What is Apache pig?

It is a high-level scripting language used to analyze big data. It was developed by Yahoo in 2006 with the aim of creating and executing MapReduce jobs on all the datasets.
It's name is "pig" as it can work on all type of data types that are: structured, semi structured, unstructured.
With the help of Apache pig, the data analysts had to spend less time on writing complex MapReduce programs.

Architecture of Apache Pig

The architecture of Apache Pig contains two components: Pig Latin Language and runtime environment.
The Pig Latin Language (*procedural language*) is used to write the data analysis programs and the programmers can use it to make their own programs for the purpose of reading, writing or processing the data.
The runtime environment is responsible for executing the Pig Latin programs.
The Pig Latin programs contain a series of transformation, which are applied to the input data to produce the output.

Pig Engine Component:
Programmers must write scripts using Pig Latin language. These scripts are then internally converted to Map and Reduce tasks. The Pig engine component is responsible for converting these scripts into MapReduce tasks.

Execution Modes of Pig:

There are two execution modes of pig, local mode and map reduce mode.

Local Mode:
This mode is more suitable for analyzing small datasets. It runs on a single JavavirtualMachine and it uses the local file systme.

MapReduce Mode:
This mode combined with a fully distributed cluster can be useful for analyzing large datasets. In this mode the queries written in Pig Latin are translated to MapReduce jobs and they are executed on a Hadoop cluster. And this cluster can either be partically or fully distributed.

Why is there a need to use Apache Pig?

Programmers who were not familiar with Java found it difficult to write the complex code for the MapReduce tasks. Not only this the Java code was lengthy as well as complex and more time consuming. The Pig Latin language allowed the programmers to write the code more quickly for the MapReduce tasks. Apache Pig is based on

“Multi-request”, which means that it follows the approach that reduces the code length. For example, an operation that would take 200 lines of code in Java, will only take 10 lines in Pig. On average, Apache Pig divides the development time by 16. Pig Latin language is similar to SQL, and people familiar with SQL can easily code in it for the MapReduce tasks.

User defined functions can be created in different programming language like Java, which can be integrated into Pig scripts.

Apache Pig can analyze all types of data, and the analysis result is stored in Apache Hadoop HDFS(Hadoop Distributed File System).