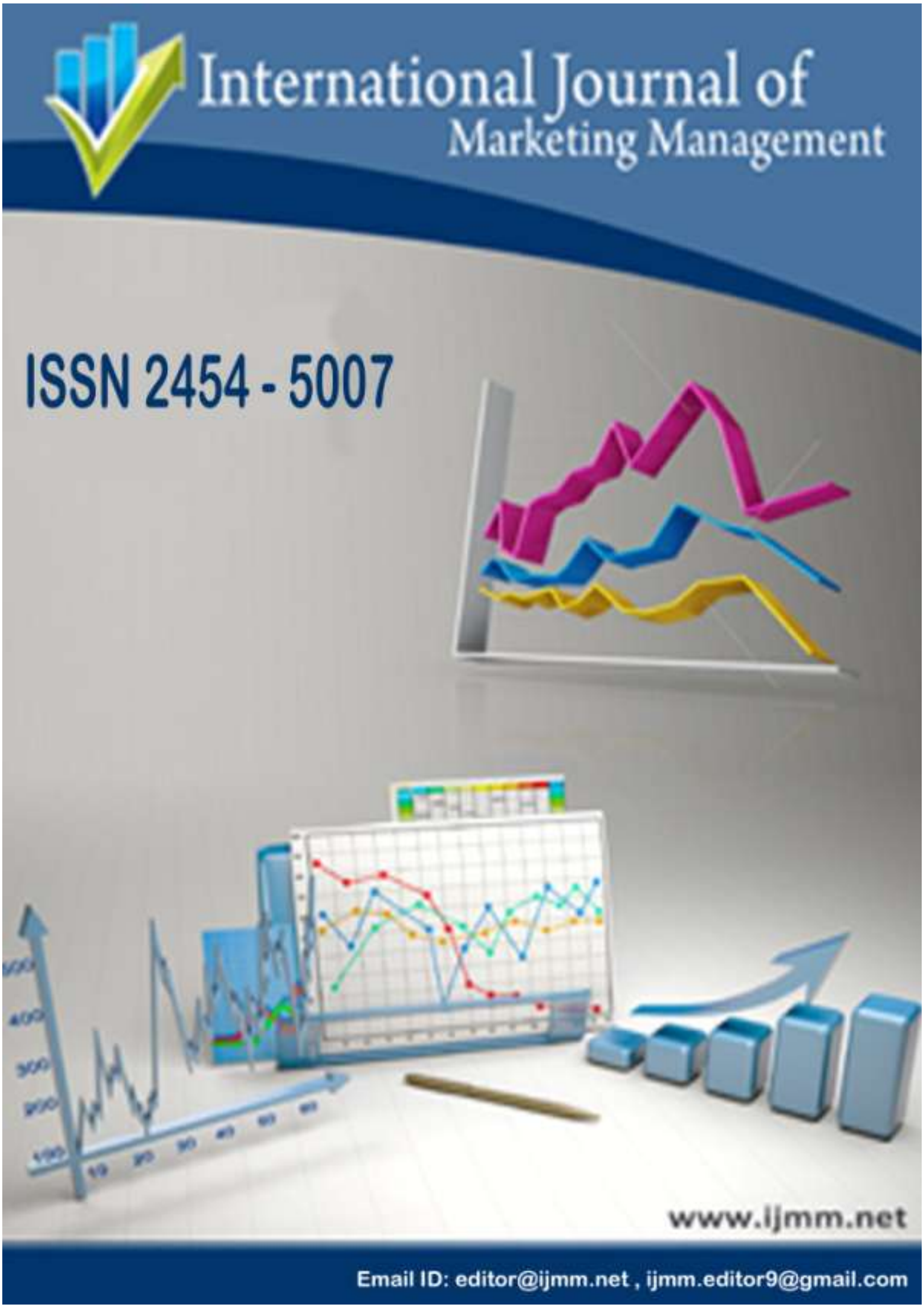

International Journal of
Marketing Management
ISSN 2454 - 5007
www.ijmm.net
Email ID: editor@ijmm.net , ijmm.editor9@gmail.com

