# Report

*by* Deeja Chhabra

**Submission date:** 17- May- 2018 10:24AM (UTC+0530)
**Submission ID:** 964924812
**File name:** Download_File.pdf (1.12M)
**Word count:** 5639
**Character count:** 29450

# Movie Review Analysis using Hive

Submitted in partial fulfilment of the requirements of the degree of

Bachelor of Technology

by

Soinaya Rao (140240)
Deeja Chhabra(140078)

under the supervision of

Dr. Anil Kumar Dahiya

MODY UNIVERSITY
LAKSHI1ANGARH • RAJASTHAN

A Leading Women's University

COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF ENGINEERING AND TECHNOLOGY

MODY UNIVERSITY OF SCIENCE AND TECHNOLOGY, LAKSHMANGARH

May.2018

Certificate

This is to certify that the report entitled Mot ie Rei ies Analysis using Hit e has been undertaken and » rittcn under my superb ision and it describes the original research iv carried out by Deeja Chhabra (140078). Somaja Rao(140240). in Computer Science and Engineering for the dcgrcc of B.Tcch. To the best of my knowledge and belief, this › ork is original and has not been submitted elsewhere for any degree from any other institution in India or abroad.

Dr. Anil Kumar Dahiya
Assistant Profcssor

Dr. Anil Kumar Dahija
Hcad of Dcpartmcnt

Apt›rox aI Certil'icate

This report entitled Movie Rei ie» Analysis using hive by Soinaja Rao, Deeja Chhabra and is approved for the degree of Bachelor of Technology.

Examiner(s)

_____

_____

Super isor(s)

_____

_____

Datc _____

## Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea / data / fact / source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or thorn whom proper permission has been taken when needed.

I also attest that this Business Director report is free of plagiarism of any kind and that I am fully aware of the Plagiarism Prevention Policy of Mody University of Science and Technology.

Somaya Rao(140240)

Deeja Chhabr(140078)

Date:

07.05.2018

# ACKNOWLEDGEMENT

Wc u ould likc to exprcss spccial thanks of gratitudc to thc Dr. V.K Jain **(Dean CET}** and **Dr.Anil Kumar Dahiy'a** (HOD CSE) for introducing Major Project in our curriculum i› li ich liclpcd us iii line ing hands on certain good tcchnologics.

We » ould also likc to thank our projcct Supcin isor Dr. Anil Kumar Dalii¿a for li is support and guidaiicc throughout the schedule of thc proiect. This project could not have been completed u'ithout his help.

Thank you all.

Dee ja Chhabra(l40078)

Somaya Rao(l40240)

# ABSTRACT

Big dota usually includes datn sets u ith sizes bGyond the ability of commonly used software tools to capture, curate, manage. and process data within a tolerable elapsed time

Hadoop is an open source, Java-based programming frameu ork that supports the processing and storage of cxtrcmcl; largc data scts in a distributcd computing cm ironmcnt. It is part of the Apnchc project sponsorcd bi thc Apache Soft› arc Foundation.
Hadoop quickly cincrgcd as a foundation for big data processing tasks. such as scicnlific analytics. businccss and salcs planning, and proccssing cnorinous › oluincs of scnsor data. including from internet of things sensors.

Hit c is a data »‘archousc infrastructure tool to proccss structurcd data in Hadoop. It rcsidcs on top of Hadeop to summarize Big Data, and makes querying and analyzing easy.
It pro idcs a siinplc qucq languagc callcd Hi c QL, » hich is bascd on SQL and u hich cnablcs users familiar u ith SQL to easily perform ad-hoc querying. summarization and data analysis

With ei ery releasing mod res u hether it is hit or flop, by generating rei ieu s to the audience u e are hcrc to transform lhc › oluminous raw data into somc mcaningful information bj applj ing Big Data Analytics

CONTENTS

# LIST OF FIGURES

# LIST OF TABLE

# CHAPTER 1
# Introduction

## 1.1 Introducion to bigdata

Big Data is a tenn that refers to dauiset » hose size or i olume. complexity.i arietj ,rate of gro» th or eracity of data » hich organixations handled hat e reached such unbelief able hi'cl thnt traditional processing and analytical tools failed to process. A data ii hich is bcjond to the storage capacity and »'liich is bejoiid to the processing po» er.that data » e can call as Big Data.

Or er the last fe» years, there has been an incredible explosion of data IBM reported that 2.5 billion gigab; tes of data ii as gcncrated ci en' day in 2012 Data is grow ing fastcr than ci'cr befor,e then one solution » as » e can use iiiultiple computers called distributed systems. but in this there are high chances of sj stein failure. programming complexity is also high as it is difficult to synchronizc daia and process . then comes thc anothcr solution known as Hadoop.

## 1.2 Ha‹Iooy:Solution for Big ‹lata

Hadoop is n fraincii ork that allow s for distribntcd processing of large daia scts across clusters of commodity computers using simple programming models It is inspired bj technical document published by Google. The » ord 'hadoop does not liar e aiij meaning long Cutting discoi'crcd Hadoop and naincd it aftcr his son's yclloii -colored toy clcphnnt

In recent years, a great proliferation has been witnessed in the amount of data generated. This rote is growing os datn is continuously growing amassing such huge omount of dota is a tough job end thus it needs some out of the box thinking os it connot be tackled with traditional tools and tcchniques . This inadequacy led to the birth of the term Big Datn and olong with it the challenges such as storage. processing. visualization and privacy. Big Data is not just about being big in size.

