**Apache Install**

Before getting into the installation part, we would first have a general overview about Apache and how it’s used in Data Science.

**What is Apache?**



Apache Web Server is an HTTP server that presents websites to visitors that come to your server. So if you want to deploy a website for a business or your organization, you would most likely use Apache for that.

There are other HTTP servers out there such as IIS, but Apache is the standard that most people use whether they’re on Linux, Windows or Mac. Apache is the default that most people go to because it’s well known, it’s very reliable, and it’s free.

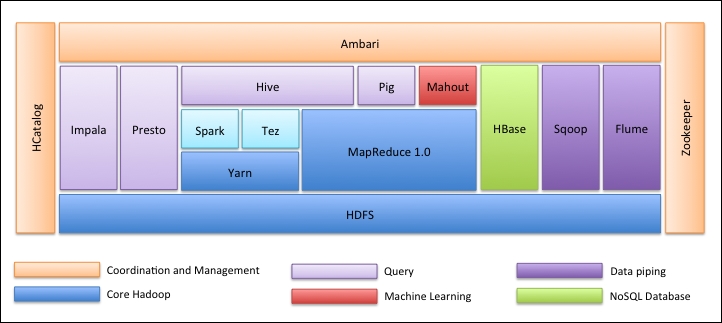
However, one thing to realize with Apache is that, as it is an HTTP server, so if you install this on Linux or Windows or Mac, all it would allow you to do is to present static websites to visitors coming to your server. Hence, if you code out an HTML website with no additional programming languages other than JavaScript, you can use that with just an Apache server. You could plug all your tags into the Apache server and present it to your visitors.

**How Apache used in Data Science?**

Data Science is the most in-demand field of study in the modern world. Data Scientist is regarded as the sexiest job in the 21st century with professionals from various disciplines wants to learn and become a Data Scientist.

Apache plays a crucial part for any Data Science enthusiast as they are required to have sufficient knowledge of Apache Hadoop Ecosystem.

**Apache Hadoop Ecosystem:-**



The very first thing is the Hadoop Ecosystem is not one tool. It’s not a programming language or a single framework. It is a group of tools which are used together by various companies in different domains for multiple tasks. We will go through each tool one by one below: -

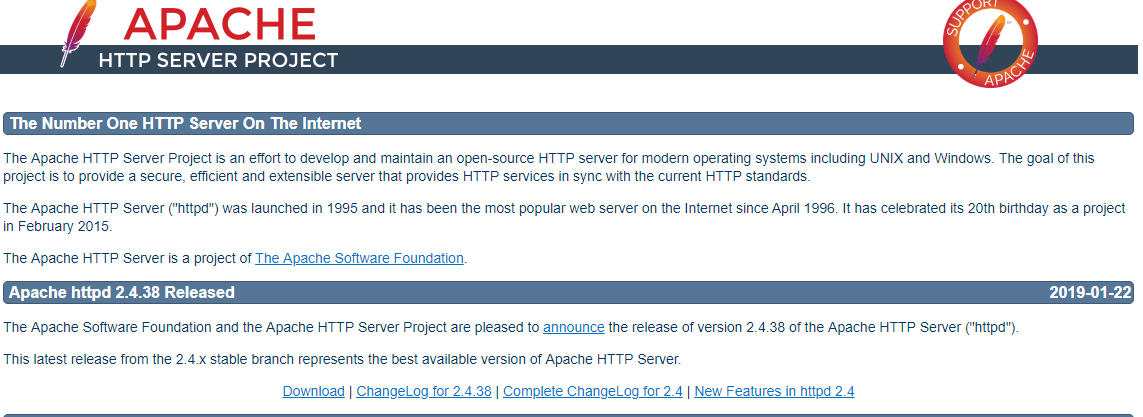
1. **Apache HDFS (Hadoop Distributed File System)** is the storage unit of Hadoop which could store structured, semi-structured and unstructured data. HDFS has a metadata which maintains the log file about the stored data. It has two components – NameNode and DataNode.
2. **Apache Yarn** is the resource negotiator which performs all processing activities like scheduling tasks, allocating resources, etc. It has two services – First is the ResourceManager who schedules applications running on top of Yarn. Second is the NodeManager who monitors resource utilization**.**
3. **Apache MapReduce** is the Data Processing component of Hadoop which processes large datasets using distributed and parallel computing based on Map, Sort and Shuffle, and Reduce functions. Map function filters the data, then sorting, and shuffling is done and at the end Reduce function aggregates and summarizes the result.
4. **Apache Pig** used mostly in ETL. It has two parts – Pig Latin and the Pig runtime. Pig Latin is the language used for data processing using a query, whereas Pig runtime is the execution environment. One line of Pig Latin is almost equal to 100 lines of Map Reduce code. The process involves first to load the data and then group, sort, filter and store it in HDFS.
5. **Apache Hive** uses a SQL-like query to analyse data in a distributed environment. It has two components – the Hive Command Line and the JDBC/ODBC server and the language used is called HiveQL.

