Transportation Of Smart Cities Using Big data

A PROJECT REPORT

Submitted by

MILIND DEV
Reg. No. 15MCA1080

in partial fulfillment for the award of the degree of

Master of Computer Applications

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
www.vit.ac.in
Vellore ▪ Chennai
CHENNAI CAMPUS
Vandalur - Kelambakkam Road, Chennai - 600127

School of Computing Science and Engineering
VIT University
Vandalur - Kelambakkam Road, Chennai - 600 127

April - 2017

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
www.vit.ac.in
Vellore ▪ Chennai
CHENNAI CAMPUS
Vandalur - Kelambakkam Road, Chennai - 600127

School of Computing Science and Engineering

DECLARATION

I hereby declare that the project entitled **Transportation Of Smart Cities Using Big data** submitted by me to the School of Computing Science and Engineering, VIT Chennai, 600 127 in partial fulfillment of the requirements of the award of the degree of **Master of Computer Applications** is a bona-fide record of the work carried out by me under the supervision of **Dr. Jagadeesh Kannan**. I further declare that the work reported in this project, has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or University.

Place: Chennai	Signature of Candidate
Date:	(MILIND DEV)

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
www.vit.ac.in
Vellore ▪ Chennai
CHENNAI CAMPUS
Vandalur - Kelambakkam Road, Chennai - 600127

School of Computing Science and Engineering

CERTIFICATE

This is to certify that the report entitled **Transportation Of Smart Cities Using Big data** is prepared and submitted by **MILIND DEV (Reg. No. 15MCA1080)** to VIT Chennai, in partial fulfullment of the requirement for the award of the degree of **Master of Computer Applications** is a bona-fide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Guide/Supervisor

Name: Dr. Jagadeesh Kannan
Date:

Program Chair

Name: Dr. Pattabiraman V
Date:

Examiner

Name:
Date:

Examiner

Name:
Date:

(Seal of SCSE)

Acknowledgement

I would like to express my gratitude to my guide Dr.Jagadeesh kannan(Dr,SCSE, VIT University) for his constant guidance, continuous encouragement and unconditional help towards the development of the project. It was he who helped me whenever I got stuck somewhere in between. The project would have not been completed without his support and confidence he showed towards me.

I would like to thank MCA Program Chair Dr. Pattabiraman V who helped me to complete the project on time by providing fixed deadlines and showed the importance of time and his support whenever required.

I would also like to thank SCSE Dean Dr. Vaidehi Vijayakumar for providing with an environment to work in and for her inspiration during the tenure of the course.

It is my pleasure to thank MCA Project Coordinator Dr. Nithyanandam P who accepted the proposed project and allowed me to continue with the work.

Lastly I would l like to thank all those who helped me directly or indirectly toward the successful completion of the project.

MILIND DEV
Reg. No. 15MCA1080

Abstract

A most valuable area come to now today is transportation which is highly increasing. Most of the country has serious issue about transportation or road accident. In some years accident frequently increases government studies maximum people loss our lives from accident. So many adminisable solution doing by government. More associated value of transport to recognize the final result. From government, some serious steps or rules for saving accident. In recent year most popular trend include long term is injuries, road accident, vehicle damage also. Every 4 to 5 minute one loss our life by road accident that means transportation controlling is more valuable issue now.

Big data is a unpredictable era that restraining a long dataset to reduce small and readable form, the intelligent transport behave like assure safety for traffic collision. Everyone has suffer in own vehicle or transport. Moreover, mostly can't imagine how much load on that route, how much vehicle going to one location to another. Cities traffic facing a lot of problem with huge traffic. Most of them accident frequently happen because of traffic controlling. Problem with the deal of those challenges to solving problems for safety risk, it will improved for handle those critical situation which are unknown even though more critical. A query processing language defines to calculate the number of vehicle visit at that traffic signal, in addition a system is proposed to efficient process high speed real time network traffic in which place traffic extremely high. Query processing focuses complex pattern with lack of entries. The entries belong to plenty dataset are conducted to show the usability and effectiveness of query language.

Contents

Declaration **i**

Certificate **i**

Acknowledgement **iii**

Abstract **iv**

1 Introduction **1**

1.1 Statement of Assumptions 1

1.2 Aim and Objectives 2

1.3 Problem statement 2

1.4 Motivation 2

1.5 Post/Related Work 3

1.5.1 Existing System 3

1.5.2 Proposed System 3

1.6 System Requirement 3

1.6.1 Software 3

1.6.2 Hardware 4

2 Overview / Literature Review **5**

2.0.1 Survey Of Road Accident 5

2.0.2 Controlling Traffic 6

2.0.3 Vechile Routing Problem 6

3 Design **7**

3.1 Data Flow Diagram 7

3.2 Acceptance 9

3.3 Validation 9
3.4 Verification 9
3.5 Entity Relationship Diagram 9

4 Implementation of System/ Methodology 11
4.1 HDFS 11
4.2 MYSQL 13
4.2.1 Mysql 13
4.3 Sqoop 18
4.3.1 Export 20
4.3.2 Import 20
4.3.3 Command 20
4.4 Hive 21
4.4.1 Partition 21
4.4.2 Bucketing 25
4.5 Pig 30
4.6 Tableau 32
4.7 System Testing 37
4.7.1 Automation Testing 37
4.7.2 Agile Testing 39
4.7.3 Spiral Model 41
4.7.4 Spiral Model Application 42
4.8 strategy cases 44

5 Results And Discussions 47

6 Conclusion And Future Work 48
6.0.1 Conclusion 48
6.0.2 Future Work 49
6.1 Coding 50

List of Tables

4.1 Transportation File 45
4.2 Extract Uid 46
4.3 Extract Signal 46

List of Figures

3.1 Data flow diagram 8
3.2 Entity Relational Diagram 10

4.1 Start Hadoop 12
4.2 Setup Wizard 14
4.3 Installation Mysql 15
4.4 Set Hostname And Port 16
4.5 Authenticate Username And Password 17
4.6 Sqoop Import And Export 19
4.7 Empty Warehouse 22
4.8 Store Hdfs Warehouse 23
4.9 Partition 24
4.10 Empty Bucket 26
4.11 Load Data To Warehouse 27
4.12 Hdfs Partition Queries 28
4.13 Bucket 29
4.14 Architect For Pig 31
4.15 Vechile Travel Day 33
4.16 Histogram Chart 34
4.17 Particular Data Status 35
4.18 Automation Testing 38
4.19 Agile Method 40
4.20 Spiral Model 42

Chapter 1

Introduction

1.1 Statement of Assumptions

Transportation is a process to move a one location to another. Actually Different types of transportation makes use to easily reach destination. Various factors mention for transportation.Vehicle increase every day with increasing human population. It's major factor to jam traffic. Traffic collision during on peek time or traffic jam increasing high rate of pollution. Both of them is not admissible for environment as well as associated with long route traveller.It's wasting few hours behind most of the people going to urgent meeting they will also late. Then what the role of transportation? the main role of transportation are all major issues comes under transport because they are not going particular way. If they will going exact timing, exact route, exact loading system. Its easily to handle those problems.

The government survey list issuing to showing different states transport then 13 states are most frequent accident in road accident. With the government under list of road accident survey state of Tamilnadu is one of the highest or top of the list and Maharashtra is second on the government list of road accident. Calculate percentage of this cities then 13.7 ratio comes.

Big data is unique era to perform reducing dataset small and readable form from that restraining long dataset, the intelligent transport behave like safety for traffic collision. Cities traffic facing a lot of problem with huge traffic. Most of them accident frequently happen because of traffic controlling problem for safety risk, its very useful part for transport. With maximum vehicle on minimum distance running on same particular route assure traffic will be increasing fastly. Now a days two types of transport public transport or private transport who gives service which is providing for use by the general public. As different modes can use such as taxi, cab, or hired buses. While vehicle moving at on the same route, it will be improved but some improvement come when to know perfectly which way most traffic are going for this problem if its not know then to handle those critical situation, which are unknown even though more critical. A query processing language define to calculate the number of vehicle visit at that traffic signal, In additional a system proposed to efficient process high speed real time network traffic in place of traffic extremely high.

Big data and Iot is important role to safe traffic collision. Smart cities have become smart transportation. Doing for smart cities growing big data of analytics which gives source of new era of smart transport. Yet smart cities analyzer think different aspects how to smartly analyze data. In particular state have also minimum population across 6.87 million. Million peoples have vehicle which is overloaded in particular route that why it's important to focus or attention on transportation.

1.2 Aim and Objectives

A main focus on to reduce long term task and easily find particular object with bucketing and partition. Some condition availability to solve those problems.

1.3 Problem statement

A big large dataset used by user but it's not easily read by user. Then user find out a particular object values. Those values which is based on some condition.

example. Transportation dataset in your hand. Then find out a particular data object just like start station mambalam and some condition also put up how many cars are going to mambaam to kelambakkam with total load 150.

1.4 Motivation

The computing revolution that began more than 2 decades ago has led to large amounts of digital data being amassed by corporations. Advances in digital sensors, proliferation of communication systems, especially mobile platforms and devices, massive scale logging of system events, and rapid movement toward paperless organizations have led to a massive collection of data resources within organizations. And the increasing dependence of businesses on technology ensures that the data will continue to grow at an even faster rate.

1.5 Post/Related Work

1.5.1 Existing System

MySQL dB means MySQL database. MySQL is a open source relational database system. MySQL sorted and retrieve information from one location to another. Its Rdms one of best web based for developing software application. MySQL stored user application data and fetch those when they enter into application login.

1.5.2 Proposed System

Proposed concept deals with providing database by using hadoop tool we can analyze no limitation of data and simple add number of machines to the cluster and we get results with less time, high throughput and maintenance cost is very less and we are using joins, partitions and bucketing techniques in hadoop..

1.6 System Requirement

1.6.1 Software

- PROCESSOR I3 PENTINUM CORE
- RAM 4GB
- MONITOR 1024 * 764
- HARD DISK 1TB

1.6.2 Hardware

- FRAMEWORK :: HADOOP
- IDE :: ECCLIPSE
- OPERATING SYSTEM :: LINUX
- DATABASE :: MYSQL

Chapter 2

Overview / Literature Review

2.0.1 Survey Of Road Accident

Road safety is a major concern in the present situation.in road safety many determination points showing that are more dangerous on that spot. The maximum significant accident including vehicle causing 300-500 accident daily. Vehicle accident involved 4564 cases, Head injury accounts for is greater than 60

During 1 year, the review of finding accident rate high. A review of safety at signal-controlled junctions and mid-block crossings was undertaken for Transports.it cover all aspects controlling traffic ,there is a conflict between safety and safe

There are different ways controlling road intersections. In the simplest cases the right-hand rule or, if the traffic is higher, a roundabout or the signal of a policeman can help steer the traffic. However, especially in big cities, in the complicated cases when the roads in the intersection have several lanes, the use of traffic lights cannot be avoided. An additional issue arises when in the intersection not only roads but also railroad tracks take part, what often occurs in suburban traffic situations. The most common way to handle this type of intersection is the conventional cyclic lights control. In more enhanced control, the traffic in different directions is monitored by sensors and the signals thus obtained control the traffic lights. In this method the control is adapting to the traffic.

2.0.2 Controlling Traffic

When properly used, traffic control signals are important devices for the control of vehicle in road. They assign right-of-way to a choice of traffic movements and thereby deeply influence traffic flow. Traffic control signals that are properly designed, located, operated and maintained. Traffic controlling is a biggest issue in traffic controlling some times it's not sustainable because when it's stuck a large volume of pollution increase.

2.0.3 Vechile Routing Problem

The problem of transport of goods, commodities and people is as relevant where it was considered as a generalization of the traveling salesman problem. Society is more aware of the environmental damage caused by human activity and is more concerned about the indiscriminate use of natural resources. With regard to the environment, transportation is one of the most visible within the supply chain aspects Transportation modes (e.g. Airplane, boat, truck, train, barge or pipeline) have different characteristics in terms of cost, transit time, accessibility and environmental performance. It must find the shortest route through all the paths and return to its starting position. Applications: garbage collection, reading utility.

Chapter 3

Design

In this module , we have to create data set for transportation dataset it contain set of information In this module we have to analysis the dataset using Hive tool which will be stored in hadoop distributed file system(HDFS). For analysis dataset hive using HiveQL language .using hive we perform tables creation, joins ,partition, bucketing concept. Hive analysis the only structure language.

Such that transportation detail for different areas in particular Chennai city. This data set will be first provide into MySQL database with the help of this dataset we analysis this project .now we are ready with dataset into hadoop(HDFS),that will be happen in this module.

Sqoop is a command-line interface application for transferring data between relational database into hadoop .In this module we fetch the dataset into hadoop(HDFS) using sqoop. we have to perform lot of the function, such that if we want to fetch the particular column or if we want to perform lot of the function to fetch the dataset with specific condition that will be support by sqoop tool and data will be stored in hadoop(HDFS)

3.1 Data Flow Diagram

data flow diagram showing where data will flow. Its explaining every step of information by diametric way. Data flow diagram explaining for user how to work flow of every steps. User easily identify the input/output of data.

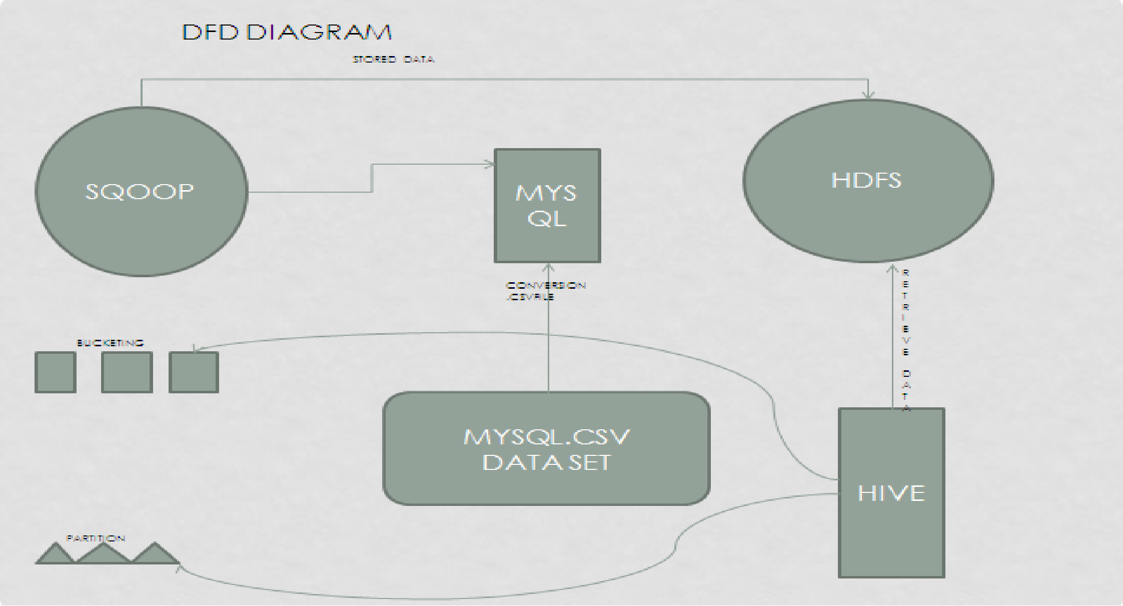


Figure 3.1: Data flow diagram

3.2 Acceptance

Acceptance testing is a technique to meet requirement specification between user and product. If its not or mismatched then Acceptance testing does not allow next step. Acceptance testing is a method to accept each and every step.

