



**SIX WEEKS SUMMER TRAINING
REPORT**

on

MODERN BIG DATA ANALYSIS WITH SQL

Submitted by

NIDAMANURI YESWANTH CHOWDARY

Registration No: 11907136

Program Name: CSE

Under the Guidance of

Glynn Durham, Ian Cook

School of Computer Science & Engineering

Lovely Professional University, Phagwara

(June-July, 2021)

DECLARATION

I hereby declare that I have completed my six weeks summer training at COURSERA from 01-06-2021 to 10-07-2021 under the guidance of Glynn Durham, Ian Cook. I have declared that I have worked with full dedication during these six weeks of training and my learning outcomes fulfil the requirements of training for the award of degree of BTECH-CSE, Lovely Professional University, Phagwara.

Yeswanth

N. YESWANTH CHOWDARY

REG.NO: 11907136

Date: 30-09-2021

ACKNOWLEDGEMENT

It is a matter of great satisfaction and pleasure to present this report on Modern bigdata analysis with SQL. I take this opportunity to owe my thanks to all those involved in my training.

Firstly, I would like to thank coursera for giving the opportunity to complete my project in the organization. I put on record my sincere thanks to my college, Lovely Professional University, Phagwara, for giving me such an opportunity.

It was a good experience for me to do my summer training with coursera. I am greatly obliged to Glynn Durham, Ian Cook my industry guides.

I express my gratitude towards staff of Coursera, those who have helped me directly or indirectly in completing the training.

Finally, I sincerely thank to my parents, family, and friends, who provide the advice and financial support.

YESWANTH

SUMMER TRAINING CERTIFICATE



3 Courses

Foundations for Big Data
Analysis with SQL

Analyzing Big Data with SQL

Managing Big Data in
Clusters and Cloud Storage



29-Jun-2021

NIDAMANURI YESWANTH CHOWDARY

has successfully completed the online, non-credit Specialization

Modern Big Data Analysis with SQL

In this Specialization, learners acquired essential knowledge and skills for data analysis with SQL using open source distributed big data systems. Through a sequence of three courses, learners gained knowledge of the fundamental concepts behind relational databases, SQL, and big data; learned how to write and run SQL queries using query engines including Apache Hive and Apache Impala; and learned how to manage large-scale data in clusters and cloud storage using the Hadoop Distributed File System (HDFS) and Amazon Simple Storage Service (S3).

The online specialization named in this certificate may draw on material from courses taught on-campus, but the included courses are not equivalent to on-campus courses. Participation in this online specialization does not constitute enrollment at this university. This certificate does not confer a University grade, course credit or degree, and it does not verify the identity of the learner.

Glynn Durham
Senior Instructor
Cloudera

Ian Cook
Staff Curriculum
Developer
Cloudera

Verify this certificate at:
coursera.org/verify/specialization/FPJC2DN3G3WQ

TABLE OF CONTENTS

CONTENT	PAGE.NO
Declaration	02
Acknowledgement	03
Summer training certificate	04
1: Introduction	07-09
2: Technologies Learnt	10-12
2.1 Technology Learnt	
2.2 Applications	
2.3 Uses of Database	
2.4 Reason for choosing technology	
3: Profile of Problem	13
3.1 Profile of problem	
3.2 Existing system	
4: Problem Analysis	14
4.1 Problem Analysis	
4.1.1 Product Definition	
4.1.2 Feasibility Analysis	
5: Software Requirements	15
5.1 Software requirement analysis	
5.1.1 Hardware Requirements	
5.1.2 Software Requirements	
6: Design	16-22
6.1 Tables	
6.2 Class Diagrams	
7: Implementation	23-24
8: Learning Outcomes	25
9: Gantt Chart	26
10: Project Legacy	27
Bibliography	28

LIST OF FIGURES

F.NO	Name Of Figure	Page.no
1.1	Coursera	8
1.2	Cloudera	9
1.3	Mobile store	9
2.1	Apache Impala	10
2.2	Applications of DBMS	11
2.3	Uses Of DBMS	12
6.1	Employee Table	16
6.2	Permission Table	16
6.3	Mobile Table	16
6.4	Accessories Table	17
6.5	Headphone Tables	17
6.6	Memory Card Table	17
6.7	Cable and Adapter	18
6.8	Customer Table	18
6.9	Employee Table	18
6.10	Permission Table	19
6.11	Mobile Table	19
6.12	Accessories Table	19
6.13	Headphone Tables	20
6.14	Memory Card Table	20
6.15	Cable and Adapter	21
6.16	Customer Table	21
6.17	Class Diagram	22
7.1	Impala	23
7.2	HUE	24
9.1	Gantt Chart	26

1: INTRODUCTION

1.1. INTRODUCTION:

Today we are producing a large amount of data. The produced must be stored. The stored data must be analysed. The information or data can be transmitted, stored and processed using modern digital technologies like the internet, disk drivers and modern computers.

