Apache HIVE

Data Warehousing & Analytics on Hadoop

Hefu Chai

What is HIVE?

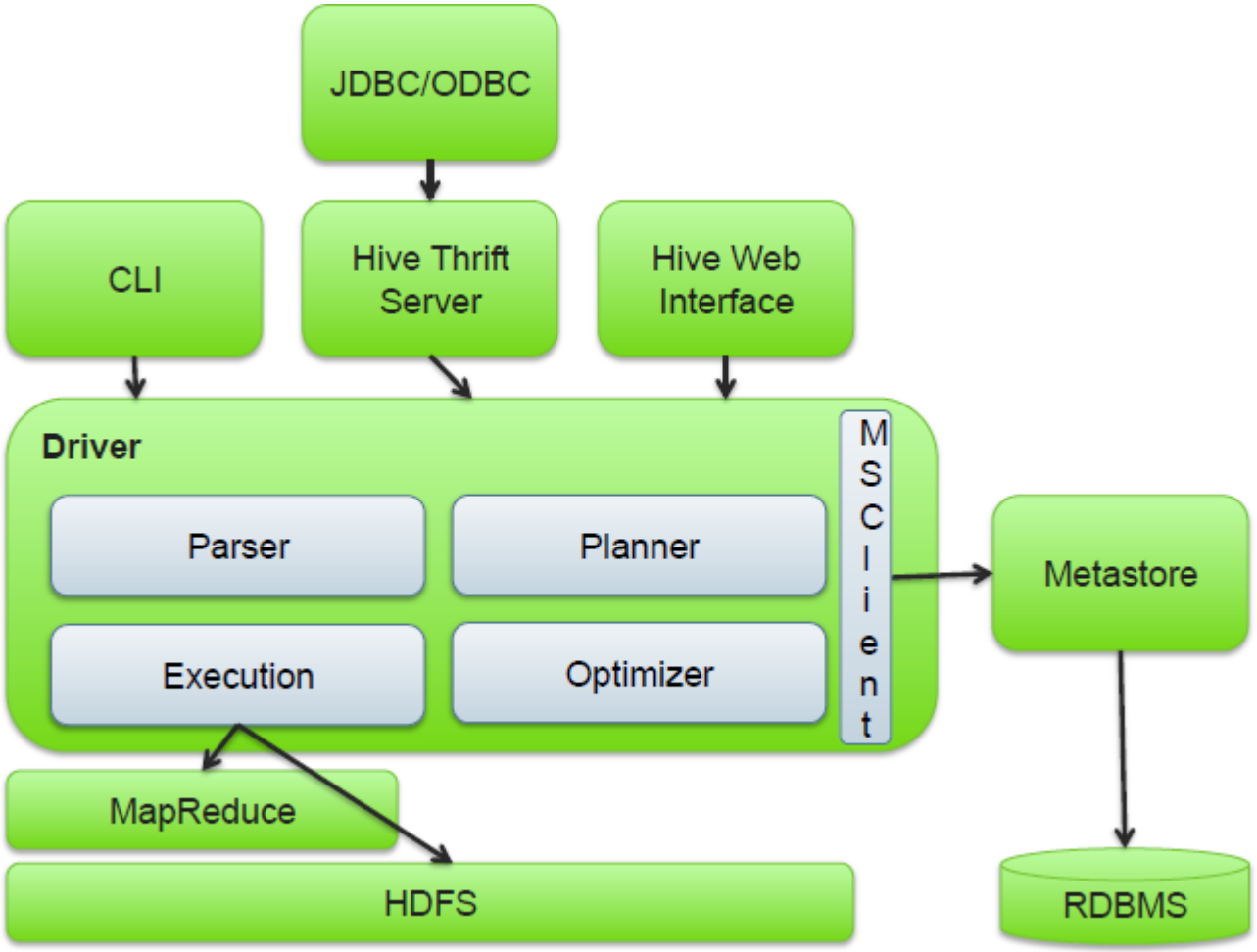
- A system for managing and querying structured data built on top of Hadoop
  - Uses Map-Reduce for execution
  - HDFS for storage
  - Extensible to other Data Repositories

