Understanding database management using AWS

Project report submitted in partial fulfillment of the requirement for the Degree

of

Bachelor of Technology

in

Electronics and Communication Engineering

by

Pranav Bansal (161048)

Under the supervision of

Mr. Karthick Selvam

Jaypee University of Information Technology,

Waknaghat, Solan-173234, Himachal Pradesh

May- 2020

Table of Contents

DECLARATION

I hereby declare that the work reported in B.Tech project report entitled **“Understanding database management using AWS”** submitted at Jaypee University of Information Technology, Waknaghat, India is an authentic record of our work carried out under the supervision of Mr. Karthick Selvam. We have not submitted this work elsewhere for any other degree or diploma.

Pranav

Pranav Bansal

161048

This is to certify that above statement made by the candidate is correct to the best of my knowledge.

Karthick Selvam
Sr. Trainer, Cognizant Academy
Cognizant Technology Solutions

Acknowledgment

I am sincerely thankful to my guide Mr. Karthick Selvam for his extraordinary and brilliant guidance, his constant and valuable feedback throughout my major final year project. I am very thankful to him for providing me this wonderful opportunity to work under his guidance. I am blessed to understand the concepts about AWS thoroughly from him.

List of Figures

ABSTRACT

Amazon has a long past of applying a devolved IT groundwork. This willing plan permitted our improvement groups to get to figure and capacity assets on-request, and it has expanded generally effectiveness and willingness. By 2004, it had gone finished lengthier than 10 years and a huge number of dollars constructing and trade with the huge scope, solid, and capable IT substance that powered one of the world's chief online retail phases.

It pushed Web Servicing with the goal that different links could profit by its vicarious and interest in seriatim a massive scope misappropriated, value-based IT outline. It has been employed since 2009, and today serves a enormous number of customers around the world. Today it runs a worldwide web stage serving a large number of clients and overseeing billions of bucks of skill each year.

Using, you can order figure force, amassing, and different managements in minutes and have the malleability to pick the improvement stage or programming model that bodes well for the issues they are attempting to explain. You pay just for what you use, with no up-front costs or long pull responsibilities, making it a financially perceptive approach to convey applications. Here are a portion of the instances of how associations, from inquire about firms to huge activities, use them today: A huge venture rapidly and financially sends new inside applications, for example, HR preparations, money requests, stock administration arrangements, and web based preparing to its took staff.

A network based commercial site suits unforeseen interest for a "hot" item brought about by the pathological buzz from Facebook and Twitter without repairing its outline. A medicinal research firm executes huge scope imitations utilizing figuring power gave by them.

CHAPTER 1

Introduction

Cloud computing presents an opportunity for the on-premises datacentre. With an on-premises datacentre, we must manage the whole lot, together with purchasing and installing hardware, virtualization, installing the running device, and any other required programs, setting up thenetwork, configuring the firewall, and setting up garage for facts. After doing all the set-up, we emerge as chargeable for preserving it through its whole lifecycle.

But if we pick Cloud Computing, a cloud vendor is liable for the hardware buy and renovation. They also offer a huge type of software program and platform as a provider. We can take any required offerings on rent. The cloud computing services may be charged based on usage.

- **Public cloud** : Server that is used as a hypervisor and has been shared to a group of companies or groups of people from different sectors, called a public cloud.

- **Private Cloud**: The cloud computing sources which are exclusively used internal a single commercial enterprise or enterprise are termed as a personal cloud. A personal cloud may additionally physically be positioned at the agency’s on-website datacentre or hosted by a 3rd-birthday party provider.

- **Hybrid Cloud**: The aggregate of public and personal clouds has bonded together by way of the generation that lets in data programs to be shared between them. Hybrid cloud affords flexibility and more deployment options to the commercial enterprise.

CHAPTER 2

Introduction to MySQL

SQL means Structured Query Language. A well-known language which is broadly used to speak with the databases.

SQL is used to perform obligations on a database. It allows interacting with the information.

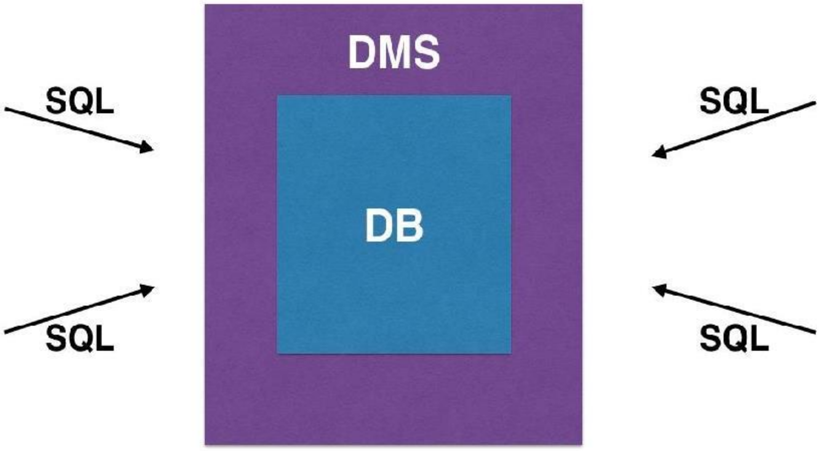


Figure 2.1: MySQL

2.1 What is MySQL

MySQL is a relational database management system. It provides a UI for users to interacts with the database.

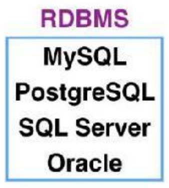
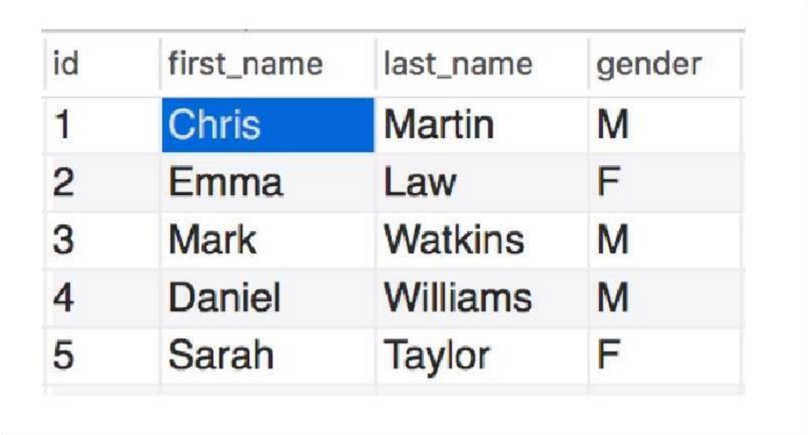


Figure 2.2: RDBMS

2.2 What is RDBMS?

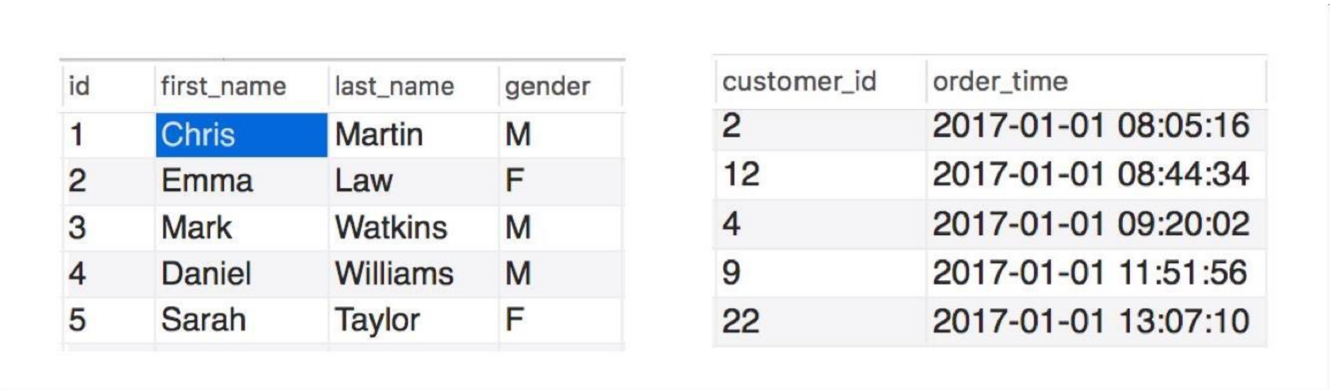
RDBMS is "Relationally Databases Managements Systemic." An RDBMS is a DBMS designedsuch that two or data pools like tables interact on the basis of relation or a constraint..



id	first_name	last_name	gender
1	Chris	Martin	M
2	Emma	Law	F
3	Mark	Watkins	M
4	Daniel	Williams	M
5	Sarah	Taylor	F

Figure:2.4

Tables in a relational database may be connected collectively. RDBMS is what we use to get admission to and engage with the relational database.



id	first_name	last_name	gender
1	Chris	Martin	M
2	Emma	Law	F
3	Mark	Watkins	M
4	Daniel	Williams	M
5	Sarah	Taylor	F

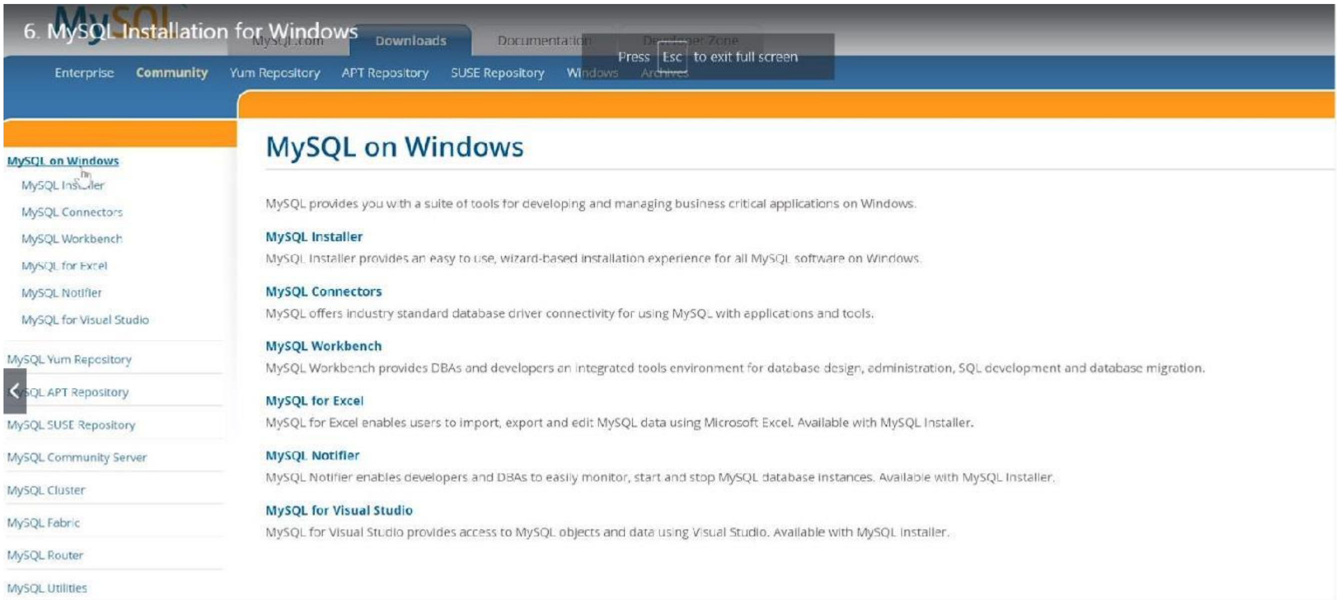
customer_id	order_time
2	2017-01-01 08:05:16
12	2017-01-01 08:44:34
4	2017-01-01 09:20:02
9	2017-01-01 11:51:56
22	2017-01-01 13:07:10

Figure 2.5: Tables Example

2.3 Installation of MySQL

Step1:

The first step is to go to the website www.dev.sql.com and install MySQL for the required operating system.



MySQL on Windows

MySQL provides you with a suite of tools for developing and managing business critical applications on Windows.

MySQL Installer
MySQL Installer provides an easy to use, wizard-based installation experience for all MySQL software on Windows.

MySQL Connectors
MySQL offers industry standard database driver connectivity for using MySQL with applications and tools.

MySQL Workbench
MySQL Workbench provides DBAs and developers an integrated tools environment for database design, administration, SQL development and database migration.

MySQL for Excel
MySQL for Excel enables users to import, export and edit MySQL data using Microsoft Excel. Available with MySQL Installer.

MySQL Notifier
MySQL Notifier enables developers and DBAs to easily monitor, start and stop MySQL database instances. Available with MySQL Installer.

MySQL for Visual Studio
MySQL for Visual Studio provides access to MySQL objects and data using Visual Studio. Available with MySQL Installer.

Figure 2.6: MySQL installation

Step2:

After installing the MySQL to your system, start setting up the workbench platform where you can write the queries.