Data Store is a collection of data of any type. The database is the organising of the data store. So, the database is the storing of data in an organised manner. Just the storing of data is not enough. The given data must be analysed.

A type of software that helps to organize data is a database management system or DBMS.

Together, the data and the DBMS, along with the applications that are associated with them, are referred to as a database system, often shortened to a just database. The DBMS should give you a way to perform at least four general activities. They are Design, Update, Retrieve and Manage.

Data within the most common types of databases in operation today is typically modelled in rows and columns in a series of tables to make processing and data querying efficient. The data can then be easily accessed, managed, modified, updated, controlled, and organized. Most databases use structured query language (SQL) for writing and querying data.

SQL is a programming language used by nearly all relational databases to query, manipulate, and define data, and to provide access control. SQL was first developed at IBM in the 1970s with Oracle as a major contributor, which led to the implementation of the SQL ANSI standard, SQL has spurred many extensions from companies such as IBM, Oracle, and Microsoft. Although SQL is still widely used today, new programming languages are beginning to appear.

A database typically requires a comprehensive database software program known as a database management system (DBMS). A DBMS serves as an interface between the database and its end-users or programs, allowing users to retrieve, update, and manage how the information is organized and optimized. A DBMS also facilitates oversight and control of databases, enabling a variety of administrative operations such as performance monitoring, tuning, and backup and recovery.

Most RDBMSs even provide ODBC (Open Database Connectivity), and JDBC (Java Database Connectivity) interfaces. These allow virtually any programming language to issue SQL commands to a database. As a result, nearly any program of any kind that needs to maintain some data from day to day can do so, using some form of RDBMS.

1.2. Introduction to company

1.2.1. Coursera:

Coursera Inc. is an American massive open online course provider founded in 2012 by Stanford university computer science professors Andrew Ng and Daphne Koller. Coursera works with universities and other organizations to offer online courses, certifications, and degrees in a variety of subjects. In 2021 it was estimated that about 150 universities offered more than 4,000 courses through Coursera.



1.1-Coursera

Coursera started offering their Stanford courses online in fall 2011 and soon after left Stanford to launch Coursera. Princeton, Stanford, the University of Michigan and the University of Pennsylvania were the first universities to offer content on the platform. Offerings have since expanded to include Specializations – collections of courses that build skills in a specific subject – as well as degrees and a workforce development product for businesses and government organizations.

In 2014 Coursera received Webby Winner (Websites and Mobile Sites Education 2014) People's Voice Winner (Websites and Mobile Sites Education).

1.2.2. Cloudera:

Cloudera, Inc. is a US-based company that provides an enterprise data cloud. Built on open-source technology, Cloudera's platform uses analytics and machine learning to yield insights from data through a secure connection. Cloudera's platform works across hybrid, multi-cloud and on-premises architectures and provides multi-function analytics throughout the edge to AI data lifecycle.

CLOUDERA

1.2- Cloudera

Cloudera was founded in 2008 by three engineers from Google, Yahoo! And Facebook (Christophe Bisciglia Amr Awadallah and Jeff Hammerbacher, respectively) joined by a former Oracle executive (Mike Olson). Olson was the CEO of Sleepycat Software, the creator of the open-source embedded database engine Berkeley DB (acquired by Oracle in 2006). Awadallah was from Yahoo!, where he ran one of the first business units using Apache Hadoop for data analysis. At Facebook, Hammerbacher used Hadoop for building analytic applications involving massive volumes of user data. Architect Doug Cutting, also a former chairman of the Apache Software Foundation, authored the open-source Lucene and Nutch search technologies before he and Mike Cafarella wrote the initial Hadoop software in 2004. He designed and managed a Hadoop storage and analysis cluster at Yahoo! before joining Cloudera in 2009. The chief operating officer was Kirk Dunn until 2015.

1.3. Introduction to Project:

A mobile store database is a database created to show how the database is defined and created. This database is useful for the people who are maintaining a mobile store. creating a database will help people in saving time as well as money. Why because the data will be stored usually in books.



1.3- Mobile Store

This may take a lot of books to maintain. If any fire accidents occurred in the store, then all data will be lost because of the burning of saved record books. If we have a database then all the data will be saved in the cloud. There will be no problem. And it is very easy to access rather than in record books. It is very easy to search specific data in the database.

2: TECHNOLOGIES LEARNT

2.1 TECHNOLOGY LEARNT: I have learnt about analysing big data.

Big data is a field that treats ways to analyse, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields (columns) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and value.

2.1.1 APACHE IMPALA:

Impala is an MPP (Massive Parallel Processing) SQL query engine for processing huge volumes of data that is stored in the Hadoop cluster. It is open-source software that is written in C++ and Java. It provides high performance and low latency compared to other SQL engines for Hadoop. In other words, Impala is the highest performing SQL engine (giving RDBMS-like experience) which provides the fastest way to access data that is stored in Hadoop Distributed File System.