ISSN 2454-5007, www.ijmm.net

Vol. 9 Issuse. 3,Aug 2021

Performance Evaluation and Optimization of Hybrid Indexes in Hive

1G.Sravanthi,2C.V.S.Satyamurty

Abstract:Big data technologies decrease corporate risks and expenses by delivering crucial and accurate analysis that leads to concrete decision-making and improved operational efficiencies. You will need a system to manage and handle massive amounts of categorized and uncategorized data in real time, as well as to safeguard sensitive information, in order for this data to be useful. Operational Big Data and Analytical Big Data are two broad categories of big data products offered by many vendors, such as Amazon, IBDM, and Microsoft. With Hive, an easy-to-use tool for query and analysis of data warehoused on Hadoop, we may recapitulate Big Data. Compression and Bitmap indexing are supported in order to increase query performance. Both indexes have been applied on the same partition in Hive and have seen significant speed improvements as a result. Analyzing Simultaneous Bitmap and Compact Indexing on a partition using text data has resulted in a significant speed improvement.

Keywords:Hive,Hadoop,Indexing,CompactIndexing,Hadoopinfrastructure,BigData,HiveQL.

1. INTRODUCTION

Data is being used by enterprise organizations to help them make better decisions and improve their performance. There has been a dramatic shift in the availability and efficacy of Organizational data over the past decade. As a result, data utilization has been transformed, and the notion of Big Data has emerged. [1]

BigData

Large amounts of data that are challenging to handle and extract from using typical storage methods [2]. For these reasons, people have welcomed online businesses (Google, eBay, Facebook, LinkedIn etc).

BigDataforSmallandBigCompanies

It's easy to see why Big Data was first adopted by online businesses. These Enterprises and startups focus on leveraging rapidly challenging data and combining unstructured data with established approaches [3]. The following are some of the concerns or challenges that face Big-Data..

1,2ComputerScienceofEngineering,InformationTechnology

1,2CVRCollegeofEngineering,Hyderabad,IndiaE-mail:sravanthi.guntakani15@gmail.com,E-mail:satyamurtycvs@yahoo.co.in

Volume:Traditional databases couldn't handle the amount of data that was available, therefore a new approach was needed.

Variety: In comparison to earlier text and table format, current available versions are in the form of pictures,videosandtweetsetc.

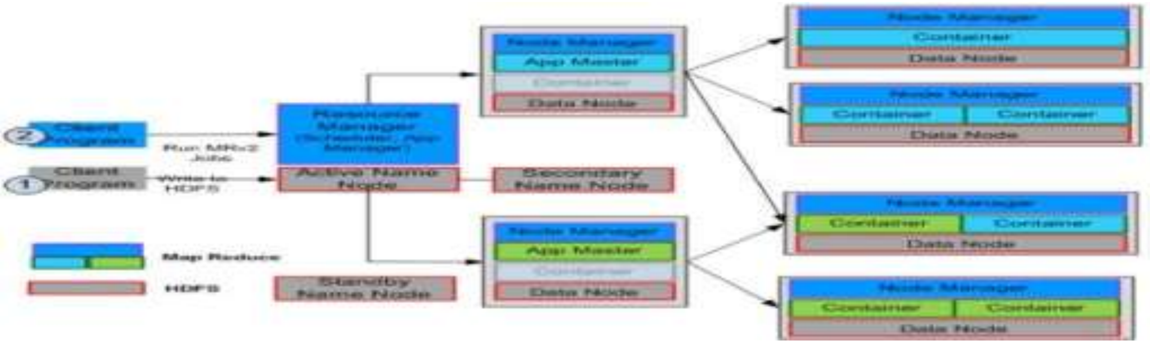
Velocity:IncreasedusageofOnlineSpaceandthatthedatawasavailablewasrapidlychangingandhavetobemadeavailableinstantlyandeffectively[4].

OpenSourceToolforBigDataAnalytics

AdistributedsoftwaresolutionHadoopisascalablefaulttolerancefordatastorage,process,andextractingarethemainuses.

i. HDFS(whichisstorage)
ii. MapReduce(Fig.1)

