# Comparison of SQL with HiveQL

**Rakesh Kumar[1] Neha Gupta[2] Shilpi Charu[3] Somya Bansal[4] Kusum Yadav[5]**
[1,2,3,4,5]Department of Information Technology & JECRC, Jaipur, India

**Abstract**—*SQL is a set based declarative programming language, keyword based language and not an imperative programming language like C or BASIC, for accessing as well as manipulating database systems. This research paper include the basic concept of SQL with its advantages, disadvantages as well as its architecture, and introduction to Apache Hive with its features, advantages, disadvantages and its architecture. Further this research paper also contains introduction to HiveQL as well as comparison of SQL with HiveQL.*

**Keywords**—Hive;HiveQL; Hadoop**;** RDBMS; SQL

## I. Introduction

Structured Query Language (SQL) comprises one of the fundamental building blocks of modern database architecture, it defines methods used to create as well as manipulate relational databases on all major platforms. HiveQL does not strictly follow full SQL-92 standard, and it offers extensions not in SQL, including multitable inserts as well as create table as select, but only offers basic support for indexes. HiveQL lacks support for transactions as well as materialized views, and only limited subquery support. Internally, a compiler translates HiveQL statements into a directed acyclic graph of MapReduce (MR) jobs, which are further submitted to Hadoop for execution.

## II. SQL

Structured Query Language (SQL) is a widely-used programming language for working with relational databases, as well as it is a computer language for storing, manipulating and retrieving data stored in relational database. The original version of SQL is called *SEQUEL* (Structured English query language) was designed by an IBM research center in 1974 and 1975, and it was first introduced as a commercial database systemin 1979 by Oracle Corporation. Relational DatabaseManagement System (RDBMS) is a database management system (DBMS) that is based on the relational model as introduced by E. F. Codd, and it is basis for SQL, as well as for all modern database systems like MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access.

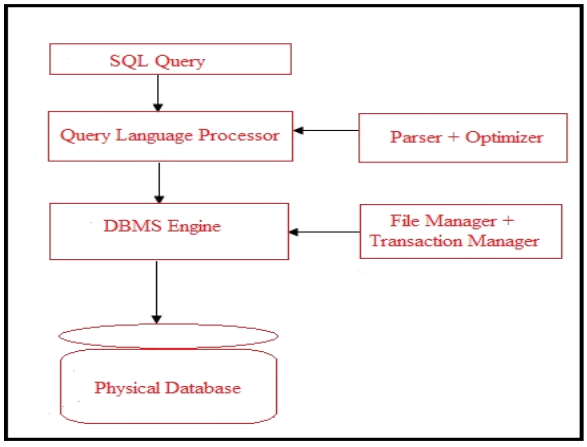


Fig. 1:SQL Architecture

## III. SQL Advantages And Disadvantages

SQL includes the ability to insert data, submit queries, update and delete data, and control the point of access.SQL is excellent for certain kinds of tasks, especially manipulating as well as retrieving sets of data.

### *A. Advantages of SQL*

The benefit of SQL database language is that it allows you to insert, update, delete, or retrieve data with simple commands. It allows users to take on administrative functions, as well as manage the database.

- Very powerful, Universal
- Easy to learn and understand
- Portable, Multiple data views
- Used with any DBMS system with any vendor
- Well Defined Standards Exist and used for relational databases
- Both as programming language and interactive language
- Complete language for a database
- Client/Server, Dynamic database language
- Supports object based programming, enterprise applications
- High Speed, Integrates with Java

### *B. Disadvantages of SQL*

SQL is declarative and it has a pseudo-natural-language style**.** SQL is hard to understand except in trivial cases, as well as it has been standardized, but too late, many vendors already developed their language extensions.

- Requires Detailed Knowledge of the Structure of the Database
- Encapsulation mechanisms are coarse
- Difficulty in Interfacing, Very specialized, geeky
- More Features Implemented in Proprietary way
- No standard SQL mechanism for hiding pieces of the code from one another or grouping them into logical units
- If you want to perform same operation on different tables, you have got to write the code twice
- Manually coding CRUD operations in SQL is repetitive as well as error-prone

## IV. Apache Hive

Apache Hive data warehouse software facilitates querying as well as managing large datasets residing in distributed storage. Hive is one of the easiest to use of the high-level MapReduce (MR) frameworks, and it also provides a SQL-like language known as HiveQL.