2.1- Apache Impala

Impala combines the SQL support and multi-user performance of a traditional analytic database with the scalability and flexibility of Apache Hadoop, by utilizing standard components such as HDFS, HBase, Meta store, YARN, and Sentry. With Impala, users can communicate with HDFS or HBase using SQL queries in a faster way compared to other SQL engines like Hive.

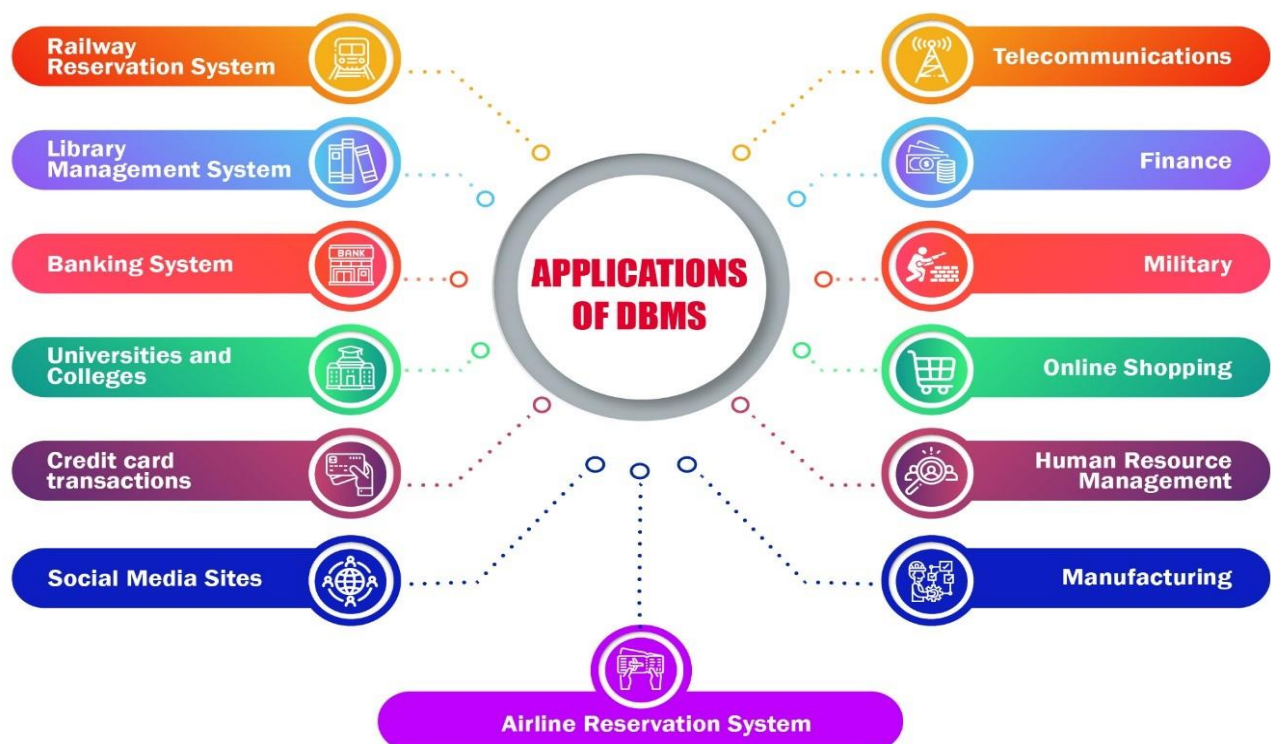
Impala can read almost all the file formats such as Parquet, Avro, RC File used by Hadoop. Impala uses the same metadata, SQL syntax (Hive SQL), ODBC driver, and user interface

(Hue Beeswax) as Apache Hive, providing a familiar and unified platform for batch-oriented or real-time queries.

2.2 APPLICATIONS:

The database management system is the most influential system in every sector is now a day. Because every industry has its own database like for example if consider a railway reservation management system in the railway reservation system, the database is required to store the record or data of ticket bookings, status about train's arrival, and departure. Also, if trains get late, people get to know it through database updates.

In these ways, every sector or industry has its own database which needs to be managed.



2.2- Applications of DBMS

2.3 USES OF DATABASE:

DBMS was designed to solve the fundamental problems associated with storing, managing, accessing, securing, and auditing data in traditional file systems.

Traditional database applications were developed on top of the databases, which led to challenges such as data redundancy, isolation, integrity constraints, and difficulty managing data access. A layer of abstraction was required between users or apps and the databases at a physical and logical level.

