BIG DATA ANALYTICS: A COMPARISON OF PROGRAMMING LANGUAGES

[1]Mr. V. KOVENDAN Assistant Professor, [2]Mrs. S. SAROJINI Assistant Professor
[1, 2]Department of Computer Science and Engineering,
Arasu Engineering College, Kumbakonam, Tamilnadu, India

Abstract— A large growth in volumes of data has affected today's large organization, were commonly used software tools to capture, manage, and process the data cannot able to handle the big data effectively. The main challenging task is that administrations must evaluate a bulky volume of big data and extract useful information for future actions in a short time. In this paper, we propose to visually analyze the big data using R statistical software, and we are going to focus on so called as big data statistical languages, such as Python, R, SQL, SAS. This growth invites the question whether R can ever unseat Python or Java as the top languages for big data. But while R has seen huge gains over the last few years, we show that R language is the best among all the other languages since it makes a processing of huge data easier.

Index Terms **— Python, R language, SAS, SQL**

I. INTRODUCTION

Big data analytics is a large volume of data, or big data. This big data is grouped from a inclusive variability of foundations, which includes including social webs, videos, digital images, sensors, and online transaction histories. The aim in analyzing all this data is to uncover patterns and connections that might otherwise be invisible, and that might the users who created it. So the businesses can make higher commercial resolutions. These big data permits the data researchers and other users to calculate bulky capacities of business data and other sources that traditional commercial arrangements would be unable to tackle [1]. Data analytics may results as conditions and technologies have emerged, including *NoSQL* databases and some other technologies make up an open-source application framework that's used to process giant data groups over grouped systems. Big data analytics is the application of advanced analytic techniques to verify big data sets. Big data is similar to small data,but bigger. But having data bigger consequently requires different approaches to solve new problems and old problems in a better way. The key enablers of growth of big data are increase of storage capacities,increase of processing power availability of data. In big data analytics, data statistics is one of the process [2]. By using big data modelling techniques any data problem can get simpler. In big data scenarios, one of the most successfull tool is R, is an open source programming language.

II. PYTHON

Python was created by Guido Van Rossem in 1991 and emphasis productivity and code readability. It is a flexible language that is great to do something novel and given its focus on readability and simplicity. Python structures a dynamic type system and programmed memory management and funds several software design patterns, including object-oriented, imperative, procedural and functional programming language. So we can use Python when integrated needed with other applications. As results a good programming language, python tool implement algorithms for manufacture use. It has a large and comprehensive standard library Python has packages as well Py Pi is the python packages index and consists of libraries to which users can contribute. Just like R, Python has great community but it is a bit more scattered. Most of these tools are unlikely to ever be implemented in Python, for the business of small data analysis, R is likely been main tool kit of data analysis. Python is fully featured programming language, which means it can be used to create real software's. R tends to be used for one-basic research analysis. If you don't already know R, learn Python and use RPy2 to access R functionality.

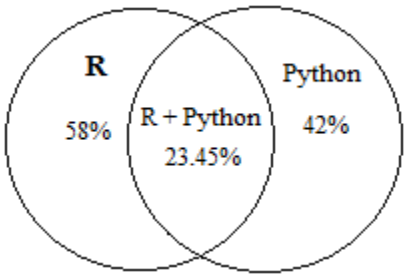

Fig1. Analysis of R and Python

Example

Program that adds two numbers

```
# two float values
val1 = 100.99
val2 = 76.15
# Adding the two given numbers
```

```
sum = float(val1) + float(val2)
# Displaying the addition result
print("The sum of given numbers is: ", sum)
```

Output:

The sum of given numbers is 177.14

Disadvantages:

- Python is slow and is not a good choice for memory intensive tasks and not supports high-graphic 3D game using Python.
- Python has limitations with database access hence they are not good for multi-processor/multi-core work.

