GOETHE UNIVERSITÄT FRANKFURT AM MAIN

www.uni-frankfurt.de

SoSe 2014: M-TANI: Big Data Analytics

Lecture 6 – 04.06.2014

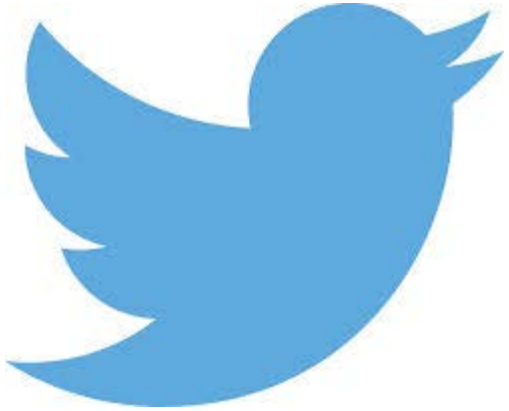
Todor Ivanov
Sead Izberovic
Dr. Nikolaos Korfiatis

Apache Hive

- **Hive** is a data warehouse system on top of Hadoop.
- Provides:
  - **ad-hoc** queries using SQL-like language called HiveQL
  - mechanism to project structure onto semi/structured data
  - **HiveQL** queries are transformed/compiled to MapReduce programs
  - users can create user defined functions (**UDFs**)
  - provide interfaces to other Hadoop tools (HCatalog, Avro etc..)

From [2]

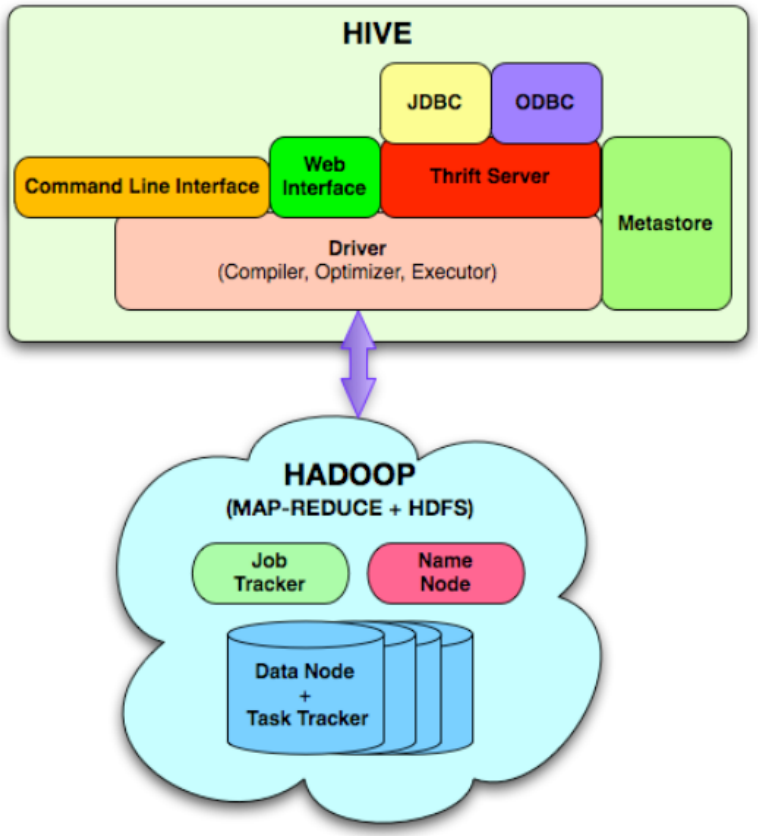
Who uses Hive?



From [5]

Hive Architecture

- **Shell (CLI)**: allows interactive queries like MySQL shell connected to database
- **Driver**: session handles, fetch, execute
- **Compiler**: parse, plan, optimize
- **Metastore**: system catalog, schema, location in HDFS



From [4]

Hive – Data Model

- **Tables** – A table is stored in a directory in hdfs.
  - **Typed columns** (int, float, string, date and boolean)
  - **Nested collection types** (array and map)

- **Partitions** – A partition of the table is stored in a subdirectory within a table's directory.
  - e.g., to range-partition tables by date

- **Buckets** – A bucket is stored in a file within the partition's or table's directory depending on whether the table is a partitioned table or not. Hash partitions within ranges (useful for sampling, join optimization).

From [3]

Physical Layout

- Warehouse directory in HDFS
  - e.g., /home/hive/warehouse
- Tables stored in subdirectories of warehouse
  - Partitions, buckets form subdirectories of tables
- Actual data stored in flat files
  - Control char-delimited text, or SequenceFiles
  - With custom formats

From [1] and [2]

HiveQL

- **DDL**
  - Create, drop table etc.
- **DML**
  - Select
  - Project
  - Join
  - Aggregate
  - Union all
  - Sub-queries in the from clause

From [1] and [2]

Hive Limitations

- Not 100% ANSI-Compliant SQL
- No „insert into“
  - Cannot insert into an exsiting table or data partition
  - Supports „insert overwrite“
- No „update“ or „delete“
- No Access Control Language supported

From [6]

Literature

1. Cloudera Inc.
   Intro to Hive
   http://blog.cloudera.com/wp-content/uploads/2010/01/6-IntroToHive.pdf
2. Capriolo, E.; Rutherglen, J. & Wampler, D. 2012
   Programming Hive - Data Warehouse and Query Language for Hadoop
   O'Reilly
3. Ashish T., Joydeep S. S., Namit J., et al. 2009
   Hive: a warehousing solution over a map-reduce framework
   Proc. VLDB Endow. 2, *2* (August 2009), 1626-1629.
4. Ashish Thusoo, Joydeep S. Sarma, Namit Jain, et al. 2010
   Hive - a petabyte scale data warehouse using Hadoop
   In ICDE '10: Proceedings of the 26th International Conference on Data Engineering (March 2010), pp. 996-1005

Literature

5. The Apache Software Foundation
   Hadoop powered by
   http://wiki.apache.org/hadoop/PoweredBy
6. Rajesh Kartha
   Big Data - An Introduction to Hive and HQL
   https://www.youtube.com/watch?v=ZSzl4mylPiw