Introducing DBMS software to manage databases results in the following benefits:

1. **Data security:** DBMS allows organizations to enforce policies that enable compliance and security. The databases are available for appropriate users according to organizational policies. The DBMS system is also responsible to maintain optimum performance of querying operations while ensuring the validity, security and consistency of data items updated to a database.
2. **Data sharing:** Fast and efficient collaboration between users.
3. **Data access and auditing:** Controlled access to databases. Logging associated access activities allow organizations to audit for security and compliance.
4. **Data integration:** Instead of operating an island of database resources, a single interface is used to manage databases with logical and physical relationships.
5. **Abstraction and independence:** Organizations can change the physical schema of database systems without necessitating changes to the logical schema that govern database relationships. As a result, organizations can upgrade storage and scale the infrastructure without impacting database operations. Similarly, changes to the logical schema can be applied without altering the apps and services that access the databases.
6. **Uniform management and administration:** A single console interface to perform basic administrative tasks makes the job easier for database admins and IT users.

Uses of DBMS



2.3- Uses of DBMS

2.4 REASON FOR CHOOSING THIS TECHNOLOGY:

I have chosen this technology because I was very interested in databases. So, I have chosen this technology. The managing of databases will show my ability in analysing. The database is one of the leading systems in each and every industry or sector. By learning this technology, it will be helpful in future jobs. It is a very interesting technology and very easy to use. The database is very much important to every industry because every industry need a database of their past to see them in future. This will help them in improving their skills and work.

3: PROFILE OF PROBLEM

3.1 PROFILE OF PROBLEM:

The main problem is every store will write their product information, staff information and customer information in books. This may take a lot of time while checking a specific item. So, if they use a database this work can be done in a fraction of seconds. And another problem is that due to any reason if any record book was missing then it will be difficult for the owner. So, this problem can be resolved by maintaining a database.

3.2 EXISTING SYSTEMS:

Existing systems in stores are manually done by maintaining records of items in the store, employees and their salaries, customers and their purchased items. This may take a lot of books because the store is started years ago and they maintain books. If the recorded books were lost due to any reason like a fire accident, etc. then there would cause a problem for the owner. To resolve these problems, we provide a database with separate tables for everything like for items in-store, employee, the customer. These will help the owner if he/she wanted the search for a specific item and it can be done in a fraction of seconds. In this way consumption of time, books will be reduced and fast working will be progressed.

4: PROBLEM ANALYSIS

4.1.PROBLEM ANALYSIS:

4.1.1. Product Definition:

A mobile store database is created in a manner to show how the database was defined and created. This will help the users who wanted to create a database for online websites. This will show how the tables are specified and make it easy for the people who wanted to create a website. The tables in the database show that what is available inventory in the store, the cost and stock of a specific inventory. And which item was purchased by the customer? This will help the store owners in finding them the information regarding the specific item.

4.1.2. Feasibility Analysis: the feasibility study consists of different types they are

- i. Economic Feasibility: the mini-project was made in an oracle virtual box which consist of a Cloudera training virtual machine which was provided during the training by Coursera.
- ii. Technical Feasibility: there were no problems or bugs in the virtual machine because these were made by the training company and well verified.
- iii. Operational Feasibility: the information for the mini-project was gathered with a general idea. Software requirement was hue which was provided during training.
- iv. Schedule Feasibility: the schedule for training and mini-project was well planned in six weeks duration. The planning had been provided below in the Gantt chart.

5: SOFTWARE REQUIREMENTS

5.1 SOFTWARE REQUIREMENTS ANALYSIS:

5.1.1 HARDWARE REQUIREMENTS:

- ✓ 64-bit operating system (32-bit operating systems will *not* work)
- ✓ 8 GB RAM or more
- ✓ 25GB free disk space or more
- ✓ Intel VT-x or AMD-V virtualization support enabled (on Mac computers with Intel processors, this is always enabled; on Windows and Linux computers, you might need to enable it in the BIOS)

5.1.2 SOFTWARE REQUIREMENTS:

- ✓ An operating system like Windows, Mac, Linux.
- ✓ Apache Impala is required for the database.
- ✓ Hue editor
- ✓ Oracle virtual machine

6: DESIGN

6.1 TABLES:

My mini-project is creating a mobile store database so I created a database mobile store. There are eight tables in it. They are

1. **Employee table:** This table will the name of the employee and details of an employee.

	Name	Type
1	i e_id	string
2	i name	string
3	i doj	timestamp
4	i contact	bigint
5	i resignation	string

6.1- Employee table

2. **Permission table:** This table will give the permission information of certain employees.

	Name	Type
1	i p_id	string
2	i e_id	string
3	i discription	string

6.2- Permission table

3. **Mobile table:** this table will give the information of the brand name of mobile and its features and also gives the information of the number of mobiles available.

	Name	Type
1	i m_id	string
2	i company	string
3	i model	string
4	i ram_with_storage	string
5	i cost	bigint
6	i in_stock	bigint
7	i color	string

6.3- Mobile table

4. **Accessories table:** This table will give information on mobile accessories like cases and tempered glasses.

		Name	Type
1	i	a_id	string
2	i	name	string
3	i	description	string
4	i	price	bigint

6.4- Accessories table

5. **Headphones table:** This table will give the information about the brand of headphone and its types and number items in stock.

