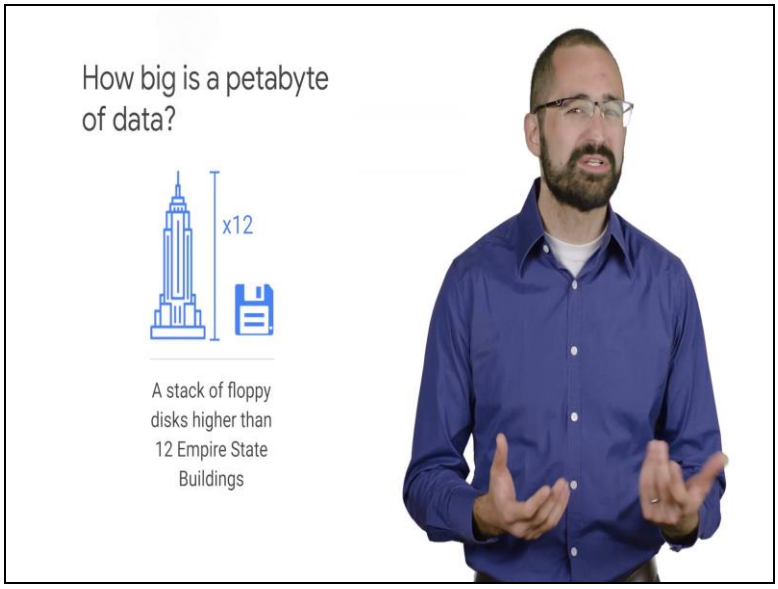
Google Cloud Computing Foundation Course
Evan Jones
Technical Curriculum Developer
Google Cloud

Lecture-70
Introduction to Big Data Managed Services in the Cloud

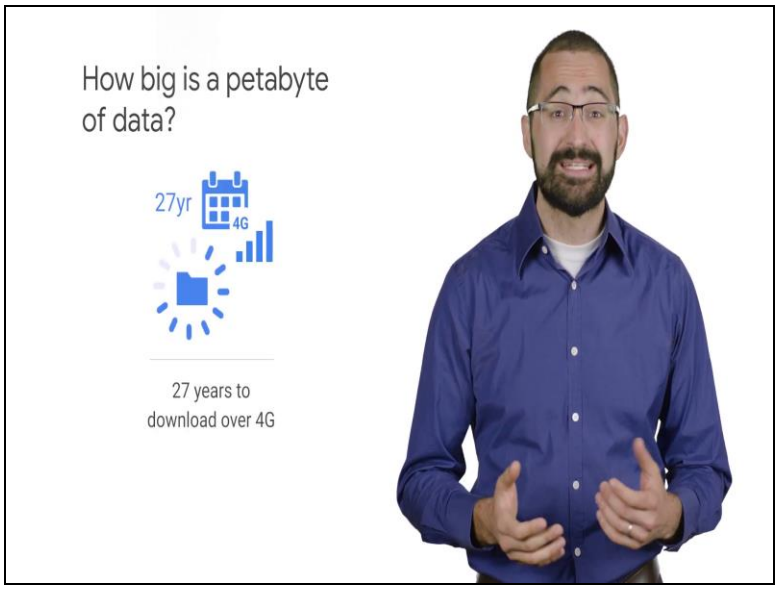
Let us start with the first topic where you will be introduced to big data and manage services in the cloud. Before we discuss big data manager services in the cloud let us take a moment to conceptualize Big Data. Enterprise storage systems are leaving the terabyte behind as a measure of data size with petabytes becoming the norm. We know that one petabyte is 1 million gigabytes or 1000 terabytes but how big is that.

(Refer Slide Time: 00:27)



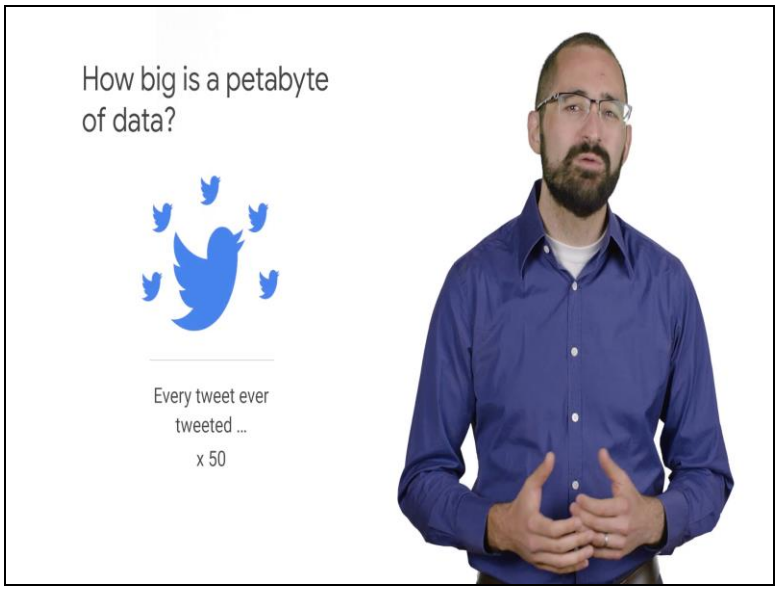
From one perspective a petabyte of data it might seem like more than you will ever need for example you will need a stack of floppy disks higher than 12 Empire State building's to store one petabyte.

