**Ans 1:- Hadoop in layman's term**

Hadoop grew out of a paper on MapReduce and GFS from Google, with the majority of its development coming from the efforts of Doug Cutting and Yahoo in the mid 2000s. At that time, the only companies capable of working with data at that scale are the Silicon Valley giants, when the rest of the world was still mostly dependent on SQL solutions In 2006, Hadoop was released as an open source project through Apache, as a result, it grew incredibly quickly in terms of related projects, ecosystem, and adoption.

Hadoop is a cornerstone to commercial data offerings today, mainly due to its huge community and ecosystem, as well as the ease of adoption. As a company, you can spend relatively little to store the same amount of data using Hadoop in comparison to traditional data warehouses. Here at [Vertafore](http://www.vertafore.com/), we found our cost of storage is about 1/5 per gigabyte after migrating to Hadoop than it was on a SQL data warehouse for our particular use case. Hadoop is also virtually infinitely scaleable both in terms of storage and performance, so there is almost no big concerns about future challenges when adopting Hadoop.

These days, Hadoop has become a go-to solution for big data in many different implementations. For commercial businesses, Hadoop can be an analytics platform to drive business decisions. For instance, vast amounts of transactional historic and real-time customer data can be stored and processed to determine benchmark certain markets in the New England area. Many of these consumers are analysts, business intelligence, industrial engineers, etc. Hadoop is slightly different for tech companies. Companies like LinkedIn, Netflix, etc also use it to store and process data in the terabytes/petabytes range. However, this data is often then re-consumed by other processes to build software features, such as Netflix’s recommendation system, or LinkedIn’s “People you may know” feature. From my experience, the two most popular use cases are for Business Analytics (tracking historical data to determine customer behavior to improve business processes), and as a distributed engine to process large amounts of data, often for tech companies (i.e. ingesting data from dozens of products in real-time, then using that to feed other software features). There’s no reason to limit Hadoop to these two cases, as it has found adoption in things like modelling stock markets, parallel computing, distributed file storage, and so much more. This gave rise to a lot of complementary projects in the ecosystem like Spark, HBase, Kafka/Storm, etc.

As for Amazon, Facebook, Google, etc., they’re not exactly using Hadoop like most of the world. Although there are many processes that run on MapReduce/HDFS inside those organizations, Amazon/FB/Google also has a lot of proprietary and cutting-edge technologies that has supplemented or replaced many components of Hadoop. For instance, Facebook uses Presto/RocksDB, Google has GFS/Beam. The idea is still the same: within your products, track everything, so that you can determine user behavior. For someone like Google/Facebook/Amazon, they know what pages you like to view, what products you like to by, what people you like to follow. All of this is critical to building products that you as a consumer like to use. Google and Facebook also offer Hadoop or Hadoop-alternatives as an infrastructure-as-a-service model (AWS, Google Data Platform), but that’s slightly unrelated to the question you’re asking, I think.

What I described above is only the tip of the iceberg, these days data has moved away from being reliant on Hadoop. It’s better to think of it as a tool, as Hadoop =/= Big Data anymore. It simply is the most popular offering for any company to begin diving into Big Data. Remember, knowledge is power, and in 2016, data is knowledge. Take everything I wrote with a grain of salt, this is simply my perception of the current state of the world.

**Ans 2 :- Explain the components of Hadoop framework**

Apache Hadoop is an open-source software framework written in Java. It is primarily used for storage and processing of large sets of data, better known as big data. It comprises of several components that allow the storage and processing of large data volumes in a clustered environment. However, the two main components are Hadoop Distributed File System and MapReduce programming.

In this article, we will first take a look at the components that make up Apache Hadoop and then some of the integrated systems and databases.

**1. Components of Apache Hadoop**

Hadoop, as a whole, consists of the following parts:

***Hadoop Distributed File System*** – Abbreviated as HDFS, it is primarily a file system similar to many of the already existing ones. However, it is also a virtual file system.

There is one notable difference with other popular file systems, which is, when we move a file in HDFS, it is automatically split into smaller files. These smaller files are then replicated on a minimum of three different servers, so that they can be used as an alternative to unforeseen circumstances. This replication count isn’t necessarily hard-set, and can be decided upon as per requirements.

***Hadoop MapReduce*** – MapReduce is mainly the programming aspect of Hadoop that allows processing of large volumes of data.

There is also a provision that breaks down requests into smaller requests, which are then sent to multiple servers. This allows utilization of the scalable power of the CPU.

***HBASE*** – HBASE happens to be a layer that sits atop the HDFS and has been developed by means of the Java programming language. HBASE primarily has the following aspects –

* Non relational
* Highly scalable
* Fault tolerance

Every single row that exists in HBASE is identified by means of a key. The number of columns is also not defined, but rather grouped into column families.

***Zookeeper*** – This is basically a centralized system that maintains –

* Configuration information
* Naming information
* Synchronization information

Besides these, Zookeeper is also responsible for group services and is utilized by HBASE. It also comes to use for MapReduce programs.

***Solr/Lucene*** – This is nothing but a search engine. Its libraries are developed by Apache and required over 10 years to be developed in its present robust form.

***Programming Languages*** – There are basically two programming languages that are identified as original Hadoop programming languages,

* Hive
* PIG

Besides these, there are a few other programming languages that can be used for writing programs, namely C, JAQL and Java. We can also make direct usage of SQL for interaction with the database, although that requires the use of standard JDBC or ODBC drivers.

**2. Systems for integrated Hadoop operations**

Most enterprise vendors have their very own Hadoop products that comprise of database as well as analytical offerings. These offerings also do not require you to source Hadoop from elsewhere, but rather provide it as a core aspect of their solutions.