		Name	Type
1	i	h_id	string
2	i	company	string
3	i	type	string
4	i	price	bigint
5	i	in_stock	bigint
6	i	color	string







6.5- Headphones table

6. **Memory card table:** This table will give information about brand name of memory card or pen drive and its details with number of items in stock.

		Name	Type
1	i	mc_id	string
2	i	comapny	string
3	i	type	string
4	i	storage	string
5	i	price	bigint
6	i	in_stock	bigint


6.6- Memory card table

7. **Cable and adapter table:** This table will give the information about company name of cable and adapter with number of items in stock.

	Name	Type
1	 c_id	string
2	 company	string
3	 type	string
4	 price	bigint
5	 in_stock	bigint
6	 color	string

6.7- Cable and adapter table

8. **Customer table:** This table will give the information of the customer's name, purchased product with date.

	Name	Type
1	 cu_id	string
2	 name	string
3	 dod	timestamp
4	 m_id	string
5	 a_id	string
6	 h_id	string
7	 mc_id	string
8	 c_id	string
9	 total_price	bigint

6.8- Customer table

6.2 Information in Tables:

1. Employee Table:

 e_id	name	doj	contact	resignation
1 e01	M.satish	2021-01-12 10:30:55.001000000	9848032165	incharge
2 e02	G.venu	2021-01-15 12:30:55.021000000	6309762501	senior seller
3 e03	A.anirudh	2021-01-15 13:30:50.051000000	9865321475	seller
4 e04	S.santosh	2021-02-06 10:30:55.031000000	9768532410	seller


6.9- Employee Table

2. Permission Table:

	p_id	e_id	discription
1	p01	e01	ALL PERMISSIONS GIVEN
2	p02	e02	ACCESS TO CUSTOMER TABLE ONLY

6.10- Permission Table

3. Mobile Table:

	 m_id	company	model	ram_with_storage	cost	in_stock	color
1	ap001	apple	iphone7	2gb+128gb	25000	3	black
2	ap002	apple	iphone8	2gb+128gb	31000	3	white
3	ap003	apple	iphoneXR	3gb+128gb	43000	4	white
4	ap004	apple	iphone11	4gb+128gb	57000	5	red
5	ap005	apple	iphone11pro	4gb+128gb	89999	5	blue
6	ap006	apple	iphone11ProMax	4gb+128gb	90000	5	black
7	ap007	apple	iphone12	4gb+128gb	80000	6	blue
8	ap008	apple	iphone12pro	6gb+128gb	115000	6	blue
9	ap009	apple	iphone12ProMax	6gb+128gb	125000	3	black
10	ap010	apple	ipadpro11	8gb+256gb	80000	3	black
11	ap011	apple	ipadpro12.9	8gb+256gb	170000	3	silver
12	ss001	samsung	galaxyS20 FE	6gb+128gb	40000	5	red and silver
13	ss002	samsung	galaxyF12	4gb+64gb	8999	5	blue
14	ss003	samsung	galaxyF62	8gb+128gb	25999	5	laser blue
15	ss004	samsung	galaxyM52	6gb+128gb	21999	3	prism dot black
16	ss005	samsung	galaxyA52	6gb+128gb	25999	3	blue
17	ss006	samsung	galaxyF41	6gb+128gb	14999	5	fusion blue
18	ss007	samsung	galaxyM51	6gb+128gb	21999	2	electric blue
19	ss008	samsung	galaxyM21	4gb+64gb	13999	4	midnight blue
20	ss009	samsang	galaxyS21Ultra	6gb+128gb	105999	3	phantom silver

6.11- Mobile Table

4. Accessories Table:

	 a_id	name	description	price
1	a01	cases	available for all mobiles	250
2	a02	temperd glass	available for all mobiles	100
3	a03	case+tempered glass	available for all mobiles	350

6.12- Accessories Table

5. Headphones Table:

	h_id	company	type	price	in_stock	color
1	h01	boult	wireless headset	1199	3	red
2	h02	jbl	wireless earphones	1699	4	black
3	h03	boat	earphones	349	6	black
4	h04	realme	wireless earphones	1799	4	black
5	h05	oneplus	wireless earphones	1999	4	black
6	h06	jbl	wireless earbuds	3999	3	black
7	h07	realme	earphones	299	5	black

6.13- Headphones Table

6. Memory Card Table:

	mc_id	comapny	type	storage	price	in_stock
1	mc01	sandisk	memory card	1TB	18999	2
2	mc02	sandisk	memory card	32gb	450	10
3	mc03	sandisk	memory card	64gb	799	8
4	mc04	samsung	memory card	64gb	700	5
5	mc05	samsung	memory card	32gb	450	10
6	mc06	samsung	memory card	256gb	2999	3
7	mc07	HP	memory card	32gb	500	5
8	mc08	HP	memory card	64gb	800	5
9	mc09	sandisk	pendrive	128gb	1699	3
10	mc10	sandisk	pendrive	32gb	399	5
11	mc11	sandisk	pendrive	64gb	699	4
12	mc12	HP	pendrive	32gb	459	4
13	mc13	HP	pendrive	64gb	799	3

