HIVE TUTORIAL

What's in it for you?

1. History of Hive
2. What is Hive?
3. Architecture of Hive
4. Data flow in Hive
5. Hive Data Modelling
6. Hive Data types
7. Different modes of Hive
8. Difference between Hive and RDBMS
9. Features of Hive

History of Hive

1. Facebook used Hadoop as a solution to handle the growing big data

2. As we know, Hadoop uses MapReduce for processing data. MapReduce required users to write long codes

3. Not all users were well versed with Java and other coding languages. This proved to be a disadvantage for them

4. Users were comfortable with writing queries in SQL

5. Hive was developed with a vision to incorporate the concepts of tables, columns just like SQL

Why Hive?

Problem

For processing and analyzing data, users found it difficult to code as not all of them were well versed with the coding languages

Processing

Analyzing

Solution

Users required a language similar to SQL which was well known to all the users

Solution

HiveQL

What is Hive?

Hive is a data warehouse system which is used for querying and analyzing large datasets stored in HDFS

Hive uses a query language call HiveQL which is similar to SQL

Architecture of Hive

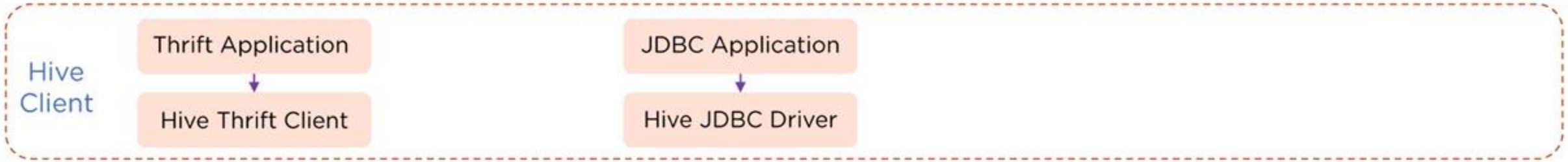
Hive
Client

Hive Client supports different types of client applications in different languages for performing queries

Architecture of Hive

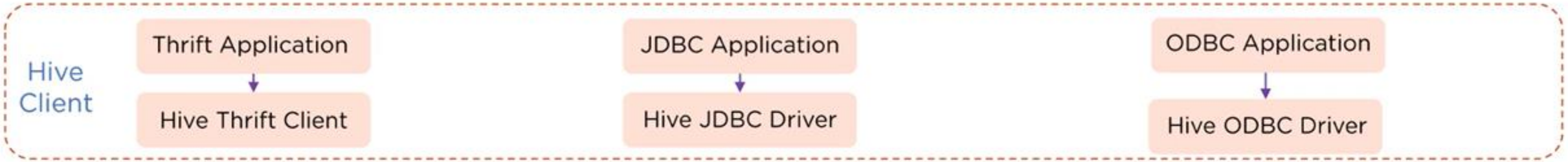
Thrift is a software framework. Hive server is based on thrift, so it can serve the request from all the programming languages that supports thrift

Architecture of Hive



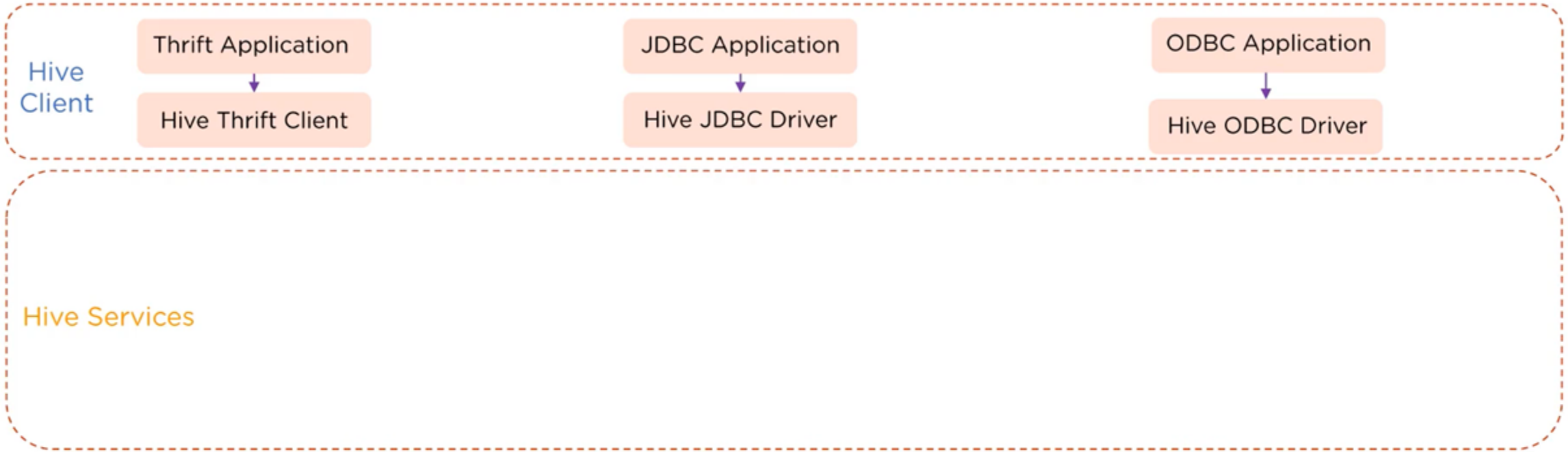
JDBC - Java Database Connectivity
JDBC application is connected through JDBC Driver

Architecture of Hive



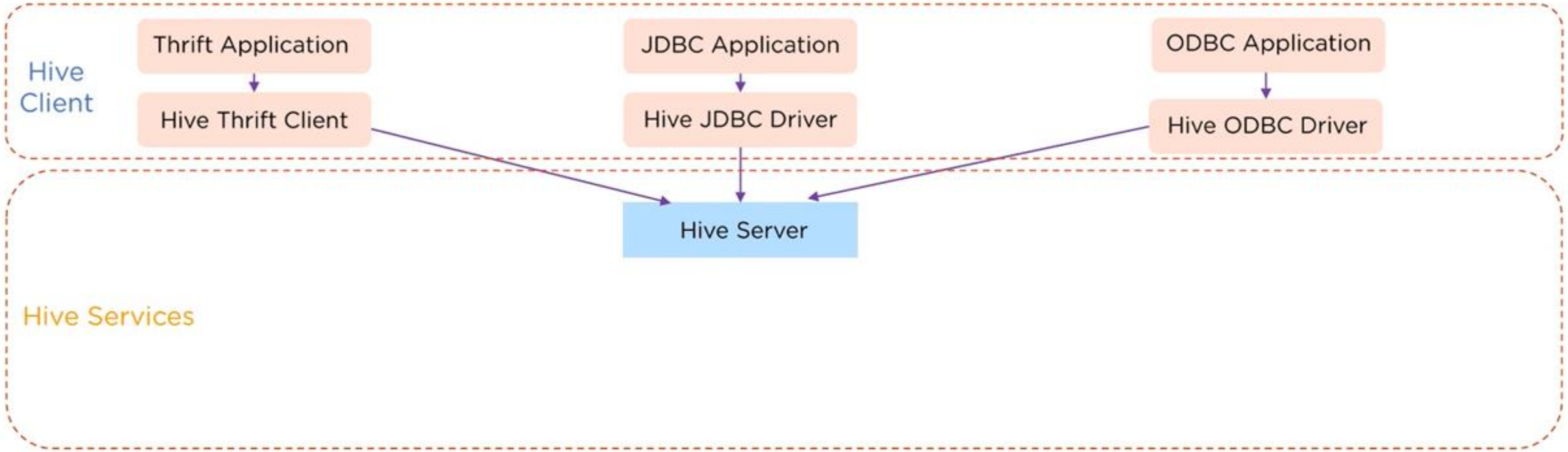
ODBC - Open Database Connectivity
ODBC application is connected through ODBC Driver