- Key Building Principles:
  - SQL on structured data as a familiar data warehousing tool
  - Extensibility (Pluggable map/reduce scripts in the language of your choice, Rich and User Defined data types, User Defined Functions)
  - Interoperability (Extensible framework to support different file and data formats)

What HIVE Is **Not**

- Not designed for OLTP
- Does not offer real-time queries

HIVE Architecture



Hive/Hadoop Usage @ Facebook

- Types of Applications:
  - Summarization
    - Eg: Daily/Weekly aggregations of impression/click counts
    - Complex measures of user engagement
  - Ad hoc Analysis
    - Eg: how many group admins broken down by state/country
  - Data Mining (Assembling training data)
    - Eg: User Engagement as a function of user attributes
  - Spam Detection
    - Anomalous patterns for Site Integrity
    - Application API usage patterns
  - Ad Optimization
  - Too many to count ..

Hive Query Language

- Basic SQL
  - CREATE TABLE sample (foo INT, bar STRING) PARTITIONED BY (ds STRING);
  - SHOW TABLES '.*s';
  - DESCRIBE sample;
  - ALTER TABLE sample ADD COLUMNS (new_col INT);
  - DROP TABLE sample;
- Extensibility
  - Pluggable Map-reduce scripts
  - Pluggable User Defined Functions
  - Pluggable User Defined Types
  - Pluggable SerDes to read different kinds of Data Formats

