Expt. No. 1 Date : 07-03-22

Experiment 1

* Aim :- To set up the necessary environment for experiments in R.

* Introduction :- The volume of data that enterprises acquire every day is increasing exponentially. It is now possible to store these vast amounts of information on low-cost platforms such as Hadoop.
The challenge these organizations now face is what to do with all this data and how to gather key insights from this data. Thus R comes into the picture. R is a very amazing tool that makes it a snap to run advanced statistical models on data, translate the derived models into colorful graphs and visualizations, and do a lot more functions related to data science.
One key drawback of R, though, is that it is not very scalable. The core R engine can process and work on a very limited amount of data. As Hadoop is very popular for Big Data processing, corresponding R with Hadoop for scalability is the next logical step.

Expt. No. Date :

Using R with Hadoop will provide an elestic data analytics platform that will scale depending on the size of the dataset to be analyzed. Hadoop's parallel processing Map/Reduce mechanism to identify pattevns in the dataset.

- Overview of R:

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zeland, and is currently developed by the R Development Core Team. This programming language was named R, based on the first letter of the first name of the two R authors, and partly a play on the name of the Bell Labs Language S. R made its first appearance in 1993. A large group of individuals has contributed to R by sending code and bug reports. Since mid-1997 there has been a core group (the "R Core Team") that can modify the R source code archive. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, windows and Mac.

Expt. No. Date :

The core of R is an interpreted computer language that allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R provides a wide variety of statistical, machine learning (linear and nonlinear modelling classic statistical tests, time-series analysis, regression, classification, clustering, recommendation, text mining) and graphical techniques, and is highly extensible. R has various built-in as well as extended functions for statistical, machine learning, and data transformation statistical analysis, predictive modelling, data visualization.

It has one of the most popular open-source statistical analysis packages available on the market today. It is cross-platform, has very vide community support, and a large and ever-growing user community that are adding new packages every day. With its growing list of package, R can now connect with other data stores, such as MySQL, SQLite, MongoDB, and Hadoop for data storage activities.

Expt. No.

Date :

• Features of R:-

The following are the important features of R:-

1. R is a well-developed, simple and effictive programming language that includes conditionals, loops, user-defined recursive functions and input and output facilities.
2. R has an effective data handling and storage facility, provides support for the relational database.
3. R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
4. R provides a large, coherent and integrated collection to tools for data analysis.
5. R provides graphical facilities for data analysis and display at the computer or printing the paper
6. R provides extensions through the vast library R packages.

• Big Data:-

Big data has to deal with larger and complex datasets that can be structured, semi-structured or unstructured and will typically not fit into memory to be processed. They have to be processed in place, which means that computation has to be done where the data resides by processing. They typically would mention the 3 Vs model of Big data, which are velocity, volume and variety.

Velocity refers to the low latency, the real time speed at which the analytics need to be applied. A typical example of this would be to perform analytics on a continuous site or aggregation of disparate sources of data.

Volume refers to the size of the dataset. It may be in KB, MB, GB, TB, or PB based on the type of application that generates or receives the data.

Variety refers to the various types of data that can exist, for example, text, audio, video, and photos.