Architecture of Hive



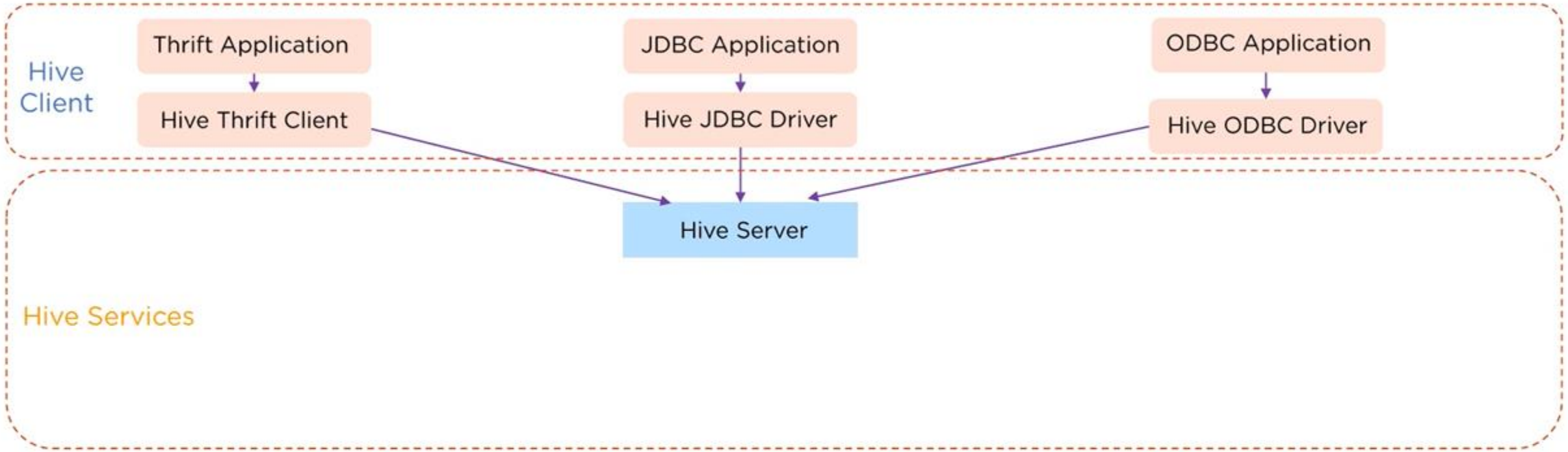
Hive supports various services

Architecture of Hive



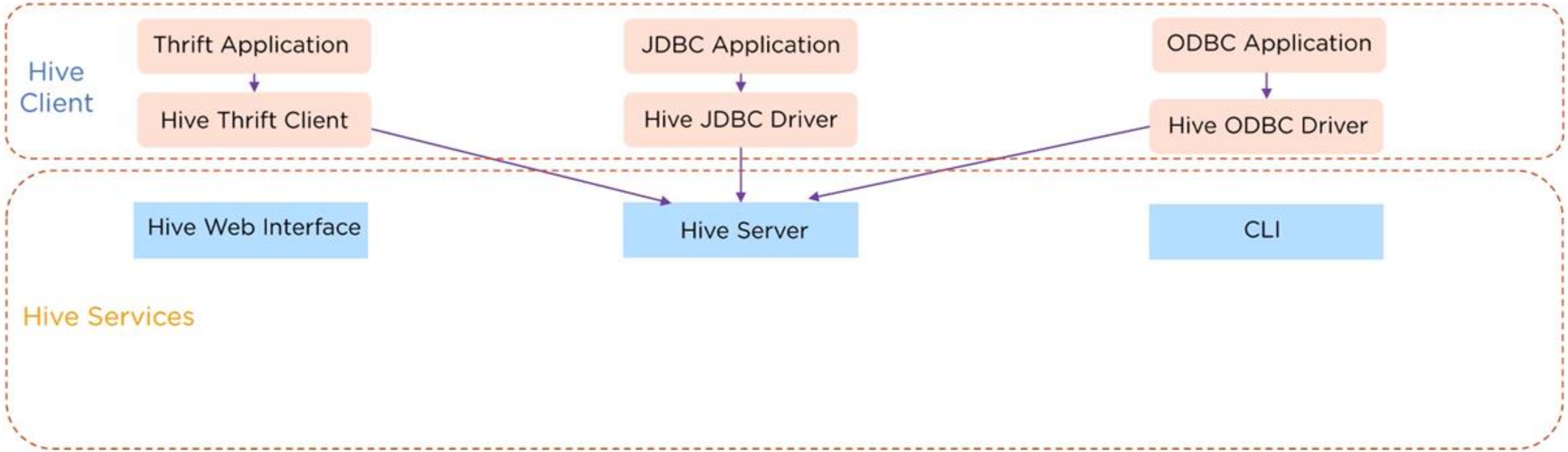
All the client requests are submitted to the **Hive server**

Architecture of Hive



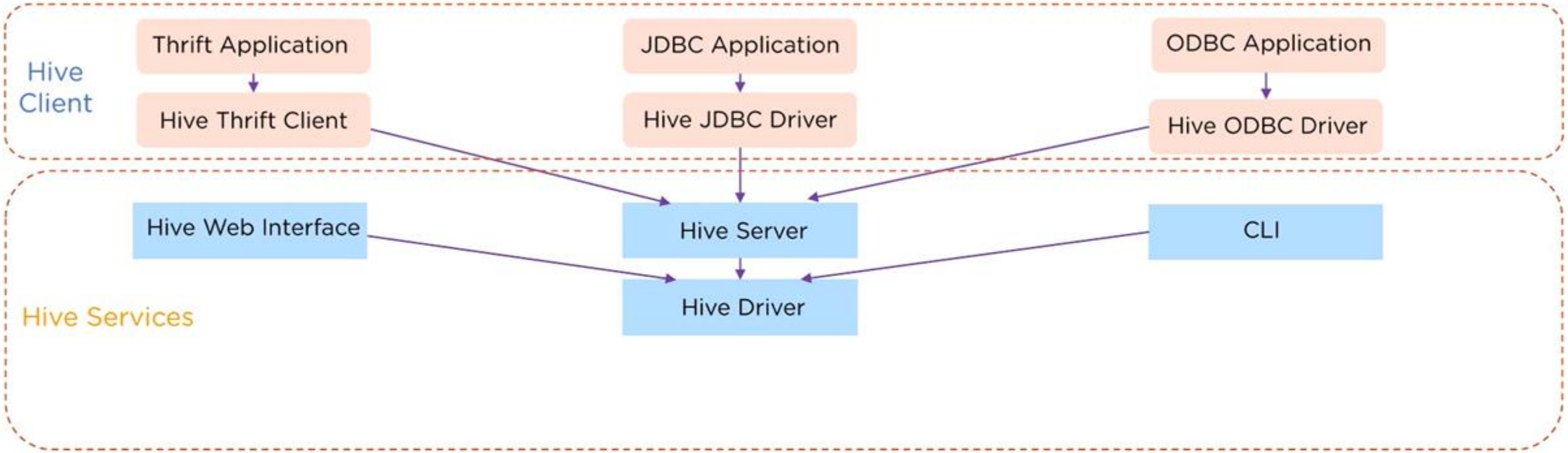
All the client requests are submitted to the **Hive server**