Some of these are –

***EMC Greenplum***

Greenplum happens to be a pretty new entrant in the enterprise business and has a reputation for being a strong provider of analytics. It comes as a Unified Analytics Platform, which consists of –

* Greenplum database – meant for use on structured data
* Greenplum HD – Its Hadoop distribution
* Chorus – A productivity layer for Data Science teams.

***IBM***

IBM’s enterprise distribution for Hadoop is known as Infosphere BigInsights. It implements an array of features for Hadoop, such as –

* Tools for management
* Tools for administration
* It also comprises of a textual data analysis tools that help in the resolution of entities, such as identifying people, phone numbers, addresses and more.

By making use of the JAQL query language, one can integrate Hadoop with various IBM products like DB2, or even Netezza. BigSheets, a spreadsheet like application working on big data is also offered. At present, BigInsights can only be used over cloud by means of Amazon, Rackspace, Rightscale, etc.

***Microsoft***

Hadoop forms the core part of Microsoft’s big data offering. Pursuing an integrated approach, it plans to make available big data over its tool suite for analytics.

Microsoft Big Data Solutions have been brought into the Windows Server platform and also to the Windows Azure platform, which is cloud-based. Integrated with Windows Systems Center and Active Directory, the company has its own distribution format of Hadoop. Further, it integrates Hadoop with its SQL Server, Visual Studio, and .NET.

***Oracle***

Oracle entered into the world of big data with an appliance based approach in the form of Big Data Appliance. This ensures easy Hadoop integration, and comes along with the new NoSQL database, which allows for analytics and also has connections to Oracle databases and the Exadata warehousing lineup. NoSQL is also known as a scalable key value-based database offering.

Oracle also happens to have the R analytical platform integrated with Hadoop, and that makes it easy to ship. Oracle’s R Enterprise product is also one that allows easy database integration, and also with Hadoop.

**Ans 3 :- Reasons to learn Big data technologies**

We are living in the age of information. The internet has made it easy for anyone to gather whatever information they need to achieve whatever end they desire. Did you know that you could even learn how to build your own car online? Did you know that you could teach yourself how to run your own Fortune 500 Company? Yes, what was only available to those who could afford Ivy League education is now available to anyone with an internet connection. The question is, how will you use this data to your benefit?

***Reason why you should learn data analytics***

**1. Data analytics is now a priority for top organizations**

With market competition stiffening, top organizations are turning to data analytics to identify new market opportunities for their services and products. As things stand today, 77% of top organizations consider data analytics a critical component of business performance. What this means is that big data professionals have a huge influence on company policies and marketing strategies.

**2. Increasing job opportunities**

As companies begin to realize that they have no capacity to comprehensively gather, interpret and use data, they are beginning to look for specialists who can do so. If you look at all the major job opportunity platforms such as Indeed and Dice, you will see that there are increasing number of job postings looking for data analysts and consultants. The demand for professionals with this particular skill set is on the rise while the supply remains low. This creates great job opportunities for individuals within this field.

**3. Increasing pay for data analytics professionals**

As the demand steadily increases and the supply remains low, data analytics professionals are getting paid more and more. In India, as it stands today, data analytics professionals are paid on average 50% more than their counterparts in other IT based professions. This trend is evident across the globe as more and more companies realize just how important these professionals are to the organization.

**4. Big data analytics is everywhere**

Just as it has become imperative to use computers in today’s workplace, the use data analytics professionals to foster growth is slowly catching on. There is practically no sector which has remained untouched from the reach of Data Analytics.

**5. You will have various job titles from which to choose**

A data analytics professional has a broad range of job titles and fields from which to choose. Since big data is used almost everywhere today, you can choose to be a:

- Metrics and Analytics Specialist

- Data Analyst

- Big Data Engineer

- Data Analytics Consultant

These are just some of the job titles you could hold in big organizations such as IBM, ITrend, Opera, Oracle, etc and the possibilities are immense.

**6. You will be at the core of decision-making in the company**

One of the leading causes of job place dissatisfaction is that most employees feel as if they do not have any decision-making power. They often feel like just another cog in the great corporate wheel. As a data analytics professional, you will be at the core of decision-making in your chosen company. In fact, you will be an integral part of business decisions and future strategies, thus giving you an important role and purpose within the organization.

**7. Adoption rate of big data analytics is high**

Just as companies started turning to social media for brand advertising and customer engagement, they have also started turning to data analytics as well. Today, it is almost impossible to find any brand that does not have social media presence. The same will be true as far as data analytics adaptation is concerned. In the very near future, every company will need [*data analytics professionals*](http://www.simplilearn.com/resources-to-learn-data-science-online-article)*.* This makes it a wise career move that actually has a future in business.

**8. Data analytics is taking over faster than expected**

A survey conducted by Nimbus Ninety shows that data analytics is taking over much faster than was projected. The survey found that data analytics tops the list of technologies to look out for in the coming 3 years.

**9. It represents perfect freelancing opportunities**

In the near future, a vast majority of the workforce will not want to be tied to just one employer. People are steadily looking for ways to diversify their sources of income and methods through which they can find the perfect work-life balance. Data analytics, being that it is a matter of studying numbers, trends, and data in general, gives you the perfect opportunity to become a well-paid freelancer or consultant for some of the biggest firms in the world. Majorly IT based, this kind of job can be done from anywhere in the world at any given time. Therefore, you do not have to be tied to a desk.

**10. Develop new revenue streams**

With your ability to analyze and put good data information to good use, you can easily identify new and unexploited streams of revenue generation. This is one of the best ways to enrich your life by increasing your earnings.