VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANASANGAMA, BELAGAVI-590 018

An Internship Report On

"Data Science with R"

Submitted in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

PRITAM RAJ

USN: 1KS15CS072

Internship Carried Out

at

iPRIMED Solutions Pvt. Ltd.

No.62B, 2nd Floor, 2nd Cross Road, Opposite Electronic City Post, Electronics City Phase 1, Electronic City, Bengaluru, Karnataka 560100

InternalGuide

Mrs. Vijayalaxmi.Mekali

AssistantProfessor

K.S. Instituteof Technology

Bengaluru-560109

External Guide

Mr. Prasad

ManagingDirector

iPrimed Solutions Pvt. Ltd.

Bengaluru-560 100

Department of Computer Science & Engineering

K.S. INSTITUTE OF TECHNOLOGY

#14, Raghuvanahalli, Kanakapura Road, Bengaluru-560 109

2018-2019

K.S. INSTITUTE OF TECHNOLOGY

#14, Raghuvanahalli, Kanakapura Road, Bengaluru-560 109

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Internship Training entitled **"Data Science with R"** presented by **PRITAM RAJ, USN: 1KS15CS072** of 8th semester in partial fulfillment of the award of Bachelor of Engineering in CSE in Visvesvaraya Technological University, Belagavi during the academic year **2018-19.** The Internship Training has been approved as it satisfies the academic requirements in respect of Internship Training work prescribed for the Bachelor of Engineering degree.

InternalGuide	**HOD**	**Principal**
----------	**----------**	**----------**
Mrs. Vijayalaxmi.Mekali	**Dr. Rekha B. Venkatapur**	**Dr.T.V.Govindaraju**
Assistant professor,	**Head of the Department,**	**Principal/Director**
Dept. of CSE, KSIT	**Dept. of CSE, KSIT**	**K.S.I.T., Bengaluru**

External Viva

Name of the Examiners **Signature with date**

1.

2.

iPRIMED Solutions Pvt. Ltd.

No.62B, 2nd Floor, 2nd Cross Road, Opposite Electronic City Post, Electronics City Phase 1, Electronic City, Bengaluru, Karnataka 560100

CERTIFICATE

Certified that the internship program was successfully completed by **Mr. PRITAM RAJ USN: 1KS15CS072** a bonafide student of **K.S. INSTITUTE OF TECHNOLOGY for 4 weeks from 9th July, 2018 to 7th August, 2018.**

It is certified that, he has completed the internship satisfactorily.

Name & Signature

iPRIMED Solutions Pvt. Ltd.

No.62B, 2nd Floor, 2nd Cross Road, Opposite Electronic City Post, Electronics City Phase 1, Electronic City

Bengaluru, Karnataka

DECLARATION

I, PRITAM RAJ(USN: 1KS15CS072) student of eight semester B.E, Computer Science and Engineering, K. S. Institute of Technology, hereby declare that the Internship Report entitled "Data Modelling" submitted to the Visvesvaraya Technological University, Belagavi during the academic year 2018-2019, is a record of original work based on the Internship carried out at iPRIMED Solutions. Under the External Guidance of Mr. Niranjan Balaji, Managing Director, iPRIMED Solutions and Internal Guidance of Mrs. Vijayalaxmi Mekali, Assistant Professor, Computer Science & Engineering, K. S. Institute of Technology. The Internship Report has been submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science and Engineering. The results embodied in this report have not been submitted to any other University or Institute for the award of anydegree.

Date: **PRITAM RAJ**

Place:Bengaluru **1KS15CS072**

EXECUTIVE SUMMARY

The internship training at iPRIMED Solutions was a storehouse of knowledge on communication related activities. The university provided us an opportunity to work in any reputed organization as a part of the curriculum to get hands on experience of practical issues in real time situations. I went through iPRIMED Solutions organization as I had an interest on communication. As a part of training, I was introduced to various departments of the company and meeting great resource persons of the organization. I was allotted a developer as my guide and I was following all the orders made by him. The guide made me more comfortable and the environment was student friendly. All the technical and non-technical staff was very helpful and veryco-operative.

I was exposed to various tasks and was made to learn all the concepts and made me a good learner. The internship program consisted of learning all the technical activities of the organization. I learnt all the basics of the software's r programming and server configuration. All the present and olden techniques involved in the organization were made to learn in the internshipprogram.

Finally, the company people let me to get a real time experience on working and study of R programming. All the people in the organization helped me to successfully complete my internshiptraining.

PRITAM RAJ

1KS15CS072

ACKNOWLEDGEMENT

The successful internship training would be incomplete without the mention of the people who made it possible and whose constant guidance crowned my effort with success.

I take this opportunity to express my sincere gratitude to our **Management** K S Institute of Technology, Bengaluru for providing the environment to present the Internship.

I would express my gratitude to **Dr. K.V.A. Balaji** C.E.O. K.S. Institute of Technology, Bengaluru, for facilitating me to present the Internship.

I would like to extend my gratitude to the Principal/Director, **Dr. T. V. Govindaraju,**K.S. Institute of Technology, Bengaluru, for facilitating me to present the Internship.

I thank **Dr. Rekha B. Venkatapur,** Professor and Head Department of Computer Science and Engineering K.S. Institute of Technology, Bengaluru for her encouragement.

I would also like to thank **Mr. K. VenkataRao**, Associate Professor, Department of Computer Science and Engineering, K.S. Institute of Technology, Bengaluru, for his constant guidance and inputs.

I thank Internship Coordinator, **Mrs.Beena K** and **Mr. Prashanth H S,** Assistant Professor, Department of Computer Science and Engineering, K.S. Institute of Technology, Bengaluru, for their constant support and guidance.

I would like to thank my internship guide, **Mrs. Vijayalaxmi.Mekali**, Assistant Professor, Department of Computer Science and Engineering, K.S. Institute of Technology, Bengaluru, for her constant support.

I would also like to thank my external guide, **Mr. Niranjan Balaji,** Managing Director, iPRIMED Solutions for his constant guidance and input.

I would like to thank all the teaching and non-teaching staff of the college for their co-operation.

Finally, I extend my heart-felt gratitude to my **family** for their encouragement and support without which I wouldn't have come so far. Moreover, I thank all my **friends** for their invaluable support andco-ordination.

