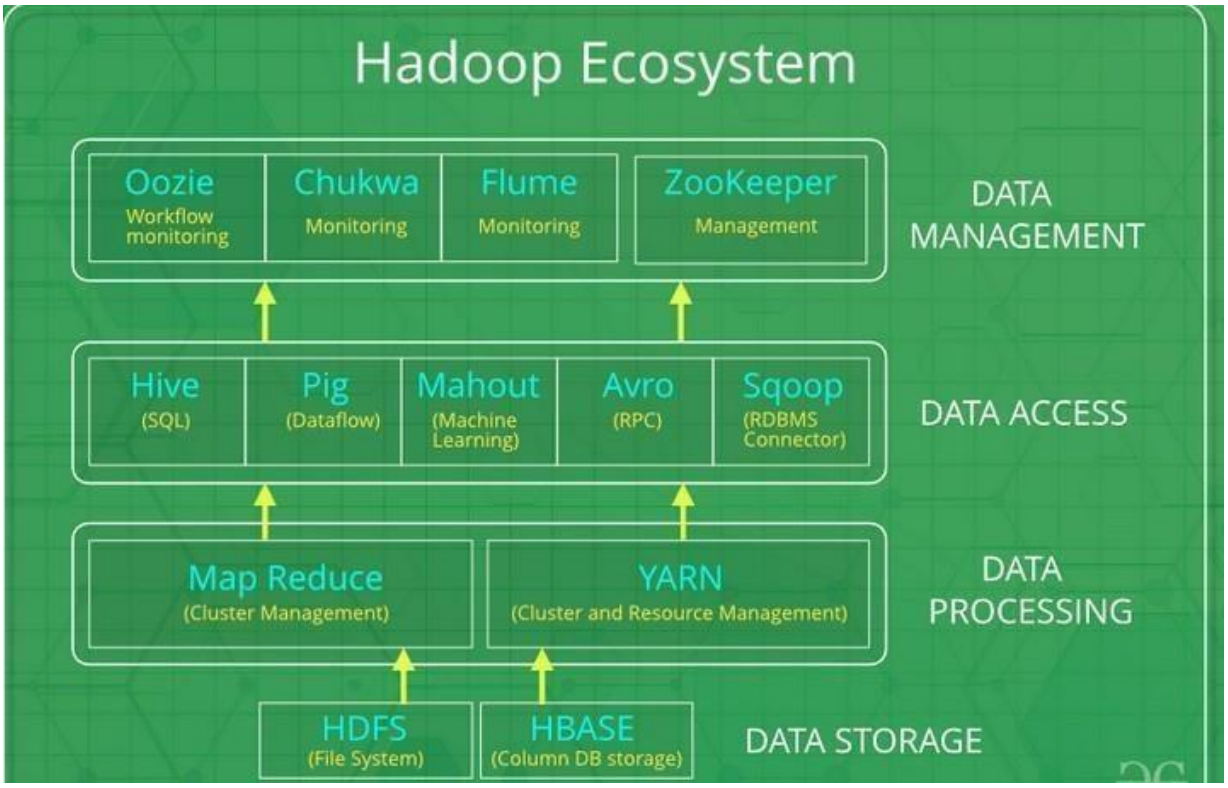
1) What are the different ecosystem components of Hadoop with diagram.



Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are *four major elements of Hadoop* i.e. **HDFS, MapReduce, YARN, and Hadoop Common**. Most of the tools or solutions are used to supplement or support these major elements. All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data etc.

2) Categorize hadoop ecosystem components based on its functionality. A)

Data Storage:

1) HDFS: HDFS is the primary or major component of Hadoop ecosystem and is responsible for storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in the form of log files.

- HDFS consists of two core components i.e.

1. Name node
2. Data Node

2) Hbase: It's a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively.

B) Data Processing:

1) MapReduce:

By making the use of distributed and parallel algorithms, MapReduce makes it possible to carry over the processing's logic and helps to write applications which transform big data sets into a manageable one.

- MapReduce makes the use of two functions i.e. Map() and Reduce() whose task is:
    1. *Map()* Map generates a key-value pair based result which is later on processed by the Reduce() method.
    2. *Reduce()*, Reduce() takes the output generated by Map() as input and combines those tuples into smaller set of tuples.

2) YARN: Yet Another Resource Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System.

C) Data Access

1) Hive: With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).

2) Mahout: Mahout, allows Machine Learnability to a system or application. Machine Learning, as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or om the basis of algorithms.

3) Pig: Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL. It is a platform for structuring the data flow, processing and analyzing huge data sets.

4) Avro: Avro is a row-oriented remote procedure call and data serialization framework developed within Apache's Hadoop project.

5) Sqoop: Sqoop is a command-line interface application for transferring data between relational tables and Hadoop. 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.

D) Data Management

1) Zookeeper: There was a huge issue of management of coordination and synchronization among the resources or the components of Hadoop which resulted in inconsistency. Zookeeper overcame all the problems by performing synchronization, inter-component based communication, grouping, and maintenance.

2) Oozie: Oozie simply performs the task of a scheduler, thus scheduling jobs and binding them together as a single unit.

3) Chukwa: Chukwa is an open source data collection system for monitoring large distributed systems.

4) Flume : Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.