Practical No 6

Aim: Integrate Python and Hadoop and perform the following operations on forest fire dataset

a. Data analysis using the Map Reduce in PyHadoop

b. Data mining in Hive

Objectives: To study the Integrate Python and Hadoop

Theory:

Hadoop Streaming :

• Hadoop streaming is a Hadoop utility for running the Hadoop Map Reduce job with executable scripts such as Mapper and Reducer.

• This is similar to the pipe operation in Linux.

• With this, the text input file is printed on stream (stdin), which is provided as an input to Mapper and the output (stdout) of Mapper is provided as an input to Reducer; finally, Reducer writes the output to the HDFS directory.

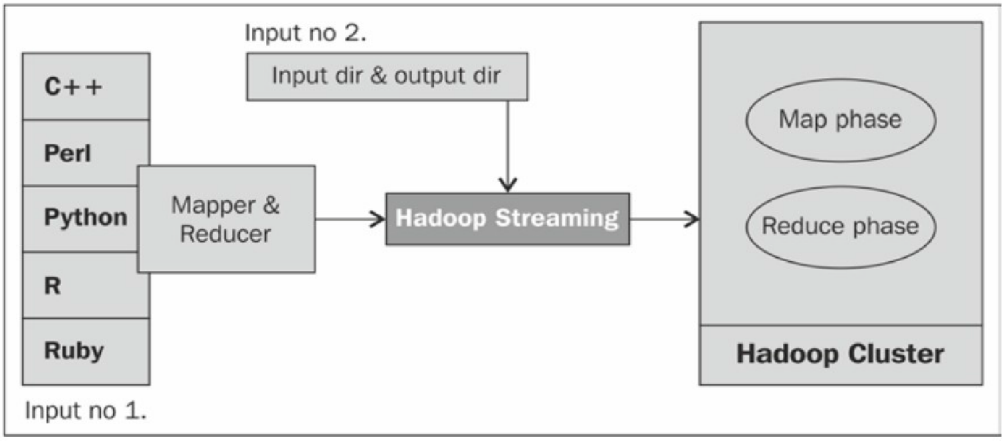
• The main advantage of the Hadoop streaming utility is that it allows Java as well as non-Java

Programmed MapReduce jobs to be executed over Hadoop clusters.

• Also, it takes care of the progress of running MapReduce jobs.

• The Hadoop streaming supports the Perl, Python, PHP, R, and C++ programming languages.

• To run an application written in other programming languages, the developer just needs to translate the application logic into the Mapper and Reducer sections with the key and value output elements.



Steps to Integrate Python with Hadoop:

Step1: Make sure that Hadoop HDFS is working correctly.

Open Terminal/Command Prompt, check if HDFS is working by using following commands:

```
start-dfs.sh
jps
```

If the output is as shown below then you are good to go.

```
14231 NameNode
14324 DataNode
14467 SecondaryNameNode
14636 Jps
```

Step2: Install libhdfs3 library

Libhdfs3 is a C/C++HDFS client which connects HDFS with python.

You cannot install it using pip, it is available on conda forge. So, make sure that anaconda or miniconda package installer is installed on your system.

I suggest you to create virtual environment first before installing.

```
conda create -n <your_env_name>
conda activate <your_env_name>
```

Use the following command to install libhdfs3 inside virtual environment

```
conda install libhdfs3
```

Step3: Install hdfs3 library

hdfs3 is small wrapper around libhdfs3. It provides user simple API to access libhdfs3 commands in a pythonic way. Use following the command to install hdfs3

```
conda install hdfs3
```

Step4: Check if connection with HDFS is successful

Make sure that HDFS is running on localhost. The default localhost port for HDFS is 9000.

Open Python command line. Enter following commands:

```
from hdfs3 import HDFileSystem
hdfs=HDFileSystem(host='localhost',port=9000)
```

If the above command works fine, then you are good to go.