The definition is broadened using five characteristics or

—V'sl, These ore:

- Volume : This characteristic signifies huge i’oluminous datn: it is in orders of terabytes end even petabytes.
- Velocity . This characteristic signifies the high i'elocity with w’hich the date is generated.
- Variety : This chnrocteristic refers to the huge variety in the big datn.
- Volue: This chorocteristic refers to the intrinsic value conlnined in big data.
- Veracity This characteristic refers to uncerlninties in big date such os missing, duplicate and incomplete entries,

Hadoop is an excellent and robust analytics platform for Big Data which can process huge data sets at o really quick speed by providing scalability. It can manoge oll aspects of Big Dota such os volume. velocity end i’oriety by storing rind processing the dnta over a cluster of nodcs . Hadoop has two major components in its architecture that is MapReduce and the Hadoop DisWibuted File System (HDFS). With the introduction of YARN (Yet Another Resource Negotiator) in later releases, Hadoop was integrated with a number of wonderful components which can be used for storing, processing and analyzing data a lot more efficiently, thus oiding in exploration of dota for undiscovered facts ct a smooth pace. Some of these components that work on top of Hadoop are Hit e.Flume, Sqoop, Hbase and Oozie.

### 1.3 Motix”ation

In todoys world. ei'ery establishment is facing ever gron'ing challenges which need to be coped up quickly and efficiently. With continually increasing Mot ie industrr and entertainment ei eryoiie ii ants to enjoj all the perks of facility. So if they tin est on idiot res orc that thci can hai’e rci’ieivs with rating and all the information relatcd to that inoi’ie.

The best place to look up to find room for improvement is the voluminous raw’ data that is generated on a regular basis from various sources by applying Big Data Analytics (BDA)|2 |. BDA refers to the tools and practices that can be used for transforming the raw' data into meaningful and crucial information The main goal with ei’ey’ releasing moi‘ies i hether it is

hit or flop. bj generating reviews to the audience iie are here to transform the i oluminous ran data into sortie meaningful information bj applying Big Data Analytics

# CHAPTER 2

## **Literature** Review

Vidyasagar S.D[2017] did a survey on Big Data and Hadoop system and found that organizations need to process and handle petabjtes of Data its in efficient and inexpensii'e manner. According to him if there is any node failure then w'e can lose some information. Hadoop is an Efficient, reliable. Open Source Apache License. Hadoop is used to deal with large data sets. Author explained its need. uses and application. Now' days, Hadoop is playing an important role in Big Data. Vidyasagar S.D concluded that

Hadoop is designed to run on cheap commodity hardsare, it automatically handles data replication and

node failure, it does the hard work — you can focus on processing data, Cost Saving and efficient and reliable data processing.

**Sujatha** .Va , **Prasanna** Devi Sb , **Vinu Kiran Sb ,ManivannanS|2016]** had analyzcd a large scale Diabetic data sets for set eral patients to find the length of time taken for treatment for each class of Diabetes and the risk of re-admission of diabetic patients performing Bigdata analj tics, the type of diabetes and its outcome is hich acted as a high risk sample of patient data sets. They have collected and integrated different sources of diabetic information for sei'eral patients, from primal' rind secondan treatment infonnation to administrative infonnation, to analj ze not el i'ie» of patient care processes such as tj pe of trcatments and ci'cq' patient behai'iors on ii'hich results multifaceted nnture of chronic care fliat they takc into their account to prcdict the sun'ii'al factors and length of staj Noii'adays by using electronic medical equipmcnts ii'ith high quality and high degrce calibrations. thcy are able to gather large amounts of rcaltimc diabetic data sets That requires the usage of distributed platforms for making Big Data analysis that results on making decisions bascd on ai'ailable data and its trcnds This type of Bigdata anah sis ;illo»'s gcographical and en ironmcntal infonnation of patients' enables the capability of

interpreting the ethnicity of data gathered and extract new analysis to identify survii al options and treatment timelines froin thein.

**Sergey V. KovalchuLl** , **Artem** V. **ltharchuLl** , **Jiaqi Liao** , Sergey V. Ivanov , Alexander V. **Boukhanovsky{2016]** presents o technology for dynamic knoo'ledge-bosed building of Domain —Specific Longunges (DSL) to describe dota-intensive scientific discoi'eq' tosks using BigDaln technology. Their proposed technology' suppons high lei'el obstruct definition of analytic end simulation parts of the task os n'ell os integration into the composite scientific solutions. Automatic translation of the obstruct task definition enables seamless integration of v'orious dota sources within single solution.

Vennila S and Priyadarshini **I[2015]** Cloud computing prox'ides flexible infrastructure and high storage capacity for BigData applications. The MapReduce frame»'ork is inost preferable for processing huge volume of unstructured data set in BigData. Increase in data volume leads to flexible and scalable privacy preservation of such dataset oi'er the MapReduce framework is BigData applications. A suiney have been taken for the MapReduce framework based big data privacy preset ation in Cloud eni'iroiunent.

Jeff Sedayaoet. AI|2014J suggested to use Hadoop to aruilyze the data and obtain useful results for the Human Factors analysts. At the saine time, the requirements of analysing were learned and anonj inized data sets need to be carefully analyzed to detennine » hether they are i'ulnerable to attack.

**Meiko Jensen et. Al[2014]** explained that the field of pri› acy in big data contexts contains a bunch of key challenges that must be addressed by research. Many of these challenges do not stem from technical issues, but merely are based on legislation and organizational matters. Ne› ertheless, it can be anticipated that it ii as feasible to ineet each of the challenges discussed here by means of appropriate technical measures.

B. Saralailevia , N. **Pazh rajaa** , **P. Victer Paula** , M.S. Saleem **Bashab** , P. Dhavachelvanc **[2013]** Their paper sho» s the big data information and characteristics used in world u ide. The issues are also mentioned to give idea about the big data issues in real time

The security issue is pointed more in order to increase the securi in big data. We can improve security in big data by using any one of the approach or by combining these three approaches in Hadoop DisWibuted File System n'hich is the bese lojer in Hodoop, where it contains large number of blocks

Mohd Rehun Ghazia and **Durgaprasad** Gangodkara[2llfl8] discuss Hadoop and its components in detail u hich comprise of MapReduce and Hadoop Distributed File System (HDFS). MapReduce engine uses JobTracker and TaskTracker that handle monitoring and execution of job. HDFS a distributed file- system u hich comprise of NameNode. DataNode and Seconds, NameNode for efficient handling of distributed storage purpose. The details proi'ided can be used for dei'eloping large scale distributed applications that can exploit computational pou er of multiple nodes for data and compute intensive applications.