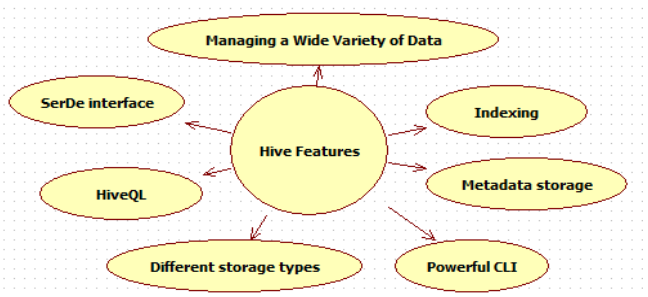


Fig. 2: Apache Hive Features

Hive was originally an internal Facebook project which eventually tenured into a full-blown Apache project, and it was created to simplify access to MapReduce (MR) by exposing a SQL-based language for data manipulation. Hive also maintains metadata in a metastore, which is stored in a relational database, as well as this metadata contains information about what tables exist, their columns, privileges, and more.

Hive is an open source data warehousing solution built on top of Hadoop, and its particular strength is in offering ad-hoc querying of data, in contrast to the compilation requirement of Pig and Cascading. Hive is a natural starting point for more full-featured business intelligence systems which offer a user friendly interface for non-technical users.

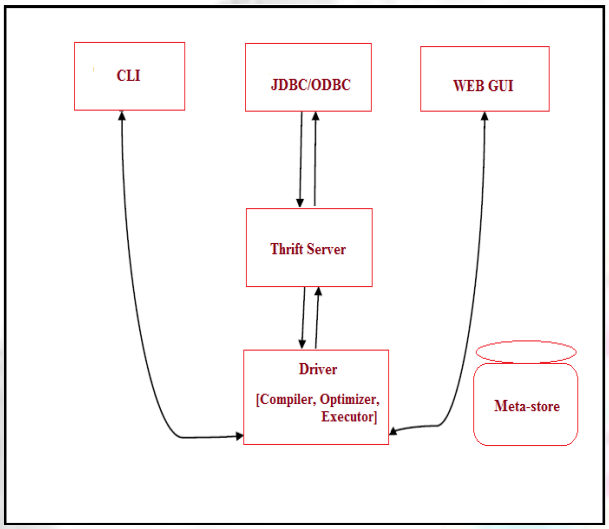


Fig. 3: Hive Architecture

## V. Apache Hive Advantages And Disadvantages

Apache Hive supports analysis of large datasets stored in Hadoop's HDFS as well as easily compatible file systems like Amazon S3 (Simple Storage Service). Amazon S3 is a scalable, high-speed, low-cost, Web-based service designed for online backup and archiving of data as well as application programs. Hive provides SQL-like language called HiveQL while maintaining full support for map/reduce, and to accelerate queries, it provides indexes, including bitmap indexes. Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, as well as analysis.

### *A. Advantages of Apache Hive*

- Perfectly fits low level interface requirement of Hadoop
- Hive supports external tables and ODBC/JDBC
- Having Intelligence Optimizer
- Hive support of Table-level Partitioning to speed up the query times
- Metadata store is a big plus in the architecture that makes the lookup easy

### *B. Disadvantages of Apache Hive*

- Not support for UPDATE & DELETE
- Not support for singleton INSERT
- Data is loaded from a file using LOAD command
- Not Access Control Language supported
- Not supported correlated subquery
- Only 1-level of partitioning available
- HiveQL is not 100% ANSI-Compliant SQL

## VI. HiveQL

Apache Hive supports a SQL-like query language known as the Hive query language over one or multiple data files located either in a local file system or in HDFS. Hive query language runs over Hadoop map-reduce framework itself, but hides complexity from the developer, and it is composed of a subset of SQL features as well as some useful extensions that are helpful for batch processing systems.

Hive query language (HiveQL) supports SQL features like CREATE tables, DROP tables, SELECT ... FROM ... WHERE clauses, Joins (inner, left outer, right outer and outer joins), Cartesian products, GROUP BY, SORT BY, aggregations, union and many useful functions on primitive as well as complex data types. Metadata browsing features such as list databases, tables and so on are also provided. HiveQL does have limitations compared with traditional RDBMS SQL. HiveQL allows creation of new tables in accordance with partitions(Each table can have one or more partitions in Hive) as well as buckets (The data in partitions is further distributed as buckets)and allows insertion of data in single or multiple tables but does not allow deletion or updating of data.

## VII. Comparison of SQL with HiveQL

| S.No | Feature | SQL | HiveQL |
|---|---|---|---|
| 1. | Select | SQL-92 | Single table or view in the FROM clause. SORT BY for partial ordering. LIMIT to limit number of rows returned. HAVING not supported. |
| 2. | Data types | Integral, floating point, fixed point, text and binary strings, temporal | Integral, floating point, boolean, string, array, map, struct |
| 3. | Updates | UPDATE, INSERT, DELETE | INSERT OVERWRITE TABLE (populates whole table or partition) |
| 4. | Functions | Hundreds of built-in functions | Dozens of built-in functions |
| 5. | Default Join | Inner Join | Equi Join |
| 6. | Multitable inserts | Not supported | Supported |
| 7. | Views | Updatable. Materialized or nonmaterialized. | Read-only. Materialized views not supported |
| 8. | Transactions | Supported | Not supported |

| | | | |
|---|---|---|---|
| 9. | Subqueries | In any clause. Correlated or noncorrelated. | Only in the FROM clause. Correlated subqueries not supported |
| 10. | Latency | Sub-second | Minutes |
| 11. | Create table as select | Not valid SQL-92, but found in some databases | Supported |
| 12. | Extension points | User-defined functions. Stored procedures. | User-defined functions. Map-Reduce scripts. |
| 13. | Indexes | Supported | Not supported |