Architecture of Hive



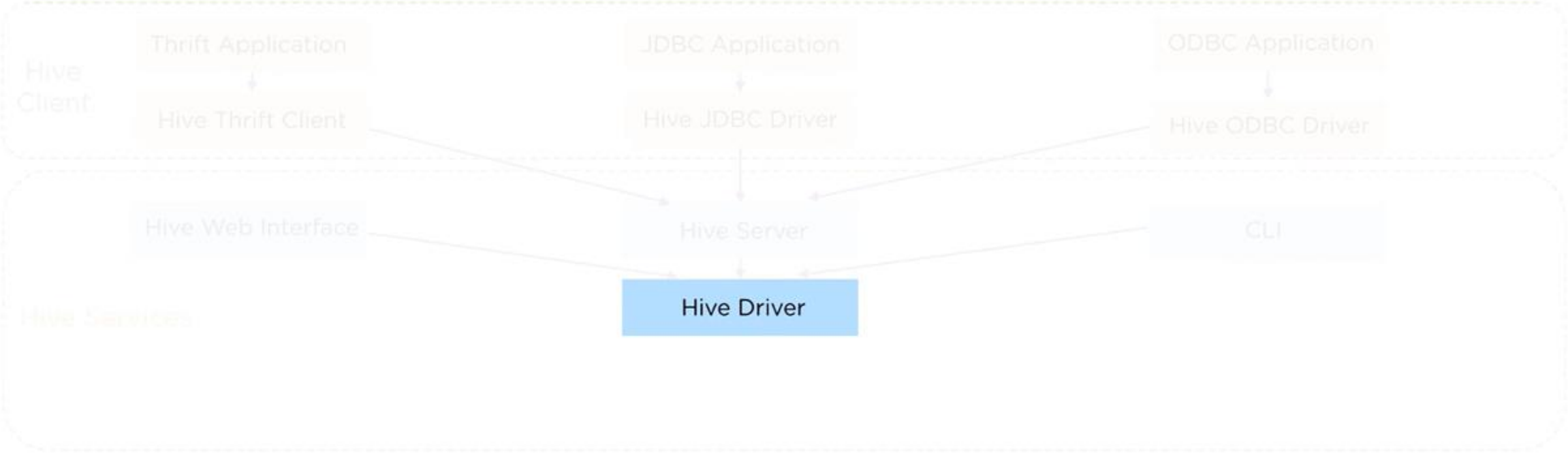
Commands are executed directly in CLI

Architecture of Hive



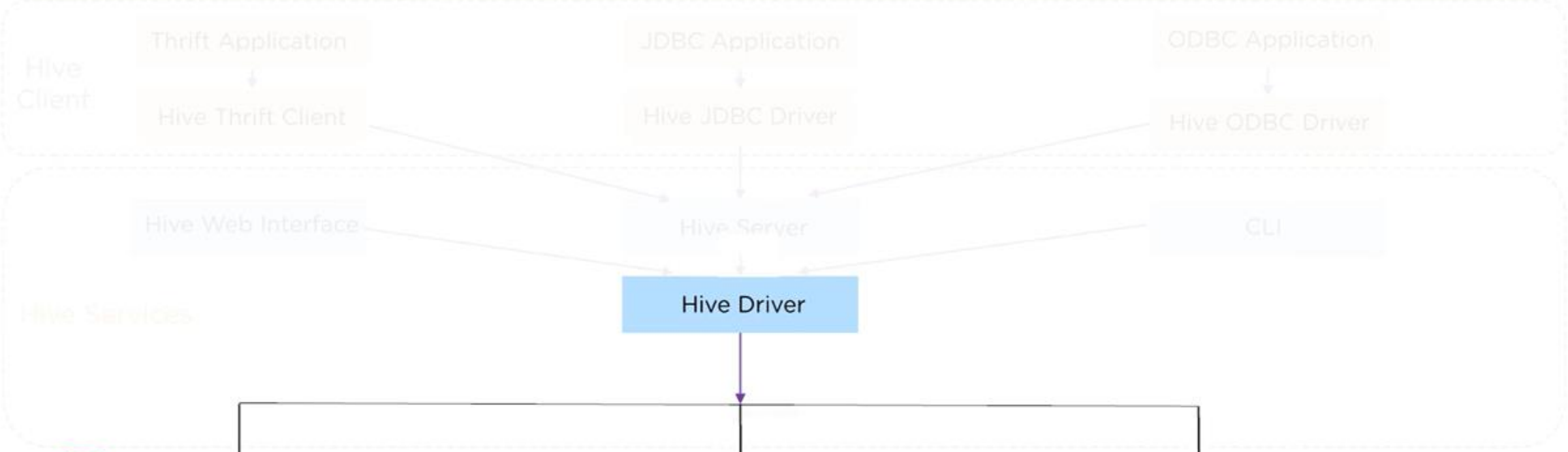
Hive driver is responsible for all the queries submitted