Fig1:HadoopHDFS,MapReduceLogicalView



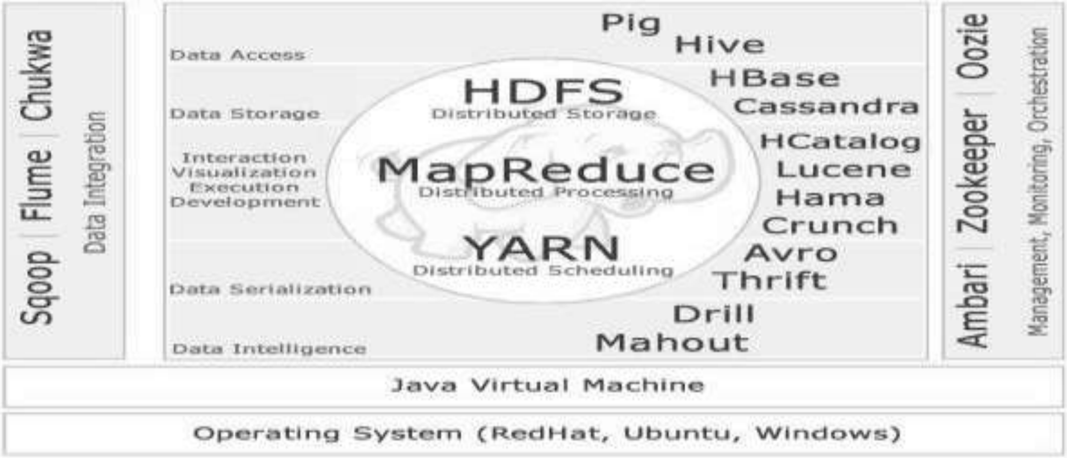


Fig2:TheHadooptechnologystack

2. RelatedWord

Hive

An OLAP data warehouse that can handle and query large amounts of data in distributed storage is Hive. Huge amounts of data are reliably stored in the Hadoop Distributed File System (HDFS)[5] ecosystem, which is impervious to hardware failures.

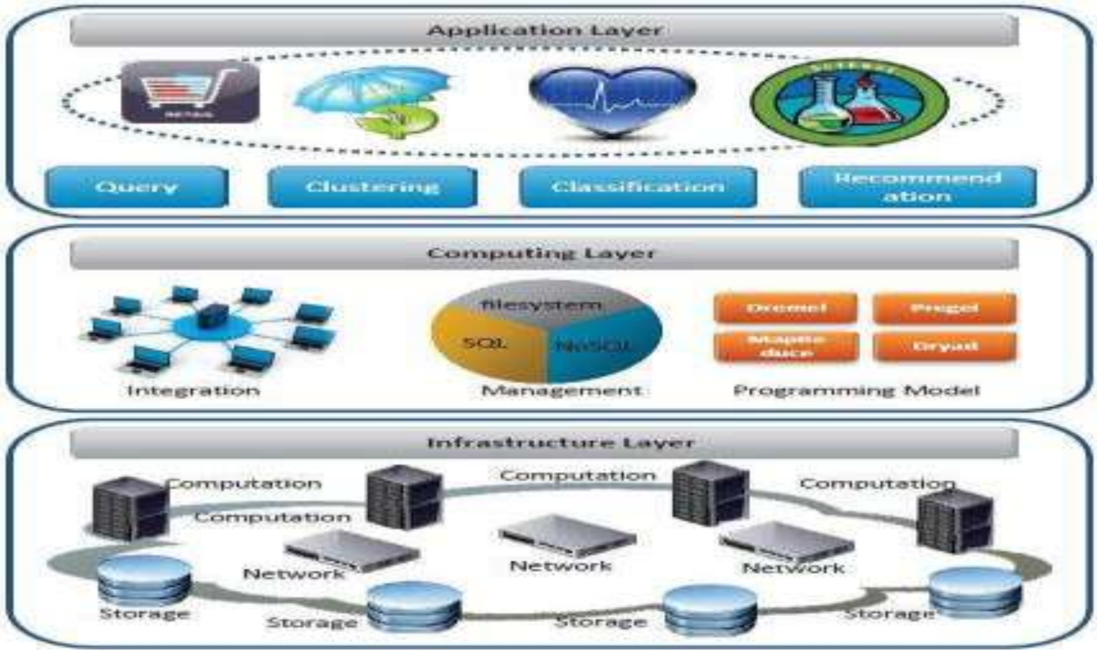


Fig3:LayeredArchitectureofBigDataSystem

As a high-level programming paradigm, Hive, a SQL-like relational data warehouse technique, is built on top of Hadoop and MapReduce, a large computation in Hadoop that enables data streaming.