III. STRUCTURED QUERY LANGUAGE

SQL is a language to request data from a database, to add, update, or remove data within a database, or to manipulate the metadata of the database. The Structural Query Language is a programming language specifically designed for storing and querying relational databases. SQL is used as part of data science workflow to connect to databases. R was especially designed for data analysis graphical representation.SQL was made especially for databases, they are allies. With R Services in SQL Server 2016, Microsoft enhances (and maintains complete compatibility with) open-source R by offering multi-threaded and parallel computing capabilities, as well as in-memory and disk-based data management.

There is an R package called sqldf that allows you to use SQL commands to extract data from a R data frame. [2] SQL is not case- sensitive whereas R is case-sensitive likely stores their data in a manner very similar to a relational database. To connect to SQL Server from R and several libraries we can use: RODBC, RJDBC, rsql server. R is the most traditional format; with data stored in a data frame (essentially a rectangular grid with rows and columns).R user might 'reshape' the data frame using basic R code or special packages like "dplyr".

Example

To get a women's club president list, to count their activities. In that case, the syntax is

```
#OPTION 1: using base R commands
    df$activities_sum<-rowSums(df[c(8:3)])
    presidents<-subset(df,club_president==1)
    president_acts<-presidents[c(2:3,14)]

#OPTION 2: using dplyr:
    Library(dplyr)
    president_acts2<-df%>%
    fitter(club_president==1)%
    mutate(activities_sum=march+petition+boycott+public_speech+club_member+sit_in)%>%
    select(name_first,name_second,activities_sum)
```

In both storage styles (data frames relational tables) and syntax types (R vs SQL) the user is doing the same thing: simply modifying the structure of the data grids without changing the underlying data itself.

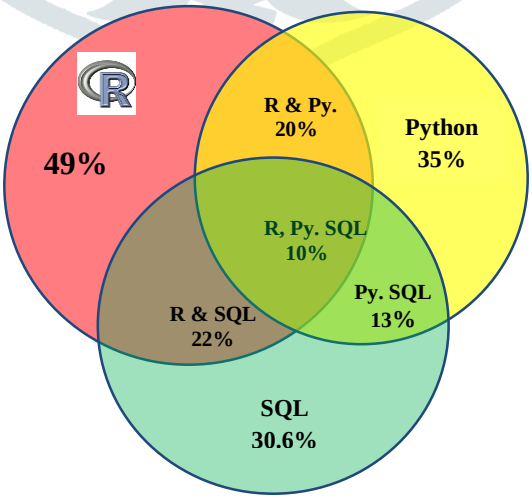

Fig.2. Analysis of R, Python &SQL

Disadvantages

- It's Got A Few Stability Issues

- It Suffers From Relatively Poor Performance Scaling
- Development Is Not Community Driven – and Hence Has Lagged
- Its Functionality Tends To Be Heavily Dependant On Addons
- Developers May Find Some Of Its Limitations To Be Frustrating
- Although the SQL databases system conforms to ANSI & ISO standards, some databases go for proprietary extensions to standard SQL to ensure vendor lock-in,
- Interfacing an SQL database is more complex.

IV. STATISTICAL ANALYSIS SYSTEM

SAS is for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics; it read input from common spreadsheets tables, graphs, and as RTF, HTML and PDF documents. It is unquestionable marketplace leader in commercial analytics space. In a comparison of all basic features for statistical software R is heads up with the best of statistical software. Statistical packages are merely to put is the open source alike of SAS, for what it's worth R can pretty much do everything SAS can't. The languages are extremely dissimilar. Since R is a true programming language it gives more flexibility and power than SAS to the programmer. R is object oriented which SAS is not. It's not like English vs Spanish it's like English vs. Mandarin. It's easier to output SAS output to Word. SAS is easy to learn and provides easy option (PROC SQL) for people who already know SQL. In terms of resources, there are tutorials available on websites of various universities and SAS has a comprehensive documentation. Even though SAS is a way to implement an analysis of big data it is not aqueous and R programming language. Compared to SAS, R is an easier statistical way to analysis the information of huge process in big data analysis and data mining techniques also uses this.