Name of the Student	**University Seat Number**
PRITAM RAJ	**1KS15CS072**

ABSTRACT

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. Data Science is one of the core subjects that were thought at iPRIMED. The company provided us with an opportunity to work with R programming language. R is a programming language and software environment for statistical analysis, graphics representation and reporting. Logistic Regression was one of the most important topics introduced. It is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and probability of particular outcome. Large datasets were used and Logistic Regression was implemented on those datasets to get various outcomes.An example of a dataset containing BreastCancer data was implemented using Logistic Regression to find the best accuracy of whether the given specimen is benign or malignant based on nine other cell features.

Keywords: Data Science, R programming, Logistic Regression, Data Modelling

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES

INTERNSHIP CERTIFICATE

The Academic Council of iPRIMED certifies

Bitan Raj from KSIT, Bangalore

for successfully completing

SAP Yuva Yuga,

Training and Internship Programme conducted for

a period of 30 days between 9th July 20 and 11th August 2018

under the aegis of NASSCOM Foundation.

We take the pleasure in recognizing the achievement with

the award of

Internship Certificate

In

Data Science with R

given on the 12th day of Jamary 2019.

COO &
Chief Operating Officer

iPRIMED™
Building Industry Ready Professionals

Head of
Academic Transformation

This certificate remains the property of iPRIMED Education Solutions Pvt. Ltd. to whom it must be returned on request.

Chapter 1

INTRODUCTION

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 Zealand, and is currently developed by the R Development Core Team.The core of R is an interpreted computer language which 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 is an implementation of the S programming language combined with lexical scoping semantics, inspired by Scheme. S was created by John Chambers in 1976, while at Bell Labs. R is an interpreted language users typically access it through a command-line interpreter. There are some important differences, but much of the code written for S runs unaltered. R is named partly after the first names of the first two R authors and partly as a play on the name of S. The project was conceived in 1992, with an initial version released in 1995 and a stable beta version in 2000. Like other similar languages such as APL and MATLAB, R supports matrix arithmetic.

Many features of R derive from Scheme. R uses S-expressions to represent both data and code. Functions are first-class and can be manipulated in the same way as data objects, facilitating meta-programming, and allow multiple dispatch. Variables in R are lexically scoped and dynamically typed. R supports procedural programming with functions and, for some functions, object-oriented programming with generic functions. A generic function acts differently depending on the classes of arguments passed to it. In other words, the generic function dispatches the function (method) specific to that class of object. For example, R has a generic printffunction that can print almost every class of object in R with a simple print(objectname)syntax.

Although used mainly by statisticians and other practitioners requiring an environment for statistical computation and software development, R can also operate as a general matrix calculation toolbox with performance benchmarks comparable to GNUOctave or MATLAB.

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand. 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") who can modify the R source code archive.

R-Programming is used for statistical analysis and business analysis. The different stages of business analytics are as follows: Descriptive Analysis, Diagnostic Analysis, Predictive Analysis, Prescriptive Analysis and Cognitive Analysis.

Data is divided into three types-

Type1- Structured Data is the type of data where the data is in the form of rows and columns. Example- Oracle.

Type2- Unstructured Data is the type of data like audio, video, graphics.

Type3- Semi-Structured Data is the type of data where the data is not structured. Example- HTML, XML, JSON.

Comparisionof R With Other Technologies

Data handling Capabilities – Good data handling capabilities and options for parallel computation.

Availability / Cost – R is an open source and we can use it anywhere.

Advancement in Tool – If you are working on latest technologies, R gets latest features.

Ease of Learning – R has a learning curve. R is a low-level programming language. As a result, simple procedures can take long codes.

Job Scenario – It is a better option for start-ups and companies looking for cost efficiency.

Graphical capabilities – R is having the most advanced graphical capabilities. Hence, it provides you with advanced graphical capabilities.

Customer Service support and community – R is the biggest online growing community.

MYSQL

MySQL is free and open-source software under the terms of the GNU General Public License, and is also available under a variety of proprietary licenses. Mysqlwas created by a Swedish company, mysqlAB, founded by David Axmark.

Original development of mysqlby Widenius and Axmark began in 1994. The first version of mysqlappeared on 23 May 1995. It was initially created for personal usage from msql based on the low-level language ISAM, which the creators considered too slow and inflexible. They created a new SQL interface, while keeping the same API as msql. By keeping the API consistent with the msqlsystem, many developers were able to use mysqlinstead of the (proprietarily licensed) msqlantecedent.

Mysqlis written in C and C++. Its SQL parser is written in yacc, but it uses a home-brewed lexical analyzer. Mysqlworks on many system platforms, including AIX, bsdi, freebsd, HP-UX, ecomstation, i5/OS, IRIX, Linux, macos, Microsoft Windows, netbsd, Novell netware, openbsd,opensolaris, OS/2 Warp, QNX, Oracle Solaris, Symbian, sunos, SCO openserver, SCO unixware, Sanos and Tru64. A port of mysqlto open VMS also exists.The mysqlserver software itself and the client libraries use dual-licensing distribution. They are offered under GPL version 2, or a proprietary license.

Data science is a field of big data geared towards providing meaningfull information based on large amount of complex data.data science, or data driven science, combines different field of work in statistics and computation in order to interpret data for the purpose of decision making.

Chapter 2

NASSCOM

The National Association Of Software And Services Companies(NASCOM)is a trade association of Indian information technology (IT) and business process outsourcing(BPO) industry. Established in 1988, NASSCOM is a non-profit organization is the apex body for the 154 billion dollar IT BPM industry in India, an industry that has made a phenomenal contribution to India's GDP, exports, employment, infrastructure and global visibility.

NASSCOM's relentless pursuit has been to constantly support the IT BPM industry in India, in the latter's continued journey towards seeking trust and respect from varied stakeholders, even as it reorients itself time and again to remain innovative, without ever losing its humane and friendly touch. NASSCOM is focused on building the architecture integral to the development of the IT BPM sector through policy advocacy, and help in setting up the strategic direction for the sector to unleash its potential and dominate newer frontiers.

NASSCOM's members, 2200+, constitute 90% of the industry's revenue and have enabled the association to spearhead initiatives at local, national and global levels. In turn, the IT BPM industry has gained recognition as a global powerhouse. In India, this industry provides the highest employment in the private sector.The vision statement"To help the IT and IT enabled products and services industry in India to be a trustworthy, respected, innovative and society friendly industry in the world".

Fig 2.1 Logo of the Company

2.1 LEADERSHIP TEAM

Navin Kumar(Founder and CEO) – He is a BE from BITS, Ranchi(1993) and holds an MBA from IIM Bangalore(PGP 1999). He worked at Tata Steel for 4 years before joining MBA program at IIM Bangalore.