HiveQL

It is a SQL-like query language, is used to query over Hive. Hive does not support SQL Commands(Update, Delete, and Insert at Row-Level) and transactions also. Hadoop Client Hive is to function with rareupdate and batch-mode data insertions are the characteristics of the Hive as a underlying system. Hive an OLAPdataForOnlineTransactionProcessing(OLTP)featureswithbigdata,oneshouldconsideraNoSQLdatabase[7].

Data in HiveHiveDataTypes

Hivecommonrelationaldatabasesdatatypesas wellasthethreecollectiontypes:STRUCT,MAP,andARRAY.

HiveFileFormats

Hive can process data coming from different data sources using Extract, Transform and Load (ETL)toolsforreadingdatafromdifferentfileformatandtheircustomization.

IndexConstructioninHive

```
CREATEINDEXindex_name
ONTABLEbase_table_name(col_name,...)AS'index.handler.class.name'

WITHDEFERREDREBUILD
[IDXPROPERTIES(property_name=property_value,...)][INTABLEindex_table_name]
[PARTITIONEDBY(col_name,...)]
[[ROWFORMAT...]STOREDAS...|STOREDBY...]
[LOCATION hdfs_path][TBLPROPERTIES(...)]
[COMMENT"indexcomment"]
```

Therearesomeparametersandkeywordsusedin thecommandlisting:

1.index_name:Theindexnamegivenbytheuserusedtoaccesstheindexbytheuseritself.Thisname isusedtoreferto theindexinALTERINDEX andDROPINDEX commands.

2.base_table_name (col_name,...): The table over which the index is to be created using the desired columnslisted.

3.'index.handler.class.name':thisspecifiesthetypeoftheindex,whichcouldbebitmaporcompact values.

4.index_table_name:TheindexnamegivenbyHive(thedefaultname)ortheuser,usedtoaccessth eindexasatablebytheuserorHive.Forexampletheusercanseethecontentofanindexusingthisparameter.Anyotherbehavioroftheindexasatablecanbeaddressedusingthisname.

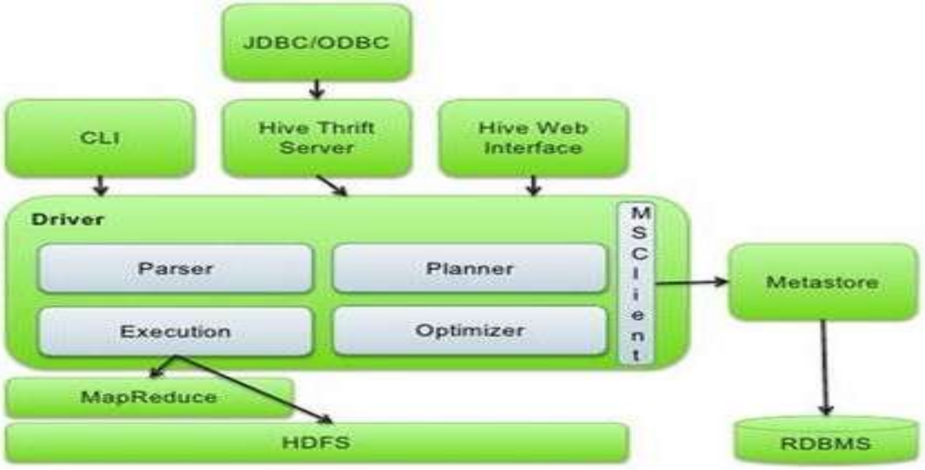
5.WITHDEFERREDREBUILD:Thismeanstheindexstartsempty.Indexwillbepopulatedusing: ALTERINDEXindex_nameONtable_name[PARTITION(...)]REBUILD

6.IDXPROPERTIES: This gives the properties of the index. For example: 'index_creator' = 'Mahsa' can beconsidereda propertyforanindex.

7.INTABLEindex_table_name:Thisisusedwhentheuserwishestobuildanindexinadifferenttablefromanalreadybuiltindex(ifany),tokeepthemseparate.

HiveArchitecture

Fig4:ApacheHiveArchitecture



MajorcomponentsoftheApacheHiveare:

1. Metastore
2. Driver
3. Compiler
4. Optimizer
5. ExecutorCLI,UI,andThriftServer

Compact Index, Aggregate Index, and Bitmap Index are all supported by Hive. In Hive, there are tables for each index dimension and the related index location arrays. Increase in the number of index dimensions results in an enormous table that takes up a lot of disk space.