Hive QL – Join

page_view

pageid	**userid**	time
1	**111**	9:08:01
2	**111**	9:08:13
1	**222**	9:08:14

X

user

userid	age	gender
111	25	female
222	32	male

=

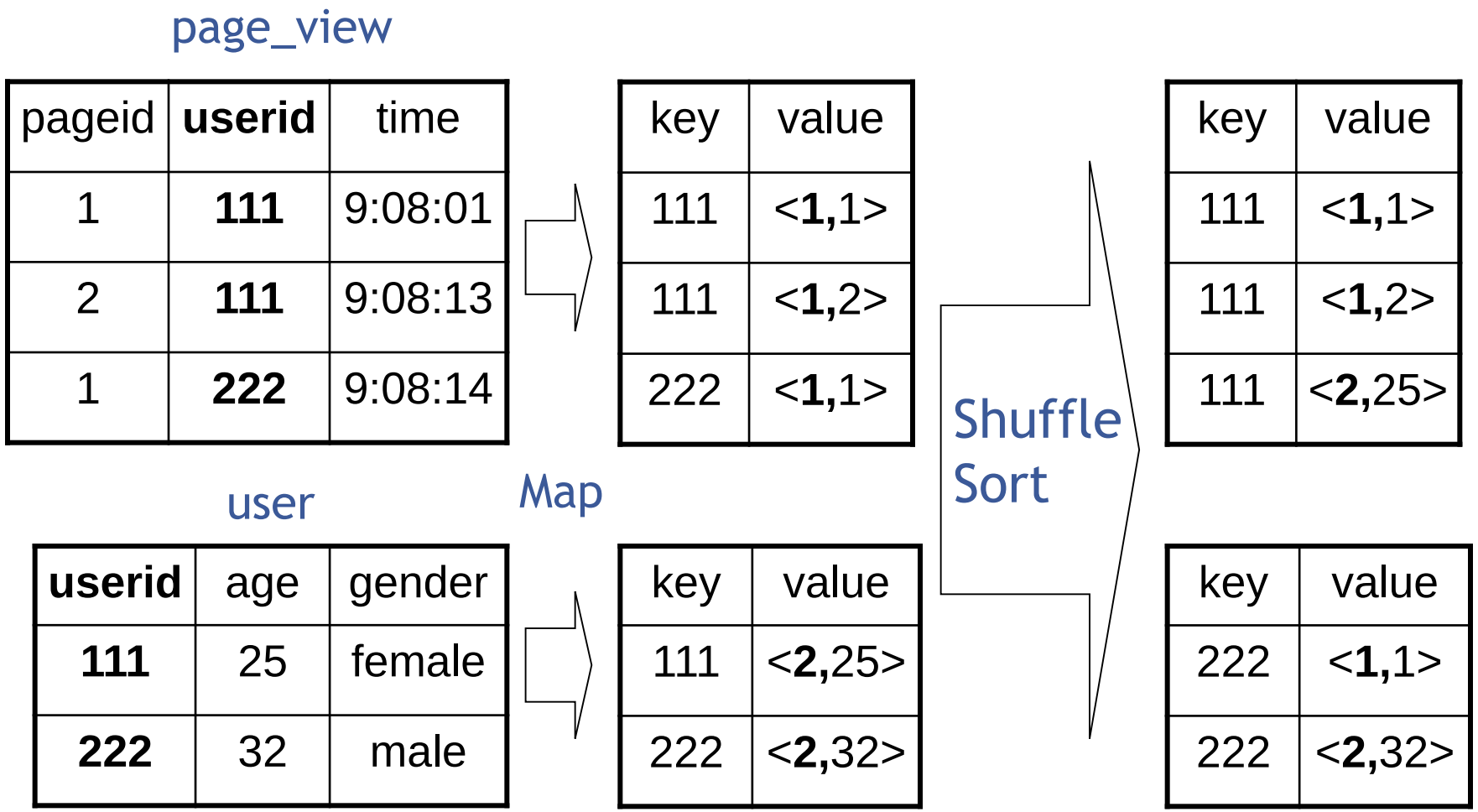
pv_users

pageid	age
1	25
2	25
1	32

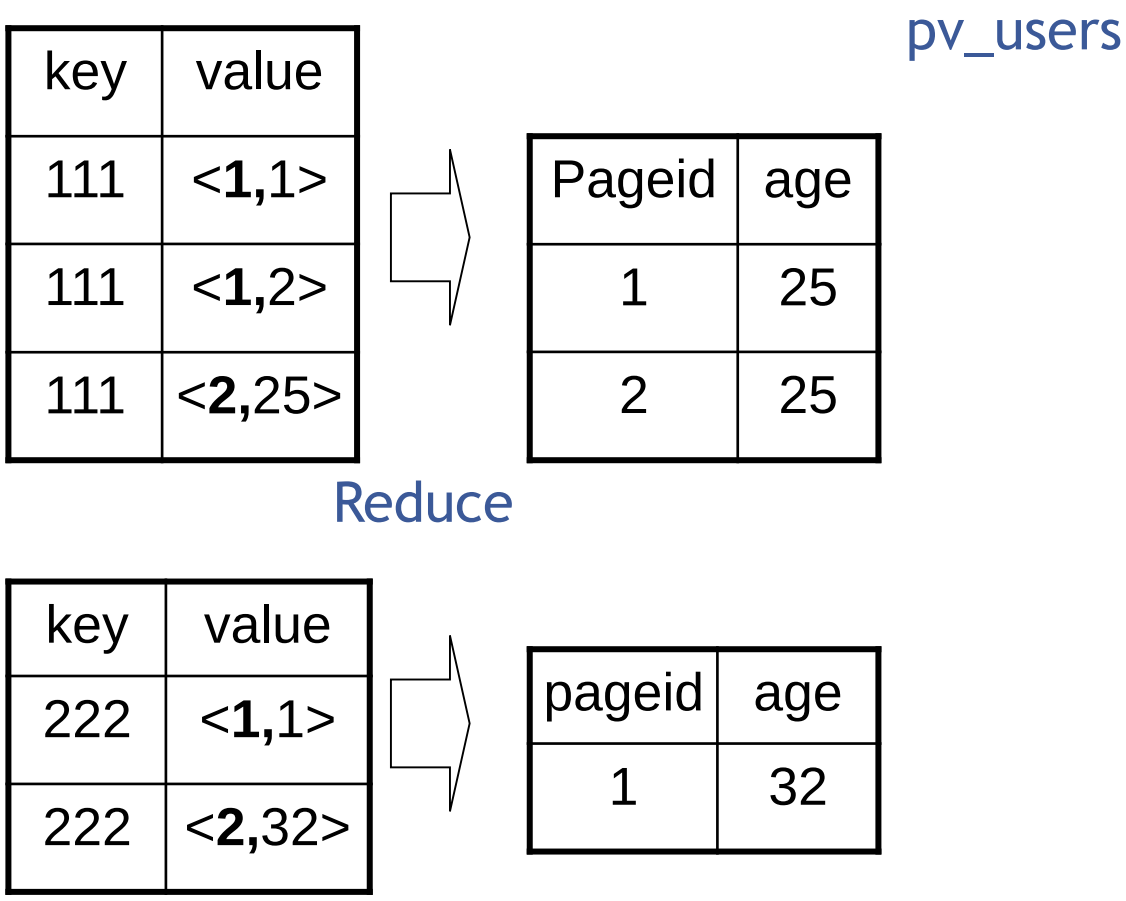
- SQL:

```
INSERT INTO TABLE pv_users
SELECT pv.pageid, u.age
FROM page_view pv JOIN user u ON (pv.userid = u.userid);
```