# CHAPTER 3

## Research Gap

20

### 3.1 Map Reduce

MapReduce »'orks by breaking the processing into two phnses: the map phase and the reduce phase. Each phase has key-value pairs as input and **output.** the types of which may be chosen by the programmer. The progmmmer also specifies m'o functions: the map function and the reduce function.

Hadoop MapReduce is a software frameii ork for easily writing applications ii hich process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity harder are in a reliable, fault-tolerantmanner.

A MapReducejob usually splits the input data-set into independent chunks ii hich are processed by the map tasks in a completely parallel manner The frames'ork sorts the outputs of the maps, ii hich are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

Ty'pically the compute nodes and the storage nodes are the same, that is. the MapReduce framework and the Hadoop Distributed File System are running on the saine set of nodes. This configuration allow's the framework to effectii'ely schedule tasks on the nodes o'here data is already present, resulting in very high aggregate bandwidth across the cluster.

The MapReduce framework consists of a single master JobTracker and one slat e TaskTracker per cluster- node. The master is responsible for scheduling the .iobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master.

## 3.2 Hive

One criticism of MapReduce is that the dex'elopment cycle is very long MapReduce allow's you. as the programmer, to specify a map function follou'ed by a reduce function, but w'orking out how to fit your data processing into th pattern, which ofien requires multiple MapReduce stages. can be a challenge. Writing the mappers and reducers. compiling and packaging the code, submitting the job(s), and retrieving the results is a time consuming business, and even u ith Streaming, which remox es the compile and package step, the experience is still inx'olved.
fi&

Hii'c raises the lei'el of abstraction for processing large datasets. With Pig. the data structures are inuch richer, typically being multivalued and nested, and the set of transfonnations you can apply to the data are much more powerful. They include joins, for example, › hich are not for the faint of heart in MnpReduce. Hii'c is made up of tn'o pieces:

• The language used to express data flo› s, called Hive Latin.

• The execution em ironinent to run hit e Latin programs. There are currently t» o eni'ironments: local execution in a single JVM and distributed execution on n Hadoop cluster

# CHAPTER 4

## Design and Implementation

### 4.1 His'e

Apache Hit c is a data u arehouse softu arc project built on top of Apache Hadoop for proc iding data siimmarization. quey' and anale sis. Hii'e gis es an SQL-liLe interface to query data stored in i arious databases and file systems that inte grate » ith Hadoop. Traditional SQL queries must be implemented in the Map reduce la a API to execute SQL applications and queries o› cr d istributed data. Hive proc ides the necessay' SQL abstraction to integrate SQL-liLe queries (His eQL) into the underlying la a u ithout the need to implement queries in the lou -let cl Jan a API. Since most data » archousing applications u ork u ith SQL-based query ing languages, Hit c aids portability of SQL-based applications to Hadoop

### 4.2 Major components of hive

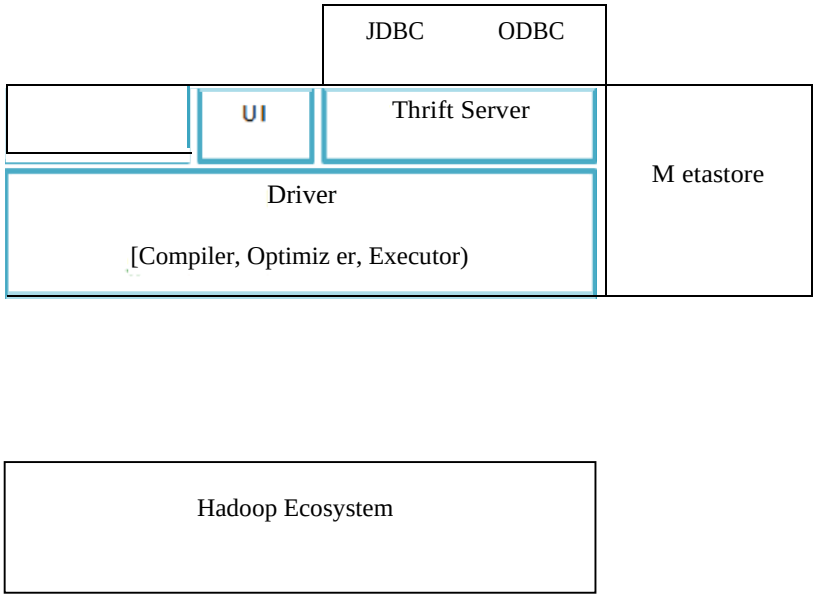


Fig.4.1 Coiiiponents of Hi e

- Metastore: Stores metadata for each of the tables such as their schema and location. It also includes the partition metadata which helps the driver to track the progress of various data sets distributed over the cluster. The data is stored in a traditional RDBMS format. The metadata helps the driver to keep a track of the data and it is highly crucial. Hence, a backup server regularly replicates the data which can be retrieved in case of data loss.
- Driver: Acts like a controller which receives the HiveQL statements. It starts the execution of statement by creating sessions and monitors the life cycle and progress of the execution. It stores the necessary metadata generated during the execution of an HiveQL statement. The driver also acts as a collection point of data or query result obtained after the Reduce operation.
- Compiler: Perfomis compilation of the HiveQL query, which converts the query to an execution plan, This plan contains the tasks and steps needed to be perfomied by the Hadoop MapReduce to get the output as translated by the query. The compiler convens the query to an abstract syntax tree (AST). After checking for compatibility and compile time errors, it converts the AST to a directed acyclic graph (DAG). The DAG divides operators to MapReduce stages and tasks based on the input query and data.
- Optimizer: Performs various transfonnations on the execution plan to get an optimized DAG. Transformations can be aggregated together, such as converting a pipeline ofjoins to a single join, for better performance. It can also split the tasks, such as applying a transfomiation on data before a reduce operation, to provide better performance and scalability. However, the loglc of transformation used for optimizanon used can be modified or pipelined using another optimizer.
- Executor: After compilation and optimization, the executor executes the tasks. It interacts with the job tracker of Hadoop to schedule tasks to be run. It takes care of pipelining the tasks by making sure that a task with dependency gets executed only if all other prerequisites are

- CLI, UI, and Thrift Server: A command-line interface (CLI) ]2IOVldes a user interface for an external user to interact with Hive by submitting queries, instructions and monitoring the

process status. Thrift scre er allow s extcnial clients to interact ii ith Hit e or er a metre ork. similar to the JDBC or ODBC protocols.