The purpose of INDEX is to increase the speed at which data may be retrieved. There are two ways to accomplish the same thing: loading the complete table or partition to process records or loading only a portion of the file to process records with an index on column name.

Compact IndexonHiveTable

PairIndexedColumn’svalueanditsbolckidofIndexColumnarestoredinCompactindexing.

CREATECOMPACTINDEX:

```
CREATEINDEXindex_name
ONTABLEbase_table_name(col_name,...)
ASindex_type
[WITHDEFERREDREBUILD]
[IDXPROPERTIES(property_name=property_value,...)][INTABLEindex_table_name]
[[ROWFORMAT...]STOREDAS...
|STOREDBY...]
[LOCATION hdfs_path][TBLPROPERTIES(...)]
[COMMENT"indexcomment"];
```

Example

```
hive>CREATEINDEXindex_studentsONTABLEstudents(id)
```

```
>AS'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'
>WITHDEFERREDREBUILD;OK
Timetaken:0.493seconds
```

ALTERCOMPACTINDEX

```
ALTERINDEXindex_nameONtable_name[PARTITIONpartition_spec]REBUILD;
```

Example

```
hive>ALTERINDEXindex_studentsONstudentsREBUILD;
QueryID=cloudera_20171111093030_6a37b92b-bae8-4fd1-91bb-d13e9d411513Total jobs = 1
...
TotalMapReduceCPUTimeSpent:4seconds180msecOK
```

DROPCOMPACTINDEX

```
DROPINDEX[IFEXISTS]index_nameONtable_name;
```

Example

```
hive>DROPINDEXIFEXISTSindex_studentsONstudents;OK
Timetaken:0.27seconds
```

Bitmap IndexonHiveTable

CombiningindexedcolumnvalueandlistofrowsarestoredasbitmapinBitmapindexing.

CREATEBITMAP INDEX:

```
CREATEINDEXolympic_index_bitmapONTABLE olympic(age)
AS 'BITMAP'
WITHDEFERREDREBUILD;
ALTERINDEXolympic_index_bitmaponolympic REBUILD;
```

IndexingAdvantages

1. Indexesimprovethequeryperformance.
2. MultipleIndexescanbecreatedonthesametable.
3. Anytypeofindexingcanbecreatedonthedata.
4. DependingondataBitmapindexesarefasterthanCompactindexesandviceversa.

3. PreliminariesandDefinitions

HDFS: HadoopDistributed File SystemYARN:YetAnotherResourceNegotiatorDN: Data Node
NN:NameNode
SN:SecondaryNameNodeRM: Resource Manager NM:NodeManager

CLI:CommandLineInterfaceUI:UserInterface
OLTP:OnlineTransactionProcessingOLAP:OnlineAnalytical Processing

4. SystemModel

Ingeneraltheindexes(CompactIndexandBitmap Index)areappliedasaseparateentityforperformanceoptimization.Intheproposedmodelboththeindexesareappliedtogetherandtestedonnumerousdatavalueswhichresultedinmoreperformanceoptimization.

5. PerformanceAnalysis

Table1:HiveQuerystatisticswithoutindexing(alltheresultsareinseconds

Rows/Tables	Table1	Table2	Table3	Table4
50	0.628	0.1	0.085	0.167
100	0.14	0.17	0.087	0.17
150	0.137	0.07	0.0204	0.098
500	0.139	0.181	0.093	0.109
1000	0.251	0.074	0.159	0.108
5000	0.161	0.125	0.089	0.196

6000	0.143	0.225	0.089	0.117
10000	0.162	0.092	0.17	0.097
11000	0.145	0.104	0.085	0.07
11500	0.14	0.098	0.086	0.176