Hive QL – Join in Map Reduce



Hive QL – Join in Map Reduce

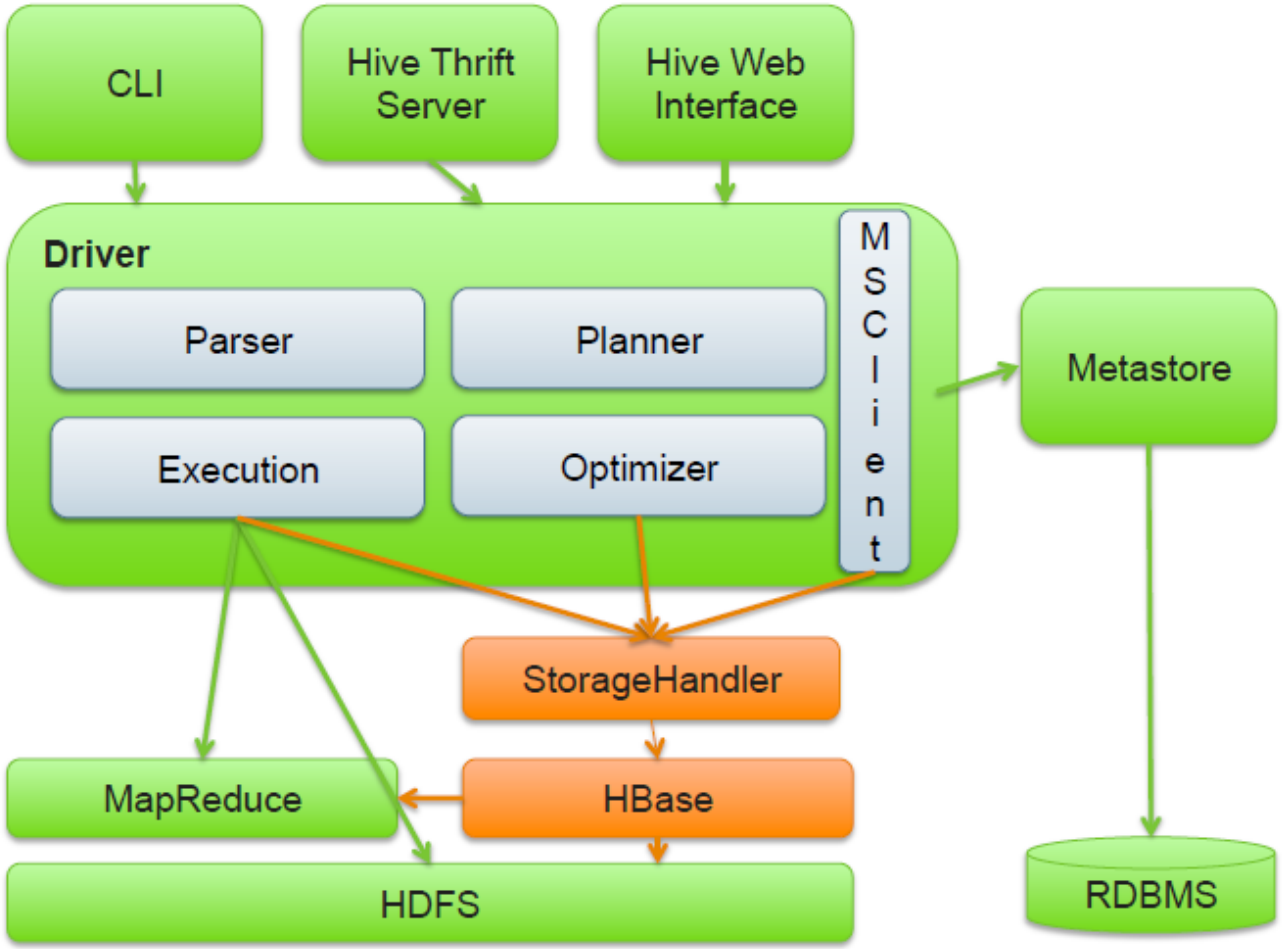


Integration with HBase

- Reasons to use Hive on HBase:
  - A lot of data sitting in HBase due to its usage in a real-time environment, but never used for analysis