NitishRaikar(COO) - Nitish holds Bachelor of Engineering from Goa University (1997) and MBA from XLRI Jamshedpur (2008). In his 20+ years of industry experience, he has been associated with companies like Infosys, IBM, Mindtree, Microland in various capacities handling large projects and managing teams.

Nitish works as the Chief Operating Officer and ensures that iPRIMED is designed for scale and ready for growth. He does this by streamlining the internal operations and manages three business verticals – T&P, Content Development, and English.

MukundJhunjhunwala(head-campus) - Mukund completed his B.Sc. from Kelly School of Business and is an entrepreneur by heart. Prior to joining iPRIMED, he has over 10+ years of experience as a self-grown entrepreneur. He has been the owner and promoter of companies in the manufacturing, power, infrastructure and renewable energy sector.

Mukund handles the digital segment of iPRIMED and focuses on expanding iPRIMED digital portfolio across campuses.

BashaShaik(head-technology) A tech evangelist by nature and having a rich 20+ years of experience in the technology industry, Basha comes with the aim of transforming iPRIMED into a technology driven organization.In his 20+ years of experience, he has lead large scale software solutions and products across industries - Banking, Supply Chain, Healthcare and IT. He has proven hands on experience in and around US, UK, Europe, Australia, Middle East and India as CIO, CTO and Delivery Head in corporates like Infosys and HCL.

Ratheesh Sebastian(head-finance) - Ratheesh is a Chartered accountant and handle the crucial role of ensuring that while we run a tight ship, finance doesn't become the bottleneck for growth. Prior to iPRIMED, he has been part of multiple startups, helping them successfully close series A round including FDI. A Aakillunisa(head-center of excellence) - Aakila comes with a rich experience of 16 years in the Learning Domain as an educator with a proven track record in Learning Methodologies. She has designed, developed and

implemented learning models that have driven effective and efficient training delivery. She has brought innovation in practices of Business, Technology, People and Process – aided by her research-driven approach and skills that have helped the organization achieve more within a short span of time. She has worked spractices. Aakila heads the Centre of Excellence at iPRIMED. She holds the focal point of knowledge and content management and brings in new practices and business intelligence to the ongoing projects.

2.2 MISSION OF NASSCOM

Recounting the adages NASSCOM to be associated with

- NASSCOM's ubiquitous raison d'être - 'Transform Business, Transform India'.
- Be a conduit of change through thought leadership, research, market intelligence and membership engagement.
- Establish India as a hub for innovation, products and start-ups.
- Work with the government to shape policy in key areas such as, skill development, trade, digital economy and business services.
- Increase the industry's outreach in its core markets and beyond, through strategic alliances.
- Be an industry platform for sharing and building best practices and collaborative engagement.
- Facilitate growth, and maintain India's leadership position as a trusted business destination.
- Expand the country's pool of relevant and skilled talent to drive inclusive and balanced growth.

2.3 DEVELOPMENT OF INDIAN IT BY NASSCOM

NASSCOM is dedicated to expanding India's role in the global IT order by creating a conducive business environment, simplifying policies, procedures, promoting intellectual capital and strengthening the talent pool. Some of the Developments in Indian IT include

2.3.1 GLOBAL TRADE DEVELOPMENT

The Global Trade Development (GTD) Initiative at NASSCOM has two broad slivers:

1. Policy Advocacy

In today's continuously evolving global regulatory environment, we work to ensure that Indian IT-BPM players remain abreast of various policy developments to try and reduce bottlenecks that have the propensity to impact business, and participate across geographies while conforming to their new laws and modified policies. NASSCOM actively works to make representation on key policy challenges faced by industry mainly in developed markets including and not limited to US, UK, EU, Australia, Canada, South Africa and Singapore.

2. Market Development

Indian IT-BPM companies have been expanding their geographic footprint for several years now. In addition to nurturing existing markets, NASSCOM is also focusing on building inroads into newer areas – geographies, verticals and customer segments. Several high growth and under-penetrated regions look promising for the IT-BPM business e.g., Nordics, Latin America, Africa, Middle East, ASEAN, China, Japan among few. While supporting member companies in creating a favourable eco-system to promote business growth, we also create suitable platform for forging Technology Partnerships & Alliances that are likely to bring long-term strategic benefit.

2.4 PROGRAMS & INTERVENTIONS

Market Development programs are conceived with an objective to provide a composite exposure to participating companies including awareness on business landscape in general and ICT in specific, exploration of market & investment opportunities, identifying suitable partnerships and one to one networking with probable clients & decision makers such as CIOs/CTOs/ Digital Leaders etc. These programs constitute of large, medium & small sized companies with a high percentage of SME participation.

Particularly for small & medium enterprises, such programs have helped shrink their learning curve and get a 360 degree view on a particular geography and plan their go-to-market strategy.

NASSCOM has been organizing market development programs in various formats such as

- Overseas business Delegations
- Participation in International Expos
- Creating region specific reports to highlight in-depth market opportunities
- Interactive learning sessions with market experts & industry leaders

- Interactive session with Trade & Investment departments of various Governments
- Roundtables with decision makers in thought leadership format
- B2B meetings

Chapter 3

MYSQL

MySQL, the most popular Open Source SQL database management system, is developed, distributed, and supported by Oracle Corporation.

• MySQL is a database management system.

A database is a structured collection of data. It may be anything from a simple shopping list to a picture Gallery or the vast amounts of information in a corporate network. To add, access, and process data Stored in a computer database, you need a database management system such as MYSQL Server. Since computers are very good at handling large amounts of data, database management systems play Acentral role in computing, as standalone utilities, or as parts of other applications.

• MySQL databases are relational.

A relational database stores data in separate tables rather than putting all the data in one big storeroom. The database structures are organized into physical files optimized for speed. The logical model, with objects such as databases, tables, views, rows, and columns, offers a flexible programming environment. You set up rules governing the relationships between different data fields, such as one-to-one, one-to-many, unique, required or optional, and "pointers" between different tables. The database enforces these rules, so that with a well-designed database, your application never sees inconsistent, duplicate, orphan, out-of-date, or missing data. The SQL part of "MySQL" stands for "Structured Query Language". SQL is the most common standardized language used to access databases.

Depending on your programming environment, you might enter SQL directly (for example, to generate reports), embed SQL statements into code written in another language, or use a language-specific API that hides the SQL syntax. SQL is defined by the ANSI/ISO SQL Standard. The SQL standard has been evolving since 1986 and several versions exist.