1. **Apache Mahout** is the Machine Learning library written in Java and used to create machine learning applications such as clustering, classification or regression. It has different algorithms inbuilt for different use cases.
2. **Apache HBase** is a NoSQL database written in Java that runs over Hadoop. It’s built based on Google’s BigTable and is capable of handling all types of data.
3. **Apache Sqoop** is one the Data ingestion tool which is used for bulk structured data transfer between RDBMS and Hadoop.
4. **Apache Flume** is another data ingestion tool which is used for semi-structured and unstructured data transfer between Hadoop and other data sources.
5. **ZooKeeper** is the coordinator which ensures coordination between various tools in the Hadoop ecosystem.
6. **Apache Ambari** is a Cluster Manager who provisions, manages Hadoop clusters and also monitors their health and status.
7. **Apache Tez** is a new tool in the Hadoop ecosystem which accelerates Hadoop's Query processing.
8. **Apache Presto** is an open source distributed SQL query engine which enables cross-platform query capability.
9. **Apache HCatalog** is a metadata and table management system for Hadoop which enables interoperability across data processing tools. It also helps users choose the best tools for their environments.
10. **Apache Spark** is the most widely used and popular framework among the Data Scientist. It is a high-speed cluster computing system which optimizes resource utilization in case of many iterative tasks. It gives flexibility for both batch processing and real-time data analysis.

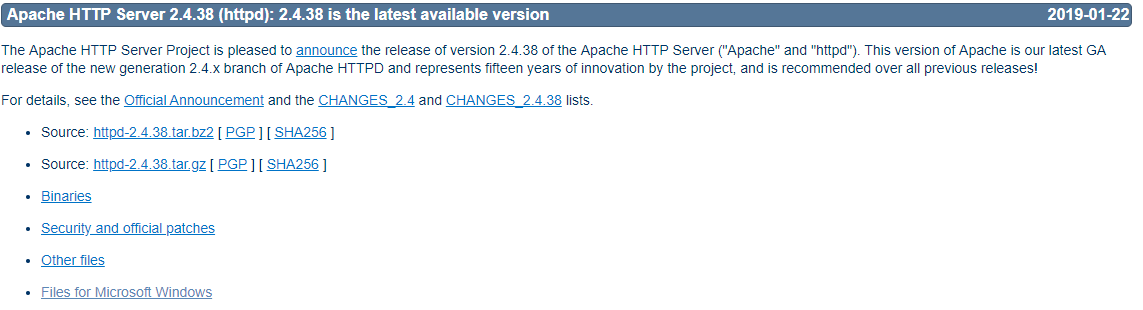
**Installation**

So far, we have learned about Apache and how it’s useful for anybody who wants to learn Data Science or Big Data Analytics. Now, we will dive down and install apache on windows based on the below steps.

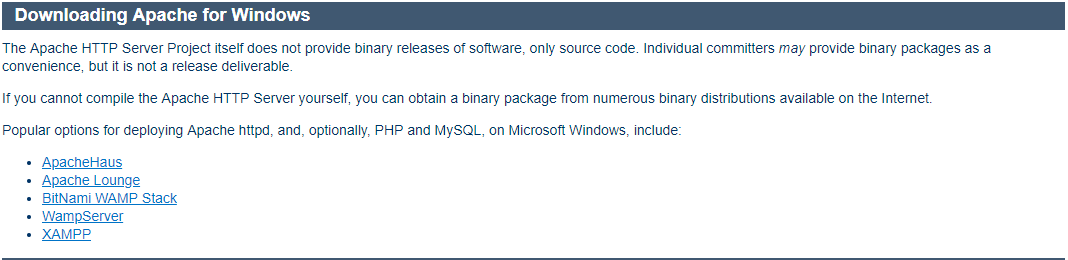
* Go to <https://httpd.apache.org/> and click on the Download link under Apache httpd 2.4.38 Released section.



* It will take you to the following page, and then click on Files for Microsoft Windows.



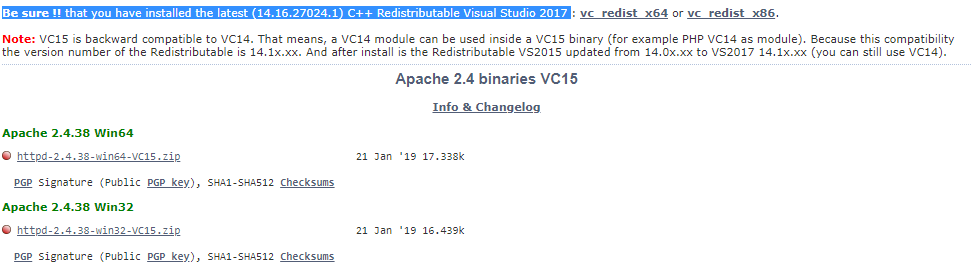
* Click on Apache Lounge.