## 4.3 Operations

**Table 4.1** Operations m Hive

| **Step** Nt›. | **Operation** |
|---|---|
| | Execute Quci y<br>The Hive interface such as Command Line or Web UI sends quey to Driver (anj database drii'er such as JDBC, ODBC, etc.) to execute |
| 2 | Get Plan<br>The drip er takes flie help of quey compiler that parses the query to check the syntax and query plan or the requirement of queq'. |
| 3 | Get Metadata<br>The compiler sends metadata requcst to Mctastorc (anj database). |
| 4 | Scntl Metadata<br>Metastorc sends inctadata as a rcsponse to thc coinpilcr. |
| | Scntl Plan<br>The compiler chccks thc requirement and resends the plan to the drii'cr. Up to |

here, the parsing and compiling of a quey' is complete

6 Execute Plan

The drn. er sends the execute plan to the execution engine.

7 **Execute Job**

Internally. the process of execution .Job is a MapReduce job. The execution engine sends the job to JobTracker, uliich is in Name node and it assigns this job to TaskTracker, u hich is in Data node. Here, the quey executes MapRcducc job,

7I Metadata Opx

Mearru hile in execution. the execution engine can execute metadata operations u ith Metnstore.

8 PctEh Result

The execution engine receii'es the results from Data nodes.

9 Scud Results

The execution engine sends those resultant x alues to the drii'er.

10 Scud Results

The dri er sends the results to Hi e Interraces

## 4.4 Security

Hii'c i'G 7.0 addcd integration ii ith Hadoop security . Hadoop be gan rising Kerberos authorization support to proi'idc scciirity Kerberos allow's for mutual autlicnticatioii bct» ccn clicnt and sco cr In this H stcin, tlic clicnt's rcqucst for a tickct is passed along iv ith thc rcqiicst The prei'ioiis i'ersions of Hadoop had sei'eral issues such as users being able to spoof their uscniamc by scttin g the hadoop.job.ugi property and also MapReduce operations being run under the same user: hadoop or inapred With Hii'e i'G.7,0's integration with Hadoop security, these issues hat c largely been fixed. TasLTrackcr jobs arc rim by the user u ho launched it and the username can no longer be spoofed by setting thc hadoop.job.ugi propcrtj Permissions for nc» lj created filcs in Hi› c arc dictated by thc HDFS Thc Hadoop distributed file s› stem aiithoriz.ation model uses ihrce entities. riser. group and others ii ith three permissions: read, » rite and execute The default permissions for nc» Ij created files can be set by changing the umask › aluc for the Hit c configuration i'ariable hive.files.umask,value

## 4.S **HIVEQL**

While based on SQL, Hix'eQL does not strictly follow' the full SQL-92 slnndard. Hix'eQL offers extensions not in SQL, including multitable inserts and create table as select, but only offers basic support for indexes. Hii'eQL lacked support for Uansactions end materialized v ievvs. and only limited subquery support. Support for insert. update. and delete with full AfiID functionali mas made available u ith release 0.14

Internally. a compiler translates HiveQL statements into a directed acjclic groph of MopReduce. Tez. or Spark jobs. which ore submitted to Hadoop for execution

4.6 Research Questions

1. Total Number of movies in 2017

It ii'ill gii'e the total number of count of inoi'ics released in the › our 2017

2. Finding maximum rating of thc mo›'ic

It ii ill short the moi'ie h;n'ing maximum r;iting.

3. **Count the numher of mo› ies hax'ing the maximum rating.**

It ii ill display the count .

4. Count number of moi'ies bctu'cen rating 1 and 2

It ii ill display the list of inoi'ies hai'ing lcast ratings

fi.Fin‹l the list of s'cars an‹l numher of mo› ics released each year.

It ii ill short thc dctails of inoi'ics.

6 Find number of movies with duration of 2 hours.

It ii ill show the duration of all the mo ies.

# Chapter S
# Technology Used

## 5.1 System Requirements: Software used: HadooJi, Hive

### 5.1.1 Hailoop:

It is a jas'a based programming frameu'ork u hich stores data and process on it and distribute it in columnar form and runs on the basis of clusters u ith the commodity liardu are.

The first component to proi'ide online access ii as H Base. a key -x alue store that rises HDFS for its underl› ing storage HBase proc ides both online read/u rite access of indii'idiial row's and batch operations for reading and u'riting data in bulk. maLing it a good solution for building applications on. The real enabler for neu processing models in Hadoop u'as the introduction of YARN (u hich stands for Yet Another Resource Negotiator) in Hadoop 2.