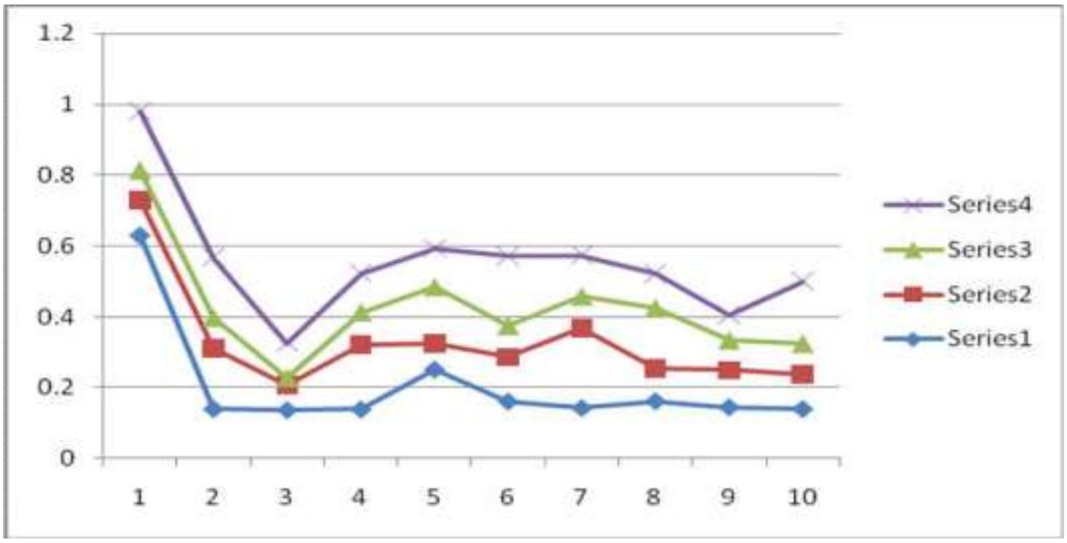


Chart1:HiveQuery statisticswithout indexing(alltheresultsareinseconds)
Table2:HiveQuerystatisticswithBitmapIndexing

Rows/Tables	**Table1**	**Table2**	**Table3**	**Table4**
50	0.129	0.935	0.085	0.086
100	0.157	0.104	0.095	0.182
150	0.104	0.09	0.181	0.096
500	0.139	0.098	0.154	0.084
1000	0.158	0.097	0.196	0.089
5000	0.123	0.175	0.173	0.096
6000	0.113	0.175	0.081	0.124
10000	0.144	0.108	0.082	0.08
11000	0.169	0.192	0.191	0.089
11500	0.121	0.096	0.095	0.185

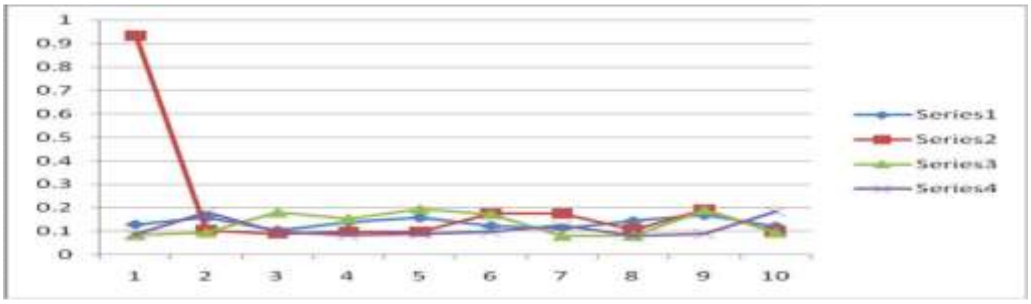
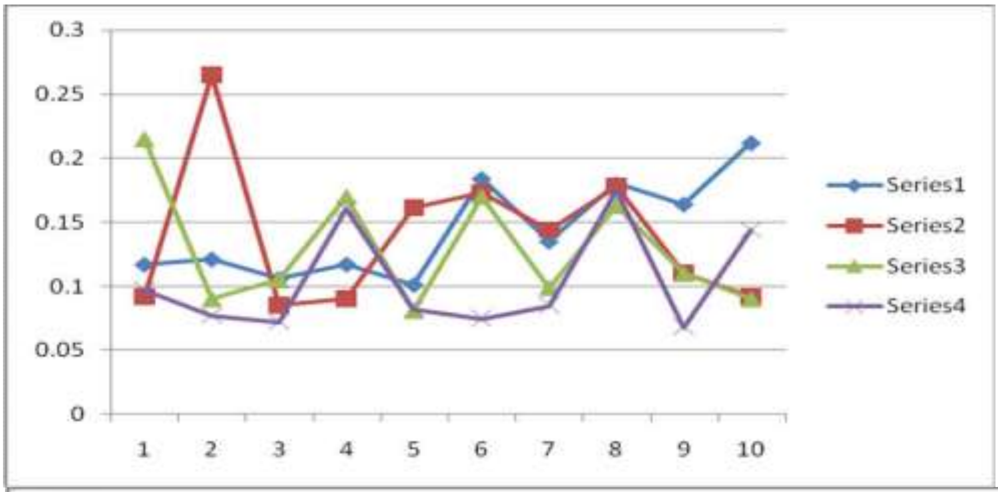