(Refer Slide Time: 00:39)



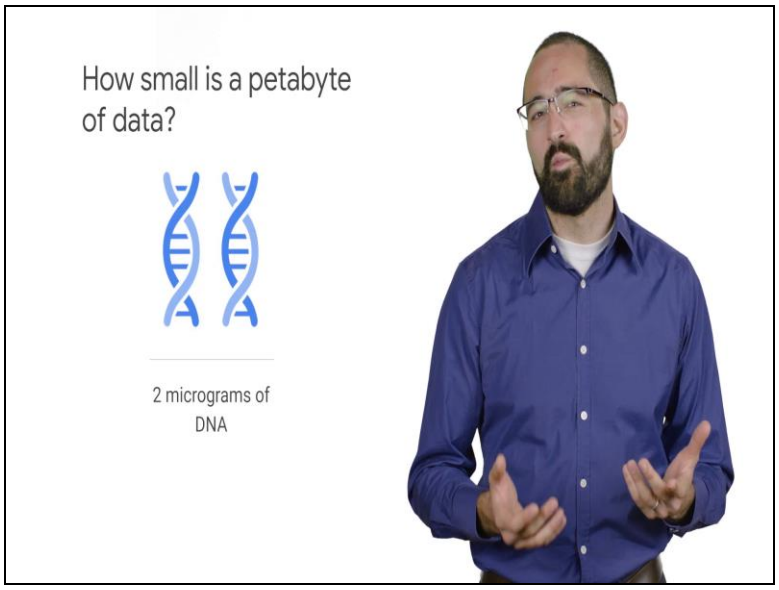
If you wanted to download one petabyte over a 4g network you will need to wait around for 27 years.

(Refer Slide Time: 00:48)



You also need 1 petabyte of storage for every tweet ever tweeted multiplied by 50 ok, so 1 petabyte is pretty big.

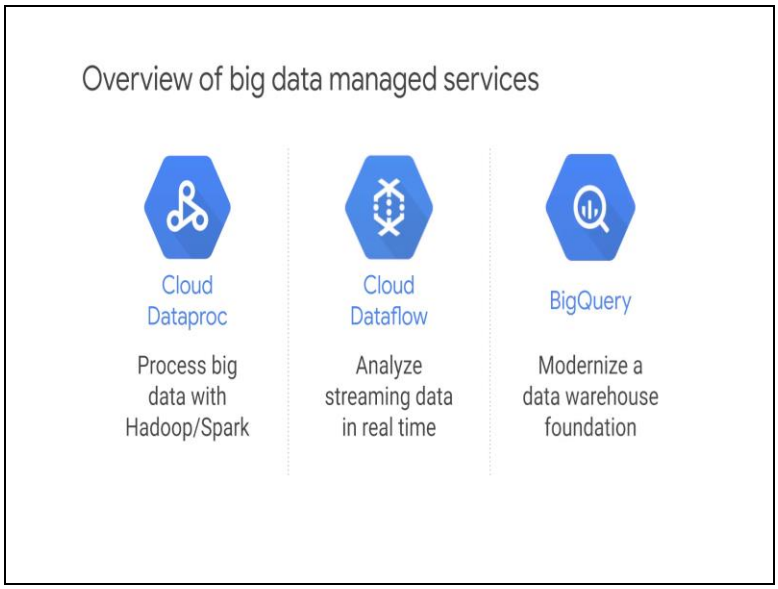
(Refer Slide Time: 00:59)



Looking at it from one other perspective though 1 petabyte is only enough to store 2 micrograms of DNA or one day's worth of video uploaded to youtube. So, for some industries a petabyte of data might not be that much at all. Every company saves data in some way 90% of data saved by companies is unstructured. With all these data available companies are now trying to gain some insight into their business based on the data that they have.

This is where big data comes in, big data architectures allow companies to analyze their save data to learn more about their business. In this module you will be focusing on 3 managed services that Google offers to process that data.

(Refer Slide Time: 01:45)



For companies that have already invested in Apache Hadoop and Apache spark and would like to continue using these tools cloud data proc provides a great way to run open source software in Google cloud. Companies looking for a streaming data solution however may be more interested in cloud dataflow as a managed service. Cloud dataflow is optimized for large-scale batch processing or long running stream processing of both structured and unstructured data.

The third managed service that we will look at is bigquery which provides a data analytic, solution optimized for getting questions answered, rapidly over a petabyte scale data sets. Bigquery allows for fast sequel or structured query language on top of your structured data.