Architecture of Hive



The Hive Driver now performs
3 steps internally

Architecture of Hive

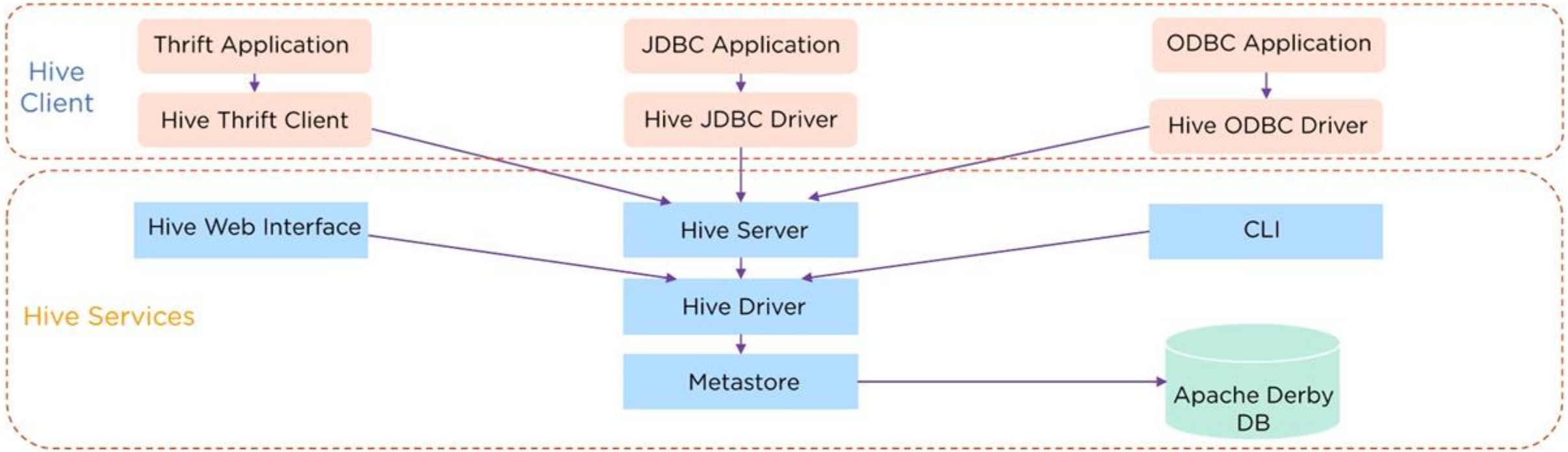


1. Compiler
Hive driver passes query to compiler where it is checked and analyzed

2. Optimizer
Optimized logical plan in the form of a graph of MapReduce and HDFS tasks is obtained

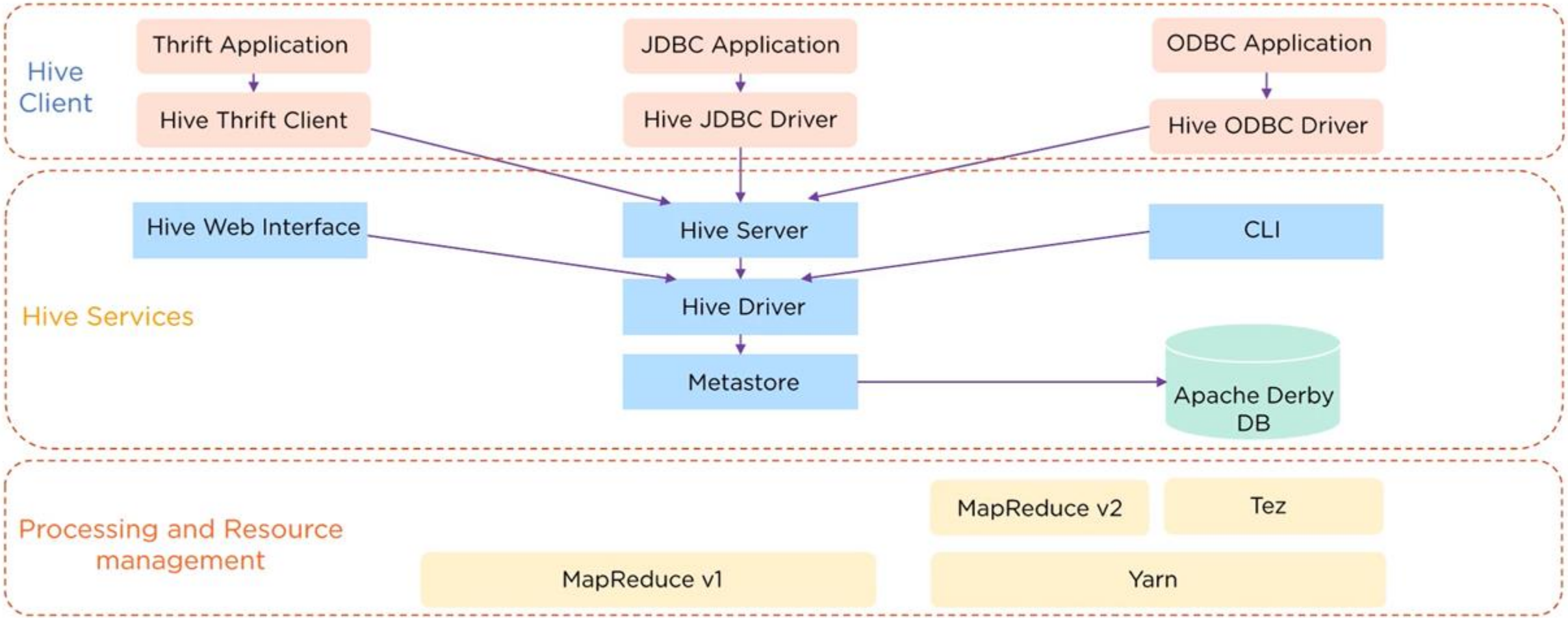
3. Executor
In the final step, the tasks are executed

