your data may already be stored in the Cloud.

But for most large organizations, a vast amount of data is still stored on-premises or buried on individual computers.

When you consider the value of your data, storing data on-premises in silos is like storing your money in a mattress (*colchón*).

And if there's data somewhere in your organization that isn't being tracked, it's vulnerable to user attacks and it's unproductive.

To get the most value from your data, you need to know what you have, find it easily, and be able to use it while keeping it secure from external threats.

This means that how data is stored and processed is critical to business success.

If you're storing your data on-premises, you'll need to start thinking about taking some or all of it to the bank, or in other words, to the Cloud.

It'll provide a greater return on investment.

How?

I'll explain.

When you store your data on-premises, you're responsible for the IT infrastructure that supports the collection, security, and processing of that data.

You're also responsible for maintaining and expanding the capacity of your IT infrastructure.

This can be expensive and time-consuming.

It's also difficult to scale quickly.

If you have limited storage space, then you risk missing opportunities to gain insights with new data.

You also risk downtime.

Imagine that you launch a mobile app and don't accurately predict demand.

A lot more people download and use it than you expect.

Your infrastructure isn't set up to absorb that much data that fast.

As a result, the app freezes or crashes, resulting in dissatisfied users.

With Cloud, you can rent space from public Cloud providers like Google Cloud.

This means that your data storage and compute powers are elastic.

It can scale up or down as the data you take increases or decreases.

Think about our money analogy.

Unlike your mattress, no matter how much your money grows, the bank can hold it.

Another way that migrating data to the Cloud provides a better return on investment is the speed at which you can ingest and use data, in particular, the speed at which diverse formats of data can be analyzed and used.

Let me explain.

In an earlier module, we discussed how data analysis was traditionally retrospective.

Business analysts would gather data from the past six months and run analytics to provide insight into what happened during those six months.

The analysis could take a few days, maybe even a few weeks.

Now businesses can ingest data in real-time.

For example, a business can use Cloud technology to track customer behavior on its website.

Each click, scroll, and back click is a data point.

Think about the millions of users viewing a website across the globe at any given time.

That's a lot of data.

But website clicks are only one kind of data.

With the right Cloud tools, businesses can unlock even more value with data.

Now a business can combine user clicks with other data to understand user intent.

For example, if a user makes regular purchases from your online store, one data point, and provides a five star rating a few days following each purchase, another data point, then those data points can be combined to determine intent.

The way you store this data provides the foundation for how you use it.

In the upcoming videos, I'll define database and data warehouse as the two main storage formats for structured and semi-structured data.

Then I'll explain how businesses typically use them and offer some Google Cloud data storage solutions.

**Cloud databases.**

In the last video, I mentioned that how data is stored is central to being able to use it.

There are many solutions for data storage. One format is a database. A database is an organized collection of data generally stored in tables and accessed electronically from a computer system. Companies typically use a database to keep track of their basic online transactions, provide information that will help the company run its business efficiently, or help managers and employees make better decisions. For example, a hotel booking site would use a database for their customer transactions. If a person books a room for a night, that data is captured in the database, and the room availability is updated in real time on all customer channels.

Another example is online banking. When someone transfers money from one account to another using their mobile app, that figure is updated in the bank's database in real time and the user is able to see the most up-to-date account balance.

Data integrity and scale are two priorities for businesses that use databases.

Data integrity or transactional integrity refers to the accuracy and consistency of data stored in a database.

Data integrity is achieved by implementing a set of rules when a database is first designed, and through ongoing error-checking and validation routines as data is collected.

Databases, therefore, also allow businesses to roll back transactions to see data history. In our banking example, suppose the customer goes to an ATM to check their account balance. It's not the same at what's displayed on the mobile app. In this case, the bank needs the ability to roll back the transactions to identify the source of the problem. Perhaps the ATM is broken, or perhaps the user didn't click the final transfer button on their app. This rollback integrity protects the bank from fraudulent claims and protects customers money.

Another priority when using databases is scalability.

Going back to our example, suppose the ATM belongs to a global bank that has a large customer base and processes high volumes of transactions every day. They need a database that can scale to meet that demand.

