

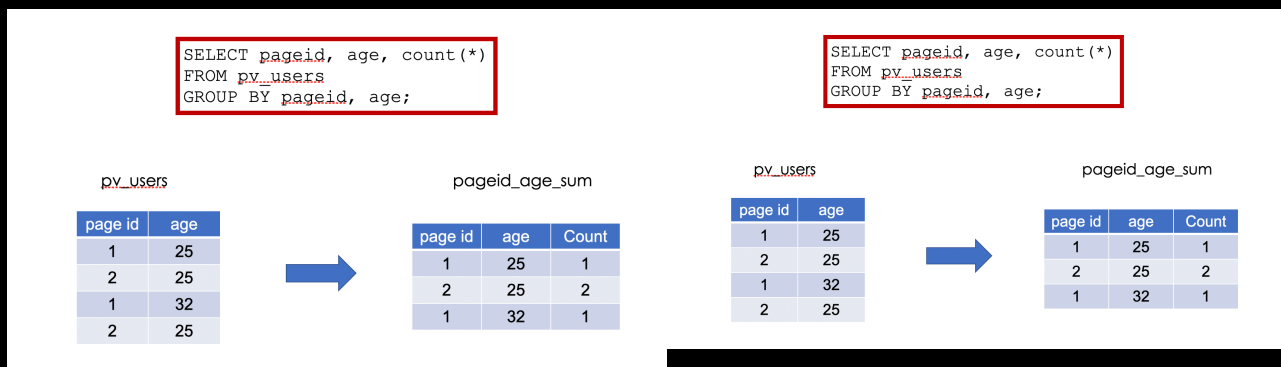
Complex types: $\text{map} \langle \text{key-type}, \text{value-type} \rangle$
 $\text{List} \langle \text{element-type} \rangle$
 $\text{struct} \langle \text{file-name}, \text{field-type}, \dots \rangle$

Hive Query Language

HQL \subseteq SQL + some extensions

↳ Support basic SQL statements
select, project, join, group by, aggregation, create table

Hive Join: Only support equality join



Hive cannot insert data is designed for data analytics.

Data is generated from outside of Hive

Pros: better concurrency control

Cons: hard for optimize, cannot ensure best storage layout

In DB's they figure out the best storage layout for fast query.

Hive Architecture

Metastore: component that store the system catalog and meta data about tables, cols, parts, . . . stored on RDBMS.

Driver: component that manages the lifecycle of a HiveSQL statement as it moves through the hive. Also, maintains a session handle and any session statistics.

Query compiler: component that compiles HiveSQL into a directed acyclic graph of map/reduce tasks.

Optimizer: chain of transformations such that the operator DAG resulting from one trans. is passed as input to the next trans.
(col. pruning, part. pruning, repartitioning of data)

Execution Engine: component that executes the tasks produced by the compiler in proper dependency order. The execution engine interacts w/ the underlying Hadoop instance.

Thrift server: component that provides a thrift interface and a JDBC/ODBC server and provides a way of integrating Hive with other apps.

Client components: Command Line, Interface (CLI)
Web UI, JDBC/ODBC driver

Some Techniques We Know and Love Are not Directly Applicable

- Indexing
 - Zone-maps
 - Co-located joins
 - Query rewrites
 - Cost-based optimization
- Databases own their storage
SQL-on-Hadoop systems do not
 - Metadata management is tricky
 - Data inserted/loaded without SQL system knowledge
 - No co-location of related tables
 - HDFS is for most practical purposes, read-only

software that
let programs talk
to the DB.