YARN is a cluster resource management sj stem. u hich allou s any distributed program (not just MapReduce) to run on data in a Hadoop cluster In the last fen years. there has been a fIou ering of different processing patterns that u ork ii ith Hadoop. Here is a sample: Interactive SQL By dispensing u itli MapReduce and u•'• s • distributed queo engine that uses dedicated —alu ays onl daemons (like Impala) or container reuse (like His e on Tez). it's possible to achiei'e lou -latency responses for SQL queries on Hadoop while still scaling up to lars• dataset sizes. Iterative processi• s Many • sorithins—such as those in machine leamin g—are iteratii e in nature, so it's much more efficient to hold each intermediate u'orking set in me more . compared to loading from disk on each iteration. The architecture of MapReduce does not allou this. but it's straightforu ard u ith Spark, for example. and it cnablcs a highly exploratory style of u orking ii ith datasets. Stream processing Streaming systems like Storm. Spark Strea •S OF Sainza make it possible to run real-time, distributed computations on unbounded streams of data and ‹unit results to Hadoop storage o.r

external systems. Search The Solr search platform can run on a Hadoop cluster, indexing documents as they are added to HDFS, and serving search queries from indexes stored in HDFS

### 5.1.2 Hive

Hi c is a data u arcliousc iiifrastructurc tool to proccss structured data in Hadoop. If rcsidcs on top of Hadoop to summarize Big Data. and inaLcs qucy iiig and analj ring easy.

Initially Hi› c i as dc clopcd bi Facebook. later the Apachc Sofiwarc Foundation took it up and dci eloped it further as an open source under the name Apache Hit e. It is used by different companies. For example, **Amazon uses it in** Amazon **Elastic** MapReducc.

Hive is not

- A rclational databasc
- A design for OnLinc Transaction Proccssing (OLTP)
- A languagc for rcal-timc qucrics and row-lc› cl updatcs

**Features of Http**

- It storcs schcma in a databasc and proccsscd data into HDFS.
- It is dcsigncd for OLAP.
- It proc idcs SQL tjpc languagc for qucw ing callcd Hit cQL or I-IQL.
- It is familiar. fast. scalable. and extensible.

5.2 Hive Job Execution Flow

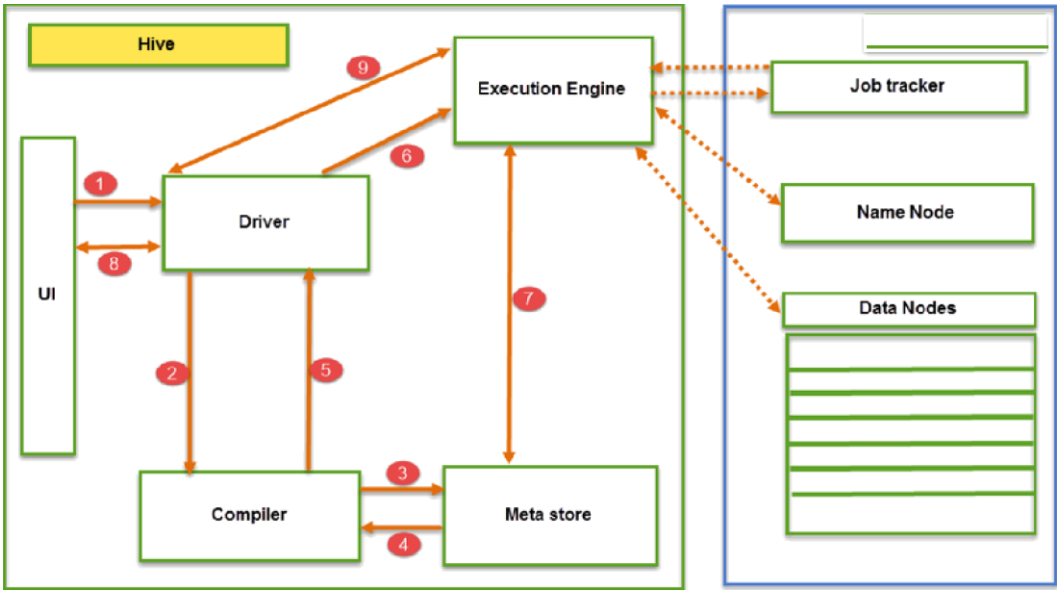


Fig 5.1 Hit e Job Execution Flow

From the above scrccnshot › c can understand the Job execution flo› in Hive with Hadoop

The data flow in Hive behaves in the follow ing pattern:

l. Executing Query from the UI( User Interface)

2. The driver is interacting with Compiler for getting the plan. (Here plan refers to query execution) process rind its related metadata information gathering
3. The compiler creates the plnn for a job to be executed. Compiler communicating with Meta store for getting metadata request
4. Meta store sends metadata information back to compiler
5 Compiler communicating with Dri› er with the proposed plan to execute the queq

fi. Driver Sending execution plans to Execution engine

7. Execution Engine (EE) acts as a bridge betu'een Hive and Hadoop to process the query. For DFS operations.

- EE should first contacts Name Node .and then to Data nodes to ,get the values stored in tables.
- EE is going to fetch desired records from Data Nodes. The.actual.data of tables resides iii.data node only. While from Name Node it only fetches the metadata information for the query.
- It collects aotual .data from data nodes related to mentioned .query
- Execution Engine (EE) communicates bi-directioiudly with Meta stnre present in Hive tn perform DDL (Data Definition Language) operafions. Here DDL operafions like KREATE, DROP atid ALTERING tables and databases are!dons. Meta store will store inforñiation about.database name, table. names and column names only. It will fetch data related to query mendoned.
- Execution Engine (EE) in turn .communicates with Haitoop ‹daemons such as Name node,, Data nodes, and job oacker tn execute the query on top of Hadoop file sysUni

8. Fetching suits from driver
9. sending re.suits to Executi.on engin.e., Once the results fetched from .data n.odes to the EE, it will seiid results.back to dfiver arid to UI ( froiit eiid)

ŀ i@ Contioufiusly in c.oust with itado.op ifile sygtem and its daemons v|a Execution engine. The .dotte‹i artow in tire Job flow ‹diagram sliows. the Execution eiigine coinmuiiicatioii with Hadoop.daemons.

- Different modes tif Hivñ

Hive can .opeiate iii two modes. ‹iependiiig oii tire size of‹iata iiodes iii Hadoop.