Example

Reads data organized in columns separated by spaces, from an external file and uses two PROCs.

Program:

```
DATA hatco;
Optionsls=79;
Optionsps=60;
INFILE '~dscbms/class/dsc8450/files/hatco';
   INPUT X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14;
LABEL X1='DELIVERY SPEED'
      X2='PRICE LEVEL'
      X3='PRICE FLEXIBILITY'
      X4='MANUFACTURER IMAGE'
      X5='OVERALL SERVICE'
      X6='SALES FORCE IMAGE'
      X7='PRODUCT QUALITY'
      X8='SIZE OF FIRM'
      X9='USAGE LEVEL'
      X10='SATISFACTION LEVEL'
      X11='SPECIFICATION BUYING'
      X12='STRUCTUREOF PROCUREMENT'
      X13='TYPE OF INDUSTRY'
      X14='TYPEOFBUYING SITUATION';
Proc means;
var x1-x14;
PROC UNIVARIATE PLOT NORMAL;
var x1-x7 x
```

Disadvantages

- R has advanced graphical capabilities that support various professional graphics templates.
- In a typical 10GB network environment the available bandwidth is split, creating multiple logical networks- the overall costs of switches and routers to achieve this is significantly more expensive.
- Expensive tool and not open to public, Huge kloc, text mining - Need to purchase SAS Enterprise / Text Miner.

V. R PROGRAMMING LANGUAGE

R was created by Ross Ihaka and Robert Gentle Man at the university of Auckland. R is for statistical analysis graphics representation & reporting. The features of R Fig.3 is designed for statistical analytics with abundant mathematics function library, suitable for complex mathematical algorism programmers need strong expertise and knowledge to learn. R relies on other systems to save the analyzed data. If a new methodology developed that R almost results immediately. Every language generally has a guiding principle, R is built for statistical computing and freely available under the GNU Genaeral Public Lincence and pre-compiled binary versions are provided for various operating systems like LINUX, WINDOWS and MAC. Many of the users

think R as a statistics system. In R, the fundamental unit of sharable code is the packages. R can be extended via packages, there are about eight packages supplied with the R distribution and many more are available through the CRAN family of internet sites covering a very wide range of modern statistics.

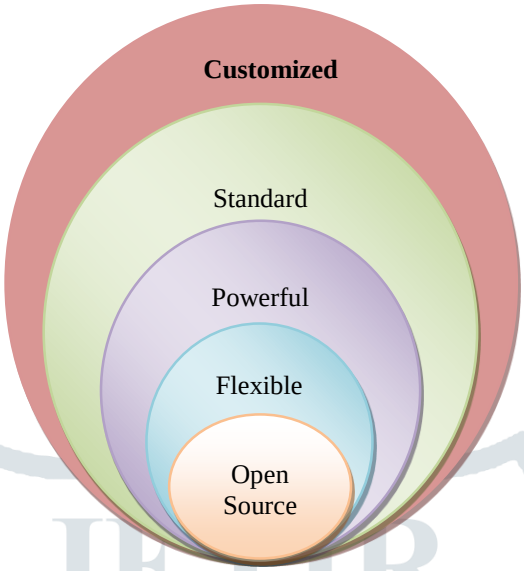

Fig.3. Feautures of R language

Example

Save/Load R Data

Data in R can be saved as R data files with functions save. Then they can be loaded into R with load.

Save multiple variables

```
# save multiple variables (objects)
x = 23
y = 14
save (x, y, file="data_xy.RData")

# load data into new R environment "ev" to avoid overwriting variables having the same name
x = 44
load("data_xy.RData", envir = (ev = new.env()) )
ev$x
 [1] 23
ev$y
 [1] 14
x
 [1] 44

# load data (without new environment): overwrites all variables in R that have the same name
x = 44 # will be overwritten by identical variable "x" in data_xy.RData
load("data_xy.RData")
x
 [1] 23
y
 [1] 14
```