Or think about our hotel booking example. Websites that process global holiday bookings have millions of transactions happening every day that require transactional integrity at scale.

Different cloud providers offer a range of database solutions. Let's look at a couple of the most common Google Cloud database services and their benefits.

We'll start with **Cloud SQL**.

**Cloud SQL is a fully-managed Relational Database Management Service, or RDBMS**.

It's easily integrates with existing applications and Google Cloud services like Google Kubernetes Engine and BigQuery and built on the performance innovation in Compute Engine.

Cloud SQL is compatible with common database management systems and methodologies.

It offers security, availability, and durability, and storage scales automatically when enabled.

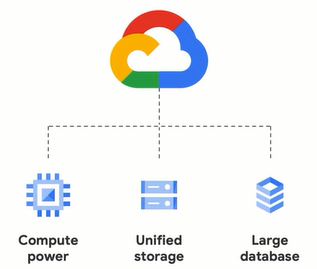
This makes it easy for organizations to set up, maintain, manage, and administer databases in the Cloud.

You might want to use Cloud SQL for databases that serve websites, for operational applications for e-commerce, and to feed into report and chart creation that informs business intelligence.

Let's look at a specific customer example headquartered in Mumbai, India. Living Consumer Products runs two flagship products: a casual dating mobile app called iCrushiFlush and a contextual digital platform, CDP, that provides digital marketing services to clients. By signing on to iCrushiFlush through Facebook, users provide details such as gender, location, and interest, as well as headshot images. iCrushiFlush stores this information in a database and displays it to other iCrushiFlush users through an algorithm depending on compatibility.

Due to the large data volumes generated by iCrushiFlush and CDP, as well as the need to allocate scarce personnel and financial resources to business projects, Living Consumer Products decided to operate in the Cloud from the beginning. However, testing and early experiences with public cloud services didn't meet the company's cost requirements. In addition, they felt that availability and scalability were non-negotiables. And they couldn't afford downtime that would drive users away. Living Consumer Products migrated iCrushiFlush and CDP to Google Cloud. Now they're running these services on Compute Engine to provide compute power, Cloud Storage to provide unified cloud storage, and Cloud SQL to run its relational database. This allows Living Consumer Products to both store and retrieve large volumes of data such as user images in real time.

We'll cover Cloud Storage in more detail in upcoming videos.



By handing off to Google the time-consuming tasks required to set up and run a database like applying patches and updates, managing backups, and configuring replications, you can save time and money, and keep your focus on building great applications.

**Cloud Spanner** is another fully-managed database service, and it's designed for global scale.

**With Cloud Spanner, data is automatically and instantly copied across regions**. This replication means that if one region goes offline, the organization's data can still be served from another region. It also means that queries always return consistent in ordered answers regardless of the region. For example, if someone in the London office updates information in the database, that update is immediately available for someone in the New York office. Consistency is critical to companies like Spotify. It provides users with music streaming services. If queries don't always reflect the latest change, this creates many challenges for the company. Spotify holds information about the objects it stores in the Cloud. This is known as metadata. Migrating their metadata storage to Cloud Spanner gave them strong listing consistency, so they know their queries will always reflect the latest data situation. Cloud Spanner provides strong consistency and massive scalability, which means that, for organizations, this is no longer a trade-off. Plus, it provides enterprise-grade security. This makes it ideal for organizations that want scalability for their databases, whether it's within a region or across the world. It's great for mission-critical online transaction processing, and because it's all managed, it dramatically reduces the operational overhead needed to keep the database online and serving traffic.

Cloud SQL and Cloud Spanner are examples of databases that enable customers to manage high volumes of transactional data at speed and at scale. With Google Cloud databases, businesses can build and deploy faster, deliver transformative applications, and maintain portability and control of their data.

Databases are one of the two main types of data storage systems in the Cloud. **The second is data warehouses**.

Data warehouses allow businesses to unlock insights and take intelligent action.

Watch the next video to learn more.

**Cloud data warehouses.**

In the last video, we explored Cloud databases and how they enable businesses to ingest and use high volumes of transactional data. Let's now discuss the second type of data storage system in the Cloud: data warehouses. While databases store transactional data in an online fashion, data warehouses assemble data

from multiple sources, including databases. Databases are built and optimized to enable ingesting large amounts of data from many different sources efficiently. However, data warehouses are built to enable rapid analysis of large and multidimensional datasets. For example, a dataset may capture every online sale every day of the week.