• MySQL software is Open Source.

Open Source means that it is possible for anyone to use and modify the software. Anybody can download the MySQL software from the Internet and use it without paying anything. If you wish, you may study the source code and change it to suit your needs. The MySQL software uses the GPL (GNU General Public License), to define what you may and

may not do with thesoftware in different situations.

• **The MySQL Database Server is very fast, reliable, scalable, and easy to use.**

If that is what you are looking for, you should give it a try. MySQL Server can run comfortably on a desktop or laptop, alongside your other applications, web servers, and so on, requiring little or no attention. If you dedicate an entire machine to MySQL, you can adjust the settings to take advantage of all the memory, CPU power, and I/O capacity available. MySQL can also scale up to clusters of machines, networked together. MySQL Server was originally developed to handle large databases much faster than existing solutions and has been successfully used in highly demanding production environments for several years. Although under constant development, MySQL Server today offers a rich and useful set of functions. Its connectivity, speed, and security make MySQL Server highly suited for accessing databases on the Internet.

• **MySQL Server works in client/server or embedded systems.**

The MySQL Database Software is a client/server system that consists of a multithreaded SQL server that supports different back ends, several different client programs and libraries, administrative tools, and a wide range of application programming interfaces (APIs). We also provide MySQL Server as an embedded multithreaded library that you can link into your application to get a smaller, faster, easier-to-manage standalone product.

• **A large amount of contributed MySQL software is available.**

MySQL Server has a practical set of features developed in close cooperation with our users. It is very likely that your favorite application or language supports the MySQL Database Server.

3.1 MYSQL: FEATURES & BENEFITS

3.1.1 Strong data protection

- Powerful mechanisms for ensuring only authorized users have access
- SSH and SSL supports safe and secure connection
- Powerful data encryption and decryption functions
- Sensitive data is protected from unauthorized viewing
- Backup and Recovery are provided

3.1.2 Comprehensive application development

- Support for stored procedures, triggers, funtions, views, cursors, ANSI-standards

SQL, and more

- Plug-in libraries to embed MYSQL database support into nearly any applicationManagement ease
- Use event scheduler automatically schedule common recurring SQL-based tasks to execute on the database server.
- Average time from software downloading to complete installation is less than fifteen Minutes.

3.1.3 Open source freedom and 24*7 support

- Around-the-clock support and indemnification available through MYSQL network
- Enterprise quality and enterprise ready, from installation to supportLowest total cost of ownership
- Save on database licensing costs and hardware expenditures, all while cutting system downtimes

3.2 MYSQL: ARCHITECTURE

The architecture of the world's most popular open source database system is very important for the Information Technology people. There are many reasons for MYSQL's popularity around the world, but one of the main reasons is its architecture, while there are many big players such as Oracle, Microsoft SQL and DB2, MYSQL's architecture makes it as unique and preferred choice for most of the developers. The fig 3.1 shows the internal architecture of the MYSQL relational database management system.

The MySQL architecture describes how the different components of a MySQL system relate to one another. The MySQL architecture is basically a client – server system. MySQL database server is the server and the applications which are connecting to MySQL database server are clients. The MySQL architecture contains the following major components.

3.2.1 Application Layer

This layer is the top most layers in MySQL architecture; you can see this same layer in many of the client – server architecture. This layer includes some of the services which are common to most of the client – server applications, some of the services are given below:

- Connection Handling.
- Authentication.

- Security.