Architecture of Hive

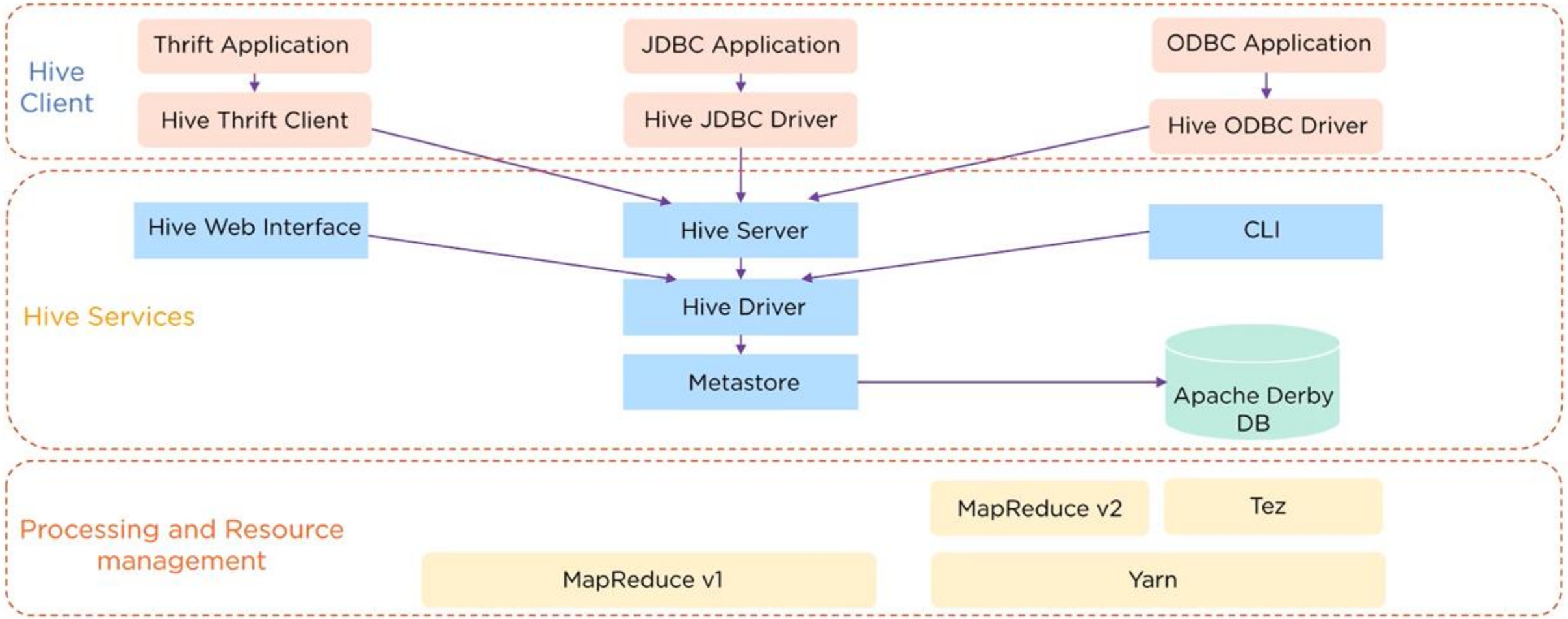


Metastore is a repository for Hive metadata. Stores metadata for Hive tables

Architecture of Hive

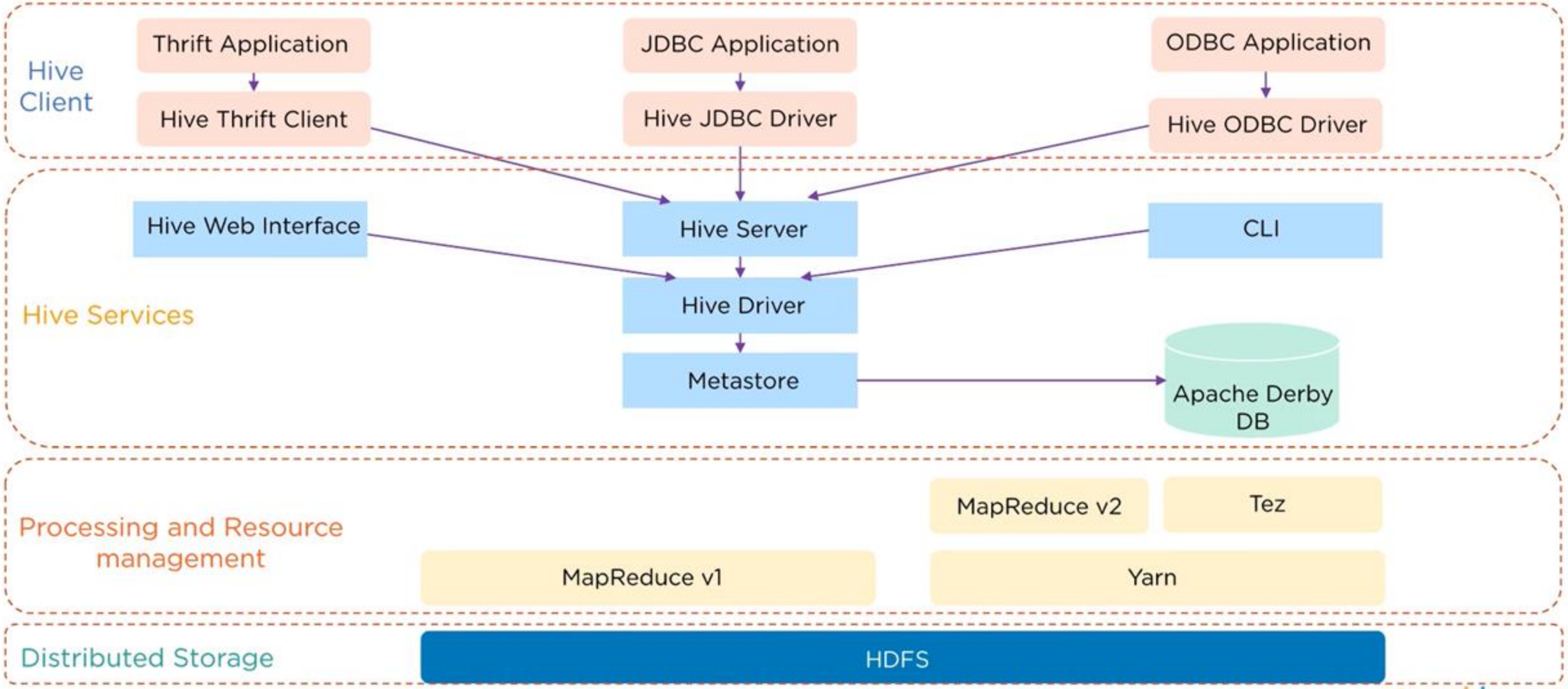