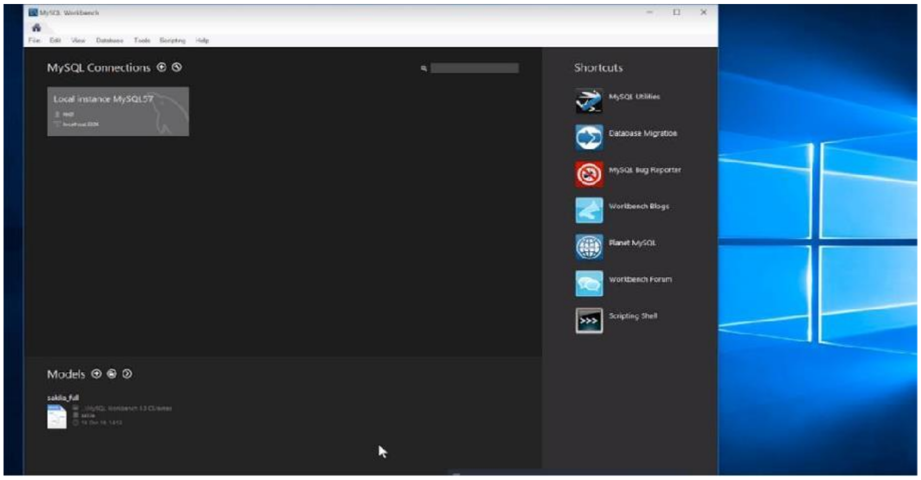
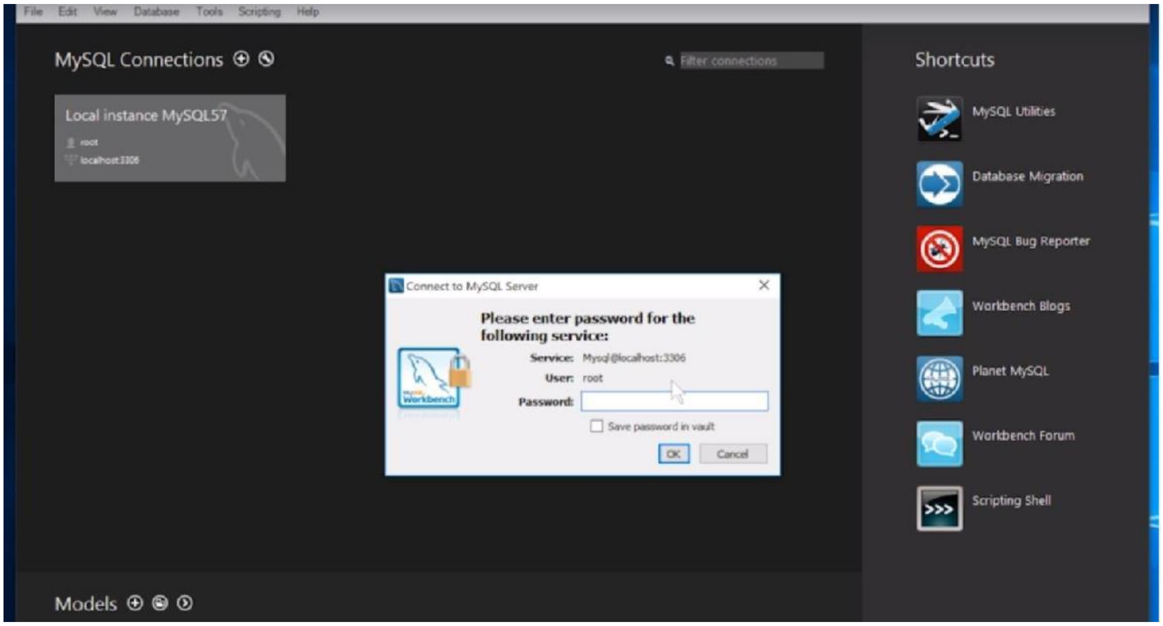


Figure 2.7: setting up the workbench

Step3:

When you open the workbench, a password prompt will pop up and you need to enter the root password that you created during the installation process.



Connect to MySQL Server

Please enter password for the following service:

Service: Mysql@localhost:3306
User: root
Password:

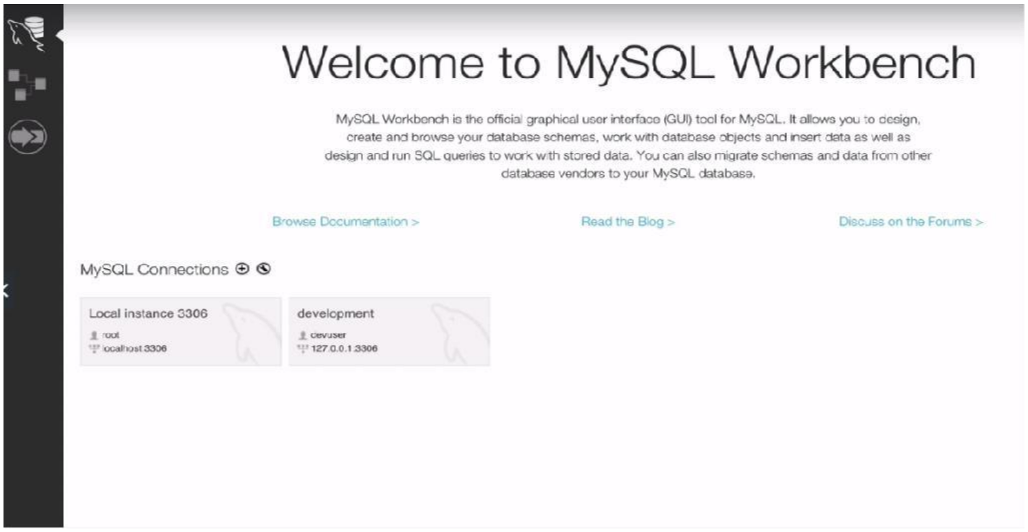
☐ Save password in vault

OK Cancel

Figure 2.8: Entering the workbench root password

Step4:

Now you can create your first database inside the workbench.



Welcome to MySQL Workbench

MySQL Workbench is the official graphical user interface (GUI) tool for MySQL. It allows you to design, create and browse your database schemas, work with database objects and insert data as well as design and run SQL queries to work with stored data. You can also migrate schemas and data from other database vendors to your MySQL database.

Figure 2.9: Opening screen of a workbench

2.4 Data Definition Language

Data definition of statistics language (DDL) is a syntax just like a laptop programming language for defining information structures, schemas in the database etc. It includes commands such as CREATEs, ALTERs, and DROPs and many more.

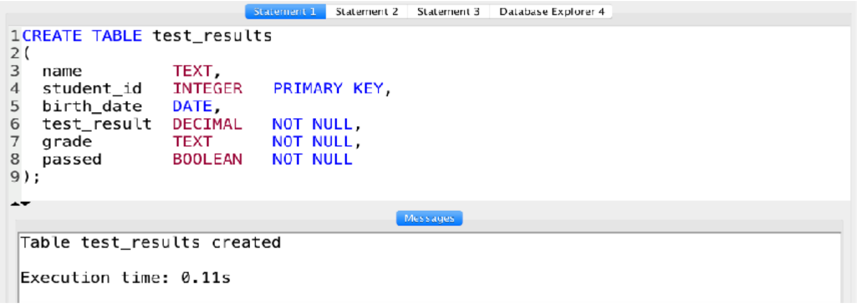
2.4.1 Data Types

1) Numeric Data type:
   - INT: Whole numbers
   - FLOAT(M, D): Decimal numbers(approx.)
   - DECIMAL(M,D): Decimal numbers(precise)
2) Non-Numeric Data type:
   - CHAR(N): Fixed length character
   - VARCHAR(N): arying length character
   - ENUM(‘M’,’F’): Value from n defined list
   - BOOLEAN: rue or alse values

2.4.2 Primary and Foreign Keys

Primary Key: It is a column or the sets of columns which very uniquely recognize a document within the given desk. Following are the situations that want to be satisfied for a column to be a number one key:

Foreign Key: It is used to hyperlink two tables collectively. An overseas secret's a column whose values suit the values of other tables’ primary keys column.

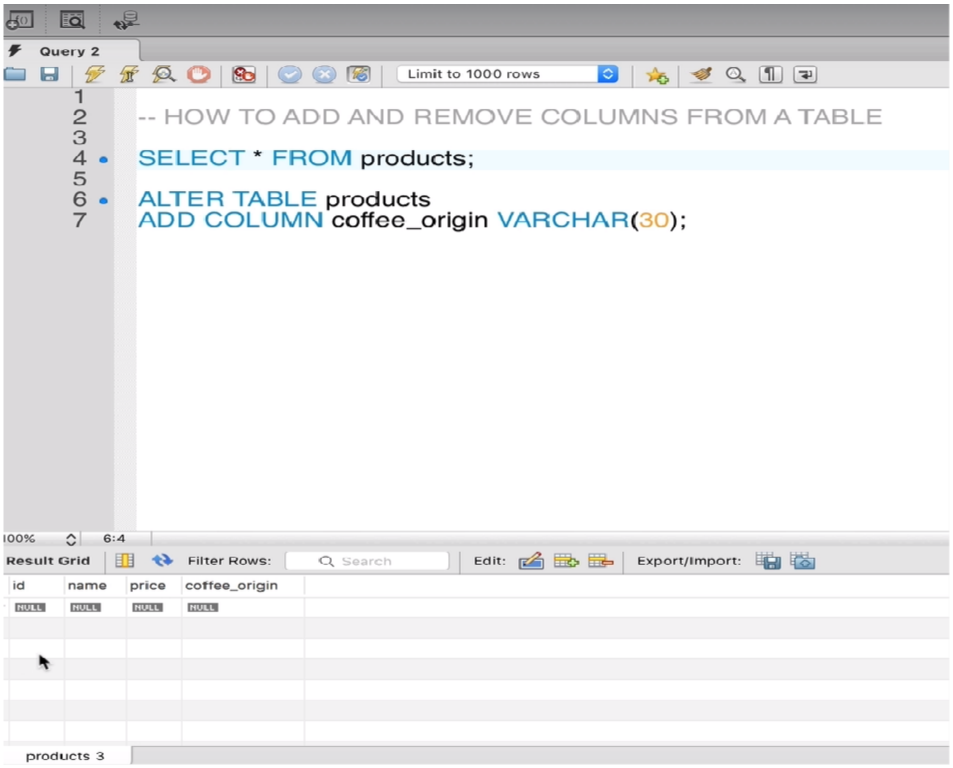


```
CREATE TABLE test_results
(
  name          TEXT,
  student_id    INTEGER    PRIMARY KEY,
  birth_date    DATE,
  test_result   DECIMAL    NOT NULL,
  grade         TEXT       NOT NULL,
  passed        BOOLEAN    NOT NULL
);
```

```
Table test_results created

Execution time: 0.11s
```

Figure 2.10: Create table command

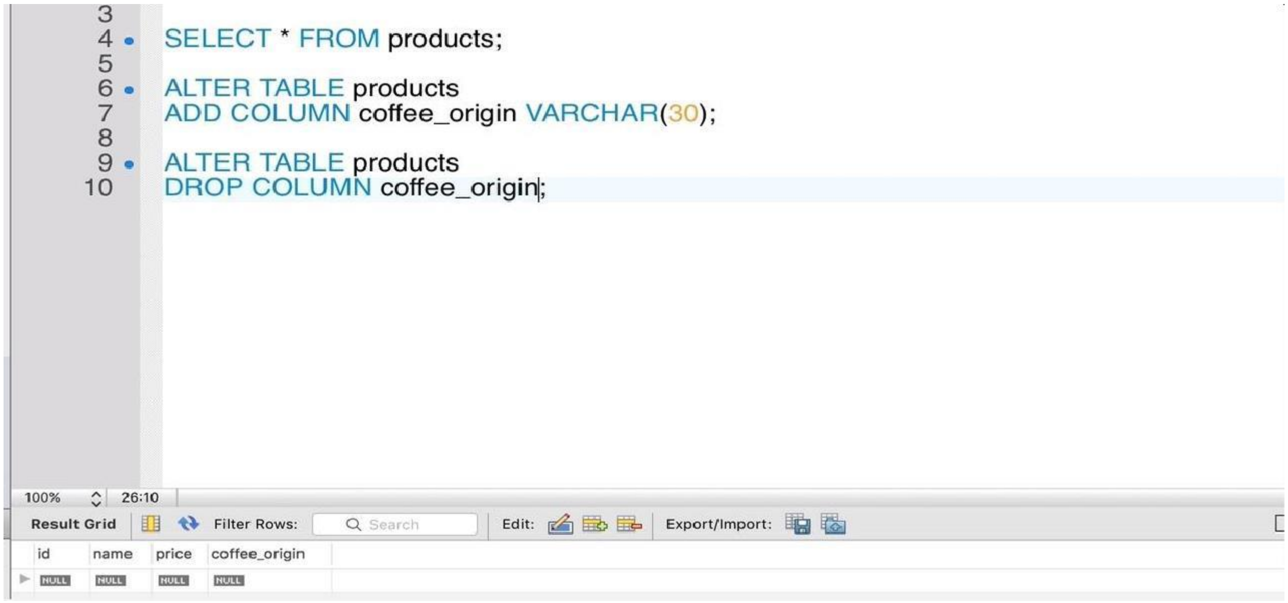
2.4.3 Modifying Tables: Adding and Removing Tables



```
-- HOW TO ADD AND REMOVE COLUMNS FROM A TABLE

SELECT * FROM products;

ALTER TABLE products
ADD COLUMN coffee_origin VARCHAR(30);
```