3.3 Validation

validation is a technique to validate product. Software testing engineering valid a one particular date that date product and customer determine expectation and requirement. Its make sure that the product stage fix in perfect place. Execution comes under validation. verification is just next step to validate. Its basically different specification and requirement. Validation follow some method are black box testing, white box testing.

3.4 Verification

verification technique to check user. Wheather it matches every key to evaluating it's correct or not. Verification satisfied every stage which is given by user. Verification satisfy make every objectives are perfectly work. Different terms check in verification meeting, Inspection, reviews.

3.5 Entity Relationship Diagram

Entity Relationship Diagram is entity diagram who connection between every interrelated point. Showing in a diagram everything is inter-related each other. Every entity attach adminisable point. Transportation are connected with different four different entity signal, distance, day, startstation, stopstation. every entity has unique identifier. Inside deep, entity inside another entity.

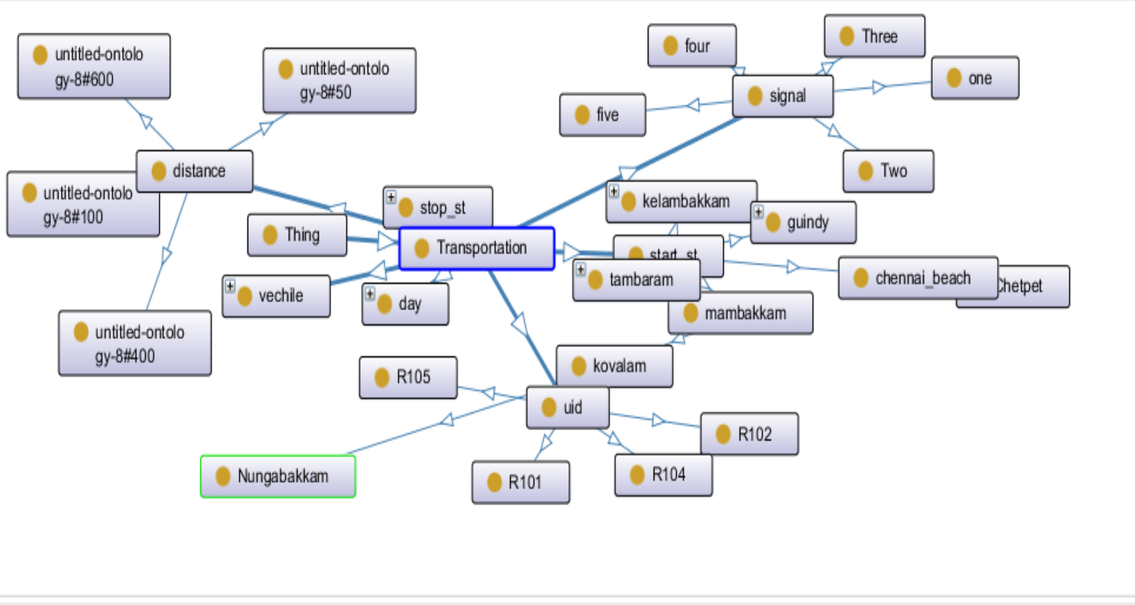


Figure 3.2: Entity Relational Diagram

Chapter 4

Implementation of System/ Methodology

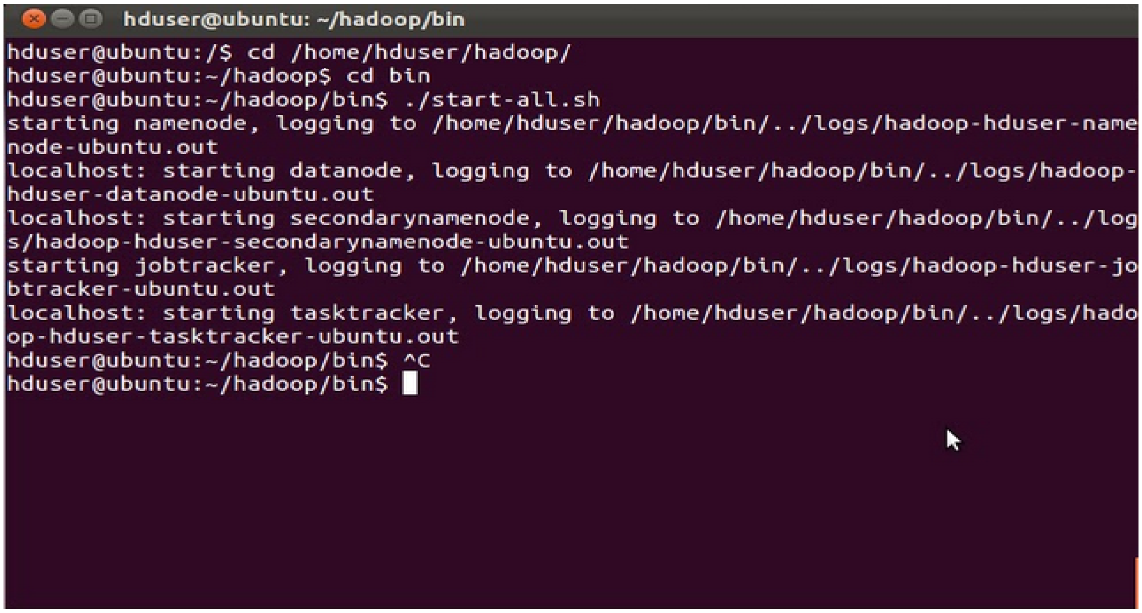
4.1 HDFS

Hadoop is distributed file system. Hadoop distributed file system is pipeline behind more than number of pipes are available to passing data or stored into local file system. Computer have own local file system that default location. Using hadoop for analytics along with big dataset are stored inside hadoop cluster and processed to loading data. Used dataset are different format Excel, Csv, Json etc. its build for single server across thousand of machine, each have computational and storage. Hadoop is supported in Linux or Gui platform, if Linux is not in your box may its other OS then you can install virtual box, In virtual box ,you can install different OS inside virtual box software. That is comparatively important for those who have not Ubunu, Linux, other version of Linux.

Hadoop is supported in Linux or Gui platform, if Linux is not in your box may its other OS then you can install virtual box, In virtual box ,you can install different OS inside virtual box software. That is comparatively important for those who have not Ubunu, Linux, other version of Linux.

Hadoop provides user permission for authentication and file permission .those file is stored useless to rescue system from perhaps data losses, when the data losses in case of failure. HDFS makes parallel processing.

Install hadoop distributed file in our computer, open terminal and type



```
hduser@ubuntu:/$ cd /home/hduser/hadoop/
hduser@ubuntu:~/hadoop$ cd bin
hduser@ubuntu:~/hadoop/bin$ ./start-all.sh
starting namenode, logging to /home/hduser/hadoop/bin/../logs/hadoop-hduser-name
node-ubuntu.out
localhost: starting datanode, logging to /home/hduser/hadoop/bin/../logs/hadoop-
hduser-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /home/hduser/hadoop/bin/../log
s/hadoop-hduser-secondarynamenode-ubuntu.out
starting jobtracker, logging to /home/hduser/hadoop/bin/../logs/hadoop-hduser-jo
btracker-ubuntu.out
localhost: starting tasktracker, logging to /home/hduser/hadoop/bin/../logs/hado
op-hduser-tasktracker-ubuntu.out
hduser@ubuntu:~/hadoop/bin$ ^C
hduser@ubuntu:~/hadoop/bin$ 
```

Figure 4.1: Start Hadoop

Start-all.sh,starting single node cluster this command start namenode,datanode,jobtrackerm and tasktracker on your machine.

Stop-all.sh ,stop singlenode cluster this command stop namenode,jobtracker,and tasktracker on your single machine

Those two command commonly used for hadoop, start and stop command running behind local host file, perhaps without this command start-all.sh hdfs cant take permission to enter local host .these are that was specific to hdfs and mapreduce. When you talking about maximum number of service that might be someplace different service on your cluster Api (Application programmer interface), Api provides user interface component or is set of protocols. Sometimes, doesnt setup location in as much as it already setup if you type exact they automatically take those script

hduser@ubuntu:-/hadoop/bin] . su hduser
hduser@ubuntu:-/hadoop/bin] . password

if start-all.sh and stop cant support than check ./ bash.sh = location ,usually a location gives you big problem,sometimes its fixed but when its not fixed then firstly fix the .bash file location

```
hduser@ubuntu:-/hadoop/bin] jps

jobtracker

namenode

tasktracker
```

4.2 MYSQL

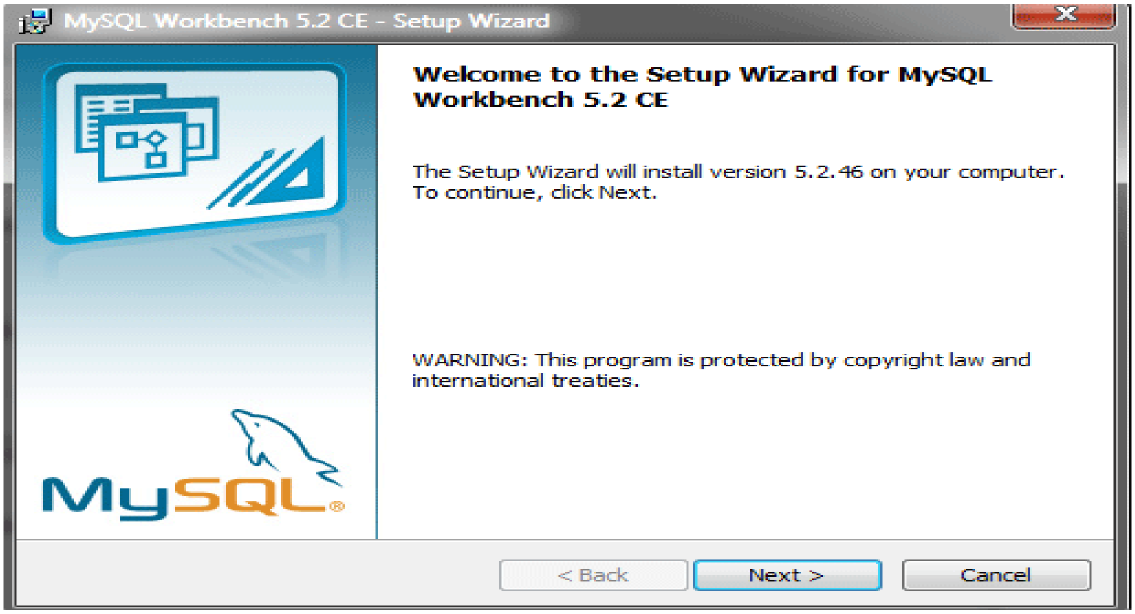
Why we use restore MySQL backup file Because hadoop can't take direct import file. We use excel sheet around with a number of maximum data stored in columns and rows. The entire column and rows. Now create one database excel sheet take dataset value from where user saved it location and save It to csv file which becomes easily export to anywhere. Csv file is comma separated value to transfer easily one particular location to another. Create MySQL database login and password. MySQL back file transfer to Ubuntu /home/Ponny.

MySQL backup file particularly stored a different format file, nearly also say that dumb file. You can say database backup by telling only ask SQL to run file. If you're storing anything in MySQL databases that you do not want to lose. Its important to make backup for securing file from regular loss. This will show you two easy ways to backup and restore the data in your MySQL database

4.2.1 Mysql

The dump file contains the SQL statements necessary to re-create the database. Here is the proper syntax:

- Your database username [username]
- The password for your database (note there is no space between -p and the password) [password]

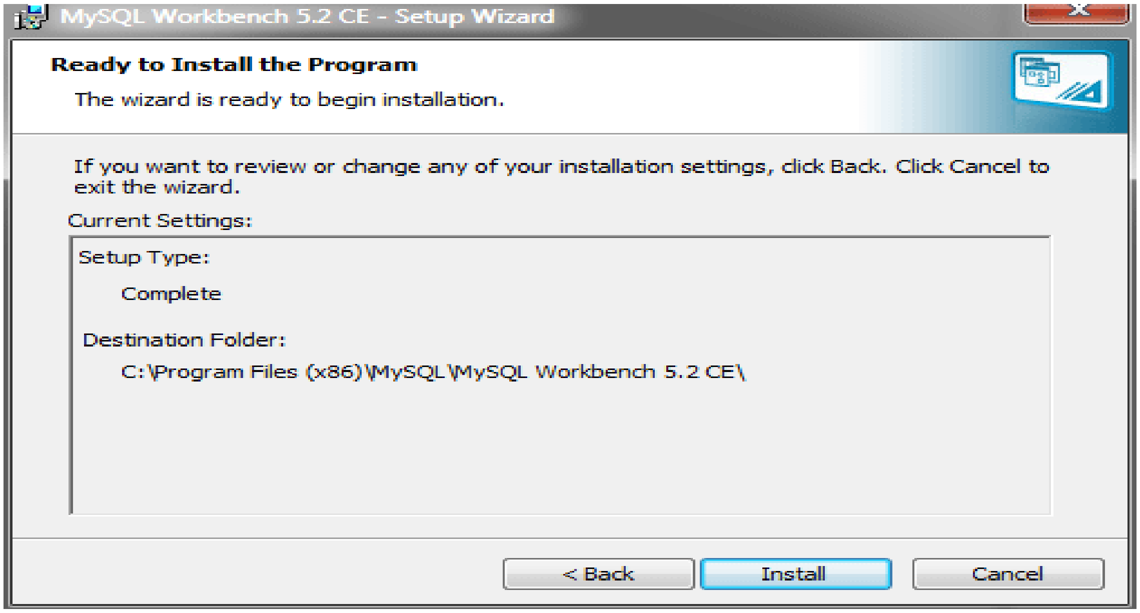

MySQL Workbench 5.2 CE - Setup Wizard

Welcome to the Setup Wizard for MySQL Workbench 5.2 CE

The Setup Wizard will install version 5.2.46 on your computer. To continue, click Next.