But if you want to analyze those sales to identify trends or even to get the sum of total sales each day, you need a data warehouse. Perhaps you also want to identify sales trends by combining data such as new product rollout dates, marketing campaign language, and operational efficiency data.

Think of the data warehouse as the central hub for all business data. Different types of data can be transformed and consolidated into the warehouse, so they're useful for analysis. In particular, a Cloud data warehouse allows businesses to consolidate data that is structured and semi-structured. Remember that unstructured data tends

to be unorganized and qualitative. In other words, it wouldn't fit in a spread sheet. When combined with connector tools, data warehouses can transform unstructured data into semi-structured data that can be used for analysis. Let's look an example. Consider the online hotel booking example we examined in the previous video.

Suppose they now want to do even more with their data. They want to use multiple types and sources of data to gain insights about hotel quality, and ultimately to improve their service to customers. They identify a list of possible data sources from their end-to-end customer journey such as:

number of bookings by type of room, number of guests, and time of year; overall satisfaction during their stay and satisfaction with hotel staff, amenities, food, and check-in and check-out processes; customer posts on social media platforms by sentiment, location, or specific event at the hotel; and customer feedback and complaints captured via the website,

mail, or in person at the customer service desk. All these different data types and formats are ingested and assembled into a data warehouse through different channels. The business can query the data warehouse quickly and at scale to derive meaningful insights. They can take their business's goal one step further

and use the source data to build machine learning models to surface personalized hotel recommendations and tailored booking experiences for customers. We'll talk more about that in the next module. For now, let's look at a Google Cloud leading data warehouse solution, BigQuery. BigQuery is a fully-managed data warehouse with downtime free upgrades

and maintenance and seamless scaling. Most of all, BigQuery allows you to analyze petabytes of data using incredibly fast speeds and zero operational overhead. This means that as an organization, you can focus on analyzing your data to find meaningful insights instead of spending time and resources on maintenance. Most data warehouse providers

link storage and compute together. So customers are charged for compute capacity, whether they're running a query or not. Importantly, BigQuery is serverless. This doesn't mean that there's no server. It means that resources such as compute power are automatically provisioned behind the scenes as needed to run your queries.

So businesses do not pay for compute power unless they're actually running a query. Ocado is one of the world's largest online-only grocery retailer. It experienced significant growth in its early years of business. As a result, it began to find that its old databases just weren't fast enough to meet business demands.

Now they've migrated to Google Cloud and use BigQuery. They found that query results are delivered 80 times faster and at 30 percent less costs. Let's look at another example. Bueno is a Software as a Service, or SaaS company, that helps businesses meet their sustainability goals by improving building systems.

Buildings are usually equipped with various networks that control and operate the facilities they contain. But air conditioning, lighting, and security systems often exist in their own silos with no link between them for communication. Bueno aims to bridge this gap by using technology to gather data and provide better transparency

between the different systems, so that the building sector can use this information for fault detection, optimization, and business intelligence. Bueno has used Cloud technology from the beginning. They decided to migrate to Google Cloud for a range of reasons, one being that they wanted to use existing data to better understand their customers and their business.

They do this using BigQuery. There are two other tools they use: Pub/Sub and Dataflow. Pub/Sub is a service for real-time ingestion of data, whereas Dataflow is a service for large scale processing of data. Remember how I described data warehouses as the central hub for data to flow into?

Well, these two different services, Pub/Sub and Dataflow, can work together to bring unstructured data into the Cloud and transform it into semi-structured data. This transformed data can then be sent directly from Dataflow to BigQuery, where it becomes immediately available for analysis. These tools enabled Bueno to unlock new insights about their customers,

such as equipment and site level data, weather data, and even the customers' last visits. Now customers can use Bueno system to look for insights and discover new things that may require action being taken. Their customers can generate a work order for maintenance to be done and log the cost of the job.