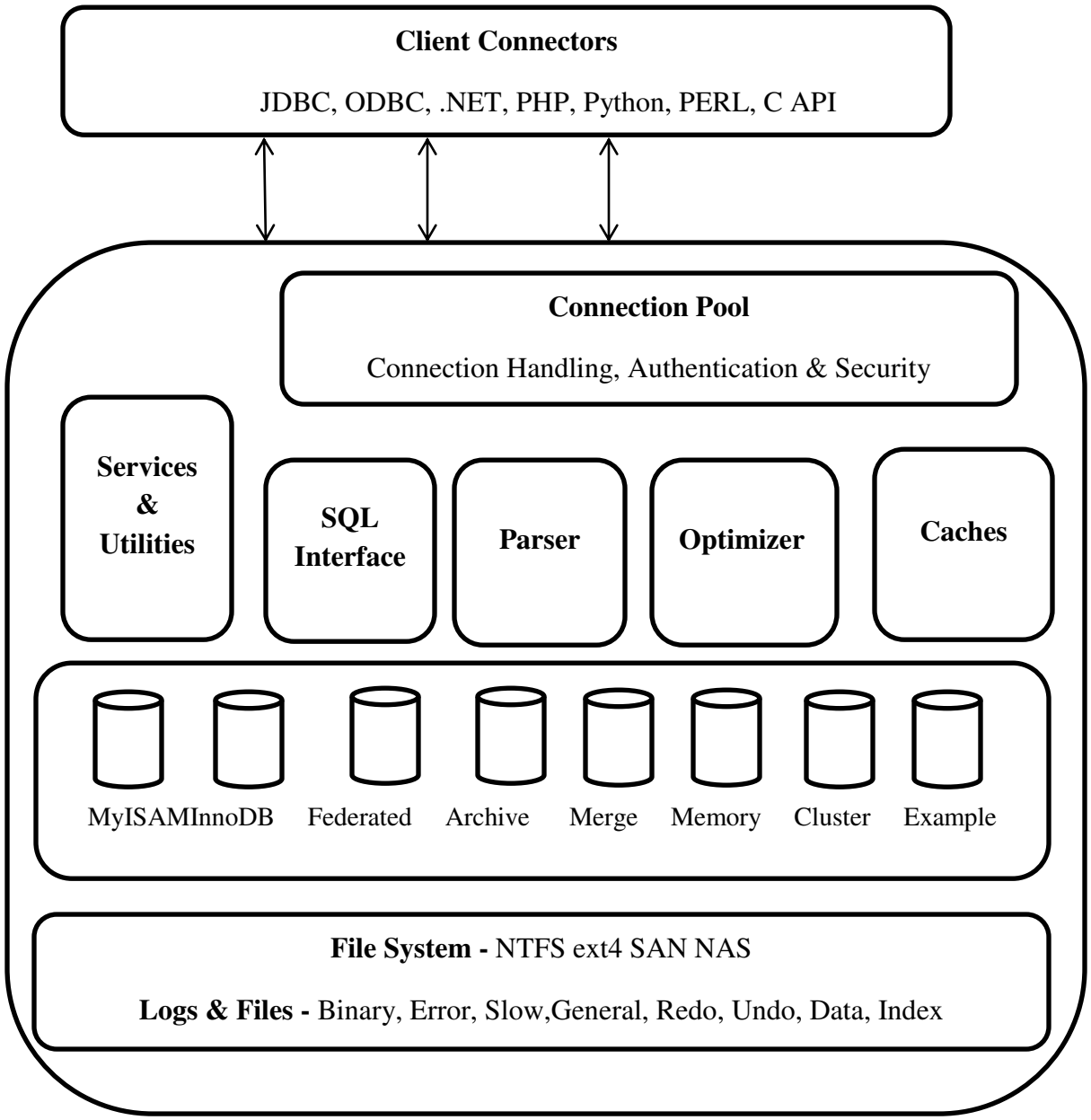

Fig 3.1 SQL architecture

3.2.2 Connection Handling

When a client connects to server, the client gets its own thread for its connection. All the queries from that client executed within that specified thread. The thread is cached by the server, so they don't need to created and destroyed for each new connection.

3.2.3 Authentication

Whenever a client connects to a MySQLserver, the serverperforms theauthenticati-

-on in the server side. The authentication is based on the username, host of the client and password of the client user.

3.2.4 Security

After the client gets connected successfully to MySQL server, the server will check whether that particular client has the privileges to issue certain queries against MySQL server.

3.2.5 MySQL Server layer

This layer takes care of all the logical functionalities of the MySQL relational database management system. The brain of the MySQL server is resides in this layer. The logical layer of the MySQL is divided into various sub components, which are given below:

- MySQL services and utilities.
- SQL Interface.
- SQL Parser.
- Optimizer.
- Caches & buffers.

3.2.6 MySQL services and utilities

MySQL comparatively provides wide range of services and utilities. This is one of the main reasons for the popularity of the MySQL. This layer provides the services and utilities for administration and maintenance of MySQL system, some of them are mentioned below:

- Backup & recovery.
- Security.
- Replication.
- Cluster.
- Partitioning.
- Workbench.

3.2.7 SQL Interface

Structured Query Language (SQL) is a query language, used to query MySQL server. It is a tool to interact between MySQL client user and server. Some of the SQL interface components are given below.

- Data Manipulation Language (DML).

- Data Definition Language (DDL).
- Stored Procedures.
- Views.
- Triggers.

3.2.8 Parser

MySQL parses queries to create an internal structure (the parse tree). The MYSQL parser behaves as a single pass compiler. As per the MySQL internals, the parser structure is given below**.**

- Lexical analysis (making words or tokens from a character stream) is implemented at first stage, when parsing regular statements.
- Syntactic analysis (making "sentences"), semantic analysis (making sure these sentences do make sense), and code generation (for compilers) – all of them – are done at once, during the phase of code.

3.2.9 Optimizer

After creating the internal parse tree, the MySQL applies a variety of optimization techniques. These techniques may include, rewriting the query, order of scanning of tables and choosing the right indexes to use. Actually you can ask the server to explain the various aspects of optimization.

3.2.10 Caches

The MySQL cache (query cache) stores complete result sets for SELECT statements. Even before parsing the query, the MySQL server consults the query cache. If any client issues a query that is identical to one already in the in the cache.

3.2.11 Storage Engine Layer

The pluggable storage engine feature makes the MySQL as unique and preferred choice for most of the developers. This is the feature which makes the MySQL to reach an edge over the big player. MySQL allows us to choose the variety storage engines for different situations and requirements.

3.3 COMMANDS

DATA DEFINITION LANGUAGE-which deals withdatabase schemes and description,

of how the data should reside in the database.

- Create: to create database and its objects like table, index, views, store, procedure, function, and triggers.
- ALTER: allows the structure of the existing database.
- DROP: delete object from the database.
- TRUNCATE: remove all records from a table, including all spaces allocated for the records are removed.
- COMMENT: add comments to the data dictionary.
- RENAME: rename an object.

DATA MANIPULATION LANGUAGE- DML deals with data manipulation and includes most common SQL statements such as select, insert, update etc and it is used to store modify, delete and update data in database.

- SELECT: retrieve data from a database
- INSERT: insert data into a table
- UPDATE: update existing data within a table
- DELETE: delete all records from a database table
- CALL: call a DL/SQL or Java subprogram
- LOCK TABLE: concurrency control

DATA CONTROL LANGUAGE- DCL includes commands such as GRANT and mostly concerned with eight, permission and other control of the database system.

- GRANT: allows users access privileges to database
- REVOKE: withdraw users access privileges given by using the GRANT command.

TRANSACTION CONTROL LANGUAGE- TCL deals with a translation within a database.

- COMMIT: commits a transaction
- ROLLBACK: rollback a transaction in case of any errors occurs
- SAVEPOINT: to rollback the transaction making points within groups
- SET TRABSACTION: specify characteristic of the transaction

3.4 CREATING AND SELECTING A DATABASE

If the administrator creates your database for you when setting up your permissions, you can begin using it. Otherwise, you need to create it yourself:

mysql>CREATE DATABASE menagerie;

Under Unix, database names are case-sensitive (unlike SQL keywords), so you must always refer to your database as menagerie, not as Menagerie, MENAGERIE, or some other variant. This is also true for table names. (Under Windows, this restriction does not apply, although you must refer to databases and tables using the same lettercase throughout a given query. However, for a variety of reasons, the recommended best practice is always to use the same lettercase that was used when the database was created.)

3.5 WHAT IS RDBMS?

RDBMS stands for Relational Database Management System. RDBMS is the basis for SQL,and for all modern database systems like MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access. A Relational database management system (RDBMS) is a database management system (DBMS) that is based on the relational model as introduced by E. F. Codd.

3.6 SQL CONSTRAINTS

Constraints are the rules enforced on data columns on a table. These are used to limit thetype of data that can go into a table. This ensures the accuracy and reliability of the datain the database. Constraints can either be column level or table level. Column level constraints are applied only to one column whereas, table level constraints are applied to the entire table. Following are some of the most commonly used constraints available in SQL:

- NOT NULL Constraint: Ensures that a column cannot have a NULL value.
- DEFAULT Constraint: Provides a default value for a column when none is specified.
- UNIQUE Constraint: Ensures that all the values in a column are different.
- PRIMARY Key: Uniquely identifies each row/record in a database table.
- FOREIGN Key: Uniquely identifies a row/record in any another database table.
- CHECK Constraint: The CHECK constraint ensures that all values in a column satisfy
- certain conditions.
- INDEX: Used to create and retrieve data from the database very quickly.