id	name	price	coffee_origin
NULL	NULL	NULL	NULL

Figure 2.11: Adding col to the table

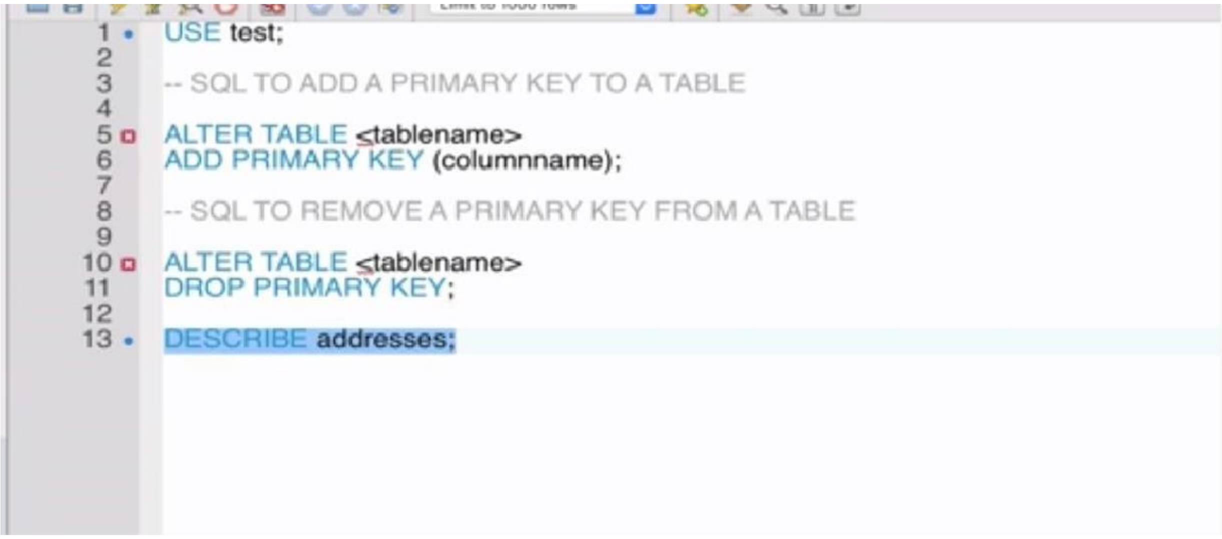


```sql
SELECT * FROM products;

ALTER TABLE products
ADD COLUMN coffee_origin VARCHAR(30);

ALTER TABLE products
DROP COLUMN coffee_origin;
```

id	name	price	coffee_origin
NULL	NULL	NULL	NULL

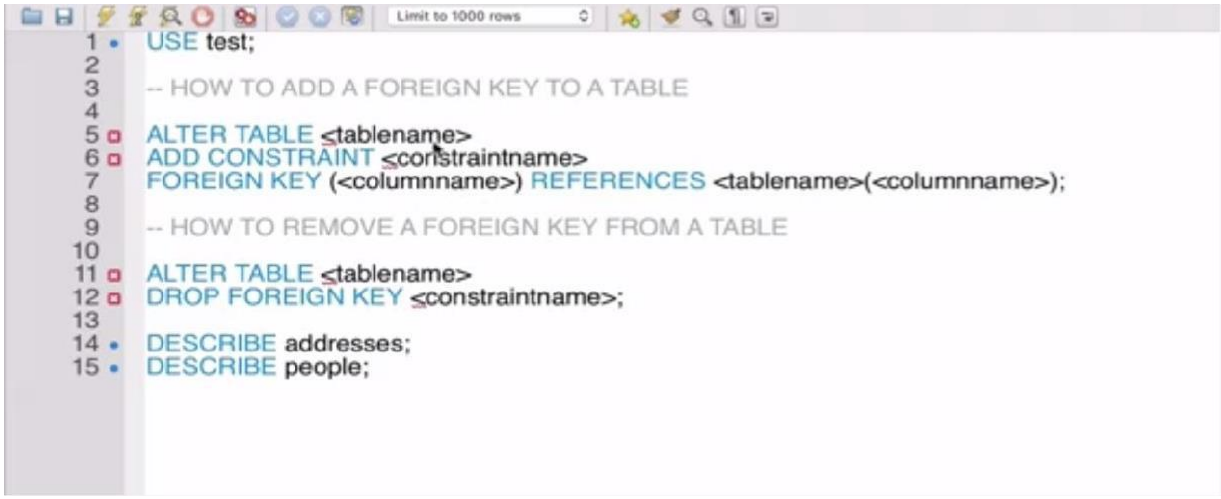
Figure 2.12: Removing a column from the table

2.4.4 Adding and Removing Primary key and Foreign key



```sql
USE test;

-- SQL TO ADD A PRIMARY KEY TO A TABLE

ALTER TABLE <tablename>
ADD PRIMARY KEY (columnname);

-- SQL TO REMOVE A PRIMARY KEY FROM A TABLE

ALTER TABLE <tablename>
DROP PRIMARY KEY;

DESCRIBE addresses;
```

Figure 2.13: Adding and removing primary key

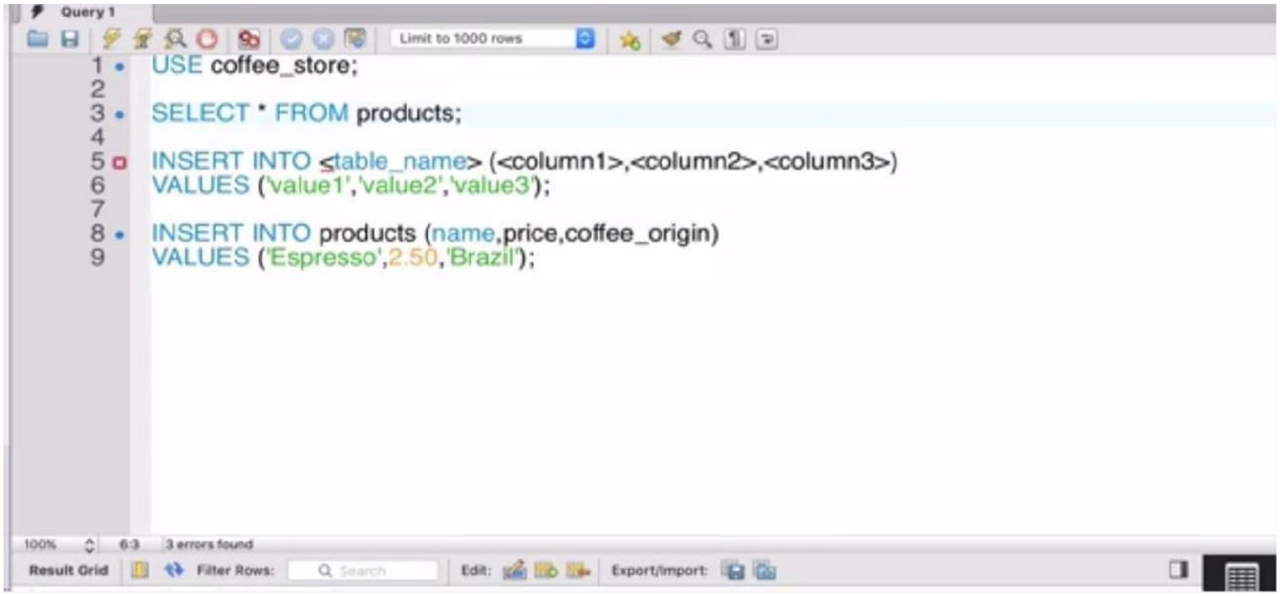


```sql
USE test;

-- HOW TO ADD A FOREIGN KEY TO A TABLE

ALTER TABLE <tablename>
ADD CONSTRAINT <constraintname>
FOREIGN KEY (<columnname>) REFERENCES <tablename>(<columnname>);

-- HOW TO REMOVE A FOREIGN KEY FROM A TABLE

ALTER TABLE <tablename>
DROP FOREIGN KEY <constraintname>;

DESCRIBE addresses;
DESCRIBE people;
```

Figure 2.14: Adding and removing foreign key

2.5 Data Manipulation Language

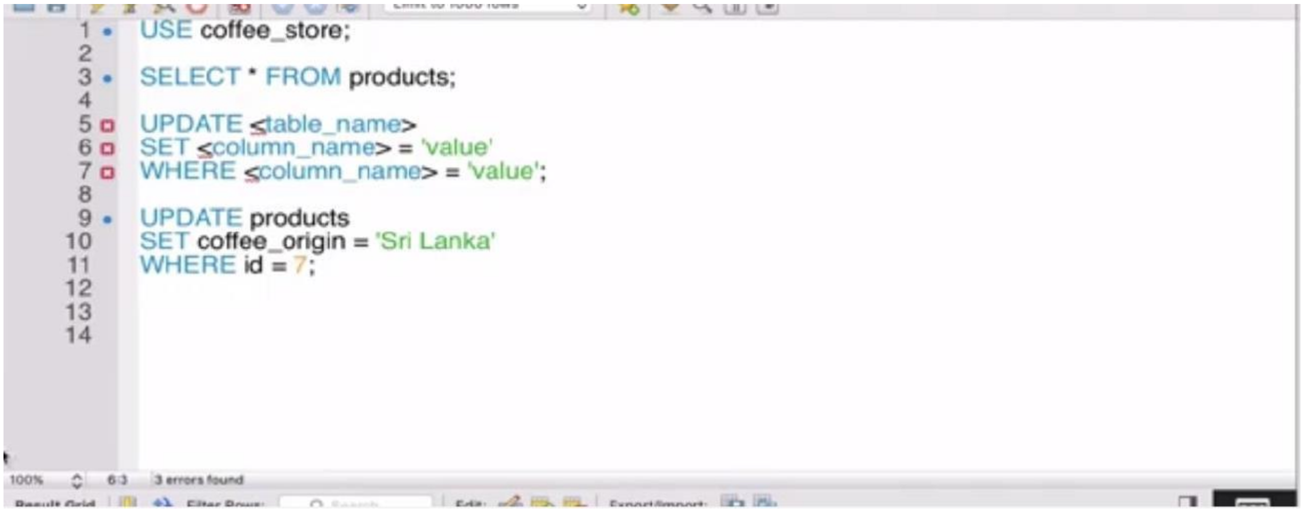
Data manipulations language is a pc language used to update, delete or modify information in a DB. It includes SQL and other languages too, with many operators and functions.



```
USE coffee_store;

SELECT * FROM products;

INSERT INTO <table_name> (<column1>,<column2>,<column3>)
VALUES ('value1','value2','value3');

INSERT INTO products (name,price,coffee_origin)
VALUES ('Espresso',2.50,'Brazil');
```

3 errors found

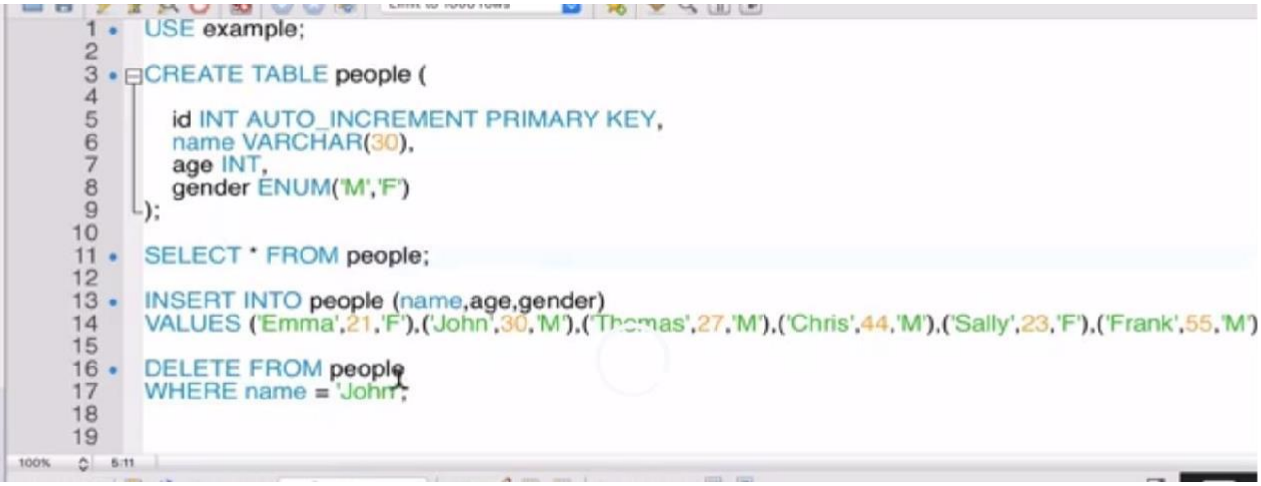
Figure 2.15: Inserting values into tables



```
USE coffee_store;

SELECT * FROM products;

UPDATE <table_name>
SET <column_name> = 'value'
WHERE <column_name> = 'value';

UPDATE products
SET coffee_origin = 'Sri Lanka'
WHERE id = 7;
```

3 errors found

Figure 2.16: Updating values into tables



```sql
USE example;

CREATE TABLE people (

    id INT AUTO_INCREMENT PRIMARY KEY,
    name VARCHAR(30),
    age INT,
    gender ENUM('M','F')
);

SELECT * FROM people;

INSERT INTO people (name,age,gender)
VALUES ('Emma',21,'F'),('John',30,'M'),('Thomas',27,'M'),('Chris',44,'M'),('Sally',23,'F'),('Frank',55,'M')

DELETE FROM people
WHERE name = 'John';
```

Figure 2.17: Deleting from the table

2.6 About Joins

A JOIN clause is used to mix rows from two or greater tables, based on a related column between them. There are many one of a kind styles of joins together with inner be part of, left be part of, proper be part of.