Save all variables

```
# save all variables (data objects)
save.image(file = "data_all.RData")
q()  # exit R

# load the saved R variables (same as above)
load("data_all.RData")

Get list of saved R data files
# list all .RData files in current directory (path=".")
```

```
list.files(path = '.', pattern=".RData")
 [1] "data_all.RData" "data_xy.RData"
```

Advantages of R

- Open source to use and easy to modify it.
- R has no license restrictions and R has over 4800 packages available from multiple repositories.
- R has excellent graphical capabilities which can able to run on different operating system and also different hardware (GNU/LINUX).
- R Language which includes conditionals, loops, and user defined recursive function and input, output facilities.
- It has an effective data handling and storage facilities.

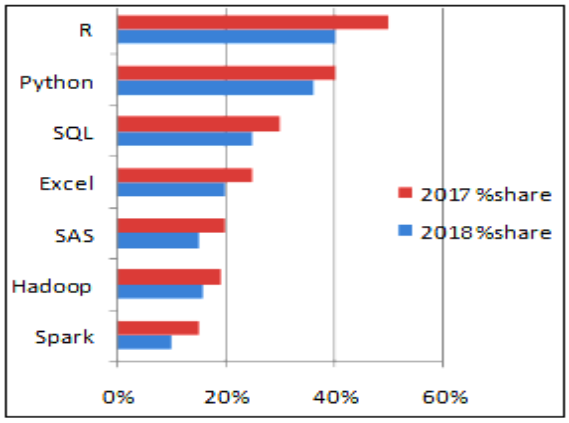

Fig.4. Performance of the R

Table5.1: Comparison of Programming Languages

CHARACTERISTICS	PYTHON	SQL	SAS	R
RELIABLE	L	A	A	E
COST	FREE	PAID	FREE	FREE
LEARNING DIFFICULTY	E	A	A	L
DATA MANIPULATION	L	L	G	A
ANALYTIC MODELING	A	L	A	G
GRAPHICAL CAPABILITY	L	L	G	G
TEXT PROCESSING	L	A	A	G
BIG DATA	A	L	G	E

Note:

L- Low, A- Average, G- Good, E- Excellent

VI. CONCLUSION

To create a powerful and reliable statistical model, data transformation, evaluation of multiple model option and visualizing the results are the most essential components for big data. This is the reason why the R programming language has proven so popular among all other programming languages for big data analytics. The R language gives the speed that allows data scientist to repeat through this process quickly. We had analyzed that the R language is a widely used programming language in big data analysis. Thus we concluded by comparing to analysis a huge amount of data was handled in easier method by using the Language R.

REFERENCES

[1] R Big R: Large-Scale Analytics on Hadoop Using R Oscar D. Lara Yejas ; Weiqiang Zhuang ; Adarsh Pannu 2014 IEEE International Congress on Big Data

[2] R-tool: Data analytic framework for big data Ayushi Malviya ; Amit Udhani ; Suryakant Soni 2016 Symposium on Colossal Data Analysis and Networking (CDAN)

[3] Web-based collaborative big data analytics on big data as a service platform Kyounghyun Park ; Minh Chau Nguyen ; Heesun Won 2015 17th International Conference on Advanced Communication Technology (ICACT)

[4] RABID -- A General Distributed R Processing Framework Targeting Large Data-Set Problems Hao Lin ; Shuo Yang ; Samuel P. Midkiff 2013 IEEE International Congress on Big Data

[5]Big R: Large-Scale Analytics on Hadoop Using R Oscar D. Lara Yejas ; Weiqiang Zhuang ; Adarsh Pannu 2014 IEEE International Congress on Big Data

[6] Fast approach for automatic data retrieval using R programming language Tran Duc Chung ; Rosdiazli Ibrahim ; Sabo Miya Hassan ; Nurfatihah Syalwiah Rosli, 2016 2nd IEEE International Symposium on Robotics and Manufacturing Automation (ROMA)

[7]http://www.infoworld.com/article/2940864/applicationdevelopment/r-programming-language-statisticaldataanalysis.html