Table. 1 Comparison of SQL with HiveQL

## VIII. Conclusions

The primary reason for moving data between SQL stores as well as Hadoop is usually to take advantage of the massive storage and processing capabilities to process quantities of data larger than you could hope to cope with in SQL alone. Increasing popularity of Hadoop as data platform of choice for many organizations, HIVE becomes must-have supplement to provide greater usability as well as connectivity within the organization by introducing high-level language support known as HiveQL.

## IX. Future Work

As future work, we plan to incorporate multiple query optimization framework and its functionality at MapReduce layer. In this way, it will be possible to eliminate even more redundant MapReduce (MR) tasks in queries as well as improve overall performance of naïve rule-based Hive query optimizer even further.

## References

[1] Ivan Tomašić, Aleksandra Rashkovska, MatjažDepolli and Roman Trobec, "A Comparison of Hadoop Tools for Analyzing Tabular Data"; Informatica 37 (2013) 131–138; Received: December 24, 2012
[2] "Introduction to Hive" © 2009 Cloudera, Inc.
[3] "Data Management in the Cloud"; Data Storage and Processing Paradigms: Hive June 26, 2013; Lecture 10
[4] Edward Capriolo, Dean Wampler, and Jason Rutherglen; "Programming Hive"; ISBN: 978-1-449-31933-5[LSI] 1347905436
[5] Tom White foreword by Doug Cutting; "Hadoop: The Definitive Guide"; ISBN: 978-0-596-52197-4 [M] 1243455573
[6] Tom White foreword by Doug Cutting; "Hadoop: The Definitive Guide"; ISBN: 978-1-449-38973-4 [SB] 1285179414
[7] AnjaGruenheid, Edward Omiecinski, Leo Mark; "Query Optimization Using Column Statistics in Hive"
[8] O'Reilly Media; "Big Data Now"; ISBN: 978-1-449-31518-4 1316111277
[9] Lars George | EMEA Chief Architect Cloudera; Hadoop Masterclass Part 4 of 4: Analyzing Big Data
[10] Introduction To Hive How to use Hive in Amazon EC2; CS 341: Project in Mining CS CS 341: Project in Mining Massive Data Sets Hyung Jin (Evion) Kim Stanford University
[11] Marissa Rae Hollingsworth; "HADOOP AND HIVE AS SCALABLE ALTERNATIVES TO RDBMS: A CASE STUDY"
[12] ALEX HOLMES "Hadoop in Practice" ISBN 9781617290237
[13] CHUCK LAM "Hadoop in Action"; ISBN: 9781935182191
[14] https://cwiki.apache.org/confluence/display/Hive/LanguageManual
[15] http://docs.treasuredata.com/articles/hive#about-apache-hive
[16] http://en.wikibooks.org/wiki/Structured_Query_Language
[17] https://cwiki.apache.org/confluence/display/Hive/Tutorial
[18] SQL Basics; Introduction to Standard Query Language
[19] "Apache Hive Review"; BHARAT SINGHVI 10305912
[20] http://www.moreprocess.com/sql/pros-advantages-of-sql-structured-query-language
[21] http://www.cs.iit.edu/~cs561/cs425/VenkatashSQLIntro/Advantages%20&%20Disadvantages.html
[22] http://en.wikipedia.org/wiki/Apache_Hive
[23] https://hive.apache.org/
[24] https://cwiki.apache.org/confluence/display/Hive/DesignDocs
[25] https://cwiki.apache.org/confluence/display/Hive/Home
[26] http://www.devx.com/Java/Article/48139/0/page/2
[27] "Apache Hive" Homework number: 3; Group number: EEDC-1
[28] "HIVE"A warehouse solution over map-reduce framework; DonyAng
[29] SQLTutorial Simply Easy Learning by tutorialspoint.com
[30] Introduction to Apache Hive; PelleJakovits
[31] Hortonworks_Tutorial_Hive_5.22.pdf
[32] http://www.journalofcloudcomputing.com/content/3/1/12
[33] http://blog.sqlauthority.com/2013/10/21/big-data-data-mining-with-hive-what-is-hive-what-is-hiveql-hql-day-15-of-21/
[34] https://www.inkling.com/read/hadoop-definitive-guide-tom-white-3rd/chapter-12/hiveql
[35] http://blog.spryinc.com/2013/10/a-list-of-subtle-differences-between.html