# Hive

**DIKSHA YADAV**[1,*,+]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[*] Corresponding authors: yadavd@umail.iu.edu
[+] HID - S17-IR-2044

*Paper-002, March 27, 2017*

**Hive is an open source data warehousing solution which is built on top of Hadoop. It structures data into understandable and conventional database terms like tables, columns, rows and partitions. It supports HiveQL queries which have structure like SQL queries. HiveQL queries are compiled to map reduce jobs which are then executed by Hadoop. Hive also contains Metastore which includes schemas and statistics which is useful in query compilation, optimization and data exploration.**

**Keywords:** Hive, Hadoop, HiveQL, SQL

https://github.com/diksha2112/sp17-i524/tree/master/paper2/S17-IR-2044/report.pdf

## INTRODUCTION

The reason behind development of Hive is making it easier for end users to use Hadoop. Map reduce programs were required to be developed by users for simple to complex tasks. It lacked expressiveness like query language. So, it was a time consuming and difficult task for end users to use Hadoop. For solving this problem Hive was built in January 2007 and open sourced in August 2008. Hive is an open source data warehousing solution which is built on top of Hadoop. It structures data into understandable and conventional database terms like tables, columns, rows and partitions. It supports HiveQL queries which have structure like SQL queries. HiveQL queries are compiled to map reduce jobs which are then executed by Hadoop. Hive also contains Metastore which includes schemas and statistics which is useful in query compilation, optimization and data exploration[1]

## ARCHITECTURE

Hive architecture includes, Database-It consists of tables created by the user. Metastore-It contains information about the system. It can be accessed by different components as and when needed. Interfaces-User interface and Application programming interface both are present in hive. Driver-manage HiveQL statements at every stage. Query compiler-It compiles HiveQL queries to acyclic graphs (directed) representing map reduce tasks. Execution Engine- It executes the tasks generated by the compiler. Hive Server- It provide JDBC/ODBC server and thrift interface[2]

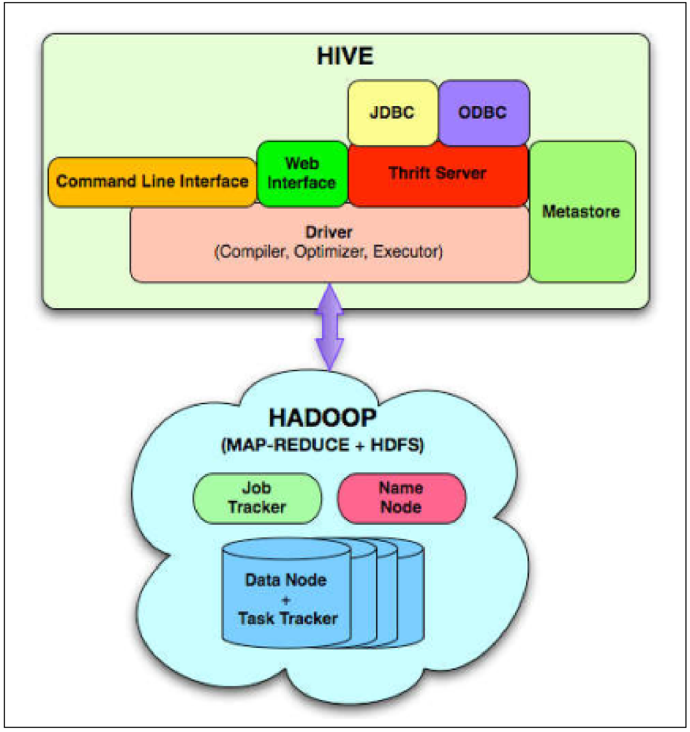

**Fig. 1.** Hive Architecture

## HIVEQL QUERY FORMAT

```
SELECT [ALL | DISTINCT] select-expr, select-expr, ...
FROM table-reference
[WHERE where-condition]
[GROUP BY col-list]
[HAVING having-condition]
[CLUSTER BY col-list | [DISTRIBUTE BY col-list] [SORT BY col-list]]
[LIMIT number]
```

[3]

## SYSTEM REQUIREMENTS

Hive is cross platform. So, It does not need any specific operating system to work.

## COMPARISON OF HIVE WITH OTHER TRADITIONAL DATABASES

Traditional databases follow schema on write approach while Hive follows schema on read approach. In schema on write, databases checks at load time if the data follows the table representation given by user while in schema on read approach, it is checked at run time only. This saves the time for hive to load the data when traditional databases takes longer time[4]

## POPULARITY OF HIVE

The popularity of hive increasing with time. This can be proved by the following plot made by DB Engines Ranking. It ranks database management systems according to their status and popularity. Following plot shows popularity of hive with time.

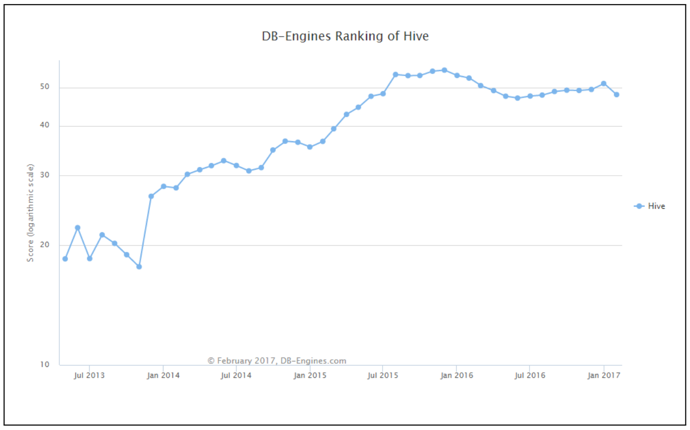

**Fig. 2.** Hive Popularity

[5]

## RESOURCES FOR LEARNING HIVE

Someone new to hive can start learning it by going through the following links in sequence: Install Hivehttps://www.edureka.co/blog/apache-hive-installation-on-ubuntu?utm_source=quora&utm_medium=crosspost&utm_campaign=social-media-edureka-ab
Hive Tutorialhttps://www.edureka.co/blog/hive-tutorial/?utm_source=quora&utm_medium=crosspost&utm_campaign=social-media-edureka-ab
Top Hive commands with exampleshttps://www.edureka.co/blog/hive-commands-with-examples?utm_source=quora&utm_medium=crosspost&utm_campaign=social-media-edureka-ab

## ACKNOWLEDGEMENT

I am also grateful to Dr. Gregor von Laszewski for providing the appropriate paper template.

## CONCLUSION

Since Hive is making the use of Hadoop easier for users, its popularity is increasing with time.

## REFERENCES

[1] “Hive,” Web Page, online; accessed 24-March-2017. [Online]. Available: https://en.wikipedia.org/wiki/Apache_Hive
[2] “Hive article,” Web Page, online; accessed 24-March-2017. [Online]. Available: https://docs.treasuredata.com/articles/hive
[3] A. T. J. S. N. J. Z. S. P. C. S. A. H. L. P. W. R. Murthy, “Hive-a warehousing solution over map reduce framework,” Paper, online; accessed 23-March-2017. [Online]. Available: http://laser.inf.ethz.ch/2013/material/breitman/additionalpercent20reading/hive.pdf
[4] ——, “Hive-a petabyte scale datawarehouse using hadoop,” Paper. [Online]. Available: http://infolab.stanford.edu/~ragho/hive-icde2010.pdf
[5] “Ranking hive,” Web Page, online; accessed 20-March-2017. [Online]. Available: http://db-engines.com/en/ranking_trend/system/Hive