These inodes. are,

- Local mode
- Map. reñuce mode

When to use! Local mode:

- If the Hadoop installed undet pseudo mode with having one data node we use Hive in tliis mode
- If the data size is smaller in term of limited to single local machine, we .can use this inode
- Proceesiiig will be: vefy lfast oii siaaller data eete present iii the: local inachihe

When to use Map reduce mode:

- If Had%p is having iiiultiple data fiodes rind data is.distributed moss different node we use Hive iñ this inode
- It will Jiel.oftii on latge aiiiouat.of data sets and query .going tn. ex.ecus. ia parallel way
- Processing of large!diita sets with better performance! can be!achieved through this

In Hive, we .cart set this property to mention which mode Hive .cati work? By default,,it works on Map U duce m.ode an.d for!oca1..p:.odel:you cân. liave: U.:following .setting..

Hive to wotk iii locd iño‹ic set

**SET mapred.job.tracker=local;**

Froin the Hive *ersion 0..7 it suirports a inotte to run inap reduce jobs. in local inode automatically

# Chapter 6
# Experimental Results

## 6.1 Movie Data Set

```
1,The Nightmare Before Christmas,1993,3.9,4568
2,The Mummy,1932,3.5,4388
3,Orphans of the Storm,1921,3.2,9062
4,The Object of Beauty,1991,2.8,6150
5,Night Tide,1963,2.8,5126
6,One Magic Christmas,1985,3.8,5333
7,Muriel's Wedding,1994,3.5,6323
8,Mother's Boys,1994,3.4,5733
9,Nosferatu: Original Version,1929,3.5,5651
10,Nick of Time,1995,3.4,5333
11,Broken Blossoms,1919,3.3,5367
12,Big Night,1996,3.6,6561
13,The Birth of a Nation,1915,2.9,12118
14,The Boys from Brazil,1978,3.6,7417
15,Blj DO ! NOtSP,19T1,}.9,fAO5
16,The Breakfast Club,1985,4.0,5823
17,The Bride of Frankenstein,1935,3.7,4485
18,Beautiful Girls,1996,3.5,6755
19,Bustin' Loose,1981,3.7,5598
20,The Beguiled,1971,3.4,6307
21,Born on the Fourth of July,1989,3.4,8646
22,Broadcast News,1987,3.4,7940
23,Swimming with Sharks,1994,3.3,5586
24,Beavis and Butt-head Do America,1996,3.4,4852
25,Brighton Beach Memoirs,1986,3.4,6564
26,The Best of Times,1986,3.4,6247
27,Brassed Off,1996,3.5,6840
28,Last Tango in Paris,1972,3.1,7732
29,Leprechaun 2,1994,3.2,5125
30,Incident at Oglala: The Leonard Peltier Story,1992,3.7,5487
31,Kalifornia,1993,3.4,7095

33,Jingle All the Way,1996,3.6,5371
34,Killing Zoe,1993,3.4,5773
35,King of Beggars,1992,3.6,6025
36,Into the Woods,1990,4.0,9077

38jO TOO D(ep,t9%,).9j8i)
39,CtA)XOO,tO7O,3,2,BO44
40 Internal Affairs 1990 3.5 6885
```

## 6.2 Research Questions

1. Total Number of movies in 2017

It will give the total number of count of movies released in the year 2017.


Fig 6.ltotal number of movies

2. Finding maximum rating of the movie

It will show the movie having maximum rating.

```
hive>  select max(rating) from movies;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be availa
spark, tez) or using Hive 1.X releases.
Query ID = shruti_20180428134423_d92d8f7a-4e69-409d-9ab1-d54f05ce
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
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_1524597057802_0006, Tracking URL = http://hadc
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_152459
Hadoop job information for Stage-1: number of mappers: 1; number
2018-04-28 13:55:03,305 Stage-1 map = 0%,  reduce = 0%
2018-04-28 13:56:04,061 Stage-1 map = 0%,  reduce = 0%
2018-04-28 13:57:04,686 Stage-1 map = 0%,  reduce = 0%
2018-04-28 13:57:32,422 Stage-1 map = 100%,  reduce = 0%, Cumulat
2018-04-28 13:58:32,520 Stage-1 map = 100%,  reduce = 0%, Cumulat
2018-04-28 13:59:19,732 Stage-1 map = 100%,  reduce = 67%, Cumula
2018-04-28 13:59:57,964 Stage-1 map = 100%,  reduce = 100%, Cumul
MapReduce Total cumulative CPU time: 5 seconds 840 msec
Ended Job = job_1524597057802_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.01 sec   HDF
Total MapReduce CPU Time Spent: 6 seconds 10 msec
OK
4
Time taken: 972.913 seconds, Fetched: 1 row(s)
```

Fig 6.2 Maximum rating of mo› ie

3. Count the number of movies hai'ing the maximum raiing.

It ii ill display the count .


Fig 6.3 Count number of inaximuin rating

4. Count number of movies between rating I and 2

It will display the list of movies having least ratings

```
hive> select distinct movie_name from movies where rating between 1 and 2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future vers
spark, tez) or using Hive 1.X releases.
Query ID = shrutl_20180505120705_8bb2b2fb-6753-4110-acfe-c6706ec0db2f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
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_1525499465151_0001, Tracking URL = http://hadoopmaster:8088/proxy/ap
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1525499465151_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-05 12:07:59,049 Stage-1 map = 0%,  reduce = 0%
2018-05-05 12:08:59,120 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 2.91 sec
2018-05-05 12:09:03,854 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.43 sec
2018-05-05 12:09:26,030 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.96 sec
MapReduce Total cumulative CPU time: 8 seconds 960 msec
Ended Job = job_1525499465151_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.96 sec   HDFS Read: 2902193 HDFS W
Total MapReduce CPU Time Spent: 8 seconds 960 msec
OK
...At First Sight
```

Fig 6.4.1 Movies having rating 1 and 2

```
.Com for Murder
10 Items or Less
10 Items or Less: Season 1
10 Items or Less: Season 2
10 Items or Less: Season 3
100 Below Zero
100 Years Of Evil
11 11 11
12 12 12
13 13 13
1313: Actor Slash Model
1313: Bermuda Triangle
1313: Bigfoot Island
1313: Billy the Kid
1313: Cougar Cult
1313: Frankenqueen
```

Fig 6A.2 Movies having rating l and 2

1.Find the list of years and number of movies released each year.