WARNING: This program is protected by copyright law and international treaties.

Figure 4.2: Setup Wizard



MySQL Workbench 5.2 CE - Setup Wizard

Ready to Install the Program

The wizard is ready to begin installation.

If you want to review or change any of your installation settings, click Back. Click Cancel to exit the wizard.

Current Settings:

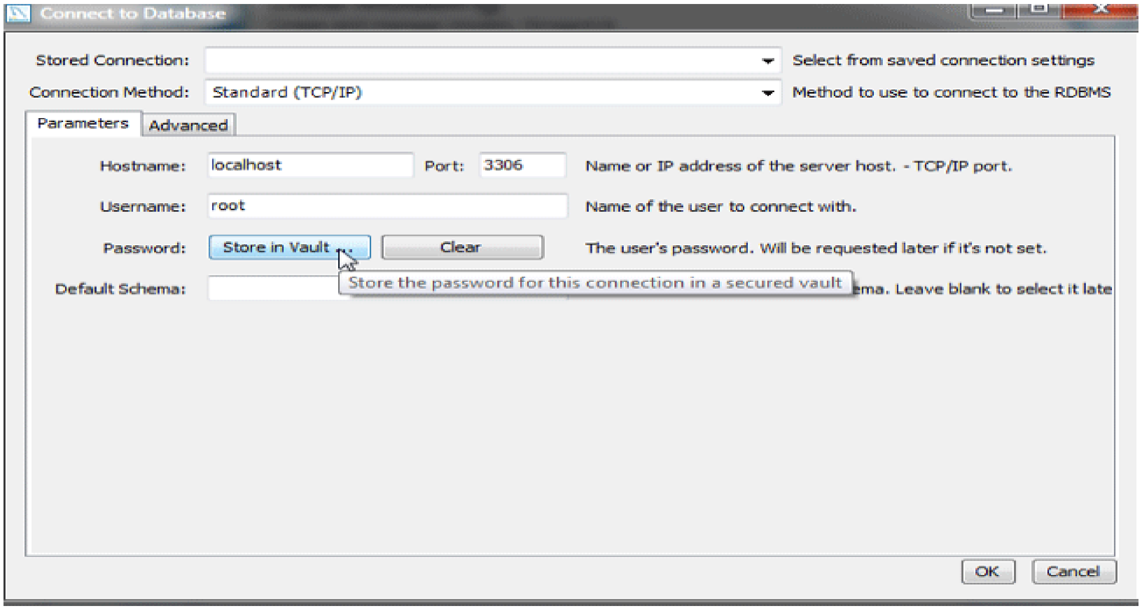
Setup Type:

Complete

Destination Folder:

C:\Program Files (x86)\MySQL\MySQL Workbench 5.2 CE\

Figure 4.3: Installation Mysql



Connect to Database

Stored Connection: Select from saved connection settings

Connection Method: Standard (TCP/IP) Method to use to connect to the RDBMS

Hostname: localhost Port: 3306 Name or IP address of the server host. - TCP/IP port.

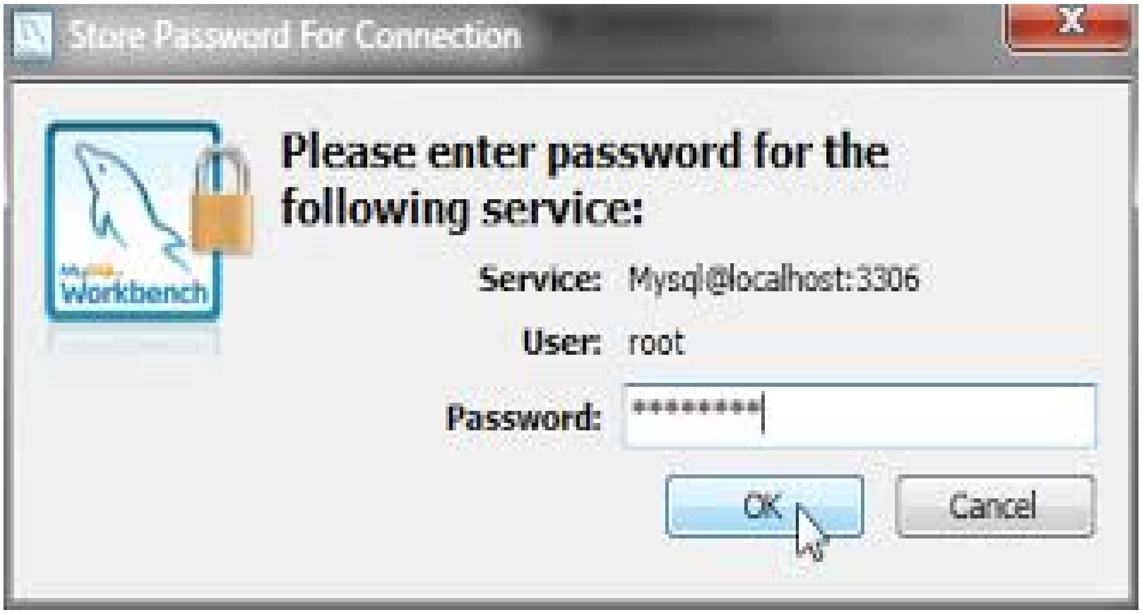
Username: root Name of the user to connect with.

Password: The user's password. Will be requested later if it's not set.

Store the password for this connection in a secured vault

Default Schema: ema. Leave blank to select it late

Figure 4.4: Set Hostname And Port



Store Password For Connection

Please enter password for the following service:

Service: Mysql@localhost:3306

User: root

Password: ********

Figure 4.5: Authenticate Username And Password

- The name of your database [databasename]
- The filename for your database backup [backupfilename.sql]
- The mysqldump option
- option [–option]

Some important steps which belongs to show common files.

```
$] mysql u username p

$] pas**ord

$] show databases;

$] use db;


$] show tab;


$] select * from t limit 2;
```

4.3 Sqoop

Sqoop is tool with relation between rational database and hdfs database. Its used to import file from relational database and export from hdfs(hadoop distributed file system).its basically tool interlink between one or more link or to fetch data from other username or password. Its allow when other user give both username and password such as oracle, MySQL and other version of sql. When big data analyzing to analyze such as hive, pig, mapreduce dataset to interactive database. its briefly clear where those database that time Sqoop determine relating to relational database. Hadoop ecosystem requiring to feasible state between relational and hadoop system. Sqoop relates with SQL to hadoop and hadoop to SQL.

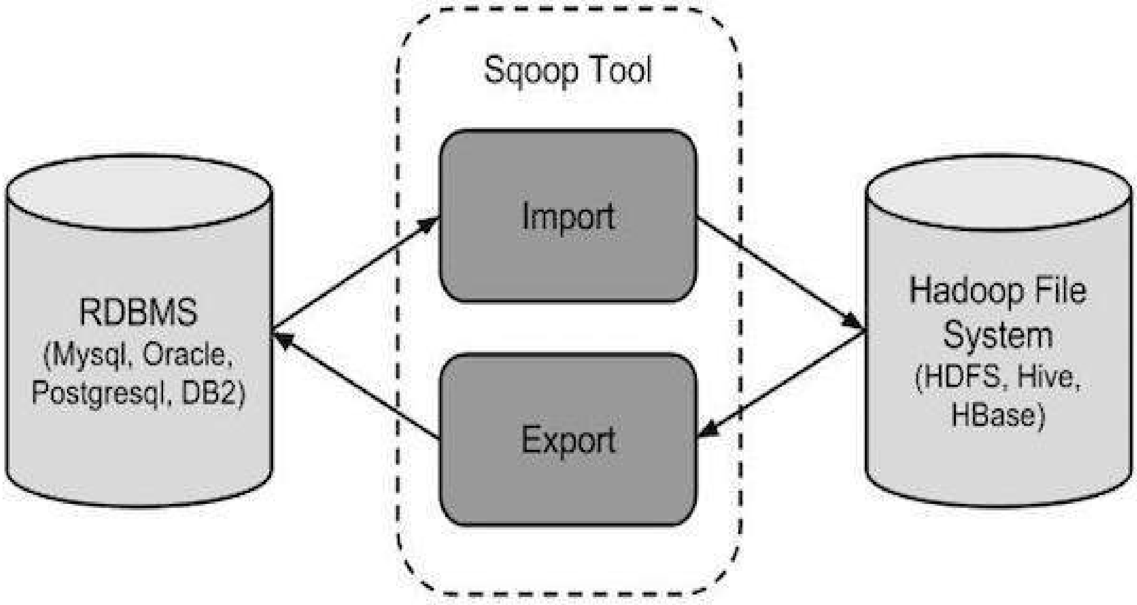


Figure 4.6: Sqoop Import And Export

4.3.1 Export

Sqoop export files exporting hdfs to rdms(Relational Database). Those file records as delimited file system

4.3.2 Import

Each indivisual file import from sqoop. Sqoop import stored a sql file which is stored as row and columb. From column work as indivisual text file or sequentioal file. Sqoop establish connection between relational database and hdfs. Both row work treated as record hdfs

4.3.3 Command

```
User Database
--------------
Name

Area

Pincode

Address

$] sqoop tra connect jdbc:mysql://localhost/userdb/ --username root
```

Sqoop importing all files from userdb database and fetching data to stored user local host a single cluster machine. Sqoop is not only import database its features are extremely different also.it is also take list of tables, evaluation of query and processing data, quering a particular data.

When using Sqoop to fetching data firstly check from database or using with mysql command showing its data are there or not

```
 select *from table name
```

otherwise its neglect sqoop command when the data is not there.

```
$]  sqoop eval \ --connect jdbc:mysql://localhost/db/
     --username u password p
     -- Query select *from  t
```

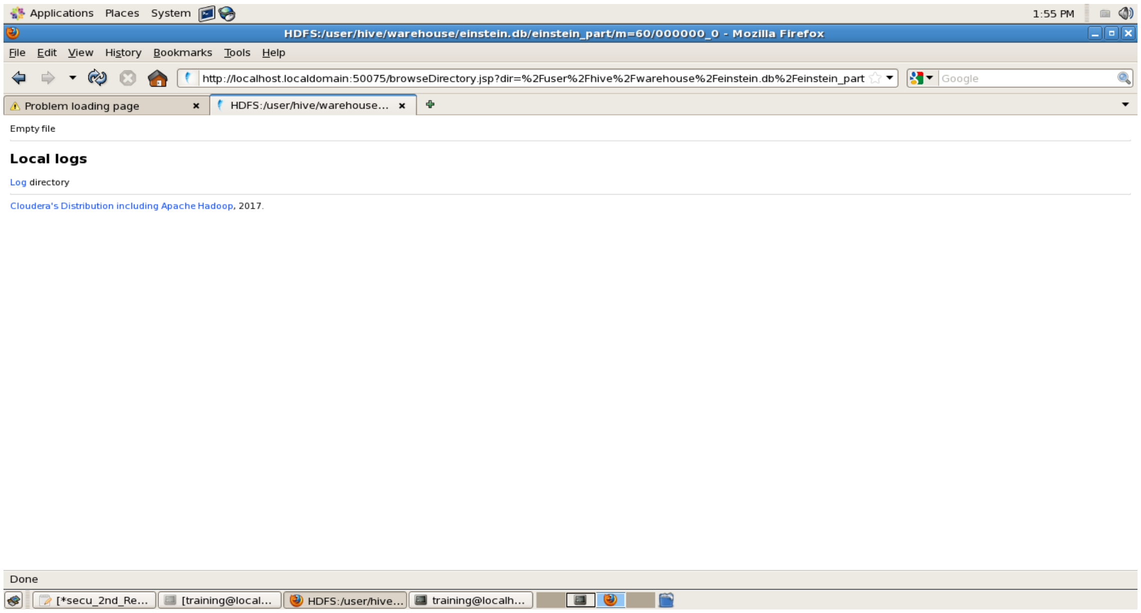
Select query in the database table to verifies how many tables, columbs, row are inside that table we use sql database with username and password to evaluate particular data u can use also insert and drop data with this Sqoop command.

4.4 Hive

Apache Hive is a datawarehouse infrastructure built for analyzing, summarizing query data. Its top of the hadoop for providing different expects of query. Its similar like sql query. Hive give query similar like sql interface stored into various database to unified with hadoop. Most of them queries support sql but hive queries little bit different may be some add-on just like terminated, clause whereas, location. Hive queries support queries direct get to hadoop. Most of the companies work in hive such as Netflix, facebook, amozon web service and other. its compact table as similar as sql, it's other also know as HIVEQL read and write with schema and convert queries into hadoop. some important features in hive which belongs to time reducing.that features are metadata stored in rdms, which provide to reducing a time and perform faster than during execution queries. Different format using or support hive are textfile, csv, RCFile. Metatore is a storage of each tables such as schema and location. it's includes partition of different metadata. with the help of metadata to know the various progress of data set of distributed cluster. Similarly data stored in traditional rdms fomat. it's higly backup server replicate those data regularly for doesn't loss that data. its retrieve because loss of data. Now, looking at different area used in hive:

4.4.1 Partition

Partition distributed parts to store extract data into different limited folders. Hive providing different tables and with those tables perform to part by hive. Partition organize tables into distribute dataset. Partition is a way to divided tables is related parts to based on columns just like as such a country, area ,code. In as much as


Applications Places System
1:55 PM
HDFS:/user/hive/warehouse/einstein.db/einstein_part/m=60/000000_0 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Feinstein.db%2Feinstein_part
Google
Problem loading page
HDFS:/user/hive/warehouse...
Done
[*secu_2nd_Re...
[training@local...
HDFS:/user/hive...
training@localh...

