Here is a list of some 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 then, query this data stored in Hadoop files.

Pig is a high level language that generates MapReduce code to analyze large data sets.

Jaql is a query language for JavaScript open notation.

ZooKeeper 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.

In that contest, Watson was successful in beating the two most winning Jeopardy players.

Approximately 200 million pages of text were input 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 geoclassification 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 due to its lack random access.

It is not good when the work cannot be parallelized or when there are dependencies within the data, that is, record one must be processed before record two.

It is not good for low latency data access.

Not good for processing lots of small files although there is work being done in this area, for example, IBM's Adaptive MapReduce.

And it is 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. Please continue with the other units in this course.

Here is a list of trademarks that may have been used in this presentation.