2.6.1 Inner Join

An inner join will retrieve information simplest while there's matching values in both the tables.

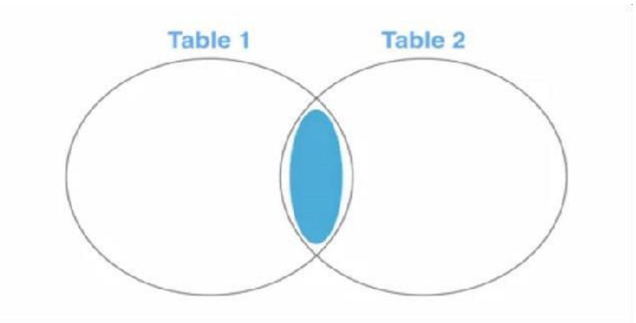
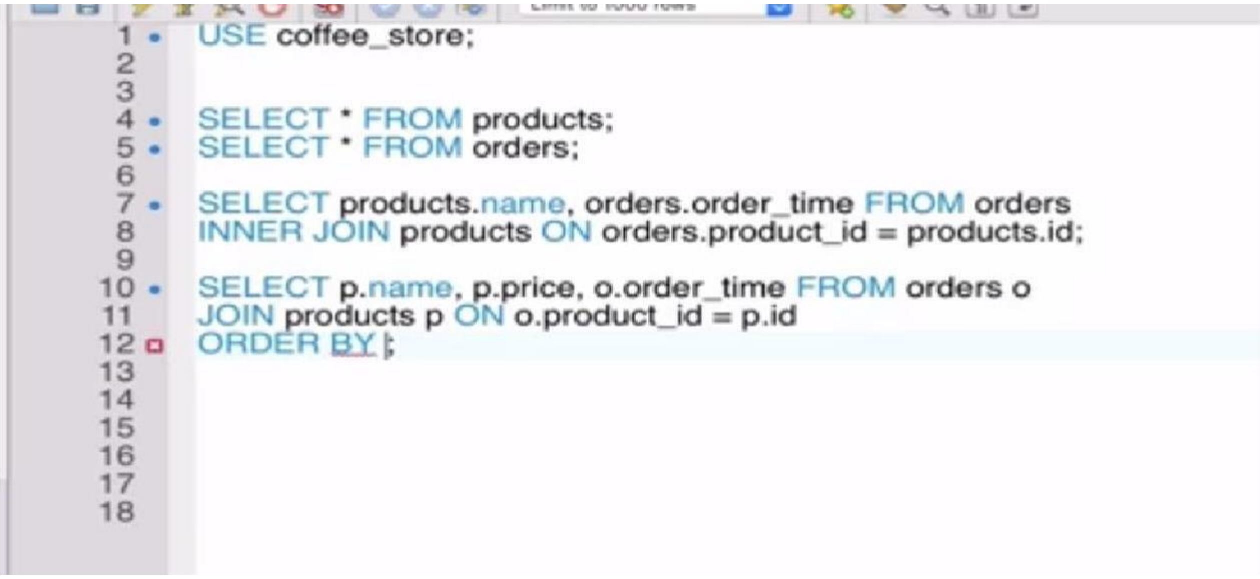


Figure 2.18: Inner joins



```
USE coffee_store;


SELECT * FROM products;
SELECT * FROM orders;

SELECT products.name, orders.order_time FROM orders
INNER JOIN products ON orders.product_id = products.id;

SELECT p.name, p.price, o.order_time FROM orders o
JOIN products p ON o.product_id = p.id
ORDER BY ;
```

Figure 2.19: Example query of inner join

2.6.2 Left Join

Left be part of will retrieve all of the statistics from the left table and the matching rows from the proper side of the desk.

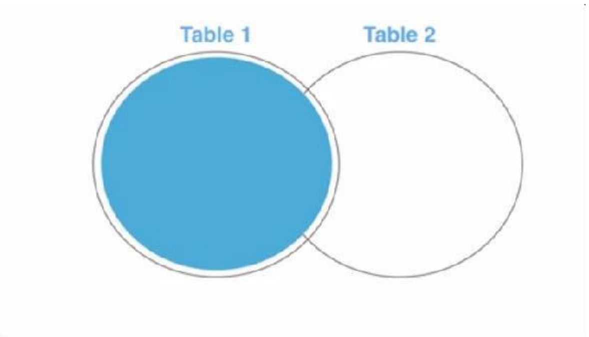
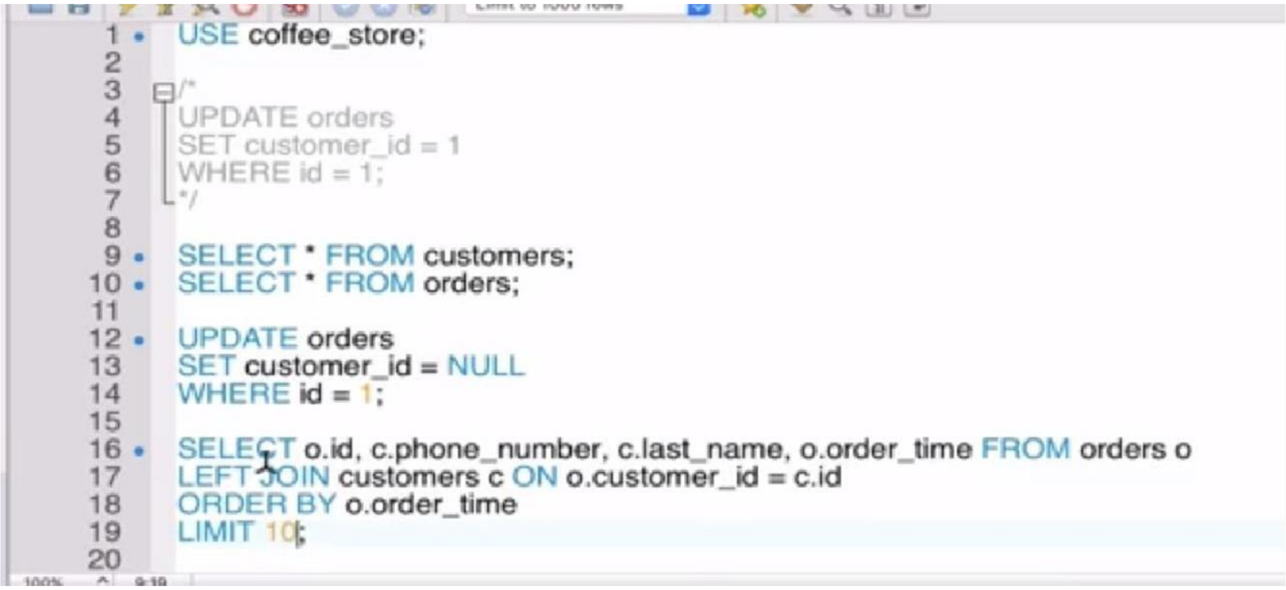


Figure 2.20: Left join example



```
USE coffee_store;

/*
UPDATE orders
SET customer_id = 1
WHERE id = 1;
*/

SELECT * FROM customers;
SELECT * FROM orders;

UPDATE orders
SET customer_id = NULL
WHERE id = 1;

SELECT o.id, c.phone_number, c.last_name, o.order_time FROM orders o
LEFT JOIN customers c ON o.customer_id = c.id
ORDER BY o.order_time
LIMIT 10;
```

Figure 2.21: Example query of left join

2.6.3 Right Join

It will retrieve the data from the right table and corresponding rows of the left table.

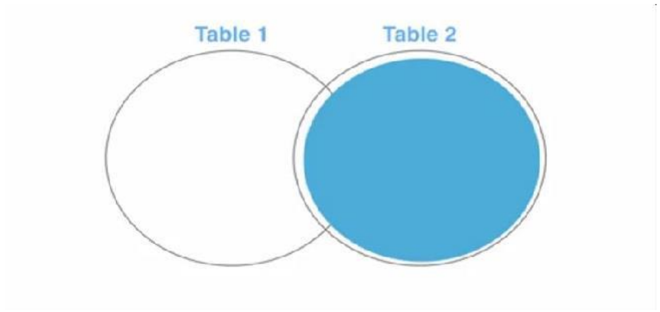
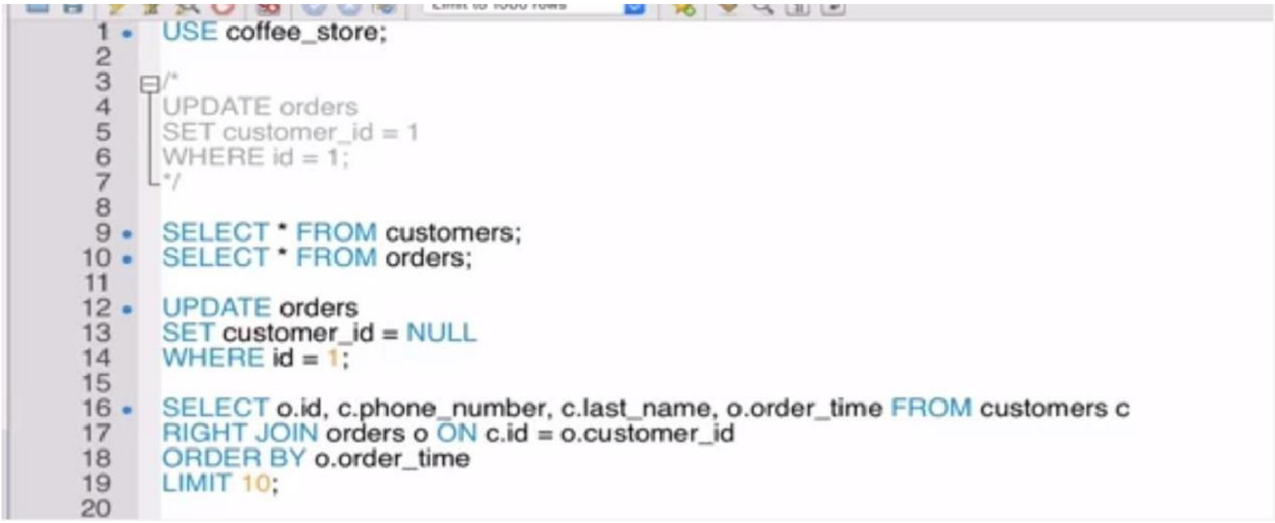


Figure 2.22: Right join



```
USE coffee_store;

/*
UPDATE orders
SET customer_id = 1
WHERE id = 1;
*/

SELECT * FROM customers;
SELECT * FROM orders;

UPDATE orders
SET customer_id = NULL
WHERE id = 1;

SELECT o.id, c.phone_number, c.last_name, o.order_time FROM customers c
RIGHT JOIN orders o ON c.id = o.customer_id
ORDER BY o.order_time
LIMIT 10;
```

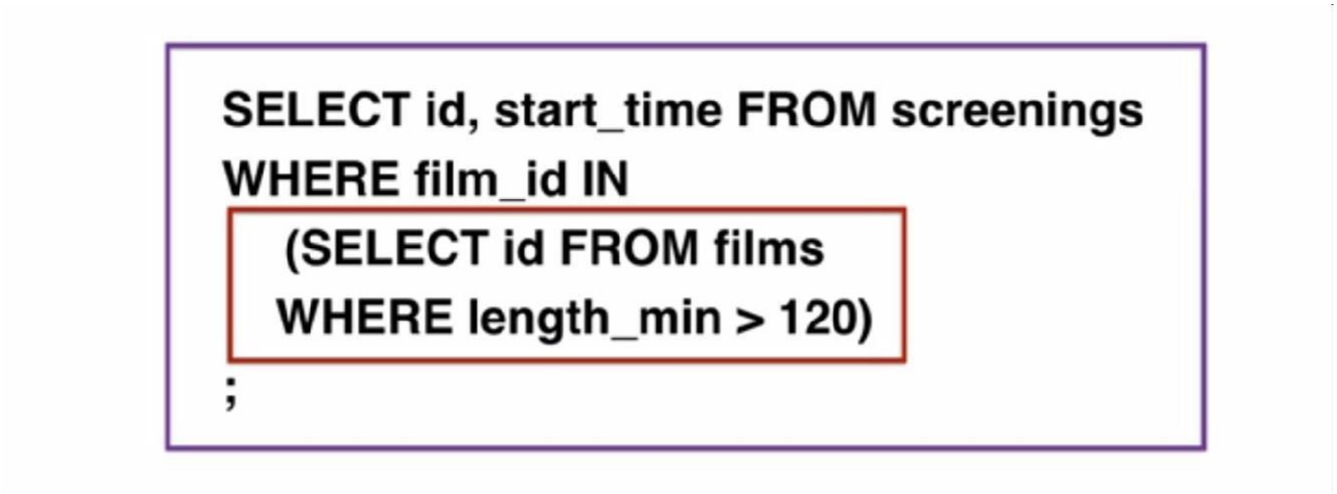
Figure 2.23: Example query of right join

2.7 Subqueries

It is a inquiry this is put in different statements, or inside every other subquery. A subquery may be used wherever an appearance is allowable.

2.7.1 Non-Correlated Subquery

In this sort of subquery, the internal query can run independently of the outer question. Inner queries run first and produces a end result set and which is then utilized by the outer query.