  - Give access to data in HBase usually only queried through MapReduce to people that don’t code (business analysts)

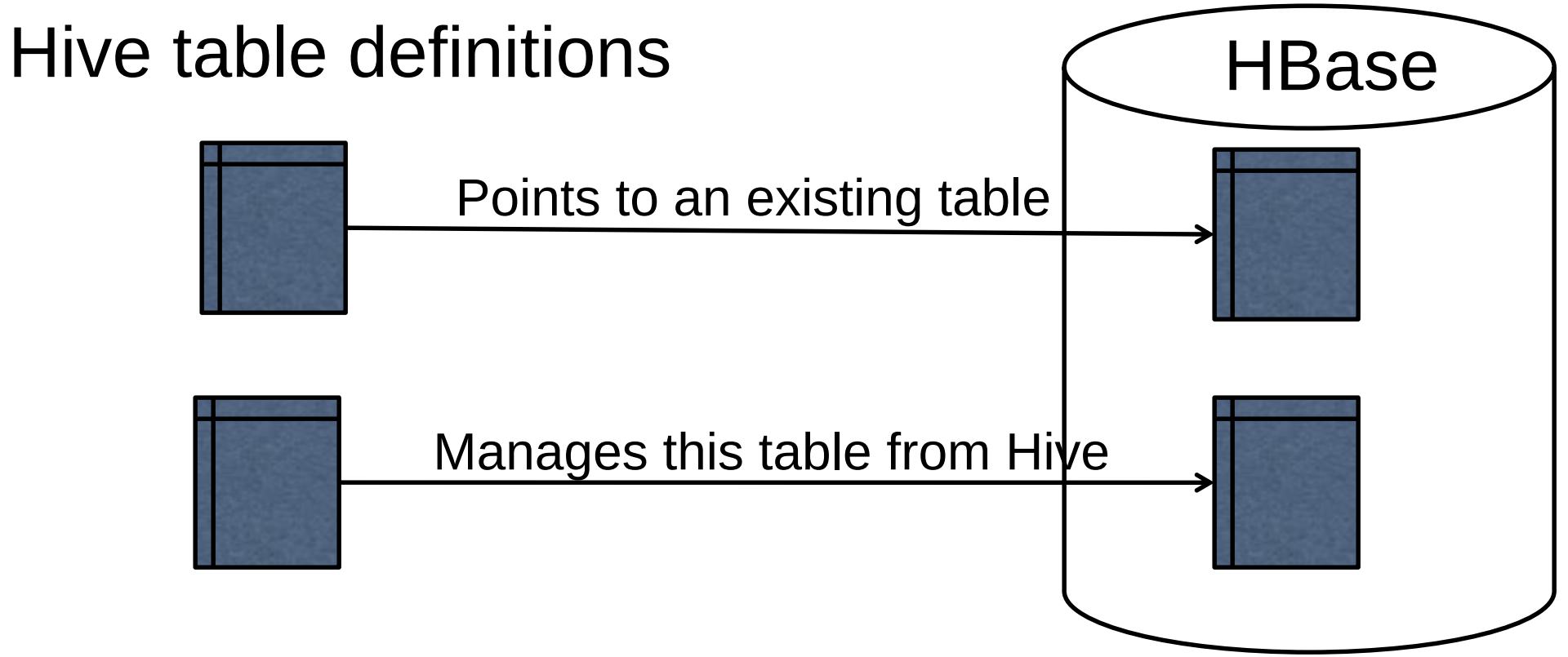
- Reasons not to do it:
  - Run SQL queries on HBase to answer live user requests (it’s still a MR job)