Architecture of Hive

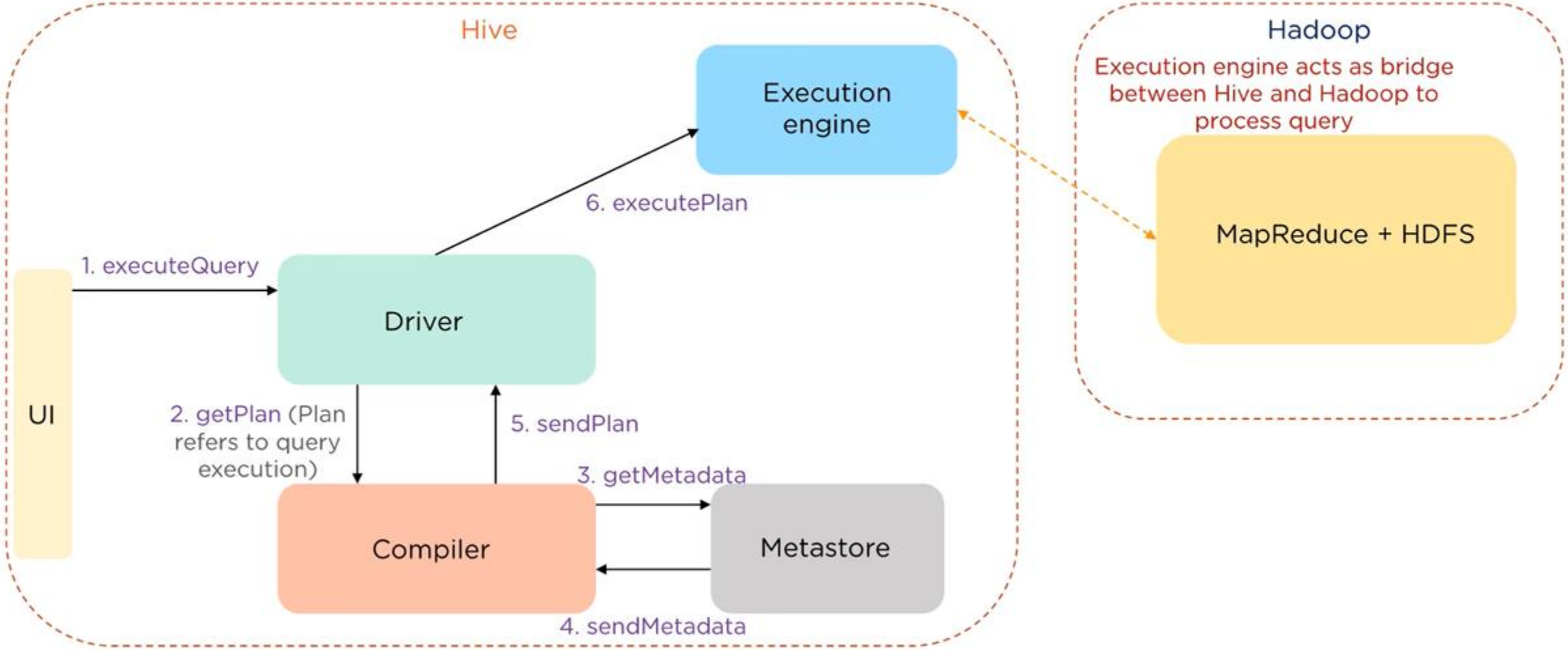


Hive uses MapReduce framework to process queries

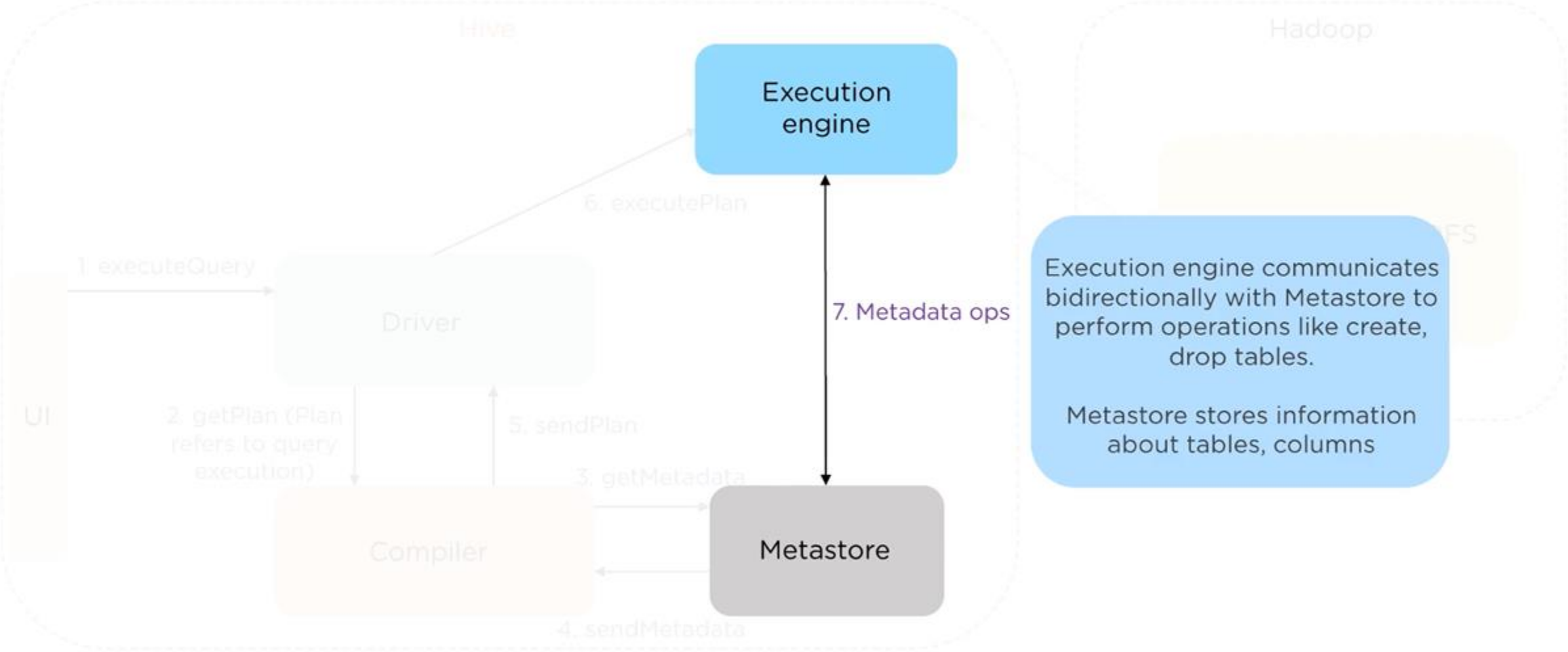
Architecture of Hive



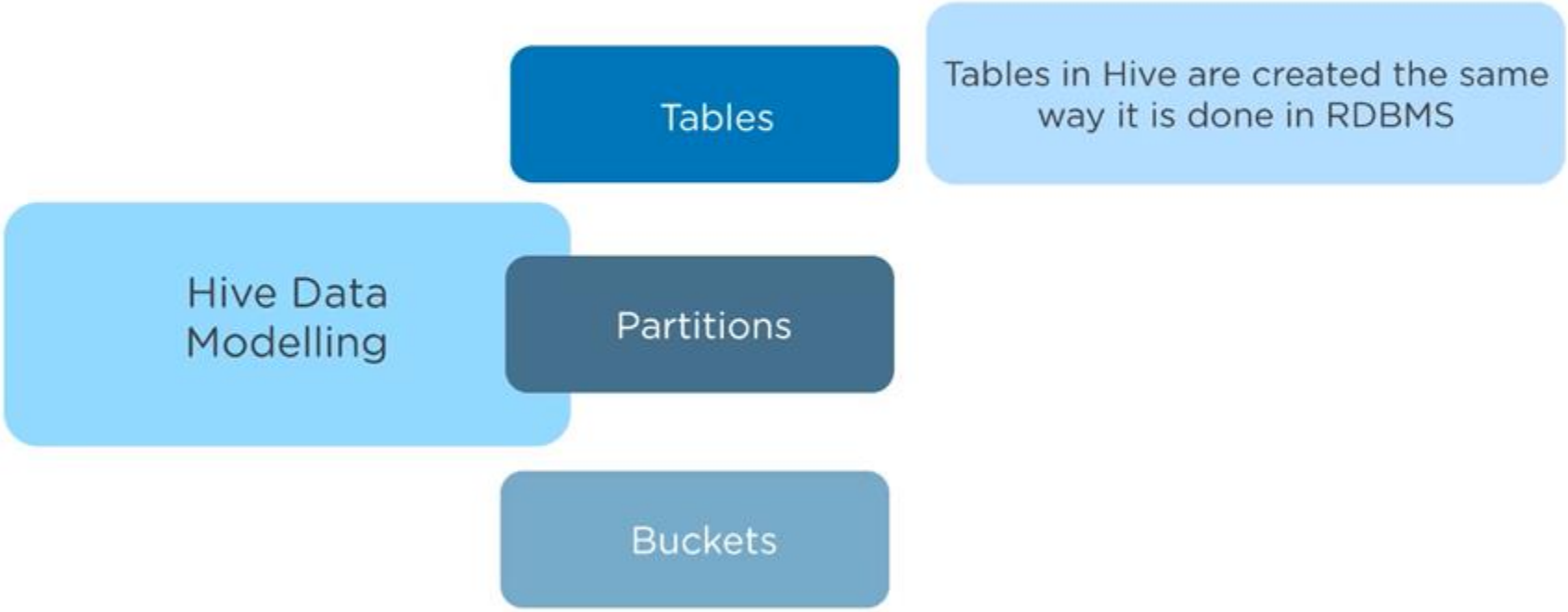