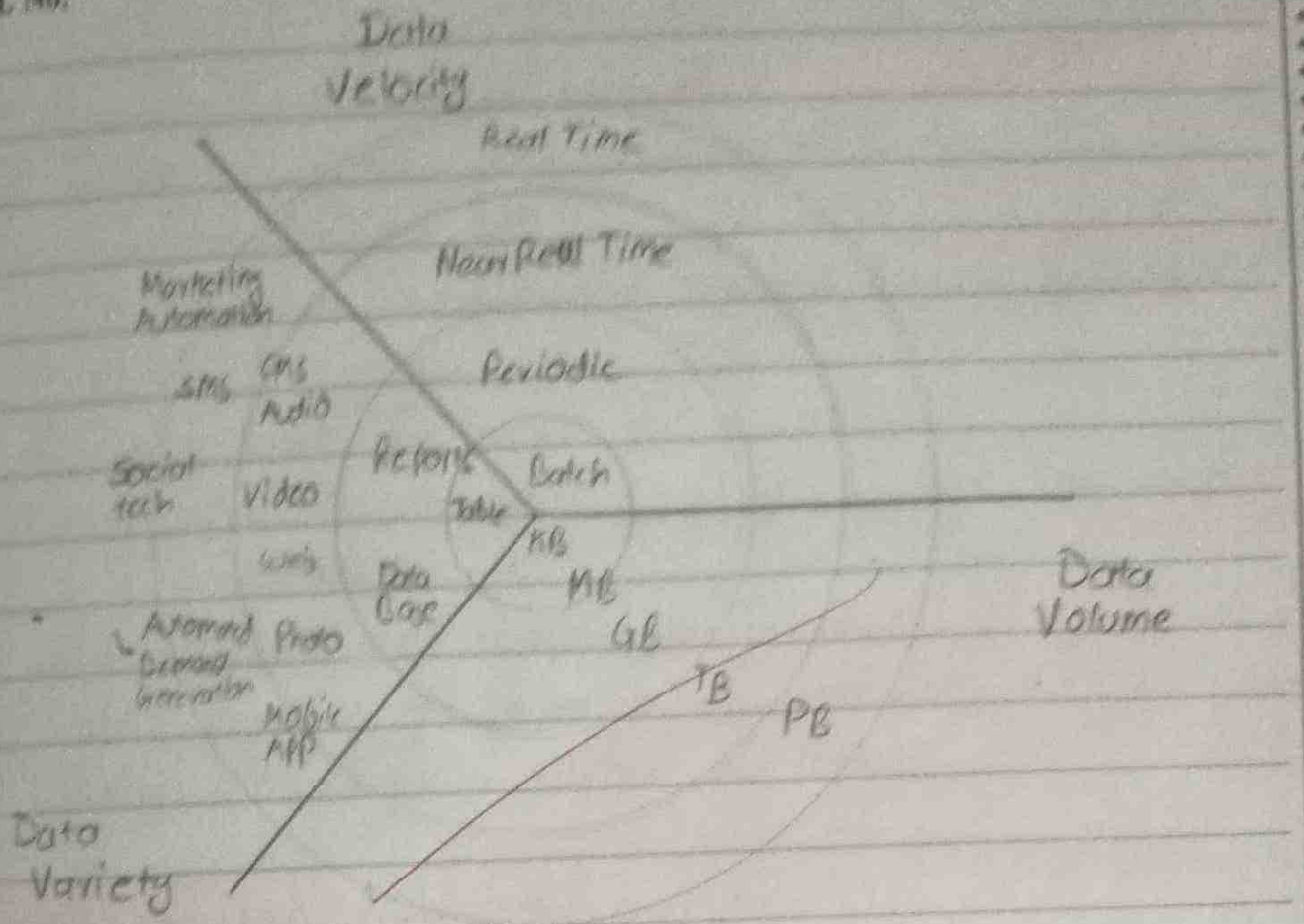
Big Data usually includes datasets with huge sizes. Such systems can't process this amount of data within the time frame mandated by the business. Big Data volumes are constantly moving target, entirely new platforms are called Big Data platforms.

Some of the popular organizations that hold Big Data are as follows:-

1. Facebook: It has 40PB of data and captures 100TB/day.
2. Yahoo!: It has 60 PB of data.
3. Twitter: It captures 8TB/day.
4. EBay: It has 40PB of data captures 50 TB/day.

Date :

Expt. No.



How much data is considered as Big Data, differs from company to company. Though true that one company's Big Data is another's small, there is something common: doesn't fit in memory, nor disk, has a rapid influx of data that needs to be processed and would benefit from distributed software software stacks. For some companies, 10 TB of data would be considered Big Data and for others, 1 PB would say that it would start in the low terabyte range.