It will show the details of movies,

```
hive> select release_year, count(distinct movie_name) from movies group by re
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the f
spark, tez) or using Hive 1.X releases.
Query ID = shruti_20180505121555_48bf360a-0f4b-4957-9d4e-c465376fb454
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
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_1525499465151_0002, Tracking URL = http://hadoopmaster:808
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1525499465151_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers:
2018-05-05 12:16:12,857 Stage-1 map = 0%,  reduce = 0%
2018-05-05 12:16:50,548 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.8
2018-05-05 12:17:28,422 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12
MapReduce Total cumulative CPU time: 12 seconds 500 msec
Ended Job = job_1525499465151_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.5 sec   HDFS Read: 2902
Total MapReduce CPU Time Spent: 12 seconds 500 msec
OK
NULL    0
1913    3
1914    20
1915    1
1916    1
1918    1
1919    3
1920    6
1921    2
1922    2
1923    4
1924    5
1925    5
1926    2
1927    4
1928    2
```

Fig 6.5.1 List of movies

```
1974    178
1975    134
1976    118
1977    136
1978    230
1979    140
1980    106
1981    112
1982    153
1983    267
1984    302
1985    333
1986    287
1987    280
1988    333
1989    419
1990    467
1991    364
1992    479
1993    561
1994    515
1995    590
1996    685
1997    783
1998    838
1999    1173
2000    900
2001    1167
2002    1109
2003    1389
```

Fig 6.5.2 List of movies

6.Find number of movies with duration of 2 hours.

It will show the duration of all the movies.

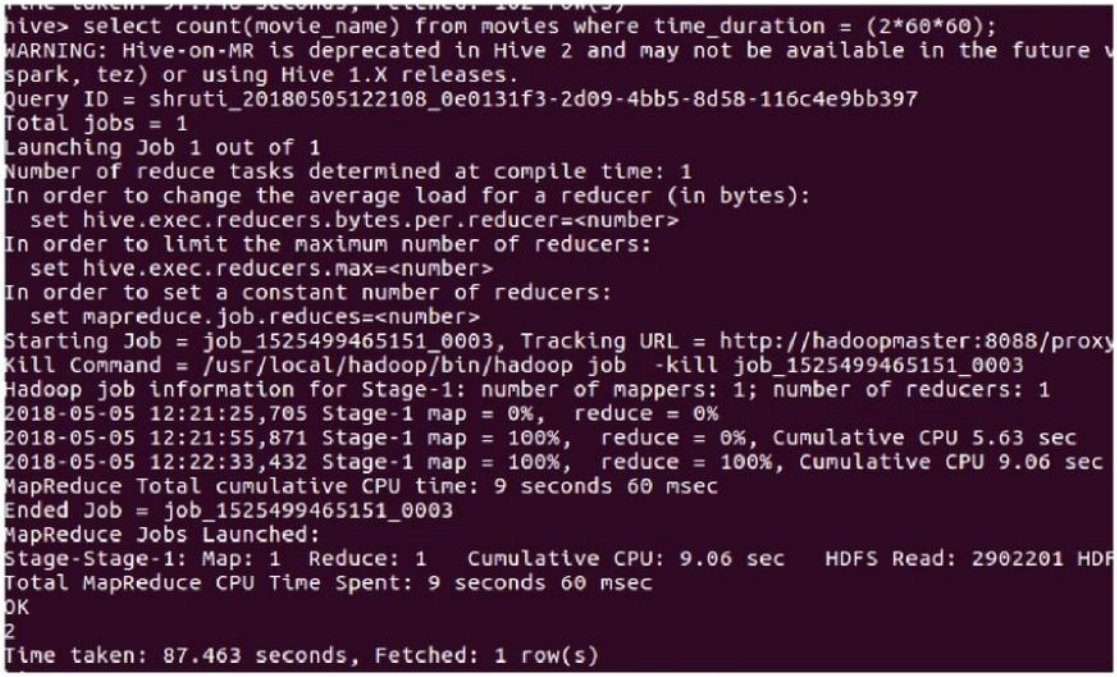


Fig 6.6 Movies with Duration of 2 hour

# Chapter 7
# Conclusion and Future Work

## 7.1 Conclusion

This is a novel approach where the user rnting decisions has been taken to purview along with inherent moi'ie attributes to model the classification approach. An experimental insight has also been proc ided for the post release aspect of the mod ie that relates initial budget with each of the ancial returns.

Big Data Analytics refers to the tools and practices that can be used for transforming this ran data into meaningful and crucial information u hich helps in forming a decision support system to make it easier for people to decide for a inoi'ie One should know' u hich moi'ie to u'atch Also. u ith the revie» it proc ides us » ith the duration of the movie and also the release date of the inoi'ie

## 7.2 **Future** Worlt

- The abundance of mod ie data in terms of rex iew, rating or ex en detail information (for example the information maintained by (IMDB) in the internet has encouraged many researches to formulate techniques to analyze the pattern in mox'ie data. Most of the researches are devoted to develop recommendation systems of movie according to user review's.
- his analysis con be further carried out on Fully Distributed cluster mode that is hadoop daemons run on a cluster of machine.
- Similar analysis can be further carried out in different sector

# REFERENCES


[1] Apache Hadoop. https://hadoop.apachc.org/

[2] Riyaz P.A., SureLiia Mariain Varghese, Lei'eraging MapReduce ii itli Hadoop for Weather Data Analytics . IOSR - Journal for Computer Science, 2015. i olume 17, issue 3. pp 6-12. [3| http.//spideropsiiet.coin/site1/blog/2014/09/18/apache-pig-tool-big-data-analytics/
[4] Andrcii Pai'lo. —A Comparison of Approaches to Large-Scale Data Analvsisl. SIGMOD. 200'3.