3.7 CREATING A TABLE

Creating the database is the easy part, but at this point it is empty, as SHOW TABLES tells you:

mysql>SHOW TABLES;

Empty set (0.00 sec)

The harder part is deciding what the structure of your database should be: what tables you need and what columns should be in each of them.

You want a table that contains a record for each of your pets. This can be called the pet table, and it should contain, as a bare minimum, each animal's name. Because the name by itself is not very interesting, the table should contain other information. For example, if more than one person in your family keeps pets, you might want to list each animal's owner. You might also want to record some basic descriptive information such as species and sex.

mysql>CREATE TABLE pet (name VARCHAR(20), owner VARCHAR(20),

species VARCHAR(20), sex CHAR(1), birth DATE, death DATE);

3.8SQL – SYNTAX

- **SQL SELECT Statement**

  SELECT column1, column2....columnNFROM table_name;

- **SQL DISTINCT Clause**

  SELECT DISTINCT column1, column2....columnNFROM table_name;

- **SQL WHERE Clause**

  SELECT column1, column2....columnN FROM table_nameWHERE CONDITION;

- **SQL AND/OR Clause**

  SELECT column1, column2....columnNFROM table_nameWHERE CONDITION-1 {AND|OR} CONDITION-2;

- **SQL BETWEEN Clause**

  SELECT column1, column2....columnN FROM table_name WHERE column_nameBETWEEN val-1 AND val-2;

- **SQL LIKE Clause**

  SELECT column1, column2....columnN FROM table_nameWHERE column_name LIKE { PATTERN };

- **SQL ORDER BY Clause**

  SELECT column1, column2....columnN FROM table_name WHERE CONDITION ORDER BY column_name {ASC|DESC};

- **SQL GROUP BY Clause**

  SELECT SUM(column_name) FROM table_name WHERE CONDITION GROUP BY column_name;

- **SQL COUNT Clause**

  SELECT COUNT(column_name) FROM table_nameWHERE CONDITION;

- **SQL CREATE TABLE Statement**

  CREATE TABLE table_name(column1 datatype, column2 datatype, column3 datatype,..... columnNdatatype, PRIMARY KEY(one or more columns));

- **SQL DROP TABLE Statement**

  DROP TABLE table_name;

- **SQL CREATE INDEX Statement**

  CREATE UNIQUE INDEX index_nameON table_name(column1, column2,...columnN);

- **SQL DROP INDEX Statement**

  ALTER TABLE table_nameDROP INDEX index_name;

- **SQL DESC Statement**

  DESC table_name;

- **SQL TRUNCATE TABLE Statement**

  TRUNCATE TABLE table_name;

- **SQL ALTER TABLE Statement**

  ALTER TABLE table_name {ADD|DROP|MODIFY} column_name {data_ype};

- **SQL ALTER TABLE Statement (Rename)**

  ALTER TABLE table_name RENAME TO new_table_name;

- **SQL INSERT INTO Statement**

  INSERT INTO table_name(column1, column2....columnN) VALUES (value1, value2....valueN);

- **SQL UPDATE Statement**

  UPDATE table_nameSET column1 = value1, column2 = value2....columnN=valueN[WHERE CONDITION];

- **SQL DELETE Statement**

  DELETE FROM table_nameWHERE {CONDITION};

- **SQL CREATE DATABASE Statement**

  CREATE DATABASE database_name;

- **SQL DROP DATABASE Statement**

  DROP DATABASE database_name;

- **SQL USE Statement**

  USE database_name;

- **SQL COMMIT Statement**

  COMMIT;

- **SQL ROLLBACK Statement**

  ROLLBACK;

Chapter 4

DATA SCIENCE

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as data mining and big data which uses the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problem.Data science is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data.It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science and computer science.

Data Science is a field that covers data cleansing, preparation, and analysis. It is an includes several scientific methods, such as mathematics, statistics, and many other tools data scientists apply to extract knowledge from data sets as shown in Figure 4.1.

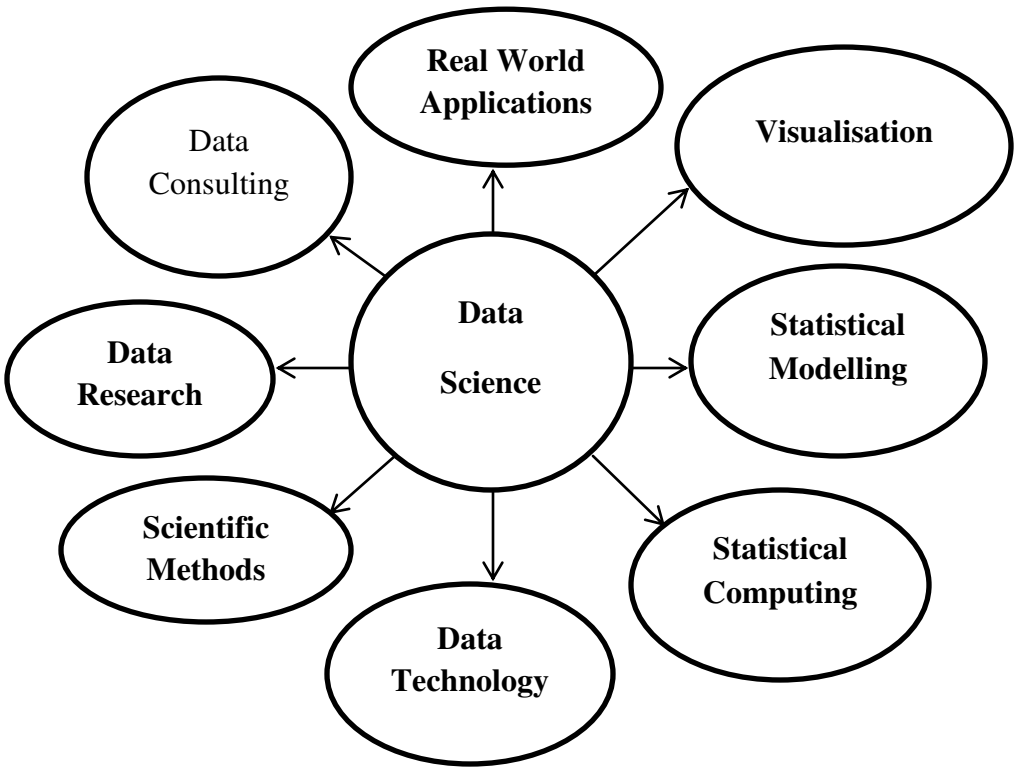

Fig 4.1.Data Science Applications

First, the Data scientist gathers datasets from multi-disciplines and compiles it. Next, he applies machine learning, predictive and sentimental analysis to analyze the data and eventually deciphers a meaningful pattern out of the huge datasets.

