



Pune District Education Association's  
**College Of Engineering**

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



DSBDAL

Assignment No: 15

- \* Aim:- Write a case study to process data driven for digital marketing or health care system with hadoop ecosystem components as shown:

HDFS: Hadoop distributed file system.

YARN: Yet another resource negotiator map-Reduce programming based data processing

spark: In memory data processing MG,  
HIVE query based processing of data services.

- \* Theory:- Apache hadoop is an open source framework intended to make interaction with big data easier.

-Hadoop Ecosystem is a platform or a suite which provides various sources to solve the big data problems.

-HDFS:- Hadoop distributed file system.

HDFS is the primary or major components 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.



- 1) Name Node
- 2) Data Node

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 the scheduling & resource allocation for the hadoop system.

- consists of three major components :-

- 1) Resource manager
- 2) Nodes manager
- 3) Application manager

MapReduce -

By making the use of distributed & parallel algorithms. MapReduce make it positive to carry over the processing logic & helps to write applications which transform big data sets into a manageable one.

mapReduce makes the use of the two functions :-

- 1) mapReduce()
- 2) Reduce()

Apache spark :-



# Pune District Education Association's **College Of Engineering**

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



It is platform that handles all the process interactive or iterative real time processing graph conversion in and visualization.

spark is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing.

PIG -

PIG was basically developed by yahoo which works on a big latin language, which is queen based language similar to SQL.

PIG does not work of executing the command & in the background all the activities of map.

Pune District Education Association

HIVE -

With the help of SQL methodology & interface HIVE performs reading & writing of large data sets.

It is highly scalable as it allows realtime processing & batch processing both.

All sql datatype are supported by the HIVE thus making the query processing easier.



\* Conclusion:-

I studied the case study to process data driven for digital marketing or health care systems with the hadoop ecosystem components.

