

DS & BDA Lab

Group A - Experiment : 1

TITLE : Single node / Multiple node  
Hadoop Installation.

#### OBJECTIVE :

1. To learn the concepts of Hadoop & Hadoop framework for Big Data.
2. To install & configure Hadoop.

#### SOFTWARE REQUIREMENTS :

1. Ubuntu 14.04 / 14.10
2. Java

#### THEORY :

##### Introduction :-

Hadoop is an open-source framework that allows to store & process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation & storage.

Due to the advent of new technologies, devices & communication like social networking sites, the amount of data produced by mankind is growing rapidly every year. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. The same amount was created in every two days in 2011 & in every ten minutes in 2013.



This rate is still growing enormously. Though all this information produced is meaningful & can be useful when processed, it is being neglected.

Big Data :

Big Data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques & frameworks. Big data involves the data produced by different devices & applications. Given below are some of the fields that come under the umbrella of Big Data.

Hadoop :

Doug Cutting, Mike Cafarella & Kam took the solution provided by Google & started an open source project called HADOOP in 2005 & Doug named it after his son's toy elephant. Now Apache Hadoop is a registered trademark of the Apache Software Foundation.

Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers & they could perform complete statistical analysis for a huge amount of data.



Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage & computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation & storage.



## \* Hadoop Architecture :

Hadoop framework includes following four modules:

**Hadoop Common :** These are java libraries & utilities required by other Hadoop modules. These libraries provide filesystem & OS level abstractions & contains the necessary java files & scripts required to start Hadoop.

**Hadoop Yarn :** This is a framework for job scheduling & cluster resource management.

**Hadoop Distributed File System (HDFS™) :** A distributed file system that provides high throughput access to application data.

**Hadoop MapReduce :** This is YARN-based system for parallel processing of large datasets.

## \* Hadoop Web Interfaces :

Let's start the Hadoop again & see its web UI:

Accessing HADOOP through browser :

<http://localhost:50070/>

Verify all applications for clusters :

<http://localhost:8088/>

## CONCLUSION :

We studied installation of Hadoop installation & configuration.



DS & BDA Lab

Group B - Experiment 4

TITLE: Perform the operations using Python  
on the Facebook metrics data sets.

AIM: Perform the following operations using Python  
PROBLEM on the Facebook metrics data sets:

STATEMENT: (a) Create data subsets (b) Merge Data  
(c) sort Data (d) Transposing Data  
(e) shape & reshape Data.

OBJECTIVE :

1. To understand & apply the Analytical concept of Big data using Python.
2. To study detailed concept Python.

SOFTWARE REQUIREMENTS :

1. Ubuntu 14.04 / 14.10
2. GNU C Compiler
3. Hadoop
4. Java
5. Python Platform.

THEORY :

- \* Python is an easy to learn, powerful programming language. It has efficient high-level data structures & a simple but effective approach to object-oriented programming. Python's elegant syntax & dynamic typing, together with its interpreted nature, make it an ideal language for



scripting & rapid application development in many areas on most platforms.

The Python interpreter & the extensive standard library are freely available in source or binary form for all major platforms from the Python web site, <https://www.python.org/>, & may be freely distributed. The same site also contains distributions of & pointers to many free third-party Python modules, programs & tools, & additional documentation.

#### \* Features of Python :

Python is a dynamic, high level, free open source & interpreted programming language.

- i) Easy to code
- ii) Object-oriented language.
- iii) Free & open source
- iv) GUI Programming support
- v) High level language
- vi) Extensible feature
- vii) Python is portable.

#### \* Data Reshaping :

Data Reshaping is about changing the way data is organised into rows & columns. Most of the time data processing in python is done by taking the input data as data frame. It is easy to extract data from rows & columns of data frame but there are situations when we need the data frame in a format that is different from format in which we received it. Python has many functions to split, merge &



& change the rows to columns & vice versa in a data frame.

### 1) Joining Columns & Rows in a Data Frame:

We can join multiple vectors using `cbind()` function.

We can merge using `rbind()` function.

### \* Merging Data frames:

- We can merge dataframes by using `merge()` function.

- We can add columns using `merge` & rows using `rbind` functions.

### 2) Subsets:

subsetting Vectors, Matrices & Data Frames.

Return subsets of vectors, matrices or data frames which meet conditions.

Arguments:-

x : object to be subsetted.

subset : logical expression indicating elements or rows to keep; missing values are taken as false.

select : expression, indicating columns to select from a data frame.

drop : passed on to indexing operators.

### 3) Melting & casting :-

The functions used to do this are called `melt()` & `cast()`.

Steps :- Melt the data, cast the Molten Data, Transpose using `t()` function.



#### 4) Sorting Data :

To sort a dataframe use `order()` function.

By default, sorting is ASCENDING. Prepend the sorting variable by a minus sign to indicate DESCENDING order.

CONCLUSION: Thus we have learnt how to perform the different reshape operations using python.