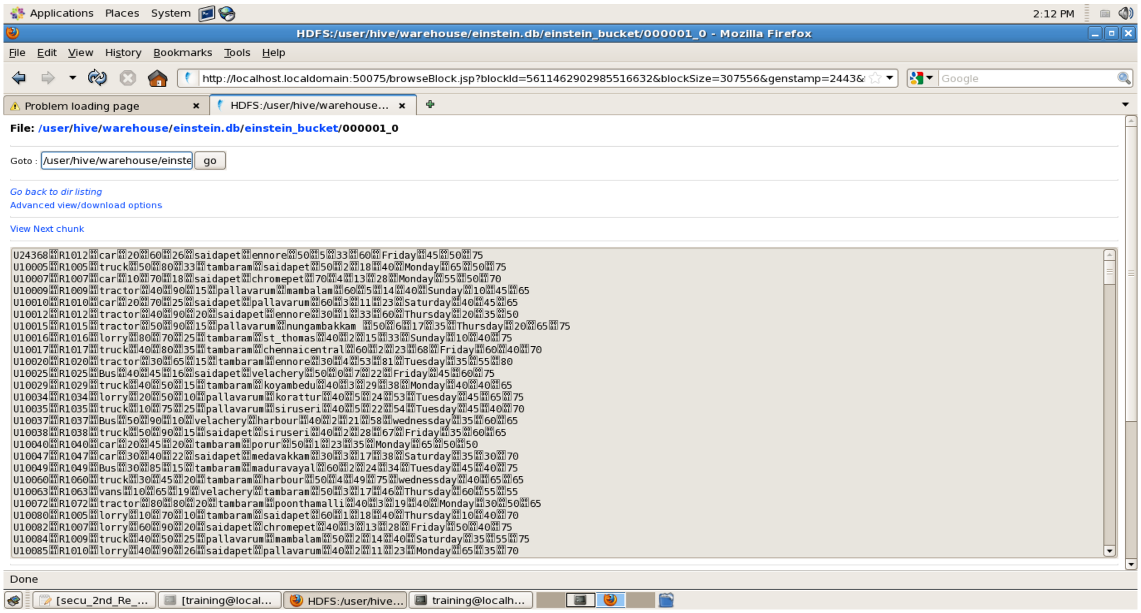
Empty file

Local logs

Log directory

Cloudera's Distribution including Apache Hadoop, 2017.

Figure 4.7: Empty Warehouse


Applications Places System
2:12 PM
HDFS:/user/hive/warehouse/einstein.db/einstein_bucket/000001_0 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075/browseBlock.jsp?blockId=5611462902985516632&blockSize=307556&genstamp=2443&
Google
Problem loading page
HDFS:/user/hive/warehouse...
Done
[secu_2nd_Re_...
[training@local...
HDFS:/user/hive...
training@localh...

File: /user/hive/warehouse/einstein.db/einstein_bucket/000001_0

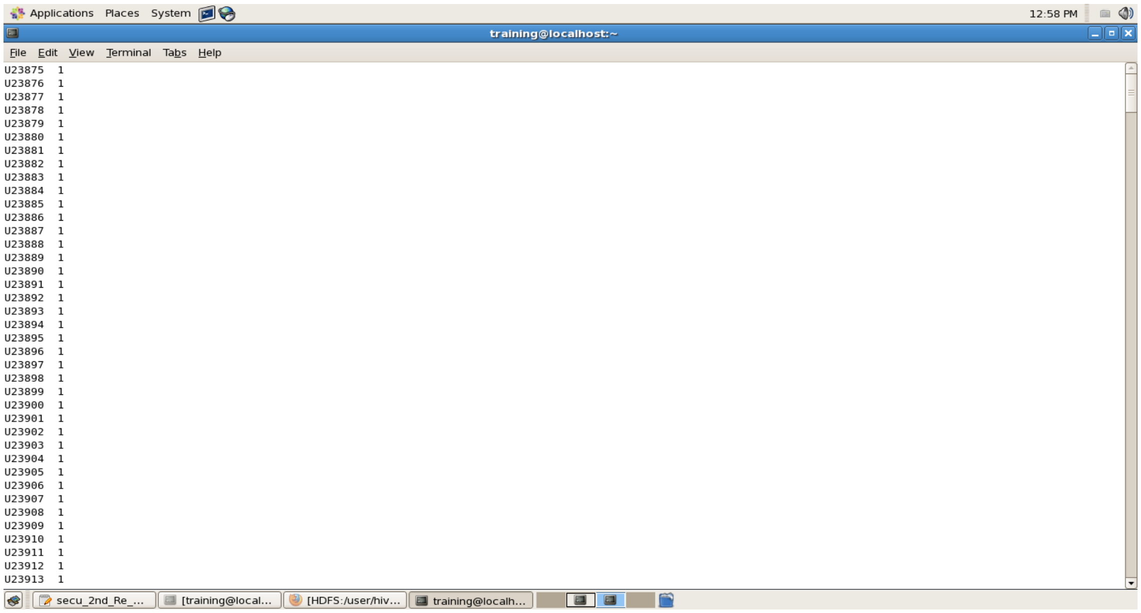
Goto : /user/hive/warehouse/einste go

Go back to dir listing
Advanced view/download options

View Next chunk

```
U24368□R1012□car□20□60□26□saidapet□ennore□50□5□33□60□Friday□45□50□75
U10005□R1005□truck□50□80□33□tambaram□saidapet□50□2□18□40□Monday□65□50□75
U10007□R1007□car□10□70□18□saidapet□chromepet□70□4□13□28□Monday□55□50□70
U10009□R1009□tractor□40□90□15□pallavarum□mambalam□60□5□14□40□Sunday□10□45□65
U10010□R1010□car□20□70□25□saidapet□pallavarum□60□3□11□23□Saturday□40□45□65
U10012□R1012□tractor□40□90□20□saidapet□ennore□30□1□33□60□Thursday□20□35□50
U10015□R1015□tractor□50□90□15□pallavarum□nungambakkam □50□6□17□35□Thursday□20□65□75
U10016□R1016□lorry□80□70□25□tambaram□st_thomas□40□2□15□33□Sunday□10□40□75
U10017□R1017□truck□40□80□35□tambaram□chennaicentral□60□2□23□68□Friday□60□40□70
U10020□R1020□tractor□30□65□15□tambaram□ennore□30□4□53□81□Tuesday□35□55□80
U10025□R1025□Bus□40□45□16□saidapet□velachery□50□0□7□22□Friday□45□60□75
U10029□R1029□truck□40□50□15□tambaram□koyambedu□40□3□29□38□Monday□40□40□65
U10034□R1034□lorry□20□50□10□pallavarum□korattur□40□5□24□53□Tuesday□45□65□75
U10035□R1035□truck□10□75□25□pallavarum□siruseri□40□5□22□54□Tuesday□45□40□70
U10037□R1037□Bus□50□90□10□velachery□harbour□40□2□21□58□wednessday□35□60□65
U10038□R1038□truck□50□90□15□saidapet□siruseri□40□2□28□67□Friday□35□60□65
U10040□R1040□car□20□45□20□tambaram□porur□50□1□23□35□Monday□65□50□50
U10047□R1047□car□30□40□22□saidapet□medavakkam□30□3□17□38□Saturday□35□30□70
U10049□R1049□Bus□30□85□15□tambaram□maduravayal□60□2□24□34□Tuesday□45□40□75
U10060□R1060□truck□30□45□20□tambaram□harbour□50□4□49□75□wednessday□40□65□65
U10063□R1063□vans□10□65□19□velachery□tambaram□50□3□17□46□Thursday□60□55□55
U10072□R1072□tractor□80□80□20□tambaram□poonthamalli□40□3□19□40□Monday□30□50□65
U10080□R1005□lorry□10□70□10□tambaram□saidapet□60□1□18□40□Thursday□10□40□70
U10082□R1007□lorry□60□90□20□saidapet□chromepet□40□3□13□28□Friday□50□40□75
U10084□R1009□truck□40□50□25□pallavarum□mambalam□50□2□14□40□Saturday□35□55□75
U10085□R1010□lorry□40□90□26□saidapet□pallavarum□40□2□11□23□Monday□65□35□70
```

Figure 4.8: Store Hdfs Warehouse



```
U23875  1
U23876  1
U23877  1
U23878  1
U23879  1
U23880  1
U23881  1
U23882  1
U23883  1
U23884  1
U23885  1
U23886  1
U23887  1
U23888  1
U23889  1
U23890  1
U23891  1
U23892  1
U23893  1
U23894  1
U23895  1
U23896  1
U23897  1
U23898  1
U23899  1
U23900  1
U23901  1
U23902  1
U23903  1
U23904  1
U23905  1
U23906  1
U23907  1
U23908  1
U23909  1
U23910  1
U23911  1
U23912  1
U23913  1
```

Figure 4.9: Partition

different columns are created around its automatically cant except how much part create. With partition is easy to read.

Partition control behind key, value by hash map function. But its one problem with partition are distributed are dislike because number of files distribute as time who will come first that allocate first as similar as first come first serve. User only sets part. Partition perform automatically.

Apart from partition to alter statement let assume neuter part as usually employee, location, date. Then we can set which location, which employee, and when that date. Partition can take individually employee table, location and date and it's stored into hdfs. Renaming or other can also do.

—-Advantage Of Partition

1.data stored as term of slice, as query is more faster than other. the best advantage is a small parts of data to process instead looking for search.

2. now, you can select a large user table select particular table which belongs to address then selecting user as address is equal to ls (address=ls)., that area whose belonging area = ls scan out a big large data set,this scan set compressed or stored in hdfs with different folder . 3. every partition scan through horizontally.

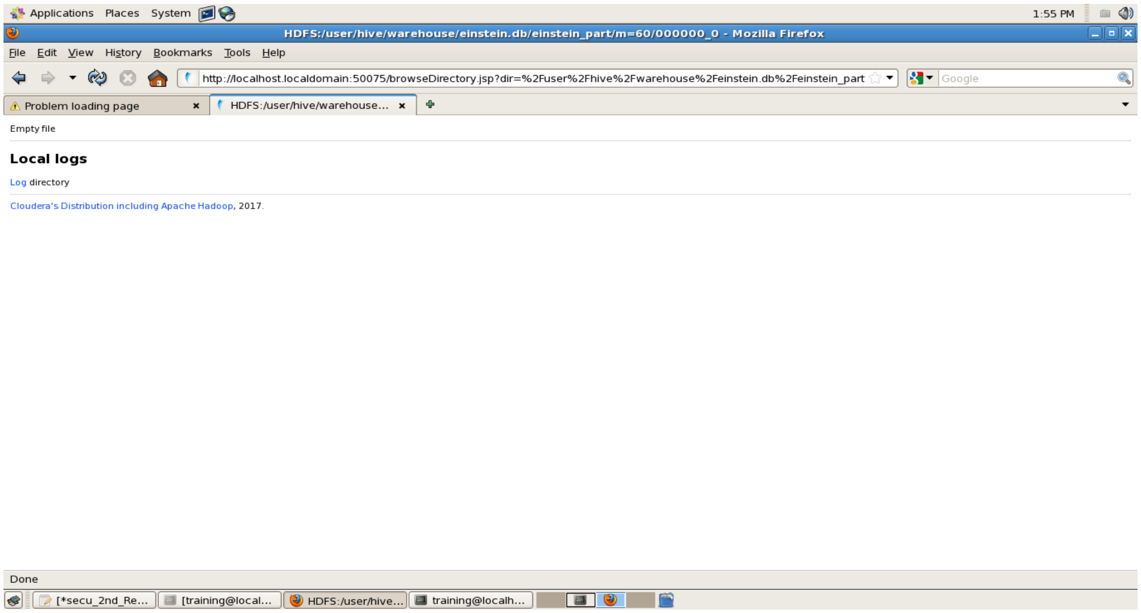
Limitation

Having a lot of partition in a single table, create a problem. Why because table create a large number of files and a lot of directories in hdfs.

Partition may repair some queries which is where clause but its less important or responsible for grouping queries .

4.4.2 Bucketing

Bucket as usual name as a small infill bucket to fill particular large volume of data fill in particular bucket. Bucket is not single bucket, when understand bucketing user should fix some limits. Used those data to stocking into bucket and given limiting wise showing in as much as thats limitation under fill all dataset. We can use two types of bucketing



HDFS:/user/hive/warehouse/einstein.db/einstein_part/m=60/000000_0 - Mozilla Firefox

http://localhost.localdomain:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Feinstein.db%2Feinstein_part

Empty file

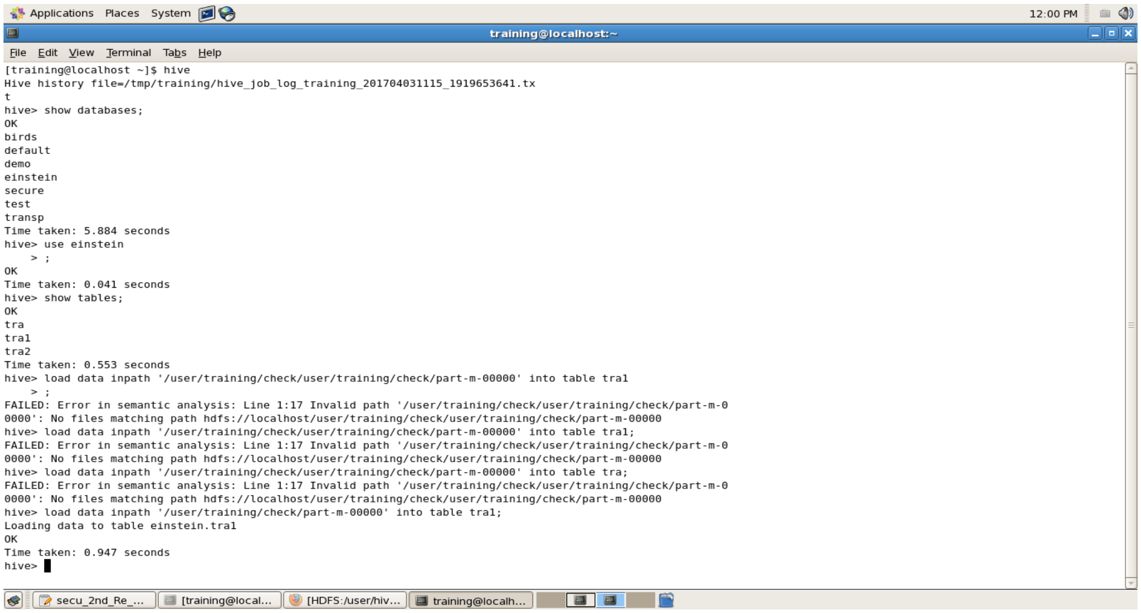
Local logs

Log directory

Cloudera's Distribution including Apache Hadoop, 2017.