Integration with HBase



Integration with HBase

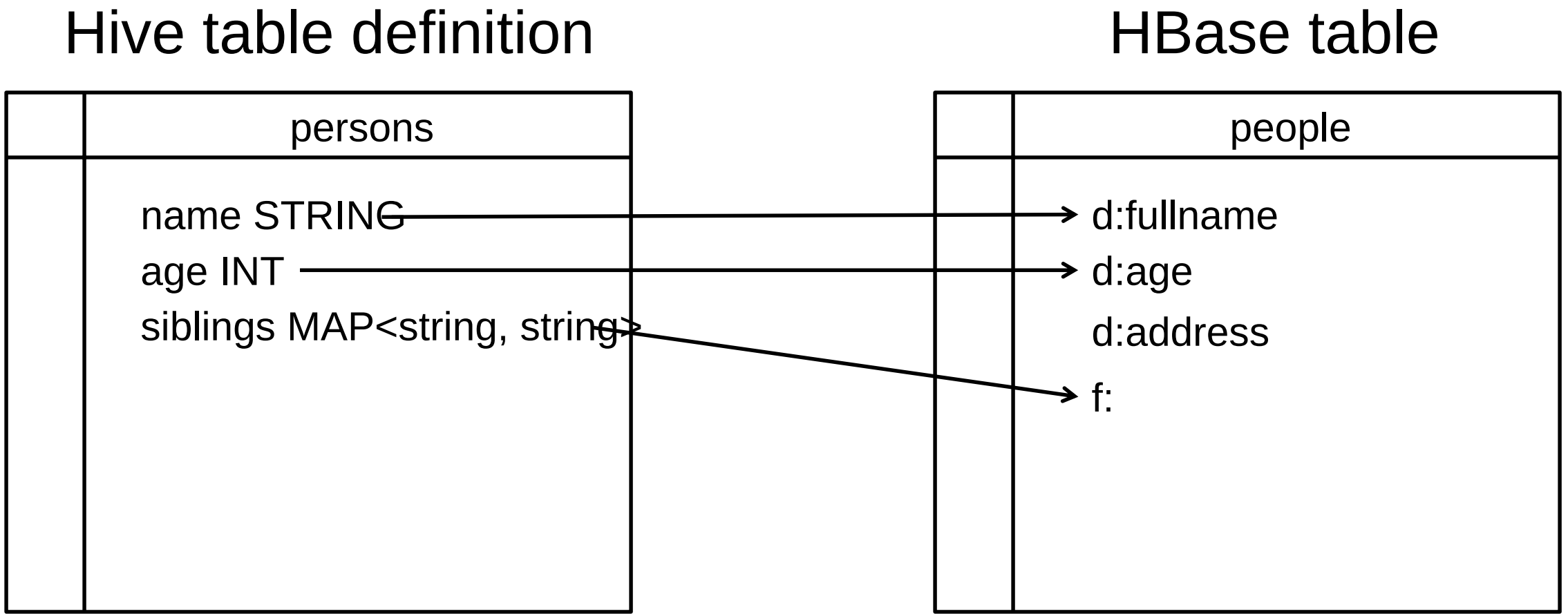
Hive can use tables that already exist in HBase or manage its own ones, but they still all reside in the same HBase instance



Integration with HBase

When using an already existing table, defined as EXTERNAL

Columns are mapped however you want, changing names and giving type



Reference

- https://cwiki.apache.org/confluence/display/Hive/Home
- Hive Facebook
- StumbleUpon

Thanks