Data flow in Hive



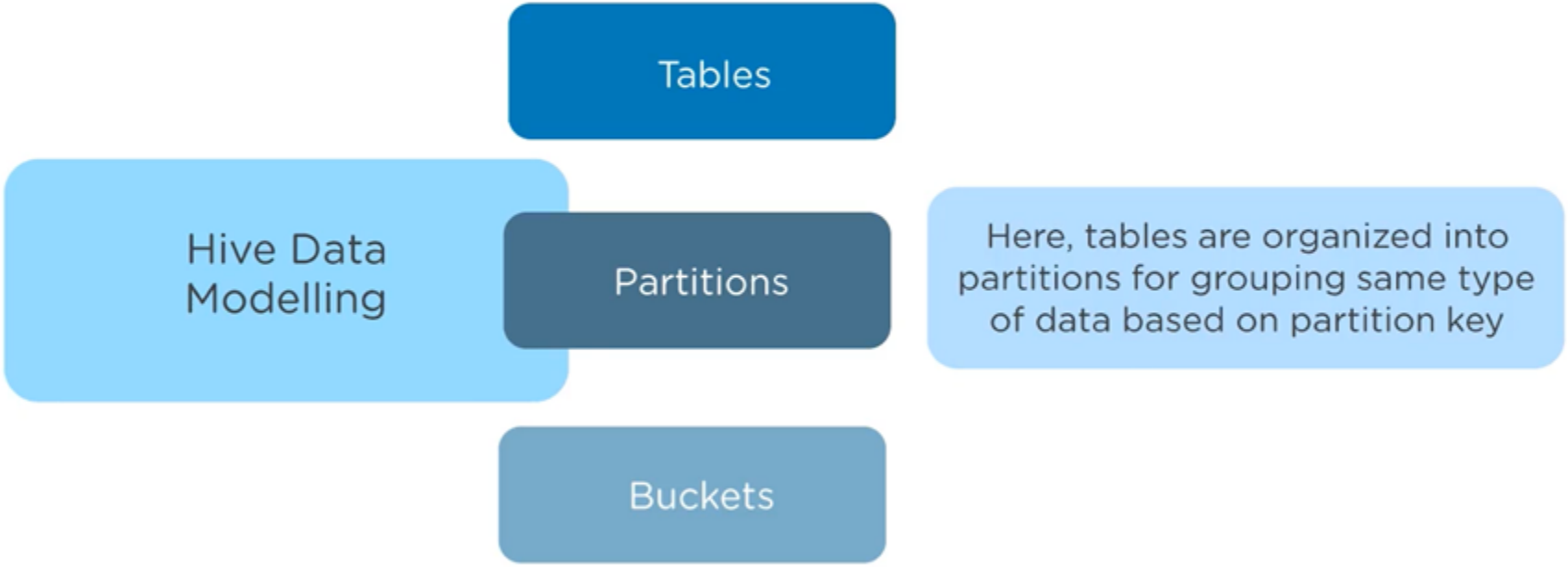
Data flow in Hive



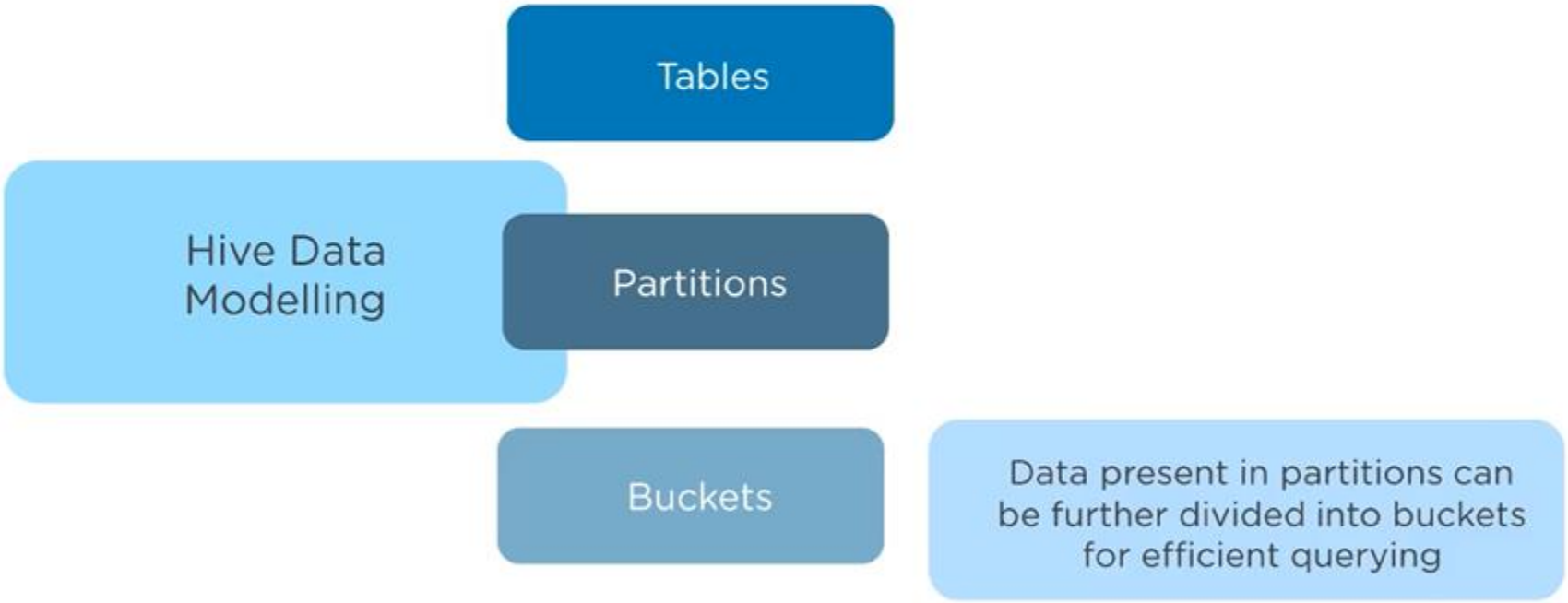
Hive Data Modelling



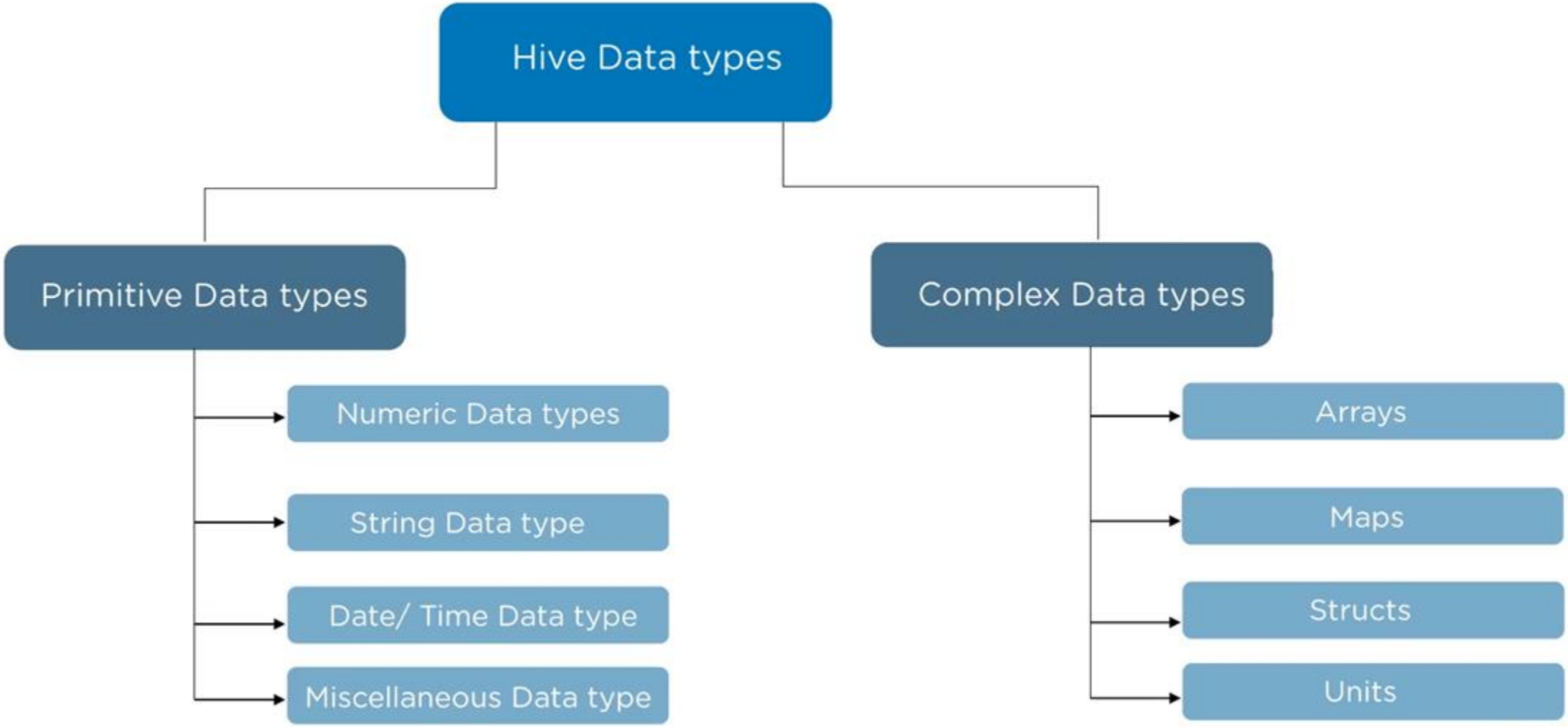