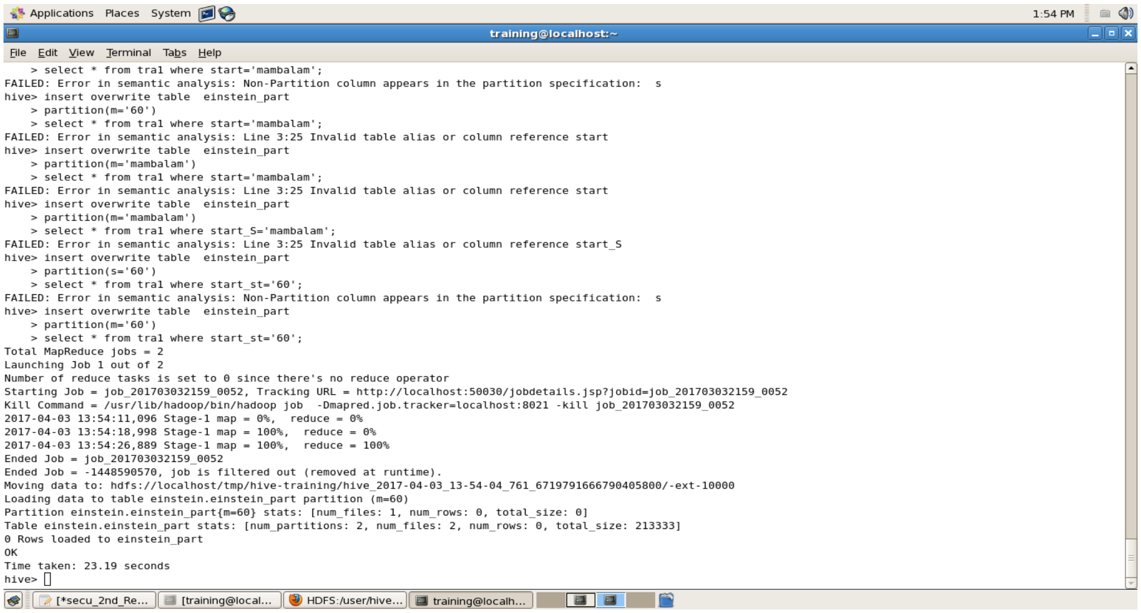
Done

Figure 4.10: Empty Bucket



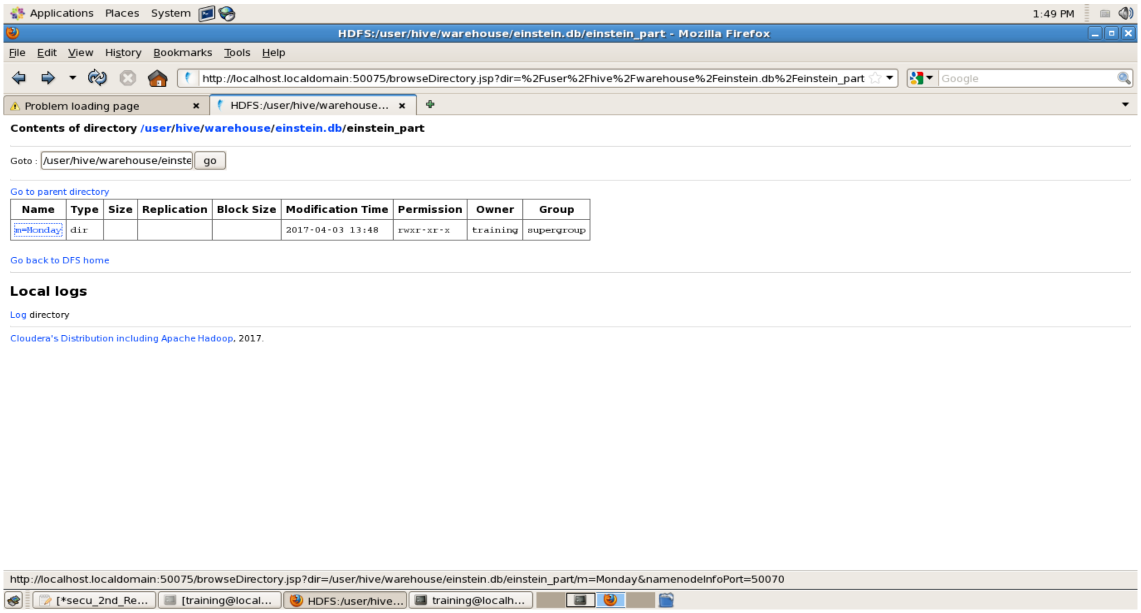
```
[training@localhost ~]$ hive
Hive history file=/tmp/training/hive_job_log_training_201704031115_1919653641.tx
t
hive> show databases;
OK
birds
default
demo
einstein
secure
test
transp
Time taken: 5.884 seconds
hive> use einstein
    > ;
OK
Time taken: 0.041 seconds
hive> show tables;
OK
tra
tra1
tra2
Time taken: 0.553 seconds
hive> load data inpath '/user/training/check/user/training/check/part-m-00000' into table tra1
    > ;
FAILED: Error in semantic analysis: Line 1:17 Invalid path '/user/training/check/user/training/check/part-m-0
0000': No files matching path hdfs://localhost/user/training/check/user/training/check/part-m-00000
hive> load data inpath '/user/training/check/user/training/check/part-m-00000' into table tra1;
FAILED: Error in semantic analysis: Line 1:17 Invalid path '/user/training/check/user/training/check/part-m-0
0000': No files matching path hdfs://localhost/user/training/check/user/training/check/part-m-00000
hive> load data inpath '/user/training/check/user/training/check/part-m-00000' into table tra;
FAILED: Error in semantic analysis: Line 1:17 Invalid path '/user/training/check/user/training/check/part-m-0
0000': No files matching path hdfs://localhost/user/training/check/user/training/check/part-m-00000
hive> load data inpath '/user/training/check/part-m-00000' into table tra1;
Loading data to table einstein.tra1
OK
Time taken: 0.947 seconds
hive> 
```

Figure 4.11: Load Data To Warehouse



```
    > select * from tra1 where start='mambalam';
FAILED: Error in semantic analysis: Non-Partition column appears in the partition specification:  s
hive> insert overwrite table  einstein_part
    > partition(m='60')
    > select * from tra1 where start='mambalam';
FAILED: Error in semantic analysis: Line 3:25 Invalid table alias or column reference start
hive> insert overwrite table  einstein_part
    > partition(m='mambalam')
    > select * from tra1 where start='mambalam';
FAILED: Error in semantic analysis: Line 3:25 Invalid table alias or column reference start
hive> insert overwrite table  einstein_part
    > partition(m='mambalam')
    > select * from tra1 where start_S='mambalam';
FAILED: Error in semantic analysis: Line 3:25 Invalid table alias or column reference start_S
hive> insert overwrite table  einstein_part
    > partition(s='60')
    > select * from tra1 where start_st='60';
FAILED: Error in semantic analysis: Non-Partition column appears in the partition specification:  s
hive> insert overwrite table  einstein_part
    > partition(m='60')
    > select * from tra1 where start_st='60';
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201703032159_0052, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201703032159_0052
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201703032159_0052
2017-04-03 13:54:11,096 Stage-1 map = 0%,  reduce = 0%
2017-04-03 13:54:18,998 Stage-1 map = 100%,  reduce = 0%
2017-04-03 13:54:26,889 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201703032159_0052
Ended Job = -1448590570, job is filtered out (removed at runtime).
Moving data to: hdfs://localhost/tmp/hive-training/hive_2017-04-03_13-54-04_761_6719791666790405800/-ext-10000
Loading data to table einstein.einstein_part partition (m=60)
Partition einstein.einstein_part{m=60} stats: [num_files: 1, num_rows: 0, total_size: 0]
Table einstein.einstein_part stats: [num_partitions: 2, num_files: 2, num_rows: 0, total_size: 213333]
0 Rows loaded to einstein_part
OK
Time taken: 23.19 seconds
hive> 
```

Figure 4.12: Hdfs Partition Queries


Applications
Places
System
1:49 PM
HDFS:/user/hive/warehouse/einstein.db/einstein_part - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
http://localhost.localdomain:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Feinstein.db%2Feinstein_part
Google
Problem loading page
HDFS:/user/hive/warehouse...

Contents of directory /user/hive/warehouse/einstein.db/einstein_part

Goto : /user/hive/warehouse/einste go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
m=Monday	dir				2017-04-03 13:48	rwxr-xr-x	training	supergroup

Go back to DFS home

Local logs

Log directory

Cloudera's Distribution including Apache Hadoop, 2017.

http://localhost.localdomain:50075/browseDirectory.jsp?dir=/user/hive/warehouse/einstein.db/einstein_part/m=Monday&namenodeInfoPort=50070

Figure 4.13: Bucket

Static

Static bucketing not much different two ways. static input data should contain column listed tables but it's not partition accordingly . Then if our input accordingly to expected layout. We are already partition for key value, if one for address and one for particular state.

Now we can loaded into partioned table with below syntax.

Where they stored address and state, This will create separate directory for partition under directory in hdfs..

Dynamic

Dynamic partition stored bulk of input file loading into hive tables. Dynamic partition support any number of partition with single execution. Hive will automatically split data with single separated file based upon key value which is present an input.

No need manual identification and its very easy coding for dynamic partition. First we need to check static key partition must come before dynamic partition. If user require dynamic partition first we need to check hive xml file.

4.5 Pig

Pig was initially to allow people using hadoop to focus more on analyzing. They allow user to focus more what they want and less how its done. As usually pig who they eat almost everything the main focus to execute everything. The pig programming language is designed to handle any kind of data. Pig is designed two component first one is pig Latin and another pig runtime itself. All data manipulation operation written on pig. Analyses large big dataset. Pig is a procedure language who works as step by step never jump another step, its basically step language never skip one of step other wise it does not proceed. Pig is a scripting language who convert script into map reduce programe.all script language support pig latin language. Internally script changes into mapreduce jobs. Pig has component which is commonly known as pig latin..

Apache pig is a latin language its simple if you are familiar with sql. it's some steps which is different because its load data into hdfs . As if, number of data are

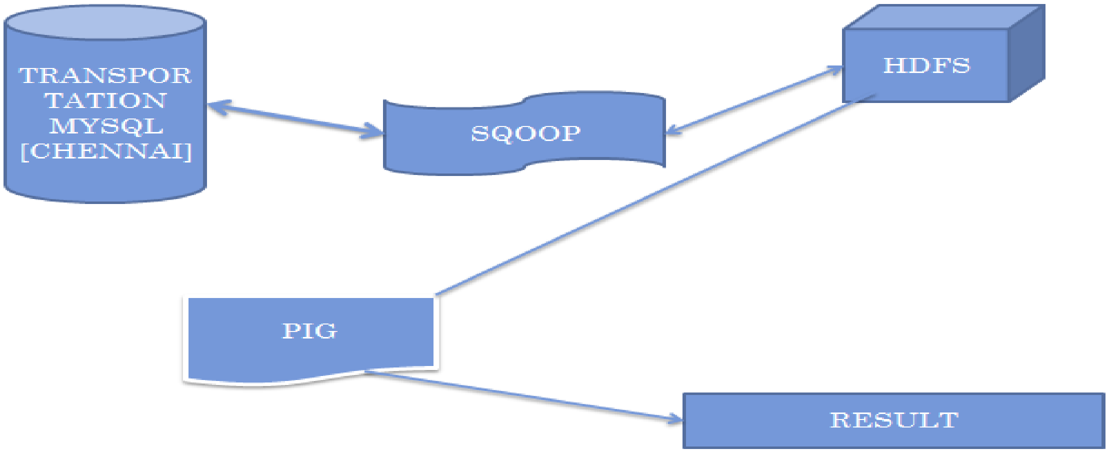


Figure 4.14: Architect For Pig

available its not possible to read data at minimum time in sql . Then pig gives query to manipulate operation with join, union ,filter ,ordering and some other condition. Pig latin language is multi programming its proceeding to reducing codes as compare to other programming code.

First step in a pig program is to load data you want to manipulate from hdfs. Then you run the data through a set of transformations which ,under the covers are translated into map and reduce tasks. Finally you dump the result to the screen or store result in a file somewhere. Just like typical map reduce job ,the data is stored in hdfs. In order to access data you first tell pig what file or files to use. Thats done through the load data file command. Data file can specify either hdfs file or directory. If a directory is specified ,all the files in the directory will be loaded into the program. if the data is stored in file format that is not accessible to pig then you can used load function which is user defined function which is read interpreted data. The transformation logic where all the data manipulation happens. All data manipulation here filter data group together with join operation and order result and much more. You can use dump command to send output to the screen, when debug program. The last step to store command. The result running your program are stored in a file further processing or analysis.

A=load data file

B=group

C=filter

Dump b

Store c into results;

Grunt is shell to run pig latin. Shell is provided utility to write script by grunt shell. From grunt shell ,we can invoke shell command. Pig latin using command sh and fs. Using sh command, we can invoke any shell commands from the Grunt shell. Using sh command from the Grunt shell. We cannot execute the commands that are a part of the shell environment. Grunt shell providing utility command. Utility program something different may its clear, help, history etc.

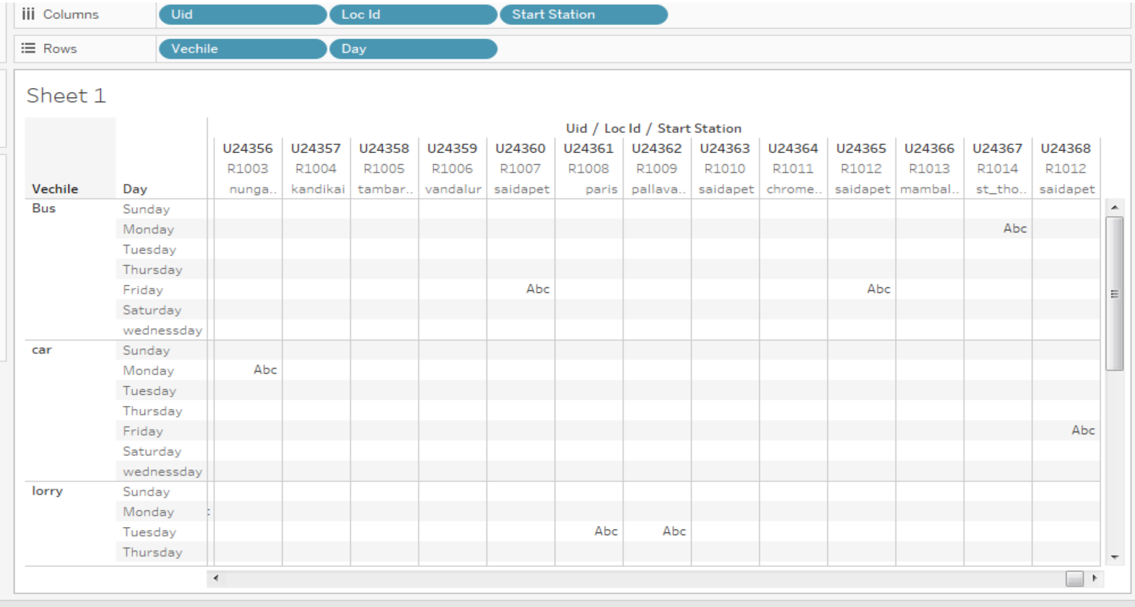
```
Grunt  >  shell-command

Grunt  >  sh ls

Grunt  >  sh clear
```

4.6 Tableau

Tableau can help to see how the data performing how much data fault or higher accuracy with the different distinct charts. the various variety of charts available.it allow to visualize perfect design or different aspect. To help out much more data visualize variety for visualizing .unlike R , both R and tableau work similar it is visually analyzing data. User can design interactive design for any intelligence as usually business, It tech.in behalf of data set, tableau interact with that dataset .its important which type of data you can used.