Bueno can then use that data to verify the value and accuracy of their analytics. Now that you understand the differences between data warehouses and databases and have learned about a few Google Cloud services, let's briefly explain how unstructured data can be stored in data lakes.

**Important data considerations.**

Capturing, storing, and analyzing vast amount of data is key to adopting Cloud technology.

But handling this volume and diversity of data comes with its own ethical considerations and requires alternative ways of thinking about security.

Google believes that capturing and managing data demands responsibility and accountability.

Not all information that can be captured should be captured.

In other words, businesses are accountable for making responsible decisions about which data they collect, store, and analyze.

This also extends the data that businesses already own.

In this case, it's essential to examine who has access to the data and how they'll be using it.

First, consider the source of the data, how it's being collected, and where it's stored.

If it's personal or sensitive data about a customer or an employee, it needs to be securely collected, encrypted when stored in the Cloud, and protected from external threats.

Additionally, only a subset of users should be granted permission to view or access the private data.

Data security and privacy becomes more complex in a global economy.

Regional or industry-specific regulations often guide data policies.

Google Cloud offers a range of solutions and best practice resources that companies can leverage.

Another consideration is whether all the data is relevant and appropriate.

Let me explain where this can be particularly important.

Suppose, for instance, you want to use thousands of lung X-ray images to train an ML model to automatically identify tumor markings in new patient X-rays.

What you need are the X-ray images.

This is the relevant data.

What's not relevant is patient's personal data.

You need to ensure that any source data about individuals such as names or addresses is omitted or redacted.

There's also some information that is not personally identifiable and should still not be included in the modeling for ethical reasons.

A good example of this is whether or not individuals have health insurance.

This is not relevant for educating the model to identify tumors.

And if the data is included, the solution could be discriminatory.

These ethical and privacy considerations are particularly complex when you're working with unstructured data.

For example, a customer support team that resolves hundreds of customer's issues a day via mail might want to use an automated tool to find patterns in the email passages and develop targeted solutions.

It's true that emails contain valuable data that can be mined to solve this challenge, but it's essential to be conscious of protecting customer privacy at the same time.

Ethical and fair considerations are particularly important and applicable when you work with artificial intelligence, AI, and machine learning.

We'll cover ML and important factors to consider in a later module.

For now, I want you to remember that human bias (*prejuicios, parcialidad*) can influence the way datasets are collected, combined, and used.

Because of this, it's always important to include strategies to remove unconscious biases as you start to leverage data to build new business value.

In this module, you learned the importance of data in digital transformation.

Unlocking the value of data enables a business to both rethink how they serve their customers and reimagine how they operate.

Ultimately, using data effectively enables any business, large or small, to better achieve its mission.

Move on to the next module to learn about data consolidation and analytics.

How you store and manage your data affects what you can do with it.

So in the next module, we'll examine the challenges, solutions, and use cases for different data consolidation and storage systems on-premises or in the Cloud.

**Quiz.**

1. What are the key benefits of using cloud technology to unlock value from data, especially for traditional Enterprises? Select the two correct answers.

Customers can collaborate with corporations to create industry trends.

Customers can now gain access to their own data instantly.

**Businesses can query their data and retrieve results instantly.**

Businesses can access open source data like never before.

**Businesses can process terabytes of data in real-time.**

2. Lucinda is creating a data map for her online learning company. Her datasets include learner demographics, their purchases, and browsing history. What data 'bucket' would these datasets fall into? Select the correct answer.

Industry data

Corporate data

Cloud data

**User data**

3. Eduardo is using a machine learning model to improve recruitment efficiency for his company. What candidate data is appropriate and relevant for training the model? Select the two correct answers.

Gender

**Years of experience**

Ethnicity

Address

**Education**

4. Mark owns a large pharmaceutical company that manufactures essential medical supplies. The production lines are required to operate efficiently at all times. How can Mark use cloud technology to ensure his production lines are meeting optimal performance requirements? Select the correct answer.

Evaluate historic data to inform new product development

Evaluate real-time data to monitor competitor landscape

Evaluate consumer feedback to identify customer sentiment

**Evaluate real-time data to predict maintenance requirements**

5.Images and videos are examples of what type of data? Select the correct answer.

Semi-structured

Structured

Organized

**Unstructured**