A data scientist makes use of tools and languages like R, MATLAB, and DB Management for data analysis and machine learning.Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education.

- Computer Science- Pattern recognition, visualization, data warehousing, Artificial Intelligence, High performance computing, Databases.
- Mathematics- Mathematical Modeling.
- Statistics- Statistical and Stochastic modeling, Probability.

4.1 IMPORTANCE OF DATA SCIENCE

Data science helps brands to understand their customers in a much enhanced and empowered manner. Customers are the soul and base of any brand and have a great role to play in their success and failure. With the use of data science, brands can connect with their customers in a personalized manner, thereby ensuring better brand power and engagement.

- One of the reasons why data science is gaining so much of attention is because it allows brands to communicate their story in such an engaging and powerful manner. When brands and companies utilize this data in a comprehensive manner, they can share their story with their target audience, thereby creating better brand connect. After all, nothing connects with consumers like an effective and powerful story, that can inculcate all human emotions.

- Big-Data is a new field that is constantly growing and evolving. With so many tools being developed, almost on a regular basis, big data is helping brands and organizations to solve complex problems in IT, human resource, and resource management in an effective and strategic manner. This means effective use of resources, both material and non-material.

- One of the most important aspect of data science is that its findings and results can be applied to almost any sector like travel, healthcare and education among others.

- Data science is accessible to almost all sectors. There is a large amount of data available in the world today and utilizing them in a proper manner can spell success and failure for brands and organizations. Utilizing data in a proper manner will hold the key for achieving goals for brands, especially in the coming times.

4.2 APPLICATIONS OF DATA SCIENCE

- **Fraud and Risk Detection**-Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyze the probabilities of risk and default.

- **Healthcare**-heath care applications include Medical Image Analysis, Genetics and Genomics, Drug Development, Virtual Assistant for Patients and Costumer Support.

- **Internet Search**-There are many other search engines like Google, Yahoo, Bing, Ask, AOL, and so on. All these search enginesmake use of data science algorithms to deliver the best result for our searched query in a fraction of seconds.

- **Targeted Advertising**-The entire digital marketing spectrum. Starting from the display banners on various websites to the digital billboards at the airports – almost all of them are decided by using data science algorithms.

- **Website Recommendations**-Internet giants like Amazon, Twitter, Google Play, Netflix, Linkedin, imdb and many more use this system to improve the user experience. The recommendations are made based on previous search results for a user.

- **Advanced Image Recognition**-When You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. In their latest update, Facebook has outlined the additional progress they've made in this area, making specific note of their advances in image recognition accuracy and capacity.

- **Speech Recognition**-Some of the best examples of speech recognition products are Google Voice, Siri, Cortana etc. Using speech-recognition feature, even if you aren't in a position to type a message, your life wouldn't stop.

- **Airline Route Planning**-Now using data science, the airline companies can Predict flightdelay**,** decides which class of airplanes to buy**,** whether to directly land at the destination or take a halt in between and Effective drive customer loyalty programs.

- **Gaming**-Games are now designed using machine learning algorithms which improve/upgrade themselves as the player moves up to a higher level. EA Sports, Zynga, Sony, Nintendo, Activision-Blizzard have led gaming experience to the next level using data science.

Chapter 5

R PROGRAMMING LANGUAGE

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 Zealand, and is currently developed by the R Development Core Team.The core of R is an interpreted computer language which 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 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.

R is free software distributed under a GNU-style copy left, and an official part of the GNU project called GNU S.Many features of R derive from Scheme. R uses S-expressions to represent both data and code. Functions are first-class and can be manipulated in the same way as data objects, facilitating meta-programming, and allow multiple dispatch. Variables in R are lexically scoped and dynamically typed. R supports procedural programming with functions and,for some functions, object-oriented programming with generic functions. A generic function acts differently depending on the classes of arguments passed to it. In other words, the generic function dispatches the function (method) specific to that class of object. For example, R has a generic printfunction that can print almost every class of object in R with a simple print(objectname) syntax.

Although used mainly by statisticians and other practitioners requiring an environment for statistical computation and software development, R can also operate as a general matrix calculation toolbox with performance benchmarks comparable to GNUOctave or MATLAB.

5.1 EVOLUTION OF R

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand. 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") who can modify the R source code archive.

5.2 FEATURES OF R

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R.

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

As a conclusion, R is world's most widely used statistics programming language.

For example, the following command will install any package which is required for 3D charts.

>install.packages("package name")

- **R Script File**

Usually, you will do your programming by writing your programs in script files and then you execute those scripts at your command prompt with the help of R interpreter called Rscript.

5.3 DATATYPES

In contrast to other programming languages like C and java in R, the variables are not declared as some data type. The variables are assigned with R-Objects and the data type of the R-object becomes the data type of the variable.R has a wide variety of data types including scalars, vectors (numerical, character, logical), matrices, data frames, and lists.

5.3.1 Vectors

When you want to create vector with more than one element, you should use **c()** function which means to combine the elements into a vector.

```
# Create a vector.

apple<- c('red','green',"yellow")

print(apple)

# Get the class of the vector.

print(class(apple))
```

When we execute the above code, it produces the following result –

```
[1] "red"    "green"  "yellow"

[1] "character"
```

5.3.2 Lists

A list is an R-object which can contain many different types of elements inside it like vectors, functions and even another list inside it. R List can also contain a matrix or a function as its elements. The List is been created using **list()** function in R.

```
# Create a list.

list1 <- list(c(2,5,3),21.3,sin)

# Print the list.

print(list1)
```

When we execute the above code, it produces the following result –

```
[[1]]

[1] 2 5 3

[[2]]

[1] 21.3

[[3]]
```

function (x) .Primitive("sin")

5.3.3 Matrices

A matrix is a two-dimensional rectangular data set. It can be created using a vector input to the matrix function.

```
# Create a matrix.

M = matrix( c('a','a','b','c','b','a'),nrow=2,ncol=3,byrow= TRUE)

print(M)
```