Columns: Uid, Loc Id, Start Station

Rows: Vechile, Day

Sheet 1

		Uid / Loc Id / Start Station												
		U24356	U24357	U24358	U24359	U24360	U24361	U24362	U24363	U24364	U24365	U24366	U24367	U24368
		R1003	R1004	R1005	R1006	R1007	R1008	R1009	R1010	R1011	R1012	R1013	R1014	R1012
Vechile	Day	nunga..	kandikai	tambar..	vandalur	saidapet	paris	pallava..	saidapet	chrome..	saidapet	mambal..	st_tho..	saidapet
Bus	Sunday													
	Monday												Abc	
	Tuesday													
	Thursday													
	Friday					Abc					Abc			
	Saturday													
	wednessday													
car	Sunday													
	Monday	Abc												
	Tuesday													
	Thursday													
	Friday													Abc
	Saturday													
	wednessday													
lorry	Sunday													
	Monday													
	Tuesday						Abc	Abc						
	Thursday													

Figure 4.15: Vechile Travel Day

Step1: Starting various data source are showing in bottom. Lets take step select anyone file ,server or saved that file as any name. different data source such as excel ,text, statically, access is as file but server as different such as sql, Microsoft server, tableau server, amazon, redfit and more server. The third and last saved data source already two file are there sample file.

Step 2: Put appropriate type of data and locating data at exact place. Select and measuring column side from analyzing dataset. When put both together they gives me proper performance of the dimension data.

Step3: Use to applying required visualization charts just like as scattered , map and other form.

Step 4: To visualize proper method or take proper result.

Software will allow specific data collaboration under corresponding different charts. In R also do same process proceed but the difference R used to be script.

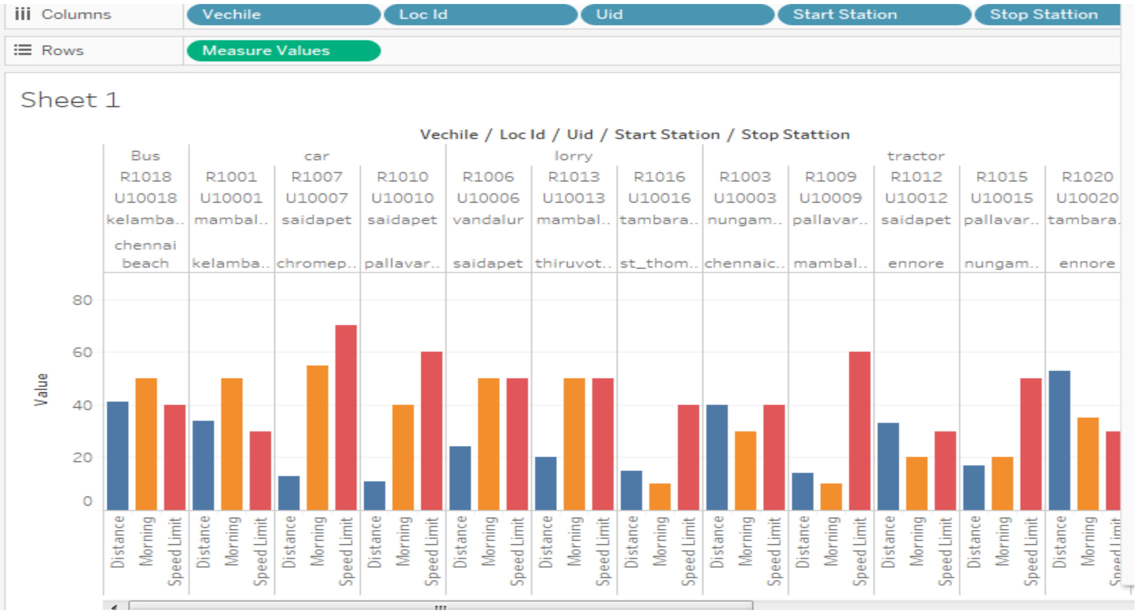
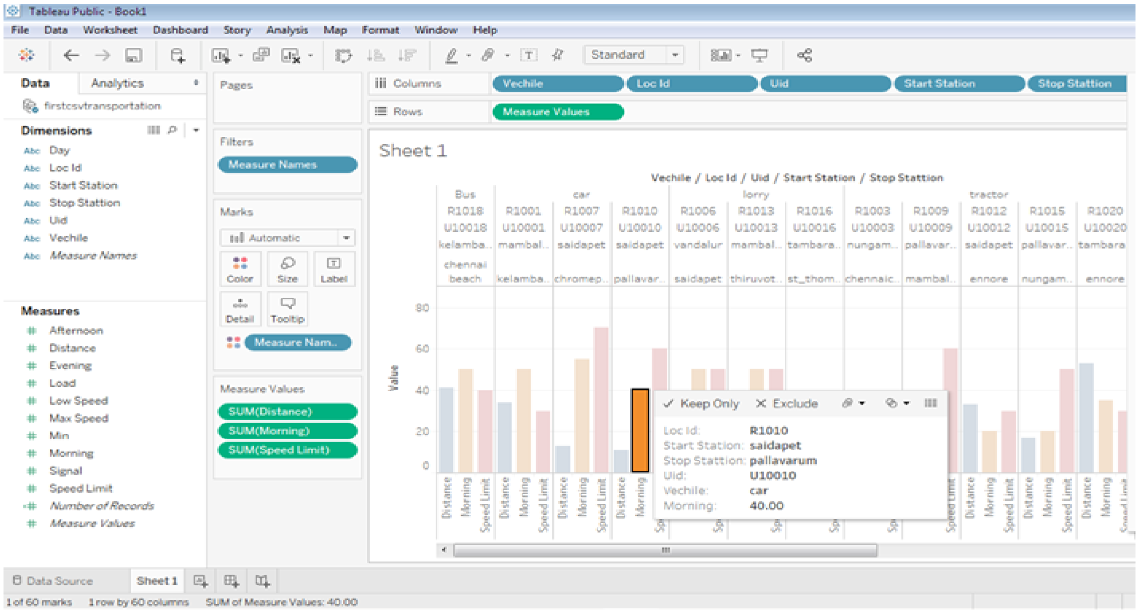


Figure 4.16: Histogram Chart



Loc Id: R1010
Start Station: saidapet
Stop Stattion: pallavarum
Uid: U10010
Vechile: car
Morning: 40.00

Figure 4.17: Particular Data Status

R script is totally different than tableau .R can take sql but firstly update R packages. Install all corresponding packages which is useful for showing result and its important package which need for sql sqldf if you not install R cant take sql data or never fetch those data which you want. But tableau it totally separate area you easily work on tableau just like drag drop or different aspect write some sparkql or sql queries its very simple than R.

4.7 System Testing

Testing is a technique to verify and identify software product.to verify each and every step by step.so many types are available v model ,spiral model ,..each step are unique work then go to next step. testing is process to check and proceed open next door. different steps are there indivisually check if any one is mistaken the whole process goes down .testing checking favour under pressure to show the capabilities of software. testing investigated the capabilities under test. testing showing quality of product inside test what type of service available its better work or not.Test technique are available to solving problem .Software testing finding bugs under test.

4.7.1 Automation Testing

Query Surgue

Query is a automation testing tool its number of allicance partenershib between data ware house and data based companies.the most popular alliance cloudera.

Query surge is hadoop automated hadoop testing will verify so many data at single time which will identify every columb.to extract all data with 100Demonstration of query surge the only test tool to automate data warehouse testing for ETL process,query search has four main modules the first being design library where you write your queries, The second being scheduling module where you can run immediately or anytime and anydate The third being of runtime dashboard where you watch execution of query Fourth being result reporting and deep dive into look at any results come back. Query based architecture based is an app server and agents launch sequel code at both the source and the target bring me information back and crunchy quickly.query search is web based and support any jdbc complaint database or data warehouse.we generally reffered to as a query pair is to logically related query test out and take data back from two sideds of an etl and test out the information that is being brought back so every query we have a source query and we have target query. Once those we are ready to collection so on one side of this we have got our source query which is againt transport database on the other sideof this we have got our target query which is going against our data warehouse now the execution in query surge is a test suite

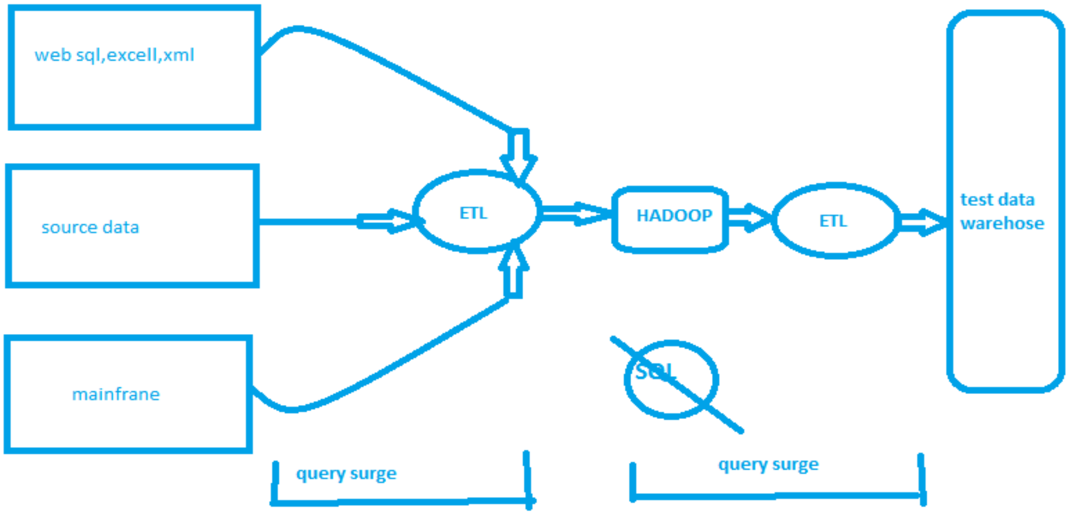


Figure 4.18: Automation Testing

4.7.2 Agile Testing

A method of software testing that follows the principles of agile software development. Individuals and interactions over processes That is, while there is value in the items on the right, we value the items on the left more and tools Working software over comprehensive documentation Customer collaboration over contract negotiation Responding to change over following a plan

This means that functional and usable software is valued over comprehensive but unusable documentation. Though this is more directed to upfront requirement specifications and design specifications, this can be true for test plans and test cases as well. Our primary goal is the act of testing itself and not any elaborate documentation merely pointing toward that goal. However, it is always best to have necessary documentation in place so that the picture is clear and the picture remains with the team if when a member leaves.

This means that the client is engaged frequently and closely in touch with the progress of the project (not through complicated progress reports but through working pieces of software). This does put some extra burden on the customer who has to collaborate with the team at regular intervals (instead of just waiting till the end of the contract, hoping that deliveries will be made as promised). But this frequent engagement ensures that the project is heading toward the right direction and not toward the building of a frog when a fish is expected.

This means accepting changes as being natural and responding to them without being afraid of them. It is always nice to have a plan beforehand but it is not very nice to stick to a plan, at whatever the cost, even when situations have changed. Lets say you write a test case, which is your plan, assuming a certain requirement. Now, if the requirement changes, you do not lament over the wastage of your time and effort. Instead, you promptly adjust your test case to validate the changed requirement. And, of course, only a FOOL would try to run the same old test case on the new software and mark the test as FAIL. This means that flexible people and communication are valued over rigid processes and tools. However, this does not mean that agile testing ignores processes and tools. In fact, agile testing is built upon very simple, strong and reasonable processes like the process of conducting the daily meeting or preparing the daily build. Similarly, agile testing attempts to leverage tools, especially for test automation, as much as possible. Nevertheless, it needs to be clearly understood that it is the testers who drive those tools and the output of the tools depend on the testers (not the other way round).

In conventional SDLC, only during the acceptance testing, the Business team

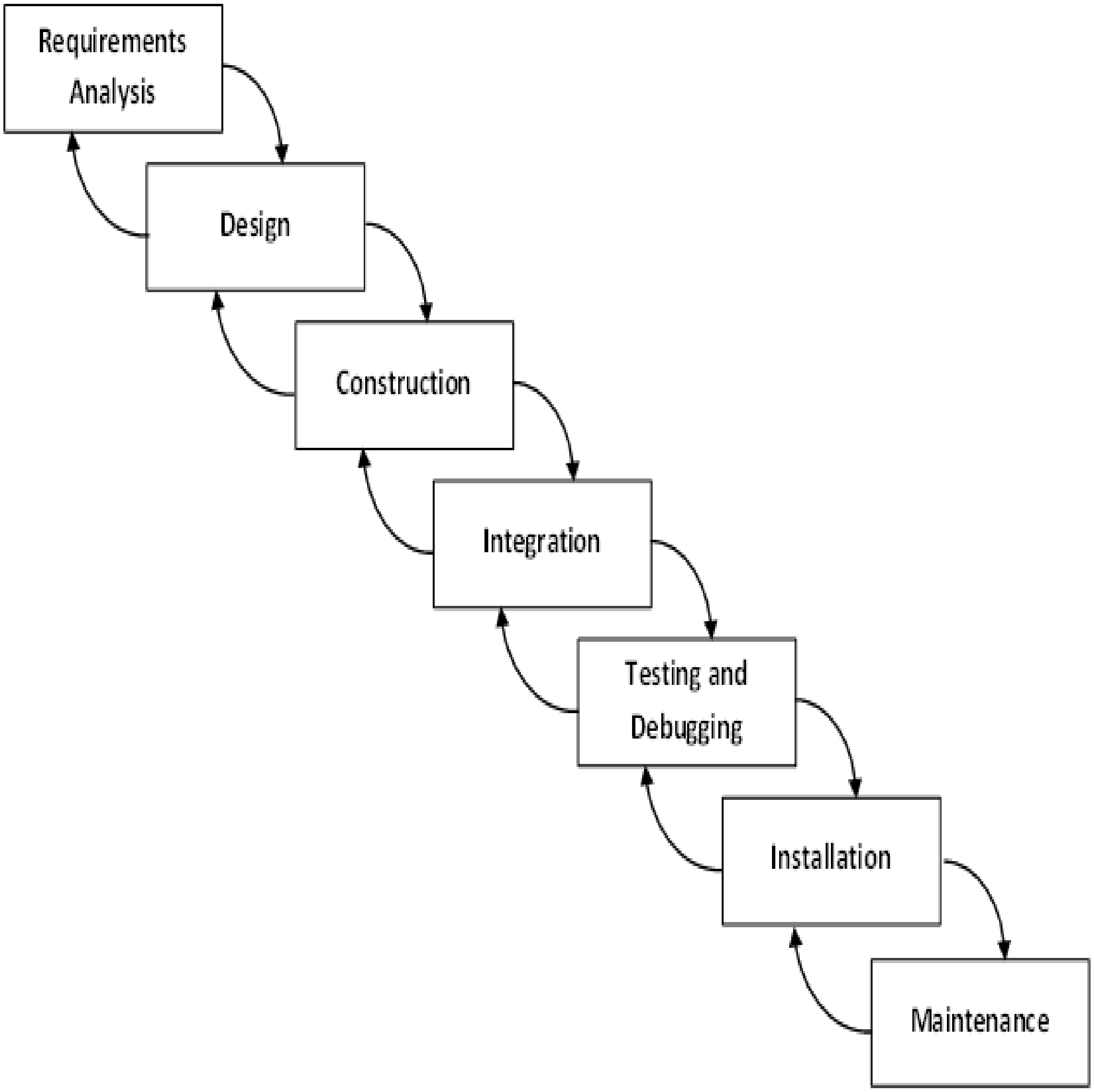


Figure 4.19: Agile Method

will get to know the product development, while in agile for each and every iteration, they are involved and continuous feedback shortens the feedback response time and cost involved in fixing is also less.

