List the Components of Hadoop 2.x and explain each component in detail.

Below are the components of Hadoop 2.x

**HDFS :** Hadoop distributed file system (HDFS) is primary storage system used by Hadoop. It is run on commodity hardware. It is provide high performance access to data across Hadoop cluster.

HDFS uses master/slave architecture with each cluster consisting of a single NameNode that manages file system operations and supporting DataNode that manages data Storage on individual compute nodes.

When HDFS takes in data, it breaks the information down into separate pieces and distributes them to different node in a cluster allowing for parallel processing.

**MapReduce**: Basically Mapreduce is part of Hadoop 1.x. It is a [programming model](https://en.wikipedia.org/wiki/Programming_model) and an associated implementation for processing and generating [big data](https://en.wikipedia.org/wiki/Big_data) sets with a [parallel](https://en.wikipedia.org/wiki/Parallel_computing), [distributed](https://en.wikipedia.org/wiki/Distributed_computing) algorithm on a [cluster](https://en.wikipedia.org/wiki/Cluster_(computing)).

A MapReduce program is composed of a

* [Map()](https://en.wikipedia.org/wiki/Map_(parallel_pattern)) [procedure](https://en.wikipedia.org/wiki/Procedure_(computing)) (method) that performs filtering and sorting
* Reduce() method that performs a summary operation

The MapReduce framework operates on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types

Input Data Reduce() Output data

Map

**Yarn:** Apache Hadoop YARN (Yet Another Resource Negotiator) is a cluster management technology. It is also called as MapReduce 2.0. The fundamental idea of YARN is to built up the functionalities of resource management and job scheduling /monitoring into separate daemons. The idea is to have a global resource management (RM) and per allocation Application Master (AM).

The Resource Manager has two main components: Scheduler and Applications Manager. The Scheduler is responsible for allocating resources to the various running applications subject to familiar constraints of capacities , queue etc. The Applications Manager is responsible for accepting job-submissions, negotiating the first container for executing the application specific Application Master and provides the service for restarting the Application Master container on failure.

**Pig:** Apache Pig is a high-level language platform developed to execute queries on huge datasets that are stored in HDFS using Apache Hadoop. It is similar to SQL query language but applied on a larger dataset and with additional features. The language used in Pig is called PigLatin. It is very similar to SQL. It is used to load the data, apply the required filters and dump the data in the required format. It requires a Java runtime environment to execute the programs. Pig converts all the operations into Map and Reduce tasks which can be efficiently processed on Hadoop. It basically allows us to concentrate upon the whole operation irrespective of the individual mapper and reducer functions.

**Hive :** Hive is a data warehousing solution developed on top of Hadoop to meet the big data challenges of storing, managing and processing large data sets without having to write complex Java based MapReduce programs. Hive is a familiar programming model for big data professionals who know SQL. Hive is not a relational database or an architecture for online transaction processing. It is particularly designed for online analytical processing systems (OLAP).

Hive compiler converts the queries written in HiveQL into MapReduce jobs so that Hadoop developers need not worry much about the complex programming code beyond the processing and they can focus on the business problem. The three important functions performed by Hive include - data summarization, data querying and data analysis. Apache Hive is extensively used by data scientists and data analysts for data exploration, building data pipelines and for processing ad-hoc queries.

**Sqoop :** Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.

**Flume:** Flume is a standard, simple, robust, flexible, and extensible tool for data ingestion from various data producers (webservers) into Hadoop.

**Oozie:** Apache Oozie is a workflow scheduler for Hadoop. It is a system which runs workflow of dependent jobs. Here, users are permitted to create **Directed Acyclic Graphs** of workflows, which can be run in parallel and sequentially in Hadoop.