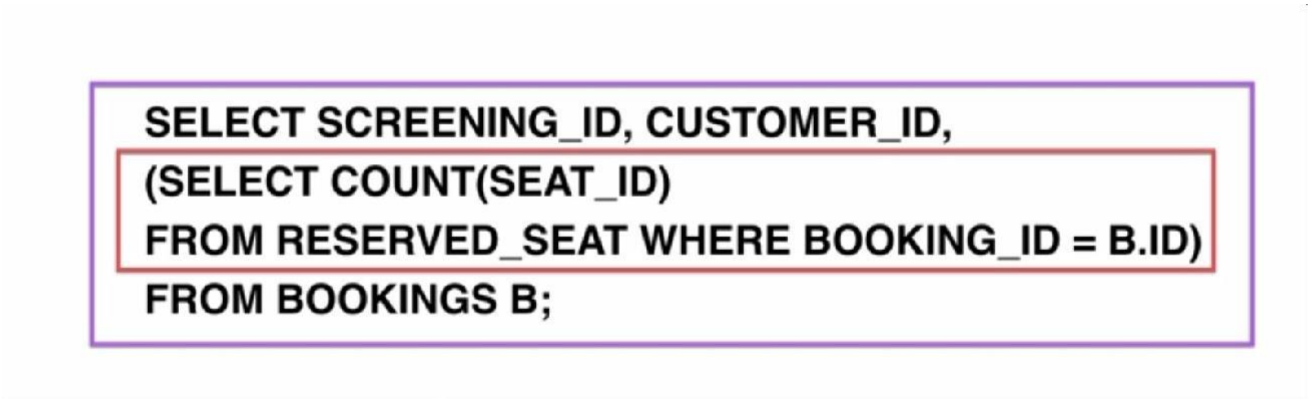


```
SELECT id, start_time FROM screenings
WHERE film_id IN
  (SELECT id FROM films
  WHERE length_min > 120)
;
```

Figure 2.24: Non-correlated query

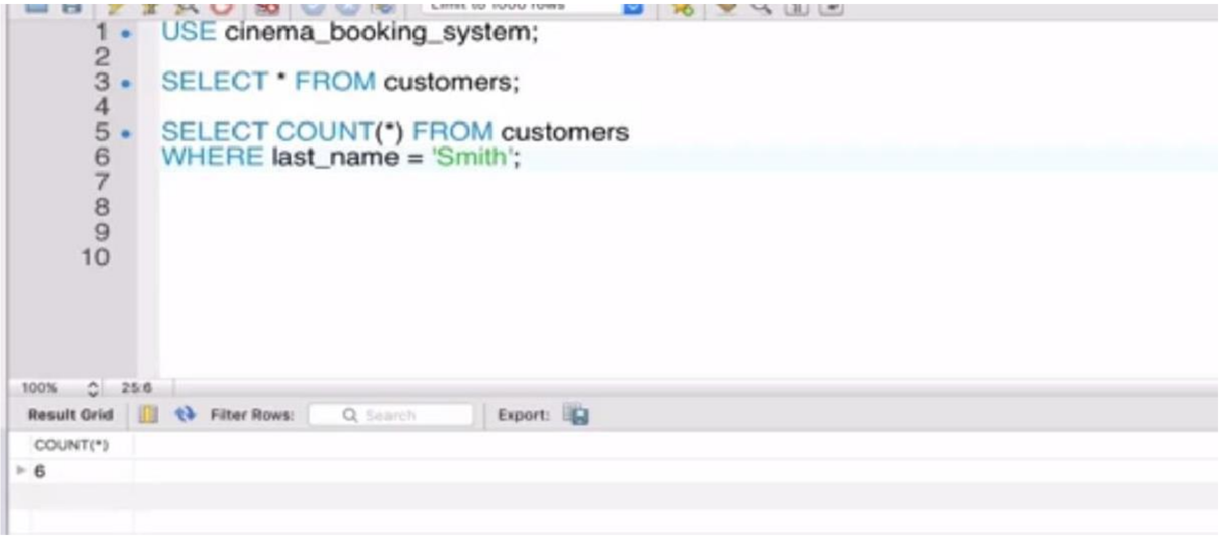
2.7.2 Correlated Subquery

In this sort of subquery, the inner question runs for each row inside the outer question. Here the internal query can't run independently of the outer question.



```
SELECT SCREENING_ID, CUSTOMER_ID,
(SELECT COUNT(SEAT_ID)
FROM RESERVED_SEAT WHERE BOOKING_ID = B.ID)
FROM BOOKINGS B;
```

Figure 2.25: Correlated subquery



```
USE cinema_booking_system;

SELECT * FROM customers;

SELECT COUNT(*) FROM customers
WHERE last_name = 'Smith';
```

```
COUNT(*)
6
```

Figure 2.26: Count function example

CHAPTER 3

CLOUD

3.1 Virtualization basics

In some systems, while booting up, some kind of error like

Keyboard, mouse not found.

Press ctrl+alt+del to reboot.

- **So is keyboard, mouse mandatory/necessary stuff to boot?**

BIOS error

- **Who is Charles Babbage?, Why is he called the father of computing, not the inventor?**

gave the concept, but did not design/invent it.

- **Is there any company who claims that we invented the computer?**
  If no one was there, how did it evolve?

Charles

ON his timeline, if they want to do any calculation, they need to use the device called abacus.

3.2 Evolution-

1. i/o unit - Punch cards, flash cards, magnetic tape,magnetic hard drive,hard drive,floppy,cd,dvd,pen drives,ssd.
2. Computing unit
3. Storage unit
4. RAM - DDR1-4

- People usually face some issues while sharing resources from apple to android, direct storage to storage, why?

1950-1973, this was the same problem on many companies' computers.

- **ISO** - came with some rules and regulations for designing the processor like the tcp/ip model.

- Why do we have a temp and perm architecture as RAM and HARDDRIVE separately**?**

Once you turn on the computer, it will look for CMOS.
All the circuits are working fine before booting.
CMOS stands for "Complementary Metal Oxide Semiconductor

BIOS- take care of hardware utilisation during the booting process and all the units are functionally working fine.

If computer satisfy above two conditions
it will enter into boot sector(where os is placed)

Copy the contents from boot to RAM(temp. memory speed up 10x tham harddrive)
Shows you the logon screen(Os content that is sitting on ram)
With your real hardware so many terms and conditions are there to boot up the process.

It allows you to run multiple operating systems on the same hardware.

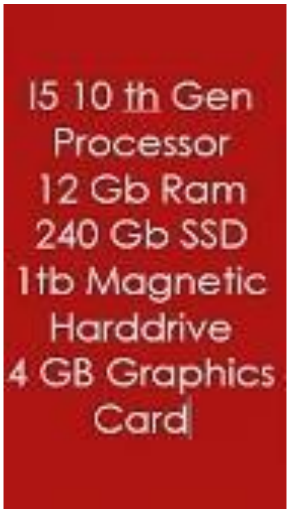
- Is it possible to work on two OS at the same time without VMware, Virtual box, directly?

no

- How will it allow multiple OS on the same hardware using the virtualization software?

Virtualization is nothing but a software that creates a fake environment and lies to the operating system those terms and conditions required by the OS are physically available there, to save the hardware cost. It has its own or kind of a BIOS.

3.3 Types of Virtualization



I5 10 th Gen
Processor
12 Gb Ram
240 Gb SSD
1tb Magnetic
Harddrive
4 GB Graphics
Card

Figure 3.1

Type 1 runs on bare metal

Type 2 runs on top of an OS.

3.3.1 Example

One to many (sharing concept)

Example-
Company - 10 employees
Working on latest Microsoft technology projects
10 laptops- lifetime of the hardware (how long your current hardware will support for all the latest software)

50k*10=5L; 5L/24=20833 you are investing only on hardware in a single shot payment.

If some company provides hardware for 10k/month on rent, which option will you choose?

Let in the company we have 5 employees, and they bought a huge server, and divided that into multiple virtual machine architecture..

3.4 Types

3.4.1 Type 2

This runs as a software, it needs an OS to install it as application and run it as an architecture.
If something goes wrong, control shifts to OS.
(First priority goes to parent OS, then hyper OS)
Ex- VM ware workstation, virtual box, QEMU

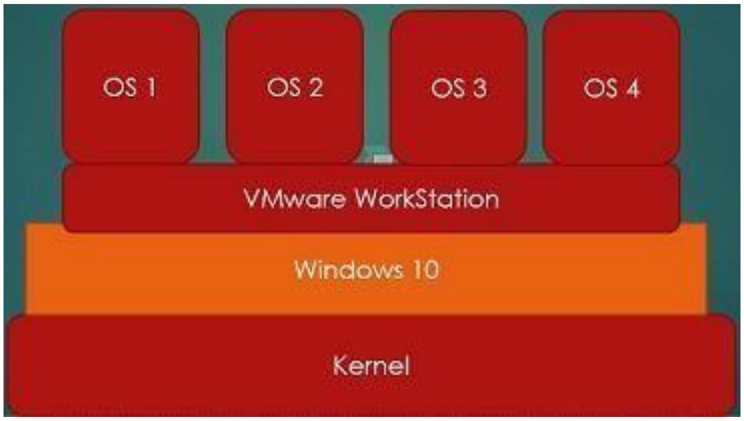


Figure 3.2

3.4.2 Type 1

Architecture itself is an OS.

Whole control goes to the hypervisor (assign resources to all virtual machines equally) as an OS.

Ex- VMware esxi, open stack

We do not need the windows layer in type1. OS especially created to manage the hypervisors or top-level layers.

It will allocate 25% to all in parallel manner.

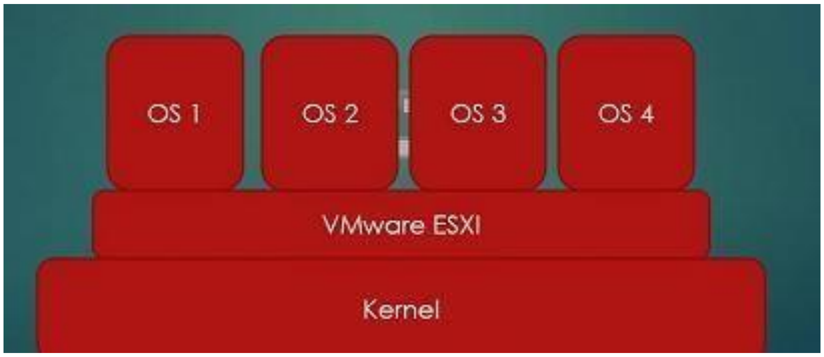
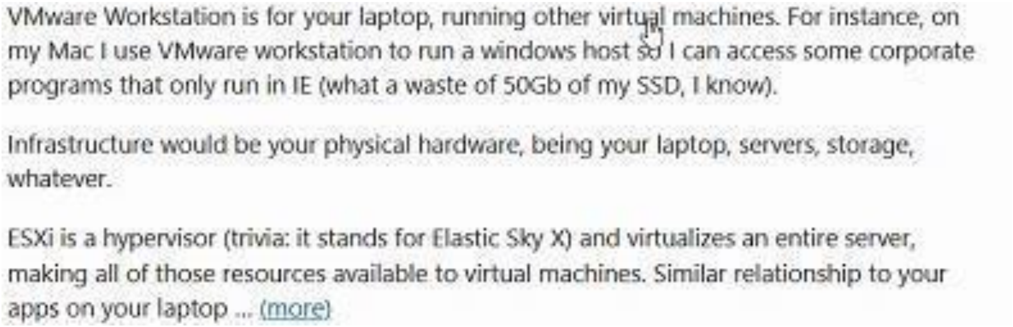


Figure 3.3

Type 1 is the best and efficient hypervisor (virtualization) technique.



VMware Workstation is for your laptop, running other virtual machines. For instance, on my Mac I use VMware workstation to run a windows host so I can access some corporate programs that only run in IE (what a waste of 50Gb of my SSD, I know).

Infrastructure would be your physical hardware, being your laptop, servers, storage, whatever.

ESXi is a hypervisor (trivia: it stands for Elastic Sky X) and virtualizes an entire server, making all of those resources available to virtual machines. Similar relationship to your apps on your laptop ... (more)

Companies offering hypervisors (type 1/2) as a service are meant to be Cloud providing vendors.

3.5 Methods via, they offer their services-

- public cloud
- private cloud
- hybrid cloud

3.5.1 Public-

Server that is used as a hypervisor and has been shared to a group of companies or groups of people from different sectors, called a public cloud.

Famous vendors-
AWS, Google, Azure, Oracle, IBM, Dell emc, Digital Ocean

3.5.2 Private-

For the same set of people/company.

All machines in the same server use the same resource. Security risk minimized.

3.5.3 Hybrid-

Same community of people sharing the same server, known as hybrid cloud. Ex- Facebook

CHAPTER 4

AWS

4.1 Basics

Offer you the services in multiple regions and technologies.
Cloudping.info to check the regions

Out of 24, two are Gov. Cloud, used strictly for govt. purposes.

- **In which circumstances, we give preference to regions?**

I am running a company in India, but all my clients are in Australia, where I will deploy my server?
Australia.

Latency, cost of Aus is less than USA.