A software testing practice that follows the principles of agile software development is called Agile Testing. Agile is an iterative development methodology, where requirements evolve through collaboration between the customer and self-organizing teams and agile aligns development with customer needs.

- Agile Testing Saves Time and Money
- Less Documentation
- Regular feedback from the end user
- Raised defects are fixed within the same iteration and thereby keeping the code clean.
- Instead of very lengthy documentation, agile testers use reusable checklist, focus on the essence of the test rather than the incidental details.
- In conventional methods, testing is performed after implementation while in agile testing, testing is done while implementation.
- Daily meetings can help to determine the issues well in advance

4.7.3 Spiral Model

The spiral model combines the idea of iterative development with the systematic, controlled aspects of the waterfall model. This Spiral model is a combination of iterative development process model and sequential linear development model i.e. the waterfall model with a very high emphasis on risk analysis. It allows incremental releases of the product or incremental refinement through each iteration around the spiral. This phase starts with gathering the business requirements in the baseline spiral. In the subsequent spirals as the product matures, identification of system requirements, subsystem requirements and unit requirements are all done in this phase.

This phase also includes understanding the system requirements by continuous communication between the customer and the system analyst. At the end of the spiral, the product is deployed in the identified market. The Design phase starts

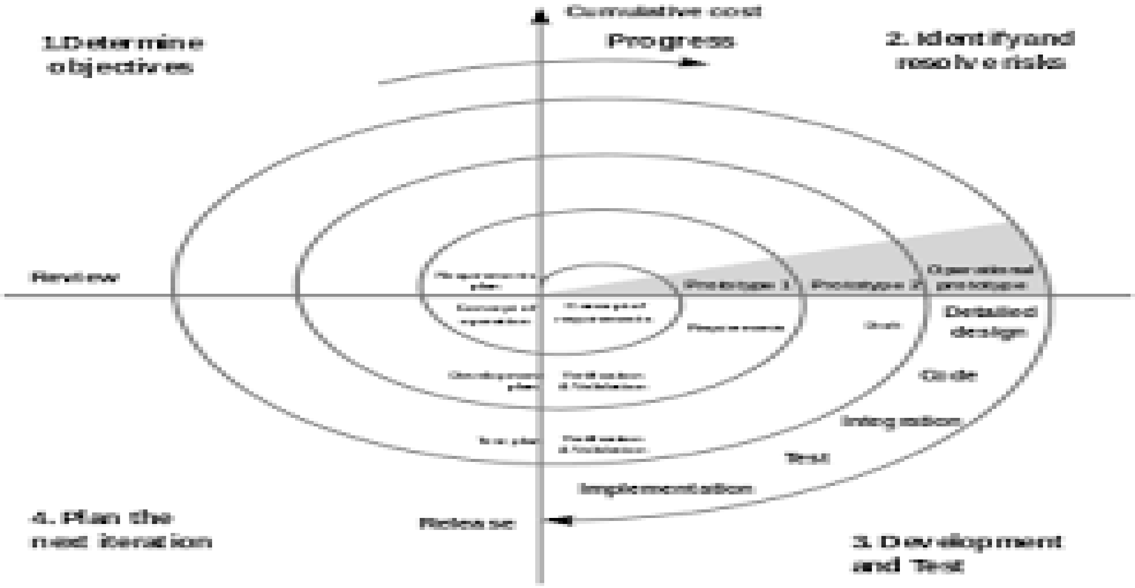


Figure 4.20: Spiral Model

with the conceptual design in the baseline spiral and involves architectural design, logical design of modules, physical product design and the final design in the subsequent spirals.

Risk Analysis includes identifying, estimating and monitoring the technical feasibility and management risks, such as schedule slippage and cost overrun. After testing the build, at the end of first iteration, the customer evaluates the software and provides feedback.

4.7.4 Spiral Model Application

The Spiral Model is widely used in the software industry as it is in sync with the natural development process of any product, i.e. learning with maturity which involves minimum risk for the customer as well as the development firms.

The following pointers explain the typical uses of a Spiral Model

When there is a budget constraint and risk evaluation is important.

For medium to high-risk projects.

Long-term project commitment because of potential changes to economic priorities as the requirements change with time.

Customer is not sure of their requirements which is usually the case.

Requirements are complex and need evaluation to get clarity.

New product line which should be released in phases to get enough customer feedback.

Significant changes are expected in the product during the development cycle.

Advantage And Disadvantage

The advantage of spiral lifecycle model is that it allows elements of the product to be added in, when they become available or known. This assures that there is no conflict with previous requirements and design.

This method is consistent with approaches that have multiple software builds and releases which allows making an orderly transition to a maintenance activity. Another positive aspect of this method is that the spiral model forces an early user involvement in the system development effort.

On the other side, it takes a very strict management to complete such products and there is a risk of running the spiral in an indefinite loop. So, the discipline of change and the extent of taking change requests is very important to develop and deploy the product successfully.

The advantages of the Spiral SDLC Model are as follows

- Changing requirements can be accommodated.
- Allows extensive use of prototypes.
- Requirements can be captured more accurately.
- with the tools specifying data manipulation or condition also define.
- Development can be divided into smaller parts and the risky parts can be developed earlier which helps in better risk management.
- The results can be tabulated, graphically presented to be displayed if any.
- showing better view by with the best tableau tool. To help out much more data visualize variety for visualizing. Tableau gives technique connect with the sql as well as sparkQl also.

The disadvantages of the Spiral SDLC Model are as follows

- Management is more complex.
- End of the project may not be known early.
- Not suitable for small or low risk projects and could be expensive for small projects.
- Process is complex
- Spiral may go on indefinitely.
- The results can be tabulated, graphically presented to be displayed if any.
- Large number of intermediate stages requires excessive documentation

4.8 strategy cases

The test strategy is a level to identifying level performance. There are so many different level unit testing, integration testing and system testing. Most of the testing are organizational, developer are responsible for strategies. it's design process to find out and fix the defects before they design.

Table 4.1: Transportation File

User id	Distance	StartStation	StopStation
uid1	70	kelambakkam	guindy.
uid3	279	tambarram	kelambakkam
uid5	250	tambaram	kovalam.
uid1	70	kovalam	guindy.
uid3	279	pallavaram	kelambakkam
uid7	601	tambaram	kovalam.
uid1	70	kovalam	guindy.
uid3	279	pallavaram	kelambakkam
uid9	250	tambaram	kovalam.
uid7	450	kovalam	guindy.
uid5	701	pallavaram	kelambakkam
uid9	250	tambaram	kovalam.
uid2	262	chrompet	kelambakkam.
uid1	70	kelambakkam	guindy.
uid3	279	tambarram	kelambakkam
uid9	250	tambaram	kovalam.
uid7	450	kovalam	guindy.
uid5	701	pallavaram	kelambakkam
user9	250	tambaram	kovalam
uid2	262	chrompet	kelambakkam.
uid1	70	kelambakkam	guindy.
uid3	279	tambarram	kelambakkam
uid9	250	tambaram	kovalam.
uid7	450	kovalam	guindy.
uid5	701	pallavaram	kelambakkam

Table 4.2: Extract Uid

Userid	Distance	StartStation	StopStation
uid5	250	tambaram	kovalam.
uid	279	pallavaram	kelambakkam
uid9	279	guindy	kelambakkam

Table 4.3: Extract Signal

Userid	Distance	Signal	Day
uid1	70	6	thursday.
uid3	279	4	monday
uid9	279	5	friday

Chapter 5

Results And Discussions

transportation dataset are different column to identify by different section. each columb specify a particular name.

- In those tool result will be perfect.
- every result has unique in as much as user gives result in particular area
- a big dataset are load into hdfs then when result comes ,its automatically create particular file for better to know what happen in result
- with the tools specifying datamanipulation or condition also define.
- some bucketing or partition technique available to finding best result
- The results can be tabulated, graphically presented to be displayed if any.
- showing better view by with the best tableau tool. To help out much more data visualize variety for visualizing.tableau gives technique connect with the sql as well as sparkQl also.
- where transport are maximum in which particular location easily know
- To identify how much distance between some particulat between those two route on those route. Whose user id start going on.
- basically, number of user id going and which one going to which location because if you want particular vehicle information then isnot easy to see whole dataset . how many vechile ,which vehicle, load and other knowing information with the tool

Chapter 6

Conclusion And Future Work

6.0.1 Conclusion

The world has become a massive interconnecting of smart device that generate data continuously and this is extremely relevant to transport. Disaster response can be simulated planned in advance option and impacts can be determined.

Smart city with big data are modern and important concepts; therefore, many started integrating them to develop smart city transportation applications that will help reach sustainability, better resilience, effective governance, enhanced quality of life, and intelligent management of smart city transportation resources. Our study explored both concepts and their different definitions and we came to identify some common attributes for each. Despite the varying definitions each concept has a number of characteristics that uniquely defines it. Relying on these common characteristics, we were able to identify the general benefits of using big data to design and support smart city transportation applications.

Intelligent Transport Systems (ITS, the transport telematics) deals with applications of information and communication technologies for enhancement of transport performance, economies, efficiency, safety, ecology and comfort of transport. Basic functions of ITS that are applicable to all modes of transport include traffic control and management, intelligent vehicles, electronic fee collection, management of rescue systems and corps, management of public mass transport, route planning, provision of traffic information, fleet management and freight transport logistics.

today more challenges to building and deploying successful smart city transport will require the following design ,its more enhance to identified several issues that may hinder big data application. There requirement and efficiently work are more compact table.

6.0.2 Future Work

Apache spark is a open source distributed system. Apache new version of hadoop which is highly popular. Its developed by apache foundation .Apache spark not using single programming language .Apache foundation built to three different language scale, java, python with minimum coding and given better result. Apache spark run fast its better than hadoop fields distributes which built parallasium which parallel work with different cluster. Quick overview of spark api. Spark api using RDD provides example: machine learning api. two most popular language recently used in IT are java and python both are have own collection who reducing size with number of collection. Data frame and machine learning organized into columns to perform with dataset, those set perform parallel vice. Using Apache spark ,the main reason for apache spark takes less time than map reduce programmer is to maintain cost effective gathering amount of less space comparison to other .its very flexible, scalable. Apache spark work on own cluster but with the help of hadoop. Hadoop gives main component because spark has no own part it storage in hadoop cluster. Hadoop gives storage of component to spark. Spark have different types of api that enables high through output for live streaming data. Spark streaming run both trained or untrained data. Already existing file which means excel, text file to change anytime under the file can change on that time and fetch ,this means trained data. Untrained data set its automatically come from other source as social sites like dislike. Spark streaming is a core spark api. From many source, data can be take like kafta, flumes, kinies or tcp socket. using complex algorithm are expressed can be processed with high level function like map, reduce, join etc.

I can take same transport dataset using with different api. Using a complex algorithm used for mlib(machine learning) and graphs. Transport dataset as input data file which ingest input stream .behind stream input file distributed or divides into batches. Which takes action by spark engine.