6.14- Memory Card Table

7. Cable and Adapter Table:

	c_id	company	type	price	in_stock	color
1	c01	amazon	C cable	399	5	white
2	c02	amazon	not C cable	599	4	silver
3	c03	boat	not C cable	299	6	black
4	c04	htc	adapter	199	3	black
5	c05	ptron	adapter	199	3	white
6	c06	boat	adapter	299	4	white

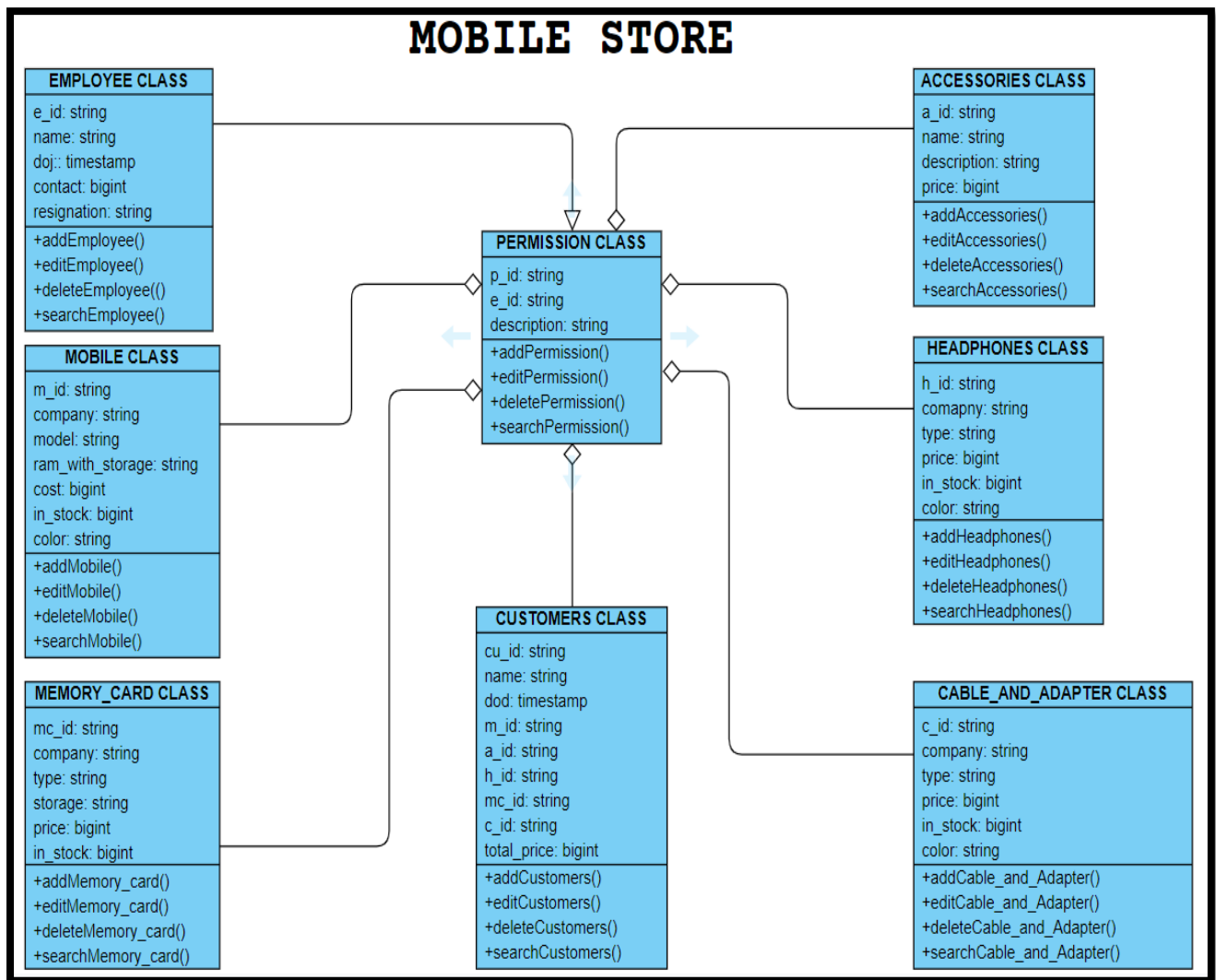
6.15- Cable and Adapter Table

8. Customer Table:

	cu_id	name	dod	m_id	a_id	h_id	mc_id	c_id	total_price
1	cu01	shyam	2021-02-03 10:30:55.001000000	ss007	NULL	NULL	NULL	NULL	21999
2	cu02	ganesh	2021-02-18 11:00:55.001000000	ss011	a03	NULL	NULL	NULL	28449
3	cu03	sundhar	2021-02-18 12:30:55.001000000	NULL	NULL	h05	NULL	NULL	1999
4	cu04	raju	2021-02-22 15:28:50.031000000	NULL	NULL	NULL	mc03	NULL	799
5	cu05	rajesh	2021-02-28 16:10:35.051000000	ap006	a01	NULL	NULL	NULL	90250
6	cu06	sai	2021-03-03 10:30:55.001000000	NULL	NULL	NULL	mc01	NULL	189990
7	cu07	gunashekar	2021-03-05 12:30:55.001000000	rl005	a01	NULL	NULL	NULL	27249
8	cu08	shashu	2021-03-10 14:30:25.001000000	NULL	NULL	NULL	NULL	c01	399
9	cu09	sailendar	2021-03-15 10:30:55.001000000	NULL	NULL	h02	NULL	NULL	1699
10	cu10	kumari	2021-03-15 16:50:55.001000000	op003	a01	NULL	NULL	NULL	18249
11	cu11	chandrashekar	2021-03-18 11:30:55.001000000	ap010	NULL	NULL	NULL	NULL	80000
12	cu12	ramesh	2021-03-22 15:50:55.001000000	NULL	NULL	NULL	mc09	NULL	1699
13	cu13	ugendar	2021-03-24 10:30:55.001000000	rm005	a03	NULL	NULL	NULL	15349
14	cu14	kalesh	2021-03-28 17:30:55.001000000	ap009	a01	NULL	NULL	NULL	125250
15	cu15	dinesh	2021-04-03 11:56:55.001000000	NULL	NULL	h03	NULL	NULL	349
16	cu16	shiva	2021-04-05 19:06:55.001000000	NULL	NULL	NULL	mc08	NULL	799
17	cu17	bhargav	2021-04-08 15:56:55.001000000	on002	a01	NULL	NULL	NULL	50249
18	cu18	sravan	2021-04-10 20:20:55.001000000	rl005	a01	NULL	NULL	NULL	11249
19	cu19	harish	2021-04-14 17:56:55.001000000	ss010	NULL	NULL	NULL	NULL	134999
20	cu20	rafeesh	2021-04-17 13:46:46.001000000	NULL	NULL	h01	NULL	NULL	1199
21	cu21	harsha	2021-04-20 18:00:55.001000000	NULL	NULL	NULL	mc011	NULL	699

6.16- Customer Table