- **Rates vary in AWS for services/regions, both.**

4.2 Services

EC2	Compute
EMR	Analytics
Kinesis	Analytics
S3	Storage
DynamoDB	Database
Lambda	Compute
Redshift	Database
Glue	Analytics
VPC	Networking & content delivery
IAM	Security, identity and Compliance
Athenna	Analytics

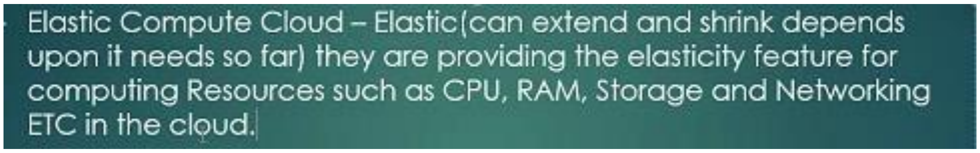
4.3 IAM

Restricted access to resources/services.

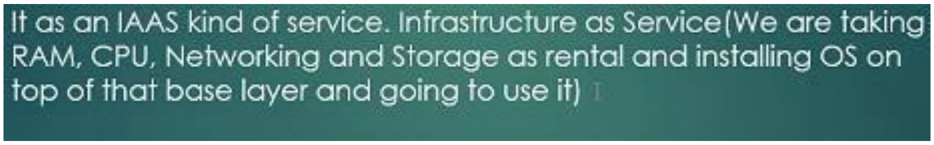
Active Directory

Creates its own network, under that put all the users on the same n/w for securely accessing the resources under the same n/w by the users belonging to the same n/w.

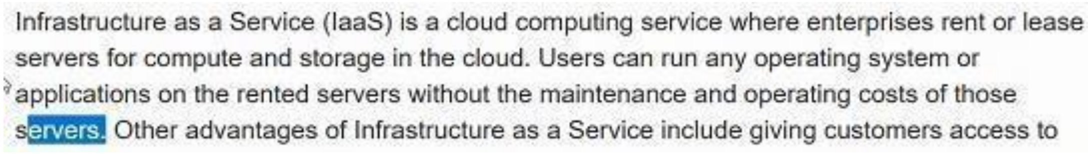
4.4 EC2



Elastic Compute Cloud – Elastic(can extend and shrink depends upon it needs so far) they are providing the elasticity feature for computing Resources such as CPU, RAM, Storage and Networking ETC in the cloud.



It as an IAAS kind of service. Infrastructure as Service(We are taking RAM, CPU, Networking and Storage as rental and installing OS on top of that base layer and going to use it)



Infrastructure as a Service (IaaS) is a cloud computing service where enterprises rent or lease servers for compute and storage in the cloud. Users can run any operating system or applications on the rented servers without the maintenance and operating costs of those servers. Other advantages of Infrastructure as a Service include giving customers access to

- Instance- virtual machine
- EBS - elastic block size

4.5 AWS CLI Commands

```
aws configure
ap-south-1
aws s3 help
aws s3 ls
aws s3 mb s3://pranav-awsdemo
aws s3 rb s3://pranav-awsdemo /bucket is not removed if it is not empty.
```

```
aws s3 rm s3://karthick1808/index.html /for file
aws s3 rm s3://karthick1808/css --recursive /for folder
aws s3 rm s3://karthick1808 --recursive
```

recursive - for all,select all kind of

dir local directories

```
aws s3 ls s3://pranav-1
aws s3 cp Desktop\aws11.txt s3://pranav-1
```

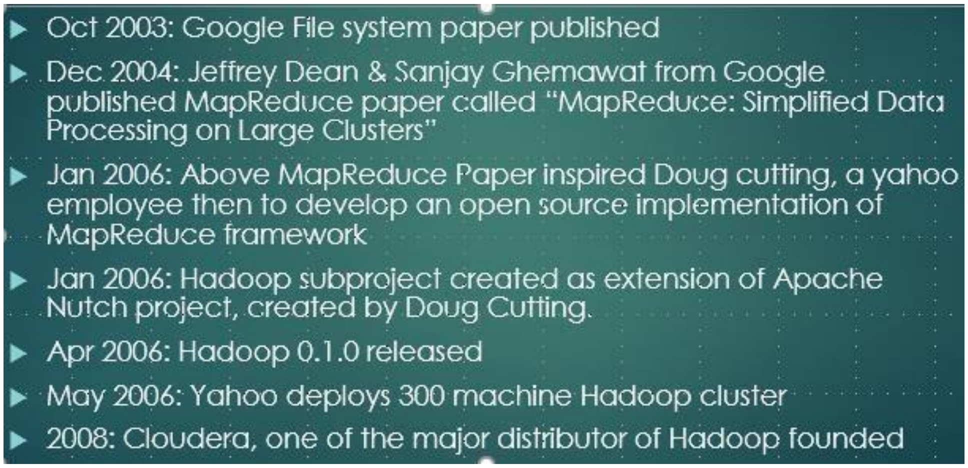
Remember- In path, \ is for windows, / for linux

use tab/enter in cli if more appears

CHAPTER 5

HADOOP

5.1 History of Hadoop



- Oct 2003: Google File system paper published
- Dec 2004: Jeffrey Dean & Sanjay Ghemawat from Google published MapReduce paper called “MapReduce: Simplified Data Processing on Large Clusters”
- Jan 2006: Above MapReduce Paper inspired Doug cutting, a yahoo employee then to develop an open source implementation of MapReduce framework
- Jan 2006: Hadoop subproject created as extension of Apache Nutch project, created by Doug Cutting.
- Apr 2006: Hadoop 0.1.0 released
- May 2006: Yahoo deploys 300 machine Hadoop cluster
- 2008: Cloudera, one of the major distributor of Hadoop founded

Figure 5.1

5.2 What is Hadoop?

Apache **Hadoop** is an open-source framework for distributed storage and processing of very large data on clusters, i.e. group of computers used for extended storage and performance.

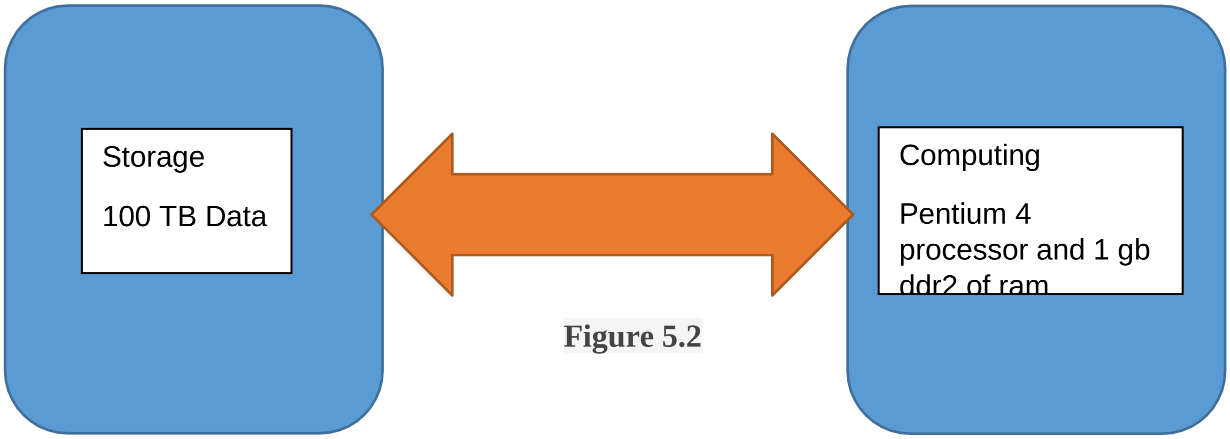


Figure 5.2

Large facts is organized records collected by the use of establishments that can be applied in AI ventures, prescient demonstrating and specific stepped forward exam packages. Frameworks that method and store large statistics have become a standard a part of facts the executives' designs in institutions. Enormous information is often portrayed thru the 3Vs: the huge extent of facts in numerous conditions, the huge collection of data sorts put away and the speed at which the records is produced, accumulated and prepared. These tendencies were first diagnosed with the resource of Gartner similarly promoted them after it acquired Meta Group in 2005. The entire extra as of overdue, some fantastic Vs have been delivered to numerous depictions of huge information, collectively with veracity, well well worth, and fluctuation.

Albeit massive facts do no longer liken to a specific extent of facts, large records corporations frequently consist of terabytes (TB), petabytes (PB), or even Exabyte (EB) of facts stuck after some time.

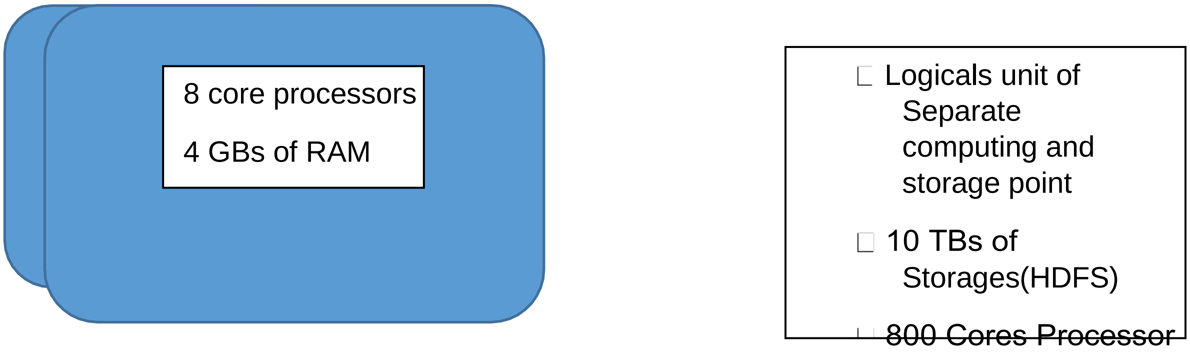

8 core processors
4 GBs of RAM
□ Logicals unit of Separate computing and storage point
□ 10 TBs of Storages(HDFS)
800 Cores Processor

Figure 5.3: block diagram

Hadoop Solves Big Data Problems

Storing Big Data was no more a Problem

And Processing did not take Years

Figure 5.4: Hadoop solves the big data

Hadoop come up with two components that are:

- Storage (HDFS)
- MapRedduce

5.3 Hadoop 1.0

MapReduuce (processing unit) for Cluster resource management and data processing:

- Master
- Slave

5.3.1 Components

Some components are as follows:

