Hello everyone and welcome to Hadoop Fundamentals – What is Hadoop. My name is Warren Pettit.

In this video we will explain what is Hadoop and what is Big Data.

We will define some Hadoop-related open source projects and give some examples of Hadoop in action.

Imagine this scenario: You have 1GB of data that you need to process.

The data is stored in a relational database on your desktop computer and this desktop computer has no problem handling this load.

Then your company starts growing very quickly, and that data grows to 10GB.

And then 100GB.

And you start to reach the limits of your current desktop computer.

So you scale-up by investing in a larger computer, and you are then OK for a few more months.

When your data grows to 10TB, and then 100TB, you are quickly approaching the limits of that computer.

Moreover, you are now asked to feed your application with unstructured data coming from sources like Facebook, Twitter, RFID readers, sensors, and so on.

Your management wants to derive information from both the relational data and the unstructured data and wants this information as soon as possible.

What should you do? Hadoop may be the answer!

What is Hadoop?

Hadoop is an open source project of the Apache Foundation.

It is a framework written in Java originally developed by Doug Cutting who named it after his son's toy elephant.

Hadoop uses Google’s MapReduce and Google File System technologies as its foundation.

It is optimized to handle massive quantities of data which could be structured, unstructured or semi-structured, using commodity hardware, that is, relatively inexpensive computers.

This massive parallel processing is done with great performance.

Hadoop replicates its data across multiple computers, so that if one goes down, the data is processed on one of the replicated computers.

It is a batch operation handling massive quantities of data, so the response time is not immediate.

Hadoop is not suitable for OnLine Transaction Processing workloads where data is randomly accessed on structured data like a relational database.

Also, Hadoop is not suitable for OnLine Analytical Processing or Decision Support System workloads where data is sequentially accessed on structured data like a relational database, to generate reports that provide business intelligence.

As of Hadoop version 2.2, updates are not possible, but appends are possible.

Hadoop is used for Big Data. It complements OnLine Transaction Processing and OnLine Analytical Processing.

It is NOT a replacement for a relational database system.

So, what is Big Data?

With all the devices available today to collect data, such as RFID readers, microphones, cameras, sensors, and so on, we are seeing an explosion in data being collected worldwide.

Big Data is a term used to describe large collections of data (also known as datasets) that may be unstructured, and grow so large and quickly that it is difficult to manage with regular database or statistical tools.

Other interesting statistics providing examples of this data explosion are:

There are more than 2 billion internet users in the world today,

and in 2014 there will be 7.3 billion active cell phones.

Twitter processes 7TB of data every day, and 500TB of data is processed by Facebook every day.

Interestingly, approximately 80% of these data are unstructured.

With this massive quantity of data, businesses need fast, reliable, deeper data insight.

Therefore, Big Data solutions based on Hadoop and other analytics software are becoming more and more relevant.