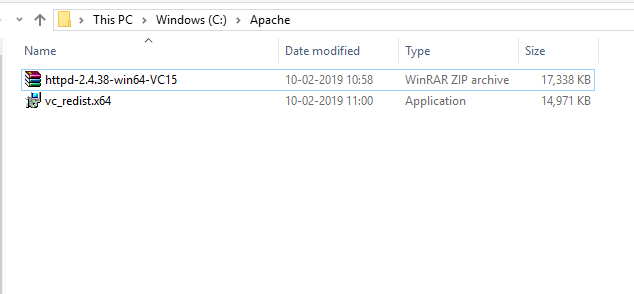
* You can download 32-bit or 64-bit of the zip file based on your windows operating system. We will download 64-bit version here. Click the corresponding .zip link to download.



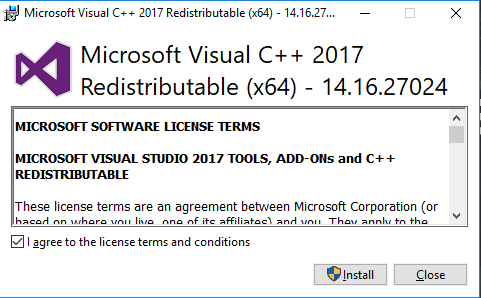
* Now, it requires C++ Redistributable Visual Studio 2017. So we will download it from the corresponding 32-bit or 64-bit link



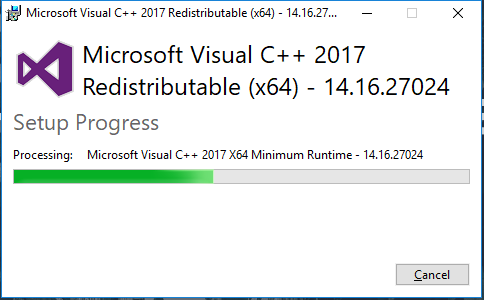
* After both the files has been download, we will go the downloaded location and install C++ Redistributable Visual Studio 2017 first. Double click on the .exe file.



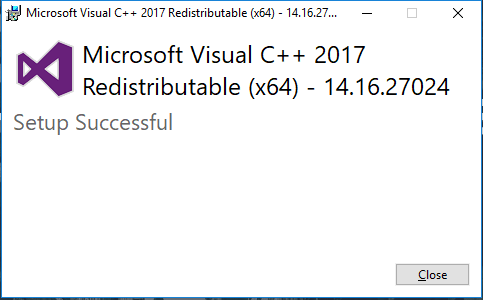
* Check ‘I agree’ and click Install.



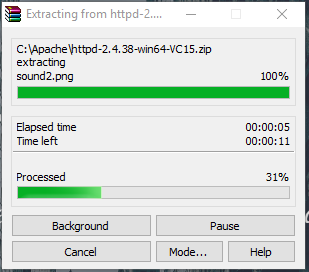
* Installation in progress.



* Once, its complete, you will get a message like this. Click Close to finish the installation.

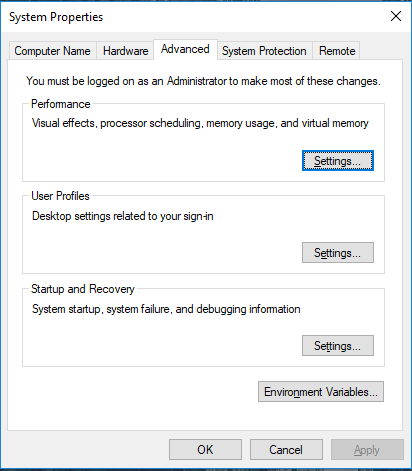


* Now, go to the folder where you download the Apache zip file. Right click on it and select extract here.

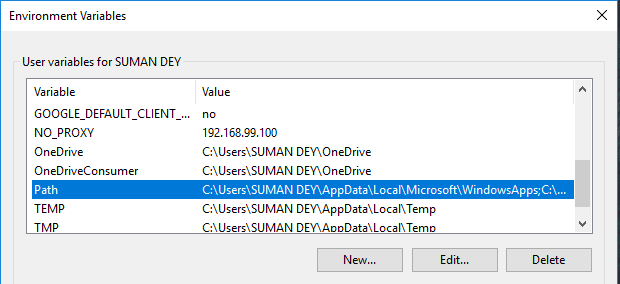


* Now, we will have an Apache24 folder created. Copy this folder to C drive, and then we will add path to system environment variables.