- HDFS(Storage)
- Name Node
- Data Node

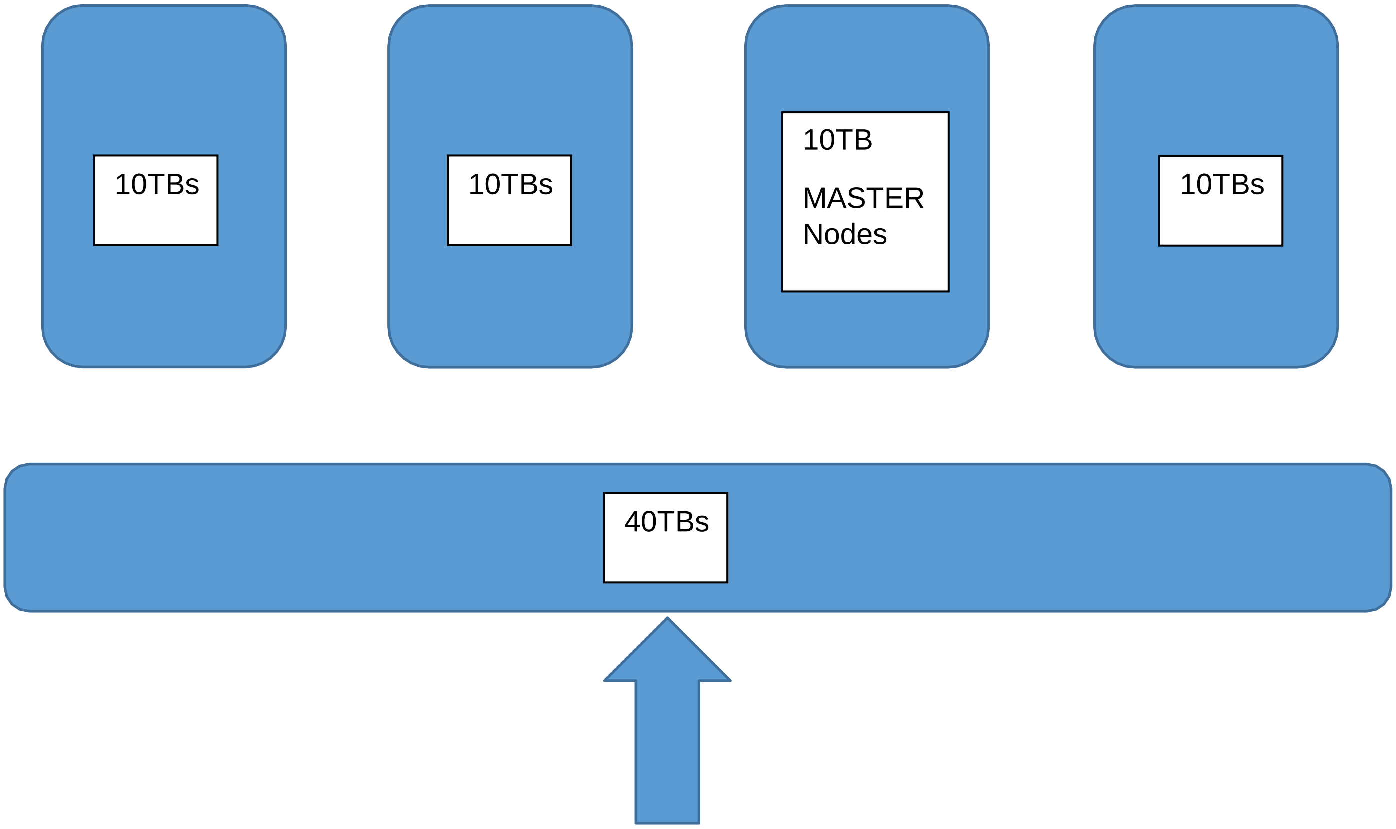
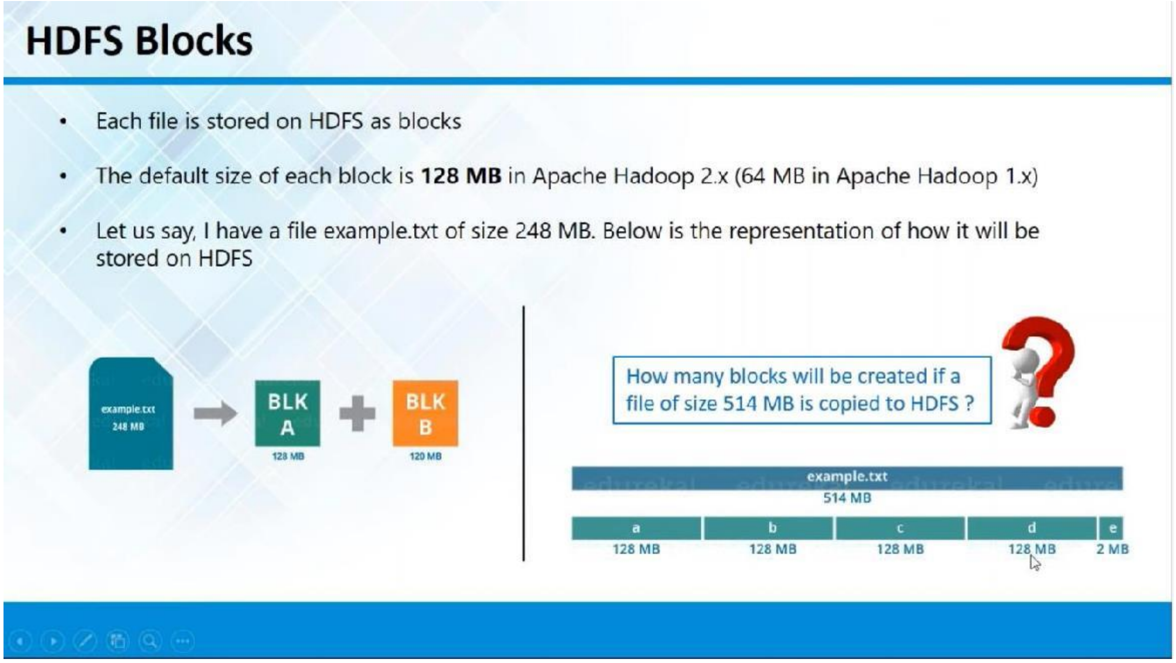


Figure 5.5: Master and slave node

All the files, which might be stored within the format of block. Each blocks can be inside the size of 64MB in Hadoop 1.0 and 128MB in Hadoop 2.Zero. For instance, I got document length will in 500 Mb method: In 1.Zero architecture first 7 blocks will 64mb last block can be in 52mb length and in 2.0 architecture first 3 can be 128MB remaining 116 MB.



HDFS Blocks

- Each file is stored on HDFS as blocks
- The default size of each block is **128 MB** in Apache Hadoop 2.x (64 MB in Apache Hadoop 1.x)
- Let us say, I have a file example.txt of size 248 MB. Below is the representation of how it will be stored on HDFS

How many blocks will be created if a file of size 514 MB is copied to HDFS ?

Figure 5.5: HDFS block

5.3.2 Rack Awareness Algorithm

Rack focus is having the knowledge of Cluster topology or greater specifically how the special records nodes are dispensed throughout the racks of a Hadoop cluster. The importance of this knowledge is based on this assumption that collocated records nodes inner a particular rack will have more bandwidth and less latency while information nodes in separate racks may have relatively less bandwidth and better latency.

The main purpose of Rack awareness is:

- Increasing the availability of data block
- Better cluster performance
- Usually Hadoops clusters of more than 100 nodes are configured in multiple racks...

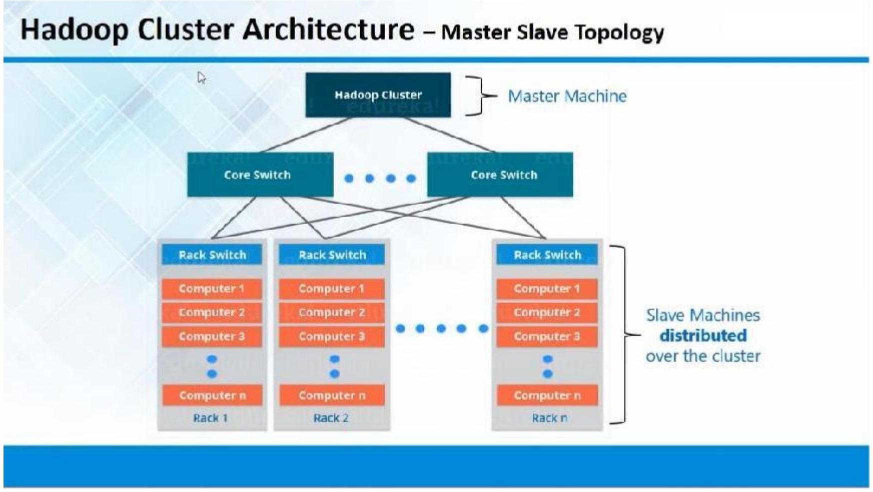
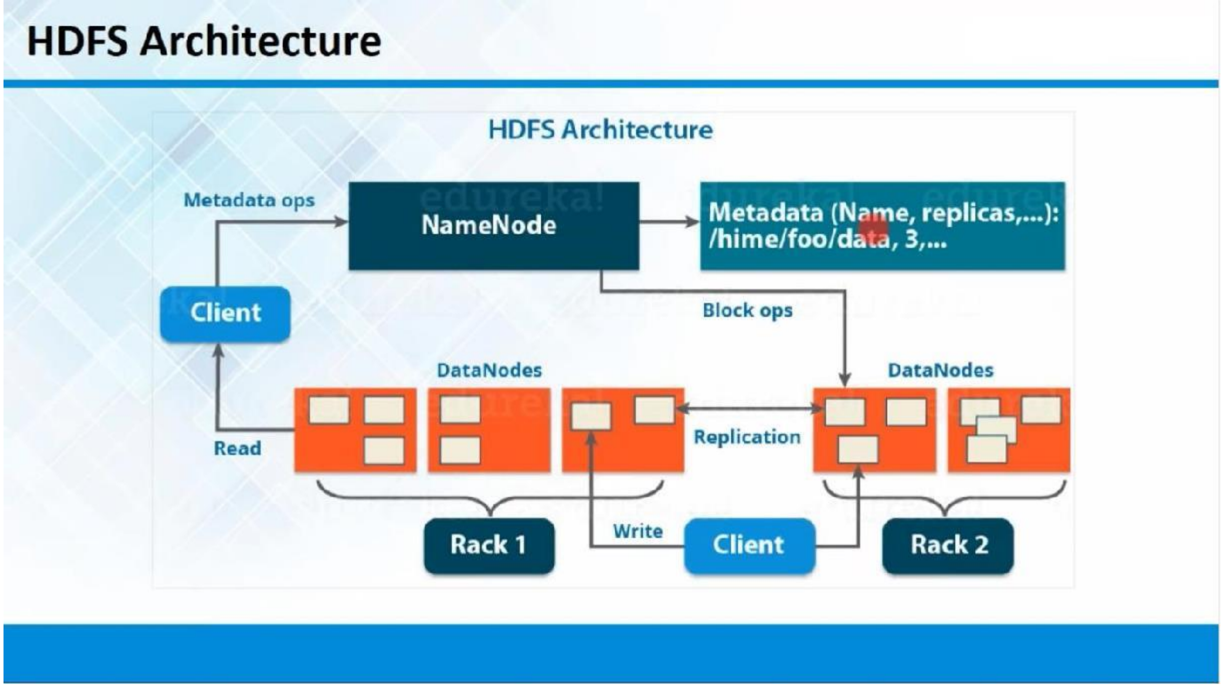


Figure 5.6

5.4 HDFS Architectural



HDFS Architecture

Figure 5.7: Block diagram of HDFS

5.4.1 Job Tracker and Task Tracker in MapReduce

- Job Tracker and Task Tracer are the components of MapReduce.
- Job Tracker always available on Master

5.4.2 Hadoop 1.0 vs 2.0

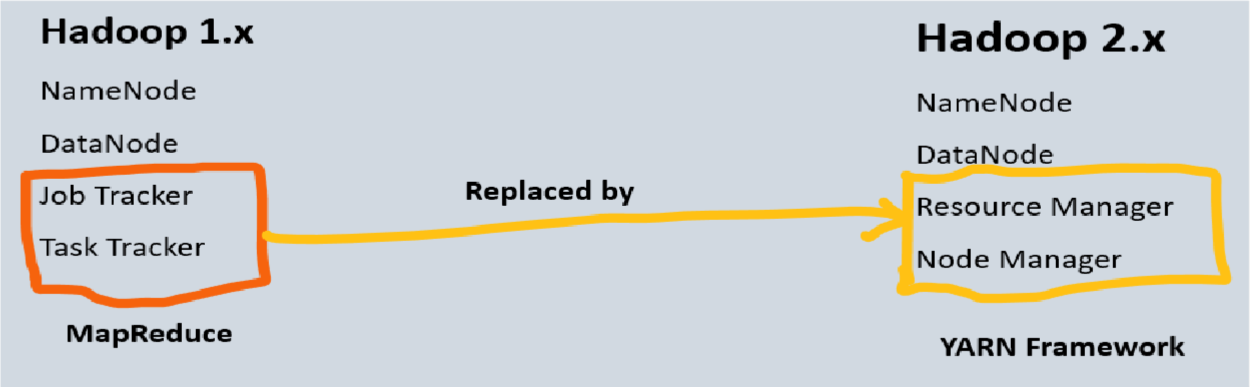


Figure 5.8

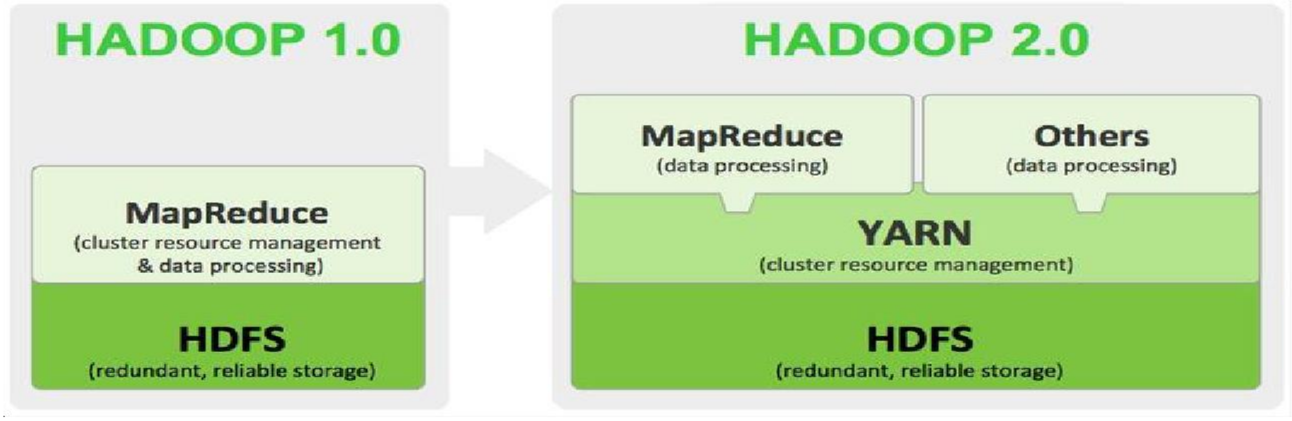

HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

Figure 5.9: Block diagram of Hadoop 1.0 and 2.0

- Because of obstacles to be able to be there in 1.Zero, it'll be given java as enter unit
- Job tracker will receive the input simplest as java language
- In Hadoop Java programming men only capable of work in this structure
- In 2.Zero YARN it will receive the input as multi programming language and software'sEx: Python/Hive

- Datacenter are group of servers avail for storing and retrieving data in same location Datacentre Rack
- rack is the electronic framework of switches assembling different servers in datacenter

5.4.3 Installation of Hadoop

Types of Hardware's used for installation:

- Manual Installation
- Ready to run Environment in Public Clouds(EMR)
- Ready to run in Private Cloud/ Virtual Machine (Cloud ERA/Horton Works)

Need of JVM

Java Virtual Machine (**JVM**) is a medium between the operating system and the Java application (**Hadoop** in this case).

