Introduction to module 2

Hi everyone. My name is Astrid and I'm a curriculum developer at Databricks. You can think of me as your tour guide through this MOOC. Anytime you see me, it'll mean that you're starting a brand new module. I'm based in Southern California, and since we are taking this MOOC journey together, it's probably good that we get to know ourselves a little bit. When I'm not working on training content at Databricks, I love to do things like go to the beach, play with my puppy, and I also really enjoy dancing. But enough about me, let's go ahead and talk a little bit about this module. You can think of this module as setting the last of the stage for the MOOC. What we'll cover in module 2 are things like: what is big data, and what characteristics define big data? How do people work with big data? Most importantly, why should you as a data analysts care about big data? What's the real point for you to take this MOOC? By the end of the module, you'll be able to describe common characteristics that define big data. Explain common struggles that data analysts face when working with big data. You'll also be able to describe the tools that data analysts use when they work with big data. Finally, we'll talk about Apache Spark and Spark SQL. You'll be able to explain how those two enable data analysts to streamline big data analysis. When you're ready, go ahead and click onto the next video to get started with this module.

# What is big data?

Hi again, I'll be leading you through a lot of the concepts we're presenting throughout this course. As we said in the introduction, all of your instructors, including me, work at Databricks, were rebuild curriculum to train data teams so that they can effectively work with spark and Delta Lake, which will learn about later. I personally have been developing statistics and computer science curriculum for over ten years. First in New York City Public Schools, and then later at the district level nationally. I currently live in New York City with my wife, my dog Charlie and my cat, Loki. And I love it here. New York City is a huge hub for tech and data science, which is fun for me personally and profesionally. But I bring it up now, mostly because I want for all of us to take a moment and just imagine New York City as a place with a lot of people who produce a lot of data, take a second and consider the number of people who live here. It's about eight million. And a huge section of that population has smartphones, and on each of those smartphones let's say they have ten or so apps. And those apps are collecting and producing data pretty much all the time. When we think about the vastness of that data well, that's big data. If the millions of people who are creating hundreds of millions of data points every day, or maybe every minute, or maybe every second. The term itself big data was going around 2005, right around the time data intensive industries were looking for a new solution. In this video, we're going to talk about the specific characteristics that define big data and uncover what does it mean when data is big? By the end of this video, you'll be able to describe the characteristics that define big data. When we start talking about big data, you'll find that it's often characterized by what people call the five v's of big data. Those v's are volume, velocity, variety, veracity and value. Will talk about what each of these words means, and what each of these words means to a data analyst.

When we talk about volume, we're talking about the massive amounts of data that's being generated every second of every day. We worked through a little bit of a thought experiment around New York City and how much data we could reasonably expect to come from all the people living in just that one city. And we said that's big data, but when we're talking about what qualifies as a massive amount, we have to consider the whole world. So lucky for us, the International Data Corporation, the IDC,

the report in 2018, where they discovered that there are currently 33 or at the time 33 zetabytes of data in existence in the world.

And based on that study, they're predicting that that is going to shoot up to 177 zetabytes of data by 2025. So, just to put that into perspective, let's try to think about how big is that zetabytes is.

My ordinary old laptop that it works well enough for me, has 256 gigabytes of storage.

If I had a really powerful computer then that would be about four times as much storage and coming in at one terabyte. One zetabyte is actually one billion terabytes, so it's vastly bigger, to say the least. And when we think about 177 zetabytes by 2025 that's a 600% increase.

So anything about volume from a data analyst perspective I've represents both a challenge and an oppurtunity. On the one hand, we've got a ton of data that's getting generated everyday, and you know more data means better decisions potentially, but

we as analysts also have to think about how are we going to access all this data? When we refer to velocity, we're referring to the speed at which new data is generated and the speed at which data moves around. But usually moved data is being generated is also moving to in from the database back to the end user. It's also moving all between databases and it's moving around really quickly. As data analysts, we can certainly appreciate the advantages of being able to gather, analyze and report on these large amounts of data, but we also have to think about how it's going to get processed and served to us for analytics. Will it be easy to query and can you create a real time report that truly represents the data? The next v is variety, and it refers to having different types and sources of data. So analysts usually work with structured and semi structured data that can generally be coerced into some kind of tabular format, something that looks kind of like a spreadsheet, but many businesses are also

collecting unstructured data like video files and social media posts. So we've got all these files of all these different types from all these different sources.