6.3 Class diagram:



6.17- Class Diagram

7: IMPLEMENTATION

7.1 IMPLEMENTATION:

This project is implemented using SQL using impala engine hue.

7.1.1 IMPALA:

Impala is an MPP (Massive Parallel Processing) SQL query engine for processing huge volumes of data that is stored in the Hadoop cluster. It is open-source software that is written in C++ and Java. It provides high performance and low latency compared to other SQL engines for Hadoop. In other words, Impala is the highest performing SQL engine (giving RDBMS-like experience) which provides the fastest way to access data that is stored in Hadoop Distributed File System. Impala combines the SQL support and multi-user performance of a traditional analytic database with the scalability and flexibility of Apache Hadoop, by utilizing standard components such as HDFS, HBase, Meta store, YARN, and Sentry. With Impala, users can communicate with HDFS or HBase using SQL queries in a faster way compared to other SQL engines like Hive.



7.1- Impala

Impala can read almost all the file formats such as Parquet, Avro, RC File used by Hadoop. Impala uses the same metadata, SQL syntax (Hive SQL), ODBC driver, and user interface (Hue Beeswax) as Apache Hive, providing a familiar and unified platform for batch-oriented or real-time queries.

Unlike Apache Hive, **Impala is not based on MapReduce algorithms**. It implements a distributed architecture based on **daemon processes** that are responsible for all the aspects of query execution that run on the same machines. Thus, it reduces the latency of utilizing MapReduce and this makes Impala faster than Apache Hive.

7.1.2 HUE:

Hue is an open-source SQL Assistant for querying Databases & Data Warehouses and collaborating. Its goal is to make self-service data querying more widespread in organizations.



7.2-HUE

The Hue team provides releases on its website. Hue is also present in the Cloudera Data Platform and the Hadoop services of the cloud providers Amazon AWS, Google cloud platform and Microsoft Azure.