Also a question well worth asking is, as well are we are not capturing problem now? In some scenarios with such platforms as Hadoop, it is possible to start capturing and storing all the data.

Expt. No.

Date :

- Installing windows 10 (64bit) :-

1. Check weather the device meets the windows 10 system requirements.

The minimum specifications needed to run windows 10 are:

CPU : 1 GHz or faster processor.

RAM : 1GB for windows 10 32-bit or 2GB for windows 10 64-bit

Storage : 32 GB of space or more.

GPU : DirectX 9 compatible or later with WDDM 1.0 driver.

Display: 800x600 resolution or higher.

2. Create USB installation media.

Visit Microsoft's windows 10 download page and select "Download tool now" under the "Create Windows 10 installation media" section. Transfer the downloaded installer tool to a USB drive.

3. Run the installer tool.

Open the installer tool by clicking on it. Accept Microsoft's terms, and then select "create installation media for another PC" on the "what do you want to do?" page. After selecting which language you want windows to run in, and which edition you want as well (32-bit or 62-bit), you'll be asked what type of media you want to use.

4. Use the installation media.

Insert your installation media into your device and then access the computer's BIOS or UEFI. These are the systems that allow you to control your computer's core hardware.

The process of accessing these systems is unique to each device, but the manufacturer's website should be able to give you a helping hand here. Generally you'll need to press the F2, F12 or Delete keys as your computer boots up.

5. Change the computer's boot order.

Once you have access to your computer's BIOS/UEFI you'll need to locate the settings for boot order. You need the Windows 10 installation tool to be higher up on the list than the device's current boot drive: this is the SSD or HDD. Now, when you restart your device the windows 10 installer should load up first.

6. Restart the device.

Save your settings in the BIOS/UEFI and reboot your device.

7. Complete the installation.

Your device should now load up the Windows 10 installation tool on restart. This will guide you through process.

Expt. No. Date :

- Installing R:

You can download the appropriate version by visiting the official R website.

Here are the steps provided for these different operating systems. We have considered windows, Linux, and Mac OS for R installation. Download the latest version of R as it will have all the latest patches and resolutions to the past bugs.

For windows, follow the given steps:-

1. Navigate to www.r-project.org.
2. Click on the CRAN section, select CRAN mirror, and select your windows OS (click to Linux; Hadoop is almost always used in a Linux environment).
3. Download the latest R version from the mirror.
4. Execute the downloaded .exe to install R.

For Linux - Ubuntu, follow the given steps :-

1. Navigate to www.r-project.org.
2. Click on the CRAN section, select CRAN mirror, and select your OS.
3. In the /etc/apt/sources.list file, add the CRAN <mirror> entry.
4. Download and update the package lists from the repositories using the sudo apt-get update command.
5. Install R system using the sudo apt-get install r-base command.

Expt. No. Date :

For Mac, follow the given steps:

1. Navigate to www.r-project.org.
2. Click on CRAN, select CRAN mirror and select your OS.
3. Download the following files: pkg, gfortran-*.dmg, and tcltk-*.dmg.
4. Install the R-*.pkg file.
5. Then, install the g fortran-*.dmg and tcltk-*.dmg file.

After installing the base R package, it is advisable to install Rstdio, which is a powerful and intuitive Integrated Development Environment (IDE) for R.

- Installing R stdio:-

To Install Rstdio, perform the following steps:

1. Navigate to http://www.rstdio.com/ide/download/desktop.
2. Download the latest version of Rstdio for your operating system.
3. Execute the installer file and install Rstdio.

The Rstudio organization and user community has developed a lot of R packages for graphics and visualization, such as ggplot2, plyr, shiny, Rpubs, and devtools.

* Conclusion:-

The necessary environments to perform experiments in R is ready.

21/03/22