- The system will deal with all the data's as byte codes
- Chrome -> YouTube video
- Bytes by byte(Streaming)
- How fast it will stream your video quality will be good
- We are going to deal with a huge amount of data, if I go by traditional processing it will consume more and more time even though very good processing unit will be there
- JVM- will deal the kernel data will be in the blocks format (it will be much faster than traditional approach)

Hadoop

Hadoop is the most effective meant for Analytical purposes. All the applications to give attention to study they gained’t concentrate on write. All the commands we execute in Hadoop structure ad not anything but predefined scripts. If I get any custom-designed requirement method what's going to I do?

Apache Sqoop is an tool designed for efficiently transferring bulk records between Apache Hadoop and based datastores which includes relational databases.

5.5 Sqoop

Apache Sqoop(TM) is a device intended for proficiently moving mass information between Apache Hadoop and organized datastores, for example, social databases.

5.5.1 Instructions to Import Sqoop

- sqoop import
- --connect
- jdbc:mysql://karthick1808.c5e69p24nuic.ap-south-1.rds.amazonaws.com/aman2910
- --username karthick1808
- --password 12345678
- --table cus_tbl1
- --incremental append
- --check-column cus_id
- --last-value 5
- --target-dir /amansql3
- -m 1

5.5.2 Hive

Hive tutorial presents fundamental and advanced standards of Hive. Our Hive tutorial is designed for novices and professionals.

Apache Hive is an information ware gadget for Hadoop. Facebook evolved the hive. It supports Data definition Language, Data Manipulation Language, and person-defined capabilities.

Our Hive academic includes all topics of Apache Hive with Hive Installation, Hive Data Types, Hive Table partitioning, Hive DDL commands, Hive DML commands, Hive type by vs order through, Hive Joining tables and many others.

- SQL like tool (Command line tool) that will run on top of the pdf file system.
- We can write the java program for every searching what we are making so far for the replace of programmatic interface they came up with an API based tool called a hive.
- Hive is the distributed tool that will run on the Hadoop cluster that will support both Mapreduce and hdfs.
- If in the case of searching large data sets also it will acquire the resource from Mapreduce.
- Import that CSV to Hadoop master machine depends on the data's going to create tables in the hive and trying to import that CSV file to the hive

Chapter 6

Big Data

6.1 Introduction

Suppose we need to install a game software, we need to have some basic hardware requirements, and if we do not meet them, we cannot install it.

Same with Big Data, we need to have some hardware.

Because of some limitations available to install the Big Data, cluster software we are approaching some vendor (hardware) to install the software.
like AWS, Google, Azure etc.

6.2 Need of Big Data

Ques1

If I have 50 MB of data, do i need the big data architecture to handle this particular data?

All the mail companies set some limitations on max. Data you can send, say 30 MB, then this 50 mb will not be supported,

We usually go for google drive and share the link.

In that case, we are looking for some alternate solution to fix this,
Big data is one kind of solution.

Ques 2

Max. Amt. of data MySQL can handle by terms of data(volume) and computing?

64 TB enterprise edition

Can handle only 64 gb, but we go for extra 10 gb
You may have to deal with the decrease in the performance of the soft.

If your system support capability exceeds by volume, so option to replace that volume is nothing but big data.

Oracle Standard edition supports v32CPUs. It's not by volume here, like generally these days we have 64 bit CPUs.

6.3 Definition

If your system/ software is incapable to handle volume/computing resource to process your data(small/large)

In those cases, Big Data will be your solution to the limitations of databases.

Ques 3

What is the difference between OLAP and OLTP?

Transactional and analytical Processing.

- **How do you categorize the data?**

- **Financial concerns are still using databases in the bank?**

Advised not to use OLAP in the data which keeps on modifying, updating etc.

- **Which one is faster in transactional and analytical?**

Depends on circumstances.

Example - we are using a database for select query, let the data is not modifiable.

OLAP is much better here,just need to retrieve the data.

Example - we are using a database for insert, delete or modify query,

OLTP is good.

6.4 Cluster

6.4.1 Definition

Facebook vs Yahoo/Orkut

They worked on the database architecture.
facebook - Apache HIve - SQL like interface to query the data

Why did they come up with this?

We are giving the servers for database management but you will need server-like hardware, it will not work on commodity hardware like computers, laptops etc. So buy one from us then use our software.

Instead of using integrated hardware, you can use your commodity hardwares to build this structure called cluster.

6.5 Apache

APACHE is an open source foundation, all the software they maintain are freewares.
Multi OS supported.
If you go for windows, these softwares are not fully supported.
Linux- 100%.

It was not easy to remember all the commands, so there were two companies what they did , they created a web page oriented software above the linux, for user friendliness

Cloudera
Hortonworks

Ubuntu machine created using aws

ubuntu@13.234.217.250
password toor to enter the machine

EC2

Web service for developers to provide them secure and resizable compute capacity in the cloud.

- easy scaling up/down
- pay for what you use
- work from home env./anywhere
- Cost eff.
- can be integrated with other services easily

Steps to use EC2:

1. **AMI** - Softwares, OS, access permissions, volume info and apps we need to run our package, template that is used to create a new instance/machine.
   Predefined and Custom AMIs

2. instance - type and size of H/w

   a. Compute optimised - apps, which require lots of processing power.
   b. Memory optimised - which requires more of in-memory cache.
   c. GPU optimised - for gaming or large gaming requirements.
   d. storage optimised - for storage
   e. General purpose - everything is equally balanced.
   These instance types are fixed and cant be altered.

3. **Configure instances VPC** - how many instances i need/ and in which subnet we need them.

   a. Do I want a public IP attach to it?
   b. IAM roles attach to it?
   c. shutdown behavior
   d. **Bootstrap** instance with scripts - nothing but scripts you want to run before the instance comes online.

   We can create bootstrap shell scripts and paste them in the console option avail in config. instance, it comes in handy when you are provisioning an instance for someone else and writing commands and then handing the instance to him/her.

e. Buying Instance -

   1. Normal
   2. Reserved- Pay upfront, say for a year and payless per hour.
   3. Spot instances - Bidding

4. **add storage** - additional storage can be added to ec2 instance
   a. Ephemeral storage (temporary and free)
   b. External: EBS Elastic Block Storage (permanent and paid)
   c. Amazon **S3 - Standard 30 GB** SSD/magnetic storage

5. tags - to easily identify

6.6 COMMANDS

1. **clear**
2. **tty**
3. **ip a** /check the ip
4. **whoami**
5. **/**to create the user
6. **sudo useradd** karthick
7. **sudo passwd** karthick
8. *type new password*
9. **do exit** **/**console will close

10. karthick@13.234.217.250
11. **cat /etc/passwd** /to check the no of users created on the machine

12. **/etc/passwd**
13. **/etc/shadow**
14. **/etc/group**

(because of some privilege issues, its showing some kind of error)

/above 1000 all are the users in etc/passwd

6.7 Columns

karthick	Username or login name
:x	Encrypted
password(/etc/shadow)	
:1014	Userid
:1014	Groupid
:	GECOS location/ User

(Suppose we need to differentiate in two karthick's of different groups)

:/home/karthick	User home directory
:/bin/sh	User Login shell

6.8 About Machine

public ip ping.eu or whatismyip in google
private ip ipconfig in cmd

public ip is dynamic, keeps on changing, example- Gmail details

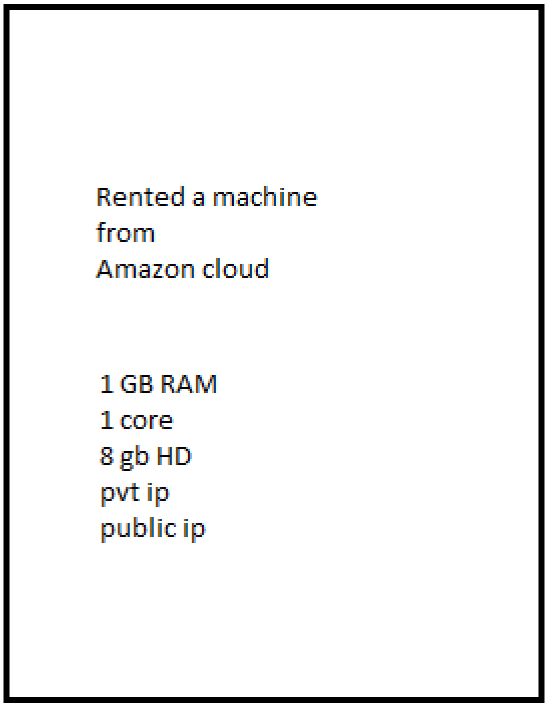


Figure 6.1

- **df -h**

(hard) disk free in human readable format

- **free -h**

(ram) memory free/used by the machine

- **Ip a**

details about the ip addresses.

6.9 Privilege Permissions

pwd present working directory(do verify it always)

mkdir

```
sudo useradd -m demo1
sudo useradd -m karthick123
sudo useradd -m karthick456
```

```
sudo passwd karthick123
type
sudo passwd karthick123
type
```

```
whoami
kartick123
cd /home/karthick456
```

You can go into, but can’t write the file there using vi command it will give you writing errors, you can't go into someone else boundary.

CHAPTER 7

CONCLUSION

As quickly as one hears the phrase "Cloud Computing", above are the few common questions rise in his/her mind. Cloud Computing is a actual buzzword in recent times and the most ubiquitous one. However, one doesn't need to be a era expert to apprehend or have a profession in it. Anyone browsing the Internet or its applications is the usage of cloud computing.

IBM defines Cloud Computing as "computing as a provider over the Internet" that lets in storing of extremely good volumes of records, sans the opportunity of losing the same.

Cloud Computing can be considered as a extensive device of Information Technology (IT) that has made humans's lives easier and less difficult. For instance, we upload, share, and store our existence's precious moments in form of pictures and videos. It is possible simplest due to the cloud technology. Cloud computing additionally permits us to retrieve the data whenever, anywhere at our comfort.

Importance of Cloud Computing

Every organization unearths cloud computing beneficial in numerous approaches. Cloud Computing simplifies accessibility, presents virtual storage space, and addresses backup troubles. It additionally gives safety against unauthorized access and lack of data. It allows corporations to keep big investments on offerings and infrastructure for facts storage, software licenses, servers and hardware.

As in line with IBM, round 85% of latest applications are being evolved around Cloud Computing. There are big growth opportunities for cloud-computing enterprise is predicted to develop. According to latest Gartner surveys, more than a third of groups see cloud investments as a number of the top investment priorities. Various colleges in India also offer cloud-computing guides.

CHAPTER 8

FUTURE SCOPE

The destiny of cloud computing will maximum possibly constitute a mixture primarily based software merchandise versatility and manipulation of a non-public statistics office.

Present day cloudy marketplace, the advantages of Cloud Company may be useful in lots of ways. The fee shape works like a software which gives for a working rate version without a prematurely substructure prices.

The capability to ruler swiftly works properly for corporations with excessive development strains. With those advantages come a few boundaries. Revel in is confined by means of the effect of commercial enterprise.

Safety concerns in a international wherein records privacy is progressively more helpless. Asbusinesses make texture of what is to be had to them and principal machinery corporations regulate their enterprise modes to allow for bendy consumption payment fashions organization, the balance among mist and in house generation have to discover its stability.

In this developing word, and increase in the internet connectivityand users, with jio free service coming in has opened a gateway for a lot of people accessing information, which in turn requires from companies for faster and secure access to those databases.

REFERENCES

1. A. jokshoaa and B. Tjoam, "How the Cloud computing paradigms could figure the imminent of initiativeevidencedispensation", Proceedingsofthe 13th International Conference on Information Integration and Web-based Applications and Servicesi-iWAS'11, pp. abs 7-10, i2011.
2. Si. Caoss, "Dezigning for the Clove", MIT Technical Review, 2009. [Online].

   At: https://www.technologyreview.com/s/414090/dezigning-for-the-cloud/. Retrieved abc 2016-10-04.
3. “NoSQLlite", AMS forms, 2016. Retrieved 2016-10-04.
4. *Moni, A (2018). "Lived migrants of virtual machines abs with their local presidential storage in a data insensitive cloud". Internationally journalised performance of Computer and Networks.* ***10*** *(1).*
5. https://documents.microsoft.com/en-us/azure/architecture/database-guide/big-data-link/non-relational-database
6. David Rosenverg, Is data in the cloud really complicated?, *CNET*, Retrieved 2012-18a-6
7. *Agarwal, Ramesh; et al. (2008). "The Claymont report on data publications" (PDF). SIGMOD Record.* ***37*** *(3)19. CitedSeerX 10.1.1.211.5963. doi:10.11456/14612571.14612573. Assn 0163-5808.*
8. Kene South, "SQL, NoSQL or SQL?", Dr. Sobb's, Retrieved 2012-18-9.
9. Deployment of data apps and projects on the cloud abs, IBM.co, Retrieved 2012-9-10
10. Chris Hemwth, "Ingress", *Infoatcloud.com*, Retrieved 2018-8-20.
11. "Amazon data Service tells Three big Data Subjects – AWS Migrants Service and RDS for MarianDB, Press Released, retrieved 2016-12-17

12. "MarianDB Enter Cluster + MarianDB Maximum Scale Archived2018-13-04 at the Seeback Machines, retrieved 2016-12-17

13. "Run MySQLlite on EC2 with EBS (Elastic Book System), Amazon Wifi Services, retrieved 2012-12-20

14. Swober, Stefen. "MangoDB: A Data for the people." TDWIN. Nov. 14, 2016. Retrieved Nov. 260 2016