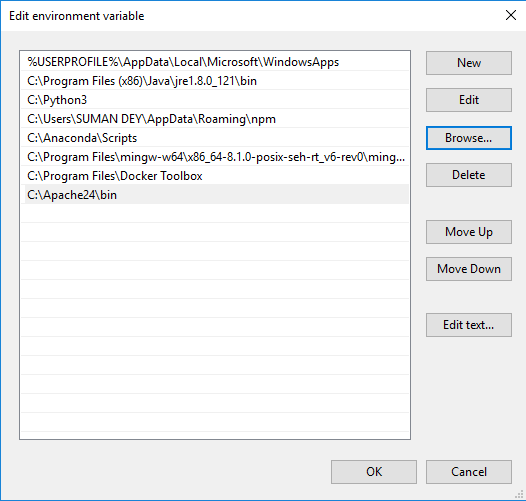
Go to System Properties -> Advanced tab -> Click on Environment Variables button below.



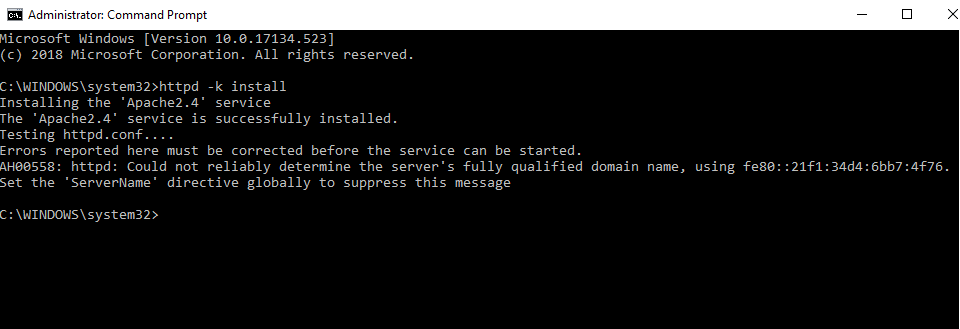
* In Variables, find Path and click Edit.



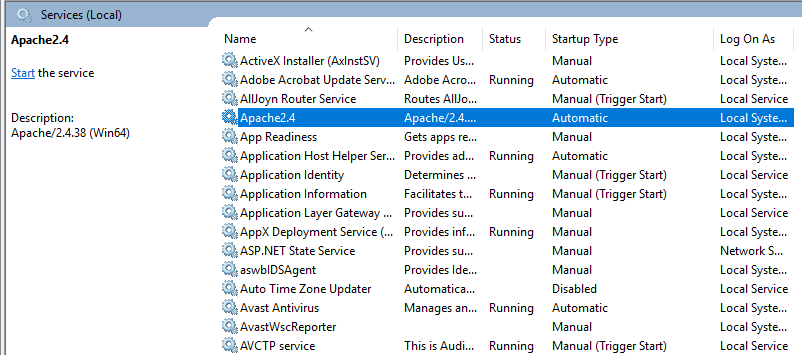
* Click Browse -> Go to C drive Apache24 folder -> Select bin folder -> Click Ok.



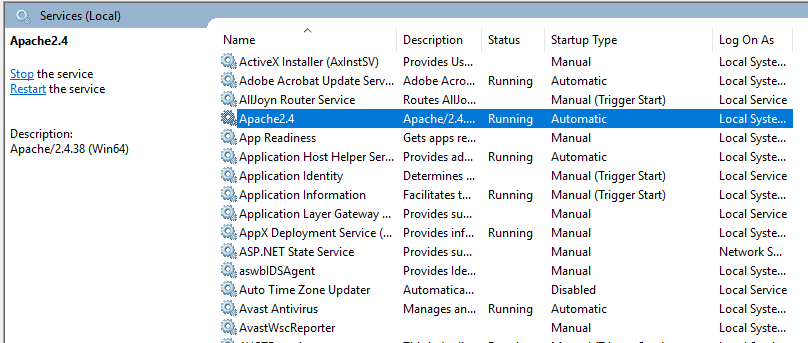
* We will install Apache as a Windows Service. Run Command Prompt as an administrator. Type httpd –k install and hit enter.



* We will check Apache service. Click on Windows icon and type services. Click on the Services app and find service with the name Apache24.



* To start the Apache server, right click on it and click start. The status will change to ‘Running’.



* We can test with browser. Open browser and navigate to <http://localhost> and hit enter. A message stating ‘It works!’ will pop up to confirm successful installation.

