So this week, we will learn about an important framework and programing model called Map Reduce, which is fundamental to the distributed computing in the big data world.

We will revisit Lambda architecture and how it is used in some of the big data scenarios.

Later, we will spend time on learning some of the key properties of the data and the data model.

And we'll look into some interesting examples. So what is MapReduce?

It is a programing model that consists of two important steps known as Map and Reduce.

It has the implementation framework for processing and generating big data sets,

with parallel and distributed algorithms running on a cluster of commodity machines.

So MapReduce program is composed of a map procedure which performs filtering and sorting, such as sorting the students by first names into a number of cues.

One cue for each name and produce method, which performs as summary operation in this case, such as counting the number of students in each cue, yeilding the name frequency of all the students.

The MapReduce system, also called the infrastructure, or the framework, orchestrates the process by marshaling the distributed servers, running the various tasks in parallel, managing all the communications between them and data transfers between the various parts of the system and providing for redundancy and fault tolerance.

The most important implementation of MapReduce is the Apache Hadoop Distributed Platform.

Think about this example.

Suppose that you need to calculate the average salary of one billion people and you have a cluster with a thousand computers, each with a processing unit and a storage memory.

The people can be divided into thousand chunks, or subsects, with the data of one million people each.

Each chunk can be processed independently by a single of the computers in parallel.

The result, which is the average salary of one million people produced by each of these thousand computers, can be averaged as a second step in the Reduce function, returning the final salary average.

That is the big advantage of MapReduce for scalability for large amount of data without sacrificing any of the performance.

Throughout this course you will learn and get to use these important concepts.

Thank you for watching.