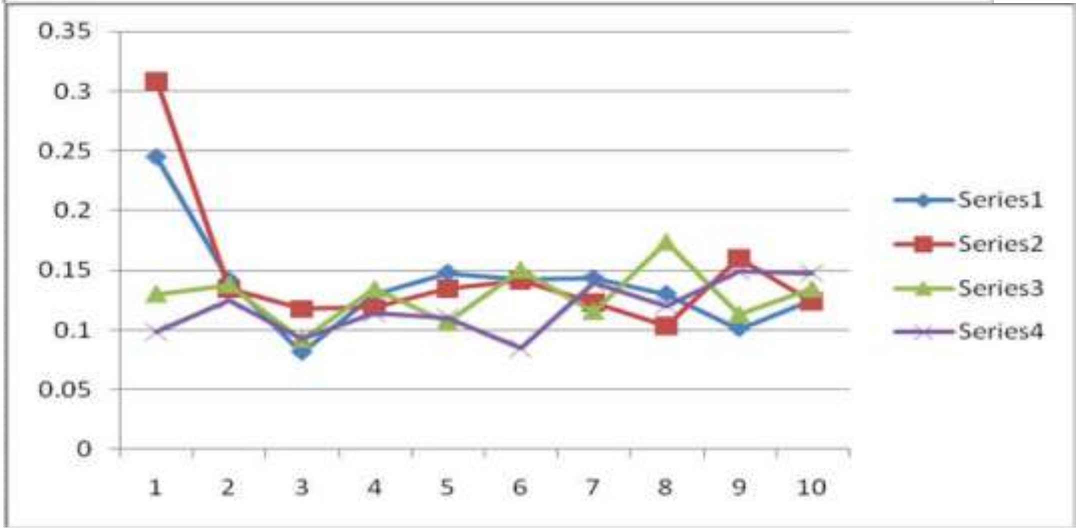
Chart2:HiveQuerystatisticswithBitmapIndexing

Table3:HiveQuerywithCompactIndexing

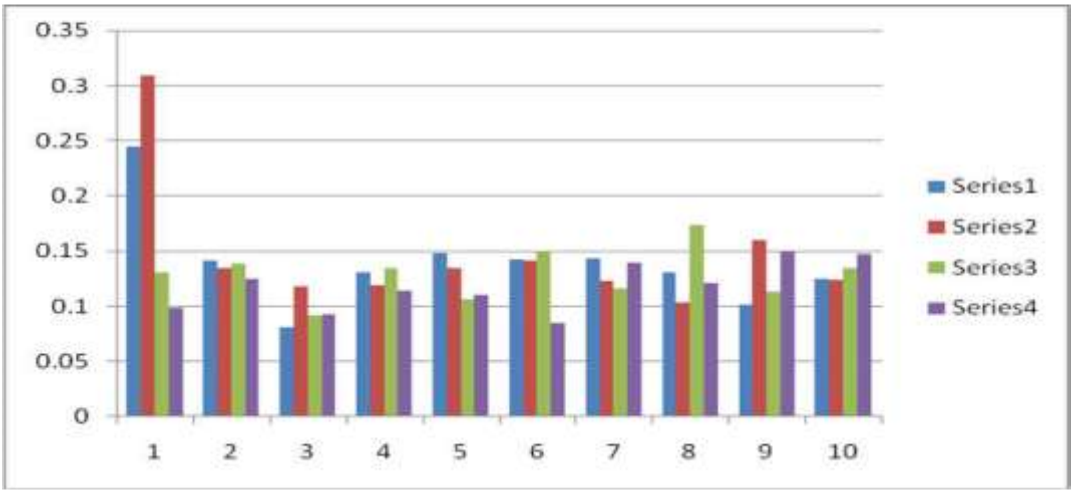
Rows/Tables	Table1	Table2	Table3	Table4
50	0.117	0.092	0.215	0.097
100	0.121	0.265	0.09	0.077
150	0.106	0.085	0.105	0.072
500	0.117	0.09	0.17	0.161
1000	0.101	0.161	0.081	0.082
5000	0.184	0.173	0.17	0.075
6000	0.135	0.144	0.099	0.085
10000	0.18	0.178	0.163	0.175
11000	0.164	0.11	0.11	0.068
11500	0.212	0.092	0.09	0.144

z





AverageHiveQuerywithnoindex,Bitmap,CompactandBothBitmapandCompactIndexingtogether