How are we as analysts going to be able to work with all those data structures and where are we going to be able to access each different kind of data? The next v, we want to talk about is veracity, which refers to the quality an accuracy of data. In this image were showing three different data sources reporting three different Q1 earnings, and we've already talked about how quickly data is coming in and how quickly data is moving around any given system, so it's not so hard to imagine that there may be some inconsistencies in that data. For data analysts, it s obviously important that we use high quality, accurate data to produce the best reports, and we always have to consider am I using the most accurate data? Can this day to be trusted? While all of these we talked about do affect how an analyst work, data value is the view that speaks directly to what an analyst brings to an organization.

Extracting value from big data can be really complicated and it needs to be transformed into shareable, actionable insights and made visible to the larger organization. Next, we're going to take a look at how these characteristics make it a challenge to work with big data.

# Common struggles with big data

Imagine for a moment that your manager asks you to produce a dashboard that hows

various levels of customer transaction data for your organization. You say okay. When you start to actually look at the data, you realize that there are hundreds of files that are stored in a wide variety of locations, and you're dealing with a bunch of different formats. How would you manage that? And over again, we're going to need that within 30 minutes. I know this request might seem exaggerated, but this is exactly the kind of big data workload that we need to consider. >> In this video, we'll explore some of the common struggles data analyst face when working with big data. By the end of the video, you should be able to explain some of the challenges and probably even let us know what we forgot to mention.

All right, so let's just revisit your task that you were just presented with.

Ultimately, we want to create a dashboard showing customer transaction data, okay, that sounds fun.

Upon closer inspection, we find that we're going to need to use hundreds of data files, okay? We're going to need to pull from a bunch of different data sources. We're going to possibly contend with lots of different file formats and we've got under 30 minutes.

Okay, well, whether or not you can complete this task might be partly to do with your skill set, and might be partly to do with this system that you're working in. In the next few slides, we'll dig into a couple of specific problems that could prevent this task from coming together. First, sometimes data analyst just have a lack of proper tools.

It is entirely possible that scalability was not part of your company's blueprint and they were focused on different things, like making the business succeed before worrying about how to make it succeed for millions of users. So, in situations like that, we might end up with some adhoc solutions in an attempt to meet big data needs. That might look like working with spreadsheets, or it could look like working with a system that's not meant to deal with big data. And it might just result in very long query processing times.

In the end if you don't have the right tools, you may just come up against a fundamental inability to meet the company's needs. Another very common pain point for data analysts is working with multiple data sources. What that might look like for you is using a wide variety of tools to access data. So for example you maybe have a bunch of CSV's and spreadsheets downloaded to your local computer. Or maybe you access data through various business intelligence platforms. Or maybe you're working with various database management systems and or SQL clients. Or maybe your data collection involves you walking down the hall as Sheila's desk and grabbing a bunch of files? I'm not sure, it looks different for a lot of different people. The point is it can be very difficult to work with multiple data sources unless you're in an organized system.

And that brings us also to having a single source of truth. It is widely recognized in the business world that having a single source of truth on which to make business decisions is pretty important. When we're working with different data sources, is it possible that that leads to working with different data? Many large systems have been built to avoid this kind of problem. But if you are working on spreadsheets in CSV's and things like that, it's very easy to be working with outdated or incorrect information.

All right, so just to review some known problems. Definitely, data analyst sometimes suffer from a lack of proper tools, or working with multiple data sources in a disconnected system. And finding that maybe there isn't a single source of truth for the company. So what can we do to make these struggles better? Data teams need a well-designed data infrastructure so that it's clear where they can go to access information.

Universal tools that work with multiple different file formats and with different file sources. A single source of truth that they can trust for reliable information. And to work on a unified data team, where we're sure that all of the parts of the team are working with the same data.

# Big Data Needs

Today, many practitioners in organizations are trying to figure out what's the best way to work with big data. What does that evaluation process look like? We can start by considering the big picture of a system to handle big data. In particular, we've developed three high level big data needs that an organization has. Collecting data, storing data, and using that data to produce actionable insights. By the end of this video, you'll be able to explain each of these needs and what it means in terms of working with big data. But column working with big data. Data coming in, in mass of volumes very quickly and in many different formats. Start simply with collecting data

about a variety of events or transactions. We'll call that high flux event data. All it means is that we've got data coming in from somewhere. From there, that data must be stored. Different companies will have different solutions for storage. Some data storage solutions might look like multiple different databases, for a particular business needs. For example, you might have one that is designated to sales data, and then another for operations data, and then finance data is stored in a database separate from those. Another choice might be a centralized data repository like a data warehouse. That's a little bit like the previous system, except it goes and brings all of

those multiple different databases under one roof, so you're still working in a single system, but you still have all that same structure where everything is very much separated out. Then, recently, data lakes have become really popular for storage. We'll talk more in a later lesson about all of these storage options, and their relative benefits. But I'll say now that data lakes are popular because they can store unstructured and semi-structured and structured data altogether, and it gives that single source of truth that we've been talking about. More on those later. But lastly, we can't forget that the whole reason to work with big data is to access insights and to make smarter decisions. We do this through real-time dashboards, periodic reporting, and artificial intelligence. In this course for data analysts, it makes sense that we're going to focus on real-time dashboards and periodic reporting. In order to enable big data workloads, we'll need to choose tools that are designed to work in a system like this.