[5] Dean. J, and Gheinau at. S..—MapRediice: a flexible data processing toolll. ACM 2010.

[6] Greenpliiin Analytics Workbench .i'isit us at u'u'u .greenpluin.coin

[7] HadoopTutorial: http:/ldcs eloper.yahoo.coin/liadoop/tutorial/module1.html

**[8]** Harslia» ardhan S Bliosale. Prof. De endra P Gadekar —A Revie» Paper on Big Data and Hadoopll HUB,i » » .ibinbigdatahub.com/infograpliic/four- **s-big-datn.**
[10] Sanjay Rathee. —Big Data and Hadoop ii ith components like Fliime. Pig. Hii'c and Jaqll,lnteniational Conference on Cloud, BigData and Tnist2013.Vol. 1.5,Noi ember 20

[1 l]Narkhede. S.. & Baraskar. T. —HMR log anal zer: anale ze Webapplication logs os'er Hadoop M;ipRcdiicc11. International Journal ofUbiComp. pp 41-51, 2013.

{l2]Hadoop — Adi'antages and Disadi'antages http://u u u j2eebrain com/,jaw a-J2ee-hadoop-advantages- and-disadi outages html Accessed on: 13-04-2017.

[ 13]Shilpa. Manjit Kaur. —BIG Data and Methodology-A rei iewl. International Journal of Adi'anccd Research inComputcr Science and Softii'arc Enginccring.Volume 3, Issue 1G. October 2013 [14] http.//ii'ii'ii iihitehoiise got'/blog/2012/03/29/big-data-big-deal/Machine learning.
[ 15] Ahmed Elda» y. Mohamed F. Mokbel — A Demonstration of
Spatial Hadoop An Efficient MapReduce Frames ork for Spatial DatnIProceedings of the VLDB Endou'ment. Vol 6. No. 12





[ l fi] MrigaiA Mridul . Akaslideep Khajuria. Sneliasish Dutta. Kumar N — Anal›'sis of Bidgata using Apache Hadoop and Map Reducel Volume 4. Issue 5. Maj 201411.27

{ 17] P D Londhe. S.S Kiimbhar. R.S Sul. and A J. Khadse. —Processing big data using liadoop fraitieu orkll. in Proceedings of 4tIi SARC-IRF International Conference. Neu Delhi. India. Apr. 27. 2014. pp 72-75

1 1 T. White. —MapReduce and the hadoop distributcd filc systcinll, in Hadoop: The dcfinitii'c guide. 1" edition. O'Rcilly Media. Inc . Yahoo prcss. 2012

1 1 D. Borthakui. "The hadoop distributed file sjstcin. architccture and dc sign," Hadoop Project Website [onlinеı A ailable: http://hadoop.apache org/ core/docs/current/hdfs design.pdf

z ›j J. L« «•d C. Djcr. "Data-intcnsii c tcxt proccssing s› ith inaprcducc". in Si nthcsis Lccturcs on Human Language Tcclinologics. ol. 3. no. 1. pp. **1-177.** 2tll().

# Report

ORIGINALITY REPORT

| 69% | 58% | 32% | 35% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | en.wikipedia.org<br>Int ernet Source | 14% |
| 2 | www.guru99.com<br>Int ernet Source | 10% |
| 3 | Jain, Arushi, and Vishal Bhatnagar. "Crime Data Analysis Using Pig with Hadoop", Procedia Computer Science, 2016.<br>Publication | 7% |
| 4 | www.tutorialspoint.com<br>Int ernet Source | 4% |
| 5 | hadoop.apache.org<br>Int ernet Source | 4% |
| 6 | V. Sujatha, S. Prasanna Devi, S. Vinu Kiran, S. Manivannan. "Bigdata Analytics on Diabetic Retinopathy Study (DRS) on Real-time Data Set Identifying Survival Time and Length of Stay", Procedia Computer Science, 2016<br>Publication | 4% |
| 7 | docs.com | |

| | | |
|---|---|---|
| | Int ernet Source | 2% |
| 8 | www.ijcsit-apm.com<br>Internet Source | 2% |
| 9 | www.uvpce.ac.in<br>Internet Source | 2% |
| 10 | S. Vennila, J. Priyadarshini. "Scalable Privacy Preservation in Big Data a Survey", Procedia Computer Science, 2015<br>Publication | 2% |
| 11 | arxiv.org<br>Internet Source | 2% |
| 12 | www.safaribooksonline.com<br>Internet Source | 2% |
| 13 | B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvan. "Big Data and Hadoop-a Study in Security Perspective", Procedia Computer Science, 2015<br>Publication | 2% |
| 14 | research2.fit.edu<br>Internet Source | 1% |
| 15 | www.oalib.com<br>Internet Source | 1% |
| | Submitted to University of Glamorgan | |

| | | |
|---|---|---|
| 16 | Student Paper | 1% |
| 17 | research.ijcaonline.org<br>Internet Source | 1% |
| 18 | Submitted to Higher Education Commission Pakistan<br>Student Paper | 1% |
| 19 | Arushi Jain, Vishal Bhatnagar. "Crime Data Analysis Using Pig with Hadoop", Procedia Computer Science, 2016<br>Publication | 1% |
| 20 | Submitted to Institute of Technology, Nirma University<br>Student Paper | 1% |
| 21 | rku.ac.in<br>Internet Source | 1% |
| 22 | Submitted to Charotar University of Science And Technology<br>Student Paper | 1% |
| 23 | Submitted to University of Strathclyde<br>Student Paper | 1% |
| 24 | www.ijcaonline.org<br>Internet Source | 1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | < 1% |
| Exclude bibliography | On | | |