Transcript name: What is Hadoop?

English

Hello everyone and welcome! My name is Akmal Chaudhri.

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

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

The data are stored in a relational database in 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.

And you are fast 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!

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

As of Hadoop version 0.20.2, updates are not possible, but appends will be possible starting in version 0.21.

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

Hadoop is not suitable for OnLine Transaction Processing workloads where data are randomly

accessed on structured data like a relational database.

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

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 statistics tools.

Other interesting statistics providing examples of this data explosion are:

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

and 4.6 billion mobile phones in 2011,

and 7TB of data are processed by Twitter every day,

and 10TB of data are 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.

This is a list of other open source projects related to Hadoop:

Eclipse is a popular IDE donated by IBM to the open source community.

Lucene is a text search engine library written in Java.

Hbase is the Hadoop database.

Hive provides data warehousing tools to extract, transform and load data, and query this data stored in Hadoop files.

Pig is a platform for analyzing large data sets. It is a high level language for expressing data analysis.

Jaql, or jackal, is a query language for JavaScript open notation.

Zoo Keeper is a centralized configuration service and naming registry for large distributed systems.

Avro is a data serialization system.

UIMA is the architecture for the development, discovery, composition and deployment for the analysis of unstructured data.

Let's now talk about examples of Hadoop in action.

Early in 2011, Watson, a super computer developed by IBM competed in the popular Question and Answer show "Jeopardy!".

Watson was successful in beating the two most popular players in that game.

It was input approximately 200 million pages of text using Hadoop to distribute the workload for loading this information into memory.

Once the information was loaded, Watson used other technologies for advanced search and analysis.

In the telecommunications industry we have China Mobile, a company that built a Hadoop cluster to perform data mining on Call Data Records.

China Mobile was producing 5-8TB of these records daily. By using a Hadoop-based system they were able to process 10 times as much data as when using their old system,

and at one fifth of the cost.

In the media we have the New York Times which wanted to host on their website all public domain articles from 1851 to 1922.

They converted articles from 11 million image files to 1.5TB of PDF documents. This was implemented by one employee who ran a job in 24 hours on a 100-instance Amazon EC2 Hadoop cluster

at a very low cost.

In the technology field we again have IBM with IBM ES2, an enterprise search technology based on Hadoop, Lucene and Jaql.

ES2 is designed to address unique challenges of enterprise search such as the use of an enterprise-specific vocabulary, abbreviations and acronyms.

ES2 can perform mining tasks to build acronym libraries, regular expression patterns, and geo-classification rules.

There are also many internet or social network companies using Hadoop such as Yahoo, Facebook, Amazon, eBay, Twitter, StumbleUpon, Rackspace, Ning, AOL, and so on.

Yahoo is, of course, the largest production user with an application running a Hadoop cluster

consisting of approximately 10,000 Linux machines.

Yahoo is also the largest contributor to the Hadoop open source project.

Now, Hadoop is not a magic bullet that solves all kinds of problems.

Hadoop is not good to process transactions because it is random access.

It is not good when the work cannot be parallelized.

It is not good for low latency data access.

Not good for processing lots of small files.

And not good for intensive calculations with little data.

Now let's move on, and talk about Big Data solutions.

Big Data solutions are more than just Hadoop. They can integrate analytic solutions to the mix to derive valuable information that can combine structured legacy data with new unstructured data.

Big data solutions may also be used to derive information from data in motion.

For example, IBM has a product called InfoSphere Streams that can be used to quickly determine customer sentiment for a new product based on Facebook or Twitter comments.

Finally, let's end this presentation with one final thought: Cloud computing has gained a tremendous track in the past few years, and it is a perfect fit for Big Data solutions.

Using the cloud, a Hadoop cluster can be setup in minutes, on demand, and it can run for as long as is needed without having to pay for more than what is used.

This is the end of this video. Thank you for watching. To learn more, visit BigDataUniversity.com.