# Introduction to Apache Spark(TM)

So where does Apache Spark fit into this big data landscape? As you're well aware, in this course we're going to be using Apache spark to process data. Now it's time to learn all about it. By the end of this video, you'll be able to explain the benefits of using Apache spark to work with big data. So what is Apache Spark? Apache spark is an analytics engine designed to unify data teams and meet big data needs. Among the big data community, it is very well known and widely used for its speed is abuse in generality. That is its ability to seamlessly integrate data. Applications that can include SQL streaming or complex analytics. In the next few slides will dig deeper into what makes it so powerful, which is the distributed computing engine. How it simplifies data access by working with a variety of sources. And how the rest of your data team can access it. Even though you may not be working in the same programming language.

Okay, so let's start by talking about Spark's distributed computing engine conceptually at a high level. It's pretty simple. Basically, if you've got a bunch of work and you're assigning it to one machine as you would in a traditional system. It's going to take that machine some length of time to finish all the work. If it's a large workload, is going to take the machine a very long time. If it's a small workload, it could take machine not very much time at all. In a distributed computing edge and what's going to happen is you're going to feed in the workload. And then that work is going to be distributed across a cluster of machines. This means that you can distribute that large workload among a few different machines. An all of them can be working on the problem simultaneously. That means that the work can get done in a fraction of the time. If you run into a very large workloads and you're hitting limits on performance. The solution may be to add more machines and when you're using, spark that simple. Another benefit to working with spark is that you can connect to multiple data stores in the eye tools. So in terms of where we're going to get data from, we can start off using Spark.

And from there we can pull in data from just regular cloud storage or we can pull in data from a traditional SQL database. Or we can pull in data from a data lake or in what we have here is a Delta Lake which will talk about later. So we can pull that data in using spark and then we can access that data from a variety of tools. It could be from a by tool or it could be from Databricks which will be using in this class. Or it could be from RE dash. Finally, one thing that smart can do is help to unify a data team. And it does im part of the way it does that is being accessible from multiple language APIs.

That means you can access spark using SQL or R or Python or Scala. And that means that all members of a data team. Who may be typically relegated to working with different tools and different systems can all work in the same platform. On the same documents and from the same data.

# Spark SQL

All right, now that we know a little bit about Spark, let's learn what Spark SQL is all about. In this video, you'll learn what Spark SQL is, why we're using it, and how you'll be working with it in this course. Spark SQL is built to work with Spark an is designed for structured data processing. It allows us to use basic SQL syntax to access Spark's unified analytics engine. SQL was developed to work with relational databases, not distributed processing engines like Spark. For the most part, we'll be using Spark SQL as a SQL interpreter, meaning you won't have to worry too much about what it's doing or how it's doing it. You can focus on writing easily readable queries that return the results you want. But Spark SQL is much more than an interpreter and we'll be investigating some of the optimizations it's applying to your code without you even having to know about it. So why use Spark SQL? Well, ease of use for one, it makes it really easy for SQL users to work with Spark. You can run exactly the same queries and commands you're used to, plus access some additional functionality all using a standard SQL syntax. And remember that means since Spark and access multiple data sources, you can easily read and write to multiple data sources and use a wide variety of data formats all from the comfort of standard SQL. Beyond that, we'll also see the benefits of Spark's super cool optimization engine. Each query you run is going to be passed through an optimization engine that is designed to give you the best possible performance. So it can adjust the order of the business logic you express or it can skip reading unnecessary data, and you can find the most cost efficient way to deliver your results. And finally, I know it doesn't help much for this particular course, but we are hoping you bring this back to your workplace, and the end result is unified analytics teams. Your ability to work with Spark SQL can bring you closer to your data teams. That is, you can access Spark through SQL and another part of the data team may be accessing Spark through Python or another part of the data team may be accessing Spark through R. These are both popular languages for data science. This means the whole data teams can work with the same tools and within the same system and easily share data and insights. Okay, well with that, we come to the end of module 2. Next, you'll take a quiz to check your understanding of the material so far. Then will start Module Three where you'll start working with Apache Spark hands on.