Appendices

6.1 Coding

```
                \*{mysql code}

[root@host]# mysql -u root -p
 Enter password: ******

Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is to server version: 5.0.9

Type ’help;’ or ’\h’ for help. Type ’\c’ to clear the buffer.

mysql> exit

mysql>show databases trans;
mysql>show *from trans;
mysql>exit;

mysql>use database trans;
mysql>INSERT INTO trans ( uid, signal,addresss,starting_location )
             VALUES ( m, 5,kelambakkam ),( m, 5,kelambakkam );


mysql>SELECT * from tutorials_tbl WHERE
       tutorial_author=’kelambakkam’;
```

```
+-------------+----------------+-----------------+------------------
| uidd      | signal |  starting location  |
+-------------+----------------+-----------------+------------------
|           1| 5         | kelambakkam       |
+-------------+----------------+-----------------+------------------
1 rows in set (0.01 sec)


mysql>ALTER TABLE testalter_tbl ADD u2 INT;

+-------------+----------------+-----------------+------------------
| uidd      | signal |  starting location       |
+-------------+----------------+-----------------+------------------
|    1        |     5         | kelambakkam       |
+-------------+----------------+-----------------+------------------
|    u2            | 5        |     tambaraam      |
+-------------+----------------+-----------------+------------------

2 rows in set (0.00 sec)
mysql>ALTER TABLE trans ADD user2 INT;


mysql> -u root -p transpotation  tansl > trans.txt
password :******


+-------------+----------------+-----------------+------------------
| uidd            | signal |  starting location        |
+-------------+----------------+-----------------+------------------
|    1             |     5    | kelambakkam              |
+-------------+----------------+-----------------+------------------
|    u2            | 5        |            tambaraam      |
+-------------+----------------+-----------------+------------------
| uidd            | signal|    kandigai                 |
+-------------+----------------+-----------------+------------------
|    1             |     5    | tambaram                   |
+-------------+----------------+-----------------+-----------------
|    u2           | 5       |     tambaraam            |
+-------------+----------------+-----------------+-----------------
| uidd            | signal|  starting location        |
+-------------+----------------+-----------------+-----------------
|    1        |     5         | kelambakkam             |
+-------------+----------------+-----------------+-----------------
|    u2           | 5       |     tambaraam            |
+-------------+----------------+-----------------+-----------------
```

```
 \*{Sqoop}
java version


java version "1.7.0_71"
Java(TM) SE Runtime Environment (build 1.7.0_71-b13)
Java HotSpot(TM) Client VM (build 25.0-b02, mixed mode)

export JAVA_HOME=/usr/local/java
export PATH=$PATH:$JAVA_HOME/bin

source ~/.bashrc

hadoop version

Hadoop 2.4.1
--
Subversion https://svn.apache.org/repos/asf/hadoop/common -r 152976
Compiled by hortonmu on 2013-10-07T06:28Z
Compiled with protoc 2.5.0
From source with checksum 79e53ce7994d1628b240f09af91e1af4
```

```
sqoop-version

$ sqoop import \
--connect jdbc:mysql://localhost/userdb \
--username root \
--table trans --tra u1

14/12/22 15:24:54 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5
14/12/22 15:24:56 INFO manager.MySQLManager: Preparing to use a MySQ
14/12/22 15:24:56 INFO tool.CodeGenTool: Beginning code generation
14/12/22 15:24:58 INFO manager.SqlManager: Executing SQL statemen
14/12/22 15:24:58 INFO manager.SqlManager: Executing SQL statemen
14/12/22 15:24:58 INFO orm.CompilationManager: HADOOP_MAPRED
14/12/22 15:25:11 INFO orm.CompilationManager: Writing jar file
-----------------------------------------------------
-----------------------------------------------------
14/12/22 15:25:40 INFO mapreduce.Job: The url to track the job: http
14/12/22 15:26:45 INFO mapreduce.Job: Job job_1419242001831_0001 run
14/12/22 15:26:45 INFO mapreduce.Job: map 0% reduce 0%
14/12/22 15:28:08 INFO mapreduce.Job: map 100% reduce 0%
14/12/22 15:28:16 INFO mapreduce.Job: Job job_1419242001831_0001 com
-----------------------------------------------------
-----------------------------------------------------
14/12/22 15:28:17 INFO mapreduce.ImportJobBase: Transferred 145 byte
14/12/22 15:28:17 INFO mapreduce.ImportJobBase: Retrieved 5 records

$ sqoop import \
--connect jdbc:mysql://localhost/userdb \
--username root \
--table city \
--m 1 \
--where city =kelambakkam \
--target-dir /u1

$ sqoop import \
--connect jdbc:mysql://localhost/userdb \
--username root \
--table trans\
--m u1 \

$ sqoop eval \
--connect jdbc:mysql://localhost/db \
--username root \
--query SELECT * FROM  trans LIMIT 3

+-------------+----------------+-----------------+------------------
| uidd              | signal              |  starting loc     |
+-------------+----------------+-----------------+------------------
|    1               |     5                 |  kelambakkam |
+-------------+----------------+-----------------+------------------
|    u2           |    7                  |  tambaraam    |
+-------------+----------------+-----------------+------------------
|    u3           |    4                  |  tambaraam    |
+-------------+----------------+-----------------+------------------


$ sqoop list-tables \
--connect jdbc:mysql://localhost/trans/
--username root

25/04/17 15:05:08 INFO manager.MySQLManager: Preparing to use a MyS


userid

vechile

stop



               \*{HiveQL}

hive> CREATE SCHEMA trans;
hive> SHOW DATABASES;
     default
     transp

hive> CREATE TABLE IF NOT EXISTS employee ( uid String, stop_st Str
start_st, signal String, location String)
COMMENT transportation
ROW FORMAT DELIMITED
FIELDS TERMINATED BY \t
LINES TERMINATED BY \n
STORED AS TEXTFILE;

OK
Time taken: 5.890 seconds

hive> ALTER TABLE tra RENAME TO travel;


bucketing
-------------
CREATE TABLE bucketed_user
( uid VARCHAR(64),
signal  VARCHAR(64),
start_st  STRING,
stop_st  VARCHAR(64),
location VARCHAR(64),
evening STRING,
morning VARCHAR(64),
afernoon    STRING,
day  STRING,
month     STRING
year      String
)
COMMENT  ’A bucketed sorted transportation table’
PARTITIONED BY (signal VARCHAR(64))
CLUSTERED BY (day) SORTED BY (city) INTO 5 BUCKETS
STORED AS SEQUENCEFILE
```

```
INSERT OVERWRITE TABLE bucketed_user PARTITION (signal)
SELECT
uid ,signal,
start_st,
location,
state,
evening,
morning ,
afternoon,
day,
month,
year,
FROM transp;
```

```
Logging initialized using configuration in jar
OK

Time taken: 12.144 seconds
OK
Time taken: 0.146 seconds
Loading data to table default.temp_user

Table default.temp_user stats: [numFiles=1, totalSize=283212]
OK

Time taken: 0.11 seconds
OK
Time taken: 0.15 seconds

Query ID = user_20141222163030_3f024f2b-e682
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 25
In order to change the average load for a reducer (in bytes):
```

```
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1419243806076_0002, Tracking URL
Kill Command = /home/user/bigdata/hadoop-2.6.0/bin/hadoop
Hadoop job information for Stage-1: number of mappers: 32;
2017-4-22 16:30:36,164 Stage-1 map = 0%,  reduce = 0%,
2017-4-22 16:31:09,770 Stage-1 map = 100%,  reduce = 0%,
2017-4-22 16:32:10,368 Stage-1 map = 100%,  reduce = 1%,
2017-4-22 16:32:28,037 Stage-1 map = 100%,  reduce = 83%,
2017-4-22 16:32:36,480 Stage-1 map = 100%,  reduce = 74%,
2017-4-22 16:32:40,317 Stage-1 map = 100%,  reduce = 89%,
2017-4-22 16:33:40,691 Stage-1 map = 100%,  reduce = 69%,
2017-4-22 16:33:54,846 Stage-1 map = 100%,  reduce = 51%,
2017-4-22 16:33:58,642 Stage-1 map = 100%,  reduce = 48%,
2017-4-22 16:34:52,731 Stage-1 map = 100%,  reduce = 56%,
2017-4-22 16:35:21,369 Stage-1 map = 100%,  reduce = 63%,
2017-4-22 16:35:22,493 Stage-1 map = 100%,  reduce = 75%,
2017-4-22 16:35:53,559 Stage-1 map = 100%,  reduce = 44%,
2017-4-22 16:36:14,301 Stage-1 map = 100%,  reduce = 100%,
MapReduce Total cumulative CPU time: 44 seconds 430 msec
Ended Job = job_1419243806076_0005
Loading data to table default.bucketed_user partition

Time taken for load dynamic partitions : 2545
Loading partition {signal=kelambakkam}
Loading partition {signal=kandigai}
Loading partition {signal=mambakkam}
Loading partition {signal=chennai_central}
Loading partition {signal=egmor}
Time taken for adding to write entity : 15

Partition default.bucketed_user [numFiles=52, numRows=510]
Partition default.bucketed_user [numFiles=52, numRows=510]
Partition default.bucketed_user [numFiles=52, numRows=510]
Partition default.bucketed_userstats: [numFiles=52, numRows=510]
Partition default.bucketed_user [numFiles=52, numRows=18]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 52   Cumulative CPU: 54.13 sec
Total MapReduce CPU Time Spent: 54 seconds 130 msec
OK


Time taken: 387.986 seconds
user@tri03ws-376:~$
```

```
A = LOAD ’input.txt’ USING tra(’,’);
               or
A = LOAD ’input.txt’ USING tra(’\t’);
```

```
DUMP A;
(vechile,signal,stop_st)
```

```
2017-4-17 23:03:04,550 [main] INFO org.apache.pig.tools

2017-4-17 23:03:04,633 [main] INFO org.apache.pig

{RULES_ENABLED=[AddForEach, ColumnMapKeyPrune
LoadTypeCastInserter, MergeFilter, MergeForEach
PartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter],

RULES_DISABLED=[FilterLogicExpressionSimplifier]}

2017-4-17 22:03:04,748 [main] INFO org.apache.pig.backend
concatenation threshold: 100 optimistic? false
```

```
2017-4-17 22:03:04,805 [main] INFO org.apache.pig.backend

2017-4-17 22:03:04,805 [main] INFO org.apache.pig.backend.hadoop.
plan size after optimization: 1

2017-4-17 22:03:04,853 [main] INFO org.apache.pig.tools.pigstats


 HadoopVersion PigVersion UserId StartedAt FinishedAt Features
1.1.2 0.12.0 hduser 2017-4-17 22:03:04 2017-4-17

Success!

Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local_0001  file:/tmp111343,

Input(s):
Successfully read records from: "/home/transportation_data.csv"

Output(s):
Successfully stored records in: "file:/tmp/temp-1826/tmp11543"


{RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter
PartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter],

RULES_DISABLED=[FilterLogicExpressionSimplifier]}

2017-4-17 22:03:04,748 [main] INFO org.apache.pig
concatenation threshold: 100 optimistic? false

2017-4-17 22:03:04,805 [main] INFO org.apache.pig.backend.
plan size before optimization: 1

2017-4-17 22:03:04,805 [main] INFO org.apache
plan size after optimization: 1
```

```
2017-4-17 22:03:04,853 [main] INFO org.apache.pig

Job DAG:
job_local_001

(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)

now, set data set into hdfs
grunt>  hadoop fs -put tra
         tra=load ’\user\training\tra’ using pigi(’,’)
         As(uid:long,loc:chararray,vec:chararray,load:int);

grunt>describe tra;

grunt>store tra1 into ’trasparent

grunt> Group_tra1 =Group_tra1 by (’loc’);

grunt>Store group_tra2 = Group_tra1 by (’loc’);

grunt> store tra_2 into ’group tra_2 result ;’




grunt>A = load ’student’ AS (name:chararray,age:int,gpa:float);

DESCRIBE A;
A: {uid: chararray,vechile: int,start_st: chararray}

DUMP A;
(uid5,car,mambakkam)
(uid15,truck,kelambakkam)
```

```
(uid6,car,tambarram)
(uid6,loader,kandigai)


filter
-------

X = FILTER A BY tra == kandigai;

(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)

(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)


grunt> tra = LOAD 'hdfs://localhost:50016/trans.txt'
         as (uid:int, vechile:chararray, load:chararray);

grunt> foreach_tra = FOREACH tra GENERATE uid,signal;

grunt> Dump foreach_tra;

grunt>(uid,4)
(uid,1)
(uid,2)
(uid,6)
(uid,1)
(uid,3)
(uid,3)
(uid,6)
(uid,1)
```

```
(uid,1)
(uid,5)
(uid,4)
(uid,5)
(uid,3)
(uid,5)
```

```
grunt> trans_greater = FILTER signal BY (float) signal>4;
grunt> DUMP trans_greater;

(uid1,car,2017,evening,5)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)
(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)
(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)
(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)
(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)
(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
```

```
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)
(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)
(uid50,car,morning,mambalam,2017,7)
(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,6)
(uid1,car,5,2017,evening)
(uid7,truck, evening,tambaram,2017,,6)
(uid20,loader,afternoon, kelambakkam,2017,5)
(uid39,tractor,evening,tambaram,7)


grunt> store trans_than_four into ’/user/hduser/trans_greater’;



word count
--------------

grunt>lword = LOAD ’/user/hadoop/trans.txt’;
grunt>word = FOREACH lword GENERATE FLATTEN(TOKENIZE(line));
grunt>groupup = GROUP words BY lword;
grunt>wordcount = FOREACH groupup GENERATE group, COUNT(word);

grunt>DUMP wordcount;

(uid20,12)
(uid39,70)
(uid50,75)
(uid1,45)
(uid7,40)
(uid20,12)
(uid39,10)
(uid50,47)
```

```
(uid1,23)
(uid7,40)
(uid20,12)
(uid39,10)
(uid50,47)
(uid1,23)
(uid7,40)
(uid1,23)
(uid7,40)
(uid20,12)
(uid39,10)
```

```
2017-4-17 22:03:04,805 [main] INFO org.apache.pig.backend.hadoop.
executionengine.mapReduceLayer.MultiQueryOptimizer
- MR plan size before optimization: 1

2017-4-17 22:03:04,805 [main] INFO org.apache.pig.backend.hadoop.
executionengine.mapReduceLayer.MultiQueryOptimizer
- MR plan size after optimization: 1

2017-4-17 22:03:04,853 [main] INFO org.apache.pig.tools.pigstats
ScriptState- Pig script settings are added to the job
```

Bibliography

[1] Apache Hive, http://hive.apache.org/. *Apache Hive Wiki: https://cwiki.apache.org/Hive/. Jeffrey Dean and Sanjay Ghemawat, MapReduce: simplified data processing on large clusters,.* IEEE Trans. on computers, vol.64, no. 2, pp.425-437, 2015.

[2] The Apache Software Foundation. Retrieved Nov 1, 2010. PLoS ONE, vol. 9, no.9, article;e106925, 2014.

[3] ”Apache Sqoop - Overview”

[4] Corman H. Thomas, Leiserson E. Charles, Rivest L. Ronald, *Stein Clifford Introduction to Algorithms Second Edition* McGrawHill Book Company

[5] cheng xueqi jin xiaolong wang zhuyuan guo on big system and analysis technolog *I jiafeng zhang liguoji*. IEEE Trans. on computers, vol.64, no. 2, pp.425-437, 2015.

[6] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: *A Distributed Storage System for Structured Data. ACM* Transactions on Computer Systems (TOCS), 26(2):4, 2008.

[7] ”Hadoop: Apache Sqoop *Identity Based Encryption with Outsourced Revocation in Cloud Computing*. IEEE Trans. on computers, vol.64, no. 2, pp.425-437, 2015.

[8] ”Hadoop: Apache Sqoop *Identity Based Encryption with Outsourced Revocation in Cloud Computing*. IEEE Trans. on computers, vol.64, no. 2, pp.425-437, 2015.

[9] Hive A Warehousing Solution Over a MapReduce Framework

[10] ”Hadoop: Apache Sqoop *Identity Based Encryption with Outsourced Revocation in Cloud Computing*. IEEE Trans. on computers, vol.64, no. 2, pp.425-437, 2015.

[11] ” Optimising Hadoop and Big Data with Text and HiveOptimising Hadoop and Big Data with Text and Hive Major technical advanced hive , Yin Huai, Ashutosh Chauhan, Alan Gates, Gunther *Hagleitner, Eric N. Hanson, Owen OMalley, Jitendra Pandey, Yuan Yuan, Rubao Lee and Xiaodong Zhang, SIGMOD* 2014

[12] Pig user defined functions. R *The Free Hive Book (CC by-nc licensed)* .

[13] pig cookbook . IEEE Trans. on computers, vol.64, no. 2, pp.425-437, 2015.

[14] Sqoop Export 2015-12-10. Archived from the original on 2015-12-10. Retrieved 2015-12-10. The Sqoop Export job allows you to export data from Hadoop into an RDBMS using Apache Sqoop. *Transactions on Computer Systems (TOCS), 26(2):4, 2008* . IEEE White, Tom (2010). Hadoop: The Definitive Guide. O’Reilly Media. ISBN 978-1-4493-8973-4.