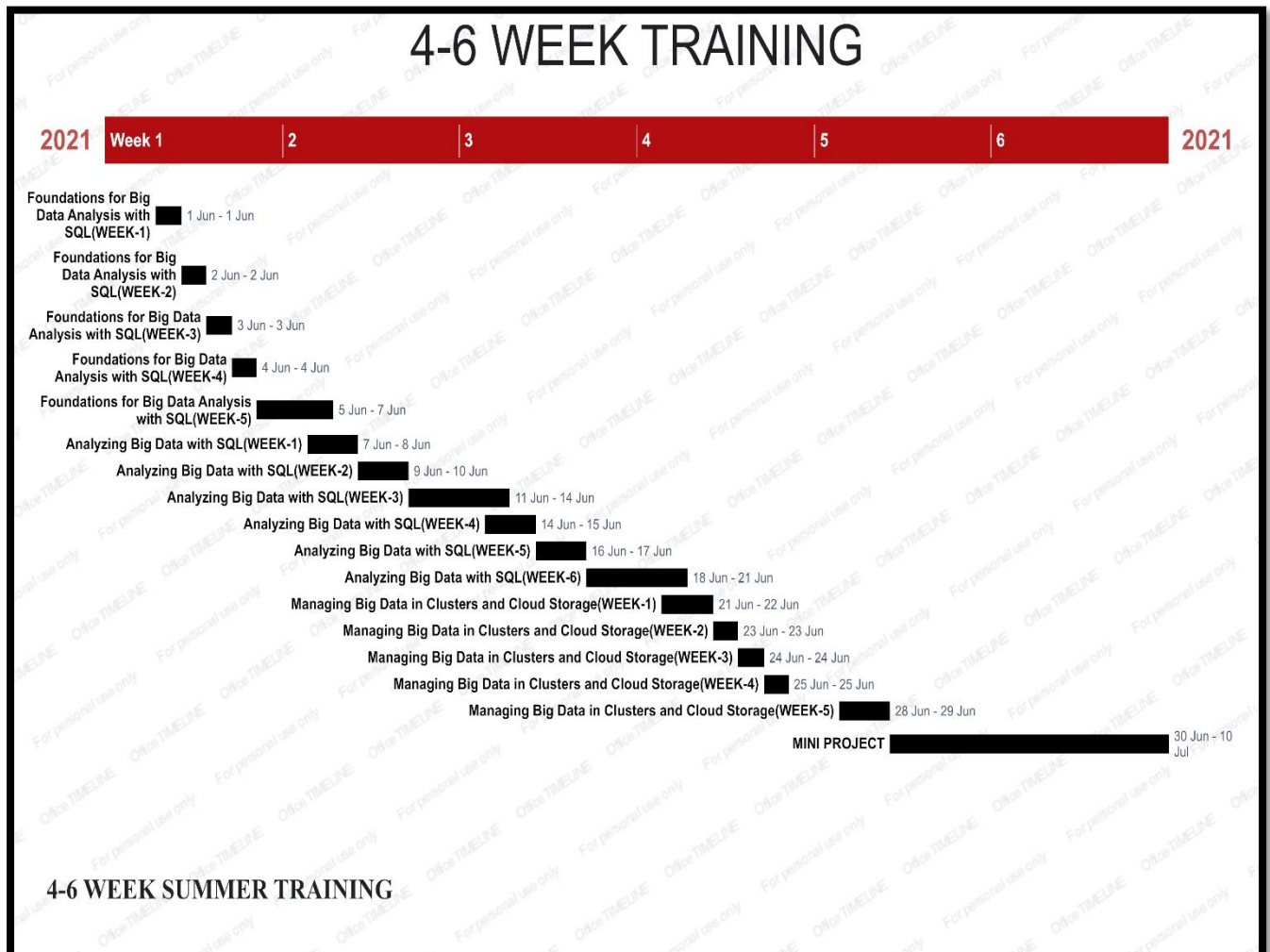
8: LEARNING OUTCOMES

8.1 LEARNING OUTCOMES:

1. I have learnt about Apache impala and hive.
2. I have learnt how to use the HUE editor.
3. I have learnt how to implement the SQL commands in impala and hive.
4. I have learnt how to create a database in impala in hue editor.
5. I have learnt how to create a mini project using impala in hue editor.
6. I have learnt how to design data.
7. I have learnt how to update data.
8. I have learnt how to retrieve data.
9. I have learnt how to manage the data.

9: GANTT CHART

GANTT CHART:



9.1- Gantt chart

10: PROJECT LEGACY

10.1 Technical lessons learnt:

- ✓ I have learnt SQL engines like Impala, Hive from training.
- ✓ I have learnt how to operate a hue editor.
- ✓ I have learnt languages and clauses in SQL during my training.
- ✓ I have learnt how to make a mini-project with help of assignments given during my training.

10.2 Managerial lessons learnt:

- ✓ I have learnt how to make a project from origin to implementation.
- ✓ I have learnt how to manage risks during the project.
- ✓ I have learnt how to map and manage the timeline.
- ✓ I have learnt how to troubleshoot problems and challenges.

BIBLIOGRAPHY

<https://www.coursera.org/learn/foundations-big-data-analysis-sql/home/welcome>

<https://www.coursera.org/learn/cloudera-big-data-analysis-sql-queries/home/welcome>

<https://www.coursera.org/learn/cloud-storage-big-data-analysis-sql/home/welcome>