When we execute the above code, it produces the following result –

```
[,1] [,2] [,3]

[1,] "a"  "a"  "b"

[2,] "c"  "b"  "a"
```

5.3.4 Arrays

While matrices are confined to two dimensions, arrays can be of any number of dimensions. The array function takes a dim attribute which creates the required number of dimension. In the below example we create an array with two elements which are 3x3 matrices each.

```
# Create an array.

a <- array(c('green','yellow'),dim = c(3,3,2))

print(a)
```

When we execute the above code, it produces the following result –, , 1

```
[,1]     [,2]     [,3]

[1,] "green"  "yellow" "green"

[2,] "yellow" "green"  "yellow"

[3,] "green"  "yellow" "green"

, , 2
```

```
  [,1]     [,2]     [,3]
[1,] "yellow" "green"  "yellow"
[2,] "green"  "yellow" "green"
[3,] "yellow" "green"  "yellow"
```

5.3.5 Factors

Factors are the r-objects which are created using a vector. It stores the vector along with the distinct values of the elements in the vector as labels. Factors are created using the factor() function. The nlevels functions gives the count of levels.

```
# Create a vector.
apple_colors<- c('green','green','yellow','red','red','red','green')
# Create a factor object.
factor_apple<-factor(apple_colors)
# Print the factor.
print(factor_apple)
print(nlevels(factor_apple))
```

When we execute the above code, it produces the following result –

```
[1] green  green  yellow red    redred    green
Levels: green red yellow
```

5.3.6 Data Frames

Data frames are tabular data objects. Unlike a matrix in data frame each column can contain different modes of data. The first column can be numeric while the second column can be character and third column can be logical. It is a list of vectors of equal length.

Data Frames are created using the data.frame() function.

```
# Create the data frame.
BMI <-data.frame(
```

```
gender= c("Male","Male","Female"),

height= c(152,171.5,165),

weight= c(81,93,78),

Age=c(42,38,26))

print(BMI)
```

When we execute the above code, it produces the following result –

```
gender height weight Age

1   Male  152.0     81  42

2   Male  171.5     93  38

3 Female  165.0     78  26
```

5.4 OPERATORS USED IN R

An operator is a symbol that tells the compiler to perform specific mathematical or logical manipulations. R language is rich in built-in operators and provides following types of operators. There are different types of operators in R programming, they are:

5.4.1 Arithmetic Operators

The R Arithmetic operators include operators like Arithmetic Addition, Subtraction, Division, Multiplication, Exponent, Integer Division and Modulus.Below Table I.shows all the arithmetic operators used in R programming.

Table I Arithmetic Operators

OPERATORS	DESCRIPTION
+	Adds two vectors.
-	Subtracts second vector from the first.
*	Multiplies both vectors.
/	Divide the first vector with the second.

x%%y	Give the remainder of the first vector with the second.
x %/% y	The result of division of first vector with second.

5.4.2 Relational Operators

Relational operators are used to compare between values are below Table II.

Table II Relational Operators

OPERATORS	DESCRIPTION
<	Checks if each element of the first vector is greater than the corresponding element of the second vector.
>	Checks if each element of the first vector is less than the corresponding element of the second vector.
==	Checks if each element of the first vector is equal to the corresponding element of the second vector.
<=	Checks if each element of the first vector is less than or equal to the corresponding element of the second vector.
>=	Checks if each element of the first vector is greater than or equal to the corresponding element of the second vector.
!=	Checks if each element of the first vector is unequal.

5.4.3 Logical Operators

Logical operators are used to carry out Boolean operations like AND, OR etc. Table III. shows the different types of logical operators used in R programming language.

Table III. Logical Operators

OPERATOR	DESCRIPTION
!	It is called Logical NOT operator. Takes each element of the vector

	and gives the opposite logical value.
&	It is called Element-wise Logical AND operator. It gives an output TRUE if both the elements are TRUE.
&&	Logical AND operator. It gives the TRUE only if both are TRUE.
|	It is called Element-wise Logical OR operator. It gives an output TRUE if one the elements is TRUE.
||	Logical OR operator. It gives the TRUE if one of them is TRUE.

5.4.4 Assignment Operator

These operators are used to assign values to vectors. The operators <- and = can be used, almost interchangeably, to assign to variable in the same environment. The Table IV shows the description of assignment operators.

Table IV. Assignment Operators

OPERATORS	**DESCRIPTION**
<-, <<-, =	Left Assignment
->, ->>	Right Assignment

CONCLUSION

Data science is important it empowers professionals with data management technologies like Hadoop, R, Flume, Sqoop, Machine lerning, Mahout etc. The knowledge and expertise of the skills is an added advantage for a better and competitive career. By migrating current database apps to MYSQL, enterprises are enjoying significant cost saving on new projects. The dependability and ease of management that accompany MYSQL save your troubleshooting time which is otherwise wasted in fixing downtime issues and performance problem. R is free and open-source, making it possible for anyone to have access to world-class statistical analysis tools. R performs a wide variety of functions, such as data manipulation, statistical modelling, and graphics. It is used widely in academia and the private sector and is the most popular statistical analysis programming language today.

REFERENCES

[1] Mattew Mayo and KDnuggets,“Data Engineering ofEssentials”, Chicago, April 28,1999

[2] VikramVaswani, “Mysql” ,Gordon, 2 Edition,2016.

[3] Ross Ihaka and Robert Gentleman at the University of Auckland,“R Language”, NewZealand,1972

[4] Jeff Leek (12 December 2013). "The key word in "Data Science" is not Data, it is Science". Simply Statistics.

[5] International Council for Science: Committee on Data for Science and Technology. (2012, April). CODATA, The Committee on Data for Science and Technology. Retrieved from International Council for Science : Committee on Data for Science and Technology: http://www.codata.org/

[6] P.C. Mahalanobis Memorial Lectures, 7th series". P.C. Mahalanobis Memorial Lectures, Indian Statistical Institute. Archived from the original *on* 29 October 2013. Retrieved 18 July 2017.

[7] Naur,Peter (1 July 1966). "The science of datalogy". Communications of the ACM. **9** (7): 485. doi:10.1145/365719.366510

[8] DataScience Journal. (2002, April). Contents of Volume 1, Issue 1, April 2002. Retrieved from Japan Science and Technology Information Aggregator, Electronic.

[9] The Journal of Data Science. (2003, January). Contents of Volume 1, Issue 1, January 2003

[10] Warden,Pete (9 May 2011). "Why the term "data science" is flawed but useful". O'Reilly Radar. Retrieved 20 May 2018.