Barchart)AverageHiveQuerywithnoindex,Bitmap,CompactandBothBitmapandCompactIndexingtogether

Note(BarChartLegend):
Series 1: No IndexSeries2:BitmapIndexSeries3:CompactIndexSeries4:BothIndex
X–CountofValuesY-AxisTimeinSeconds

6. **ConclusionandFutureScope**

We conclude that Individually Compact and Bitmap indexes work well when it comes to data analysis. It has been shown that using both indexes in Hive Partition for single-dimensional data analysis is the most efficient method. Apache Spark must be used in order to analyze various data sets in the form of audio, video, and image data.

REFERENCES

[1] The importance of 'big data': A definition, M. A. Beyer and D. Laney, Gartner Technical Report, 2012.

[2] There is a lot of data out there that can be mined with the help of big data mining, and this is what we're going to focus on in this section of the IEEE Transactions on Knowledge and Data Engineering (TKDE).

[3] "Mining of huge datasets" by Rajaraman and J. D. Ullman, Cambridge University Press, 2012.

[4] Z. Zheng, J. Zhu, and M. R. Lyu [3] are the authors. An Overview of Service-Generated Big Data and Big Data-as-a-Service," in IEEE BigData, p. 403-410, October 2013. To compare the effectiveness of social, collaborative filtering and hybrid recommenders, ACM Transactions on Intelligent Systems and Technology published an article by Abellogin, Cantador, Dez et al in volume 4, issue 1, January 2013 (pp. 1-37).

[5] In "Can Dissimilar Users Contribute to Accuracy and Diversity of Personalized Recommendation?" by W. Zeng, M. S. Shang, and Q. M. Zhang et al., International Journal of Modern Physics C, vol. 21, no. 10, June 2010, pages 1217-1227.

[6] [5] Apache Hadoop [On-line]. [6] Hadoop is currently available at http://hadoop.apache.org/.

[7] This paper is entitled "The Hadoop Distributed File System," and it was published in the Proceedings of the IEEE Conference on Mass Storage Systems and Technologies (MSST) in Incline Village, NV, in 2010.

[8] "[Online] Apache Hbase" (version 7) You can access it at http://hbase.apache.org/.

[9] (nine) Abouzeid, Bajda-Piwlikowski, D. Abadi, Silberschatz, and Rasin, all of them are from the University of Pennsylvania. Analytical workloads can benefit from Hadoopdb, an architectural combination of mapreduce and dbms technologies. IEEE

Transactions on Knowledge and Data Engineering, 2(1):926–933, 2009.

[10]

[11] Allan Jindal, Yuri Kargin, Vijay Setty, and J. Schad are listed as co-authors on a study published in the Journal of Clinical Pathology (JCP) in 2009. It's possible to speed up a yellow elephant with Hadoop++ (without it even noticing). 3(1-2):515–529. 2010 VLDB Endowment Proceedings.

[12] [20] http://dwgeek.com/hive-different-file-formats-text-sequence-rc-avro-orc-parquet-file.html/ https://snippetessay.wordpress.com/2015/07/25/hive-optimizations-with-indexes-bloom-filters-and-statistics/

[13] http://web.cse.ohiostate.edu/hpcs/WWW/HTML/publications/papers/TR-11-4.pdf

[14] https://www.semantikoz.com/blog/orc-intelligent-big-data-file-format-hadoop-hive/

[15] U. F. Minhas, F. Ozcan, and A. Floratou. It's a return to shared-nothing database architectures with SQL on Hadoop. 7(12), 2014, Proceedings of the VLDB Endowment

[16] AshwiniSomnath, A. Thusoo, J. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy are the authors. Hadoop-based data warehouse on the order of a petabyte in size. IEEE 26th International Conference on Data Engineering (ICDE), pages 996–1005. IEEE,2010.