Hive Data Modelling



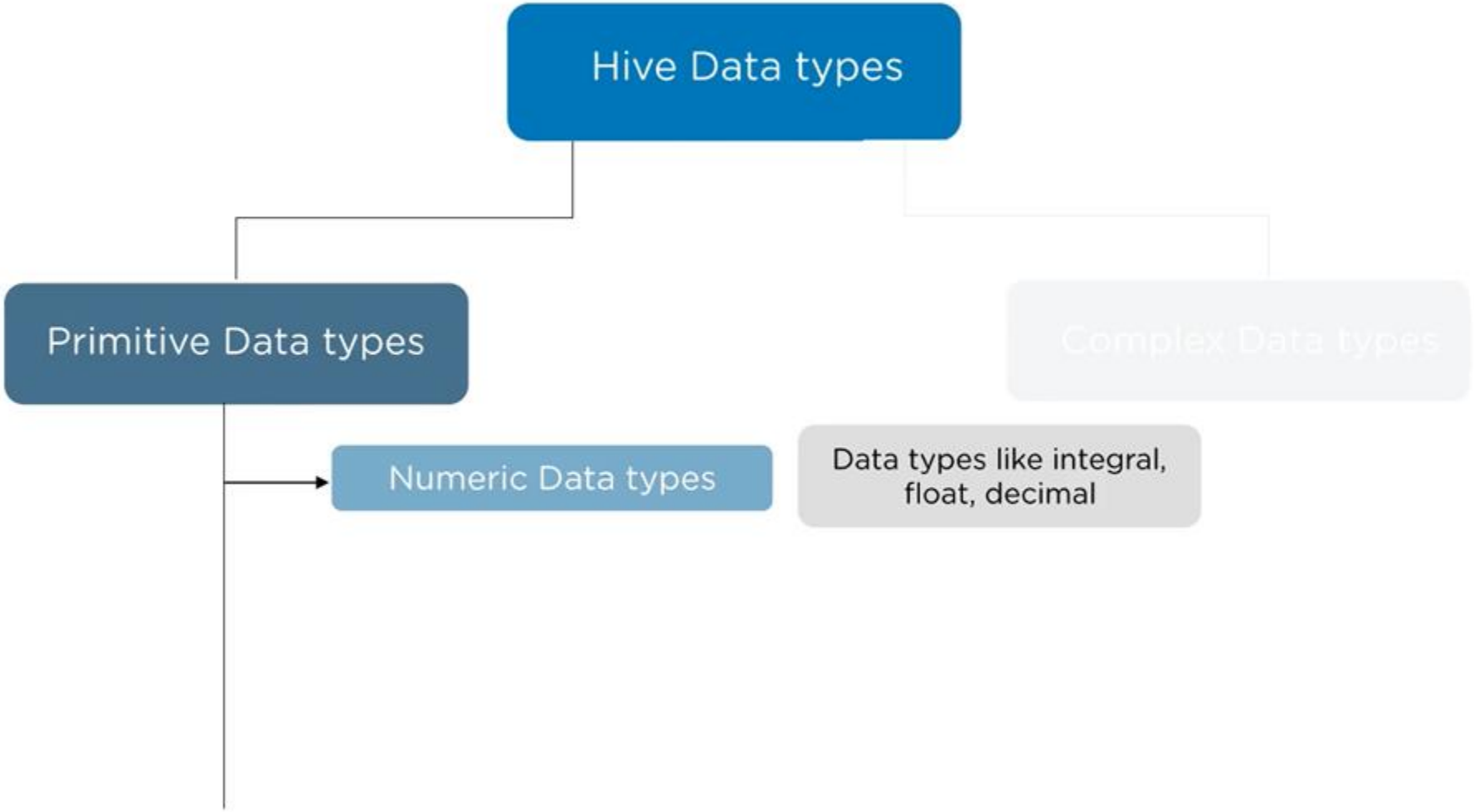
Hive Data Modelling



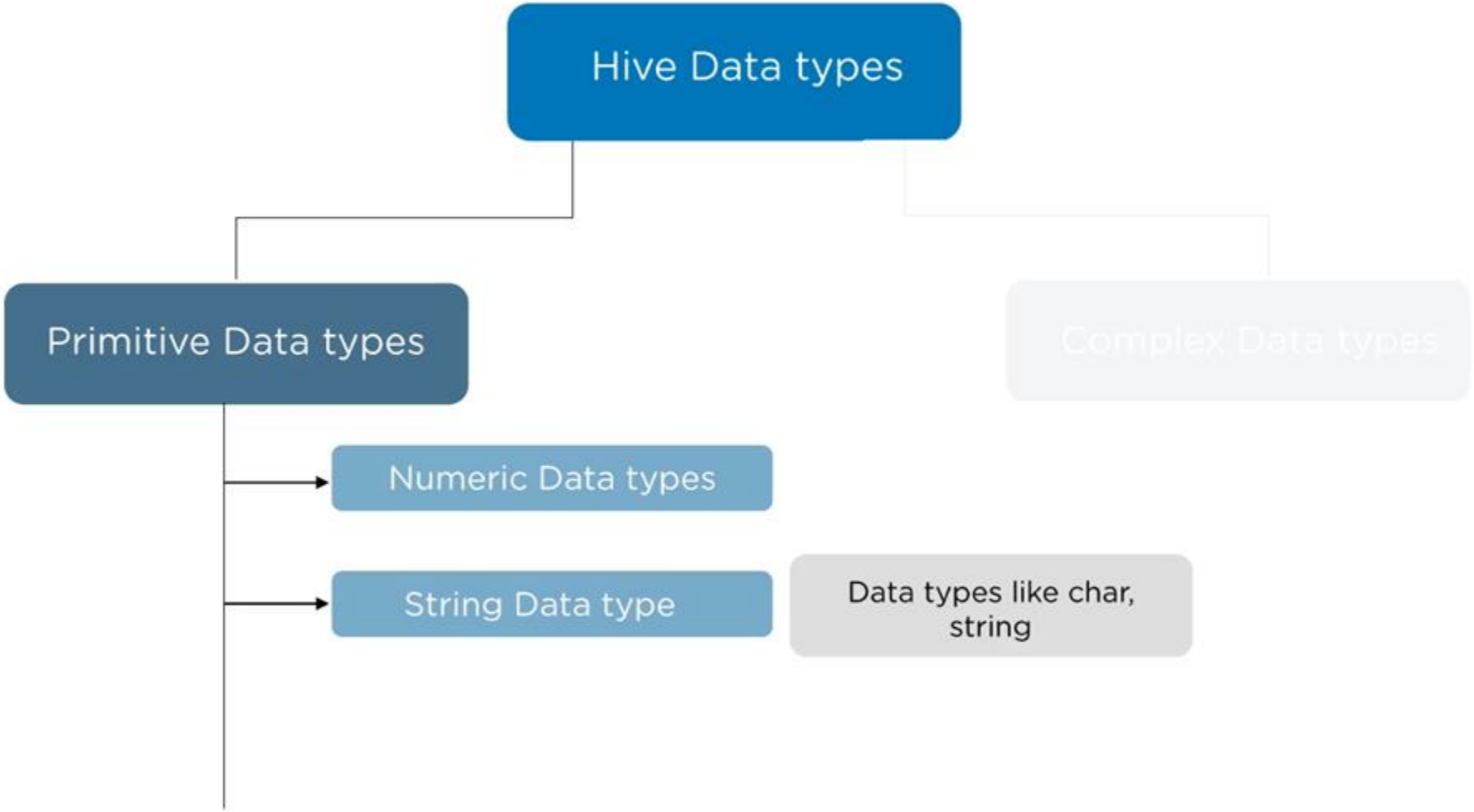
Hive Data types



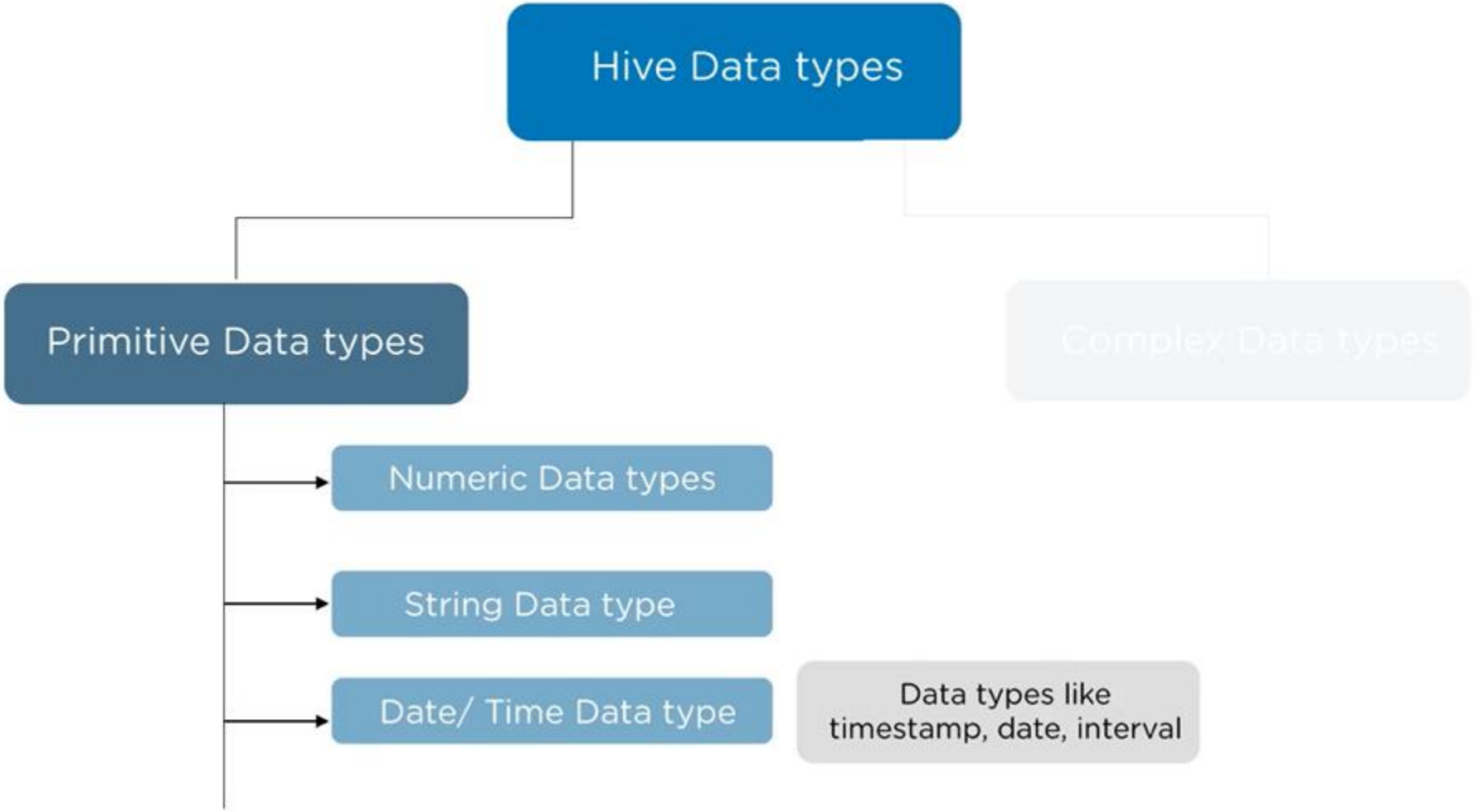
Hive Data types



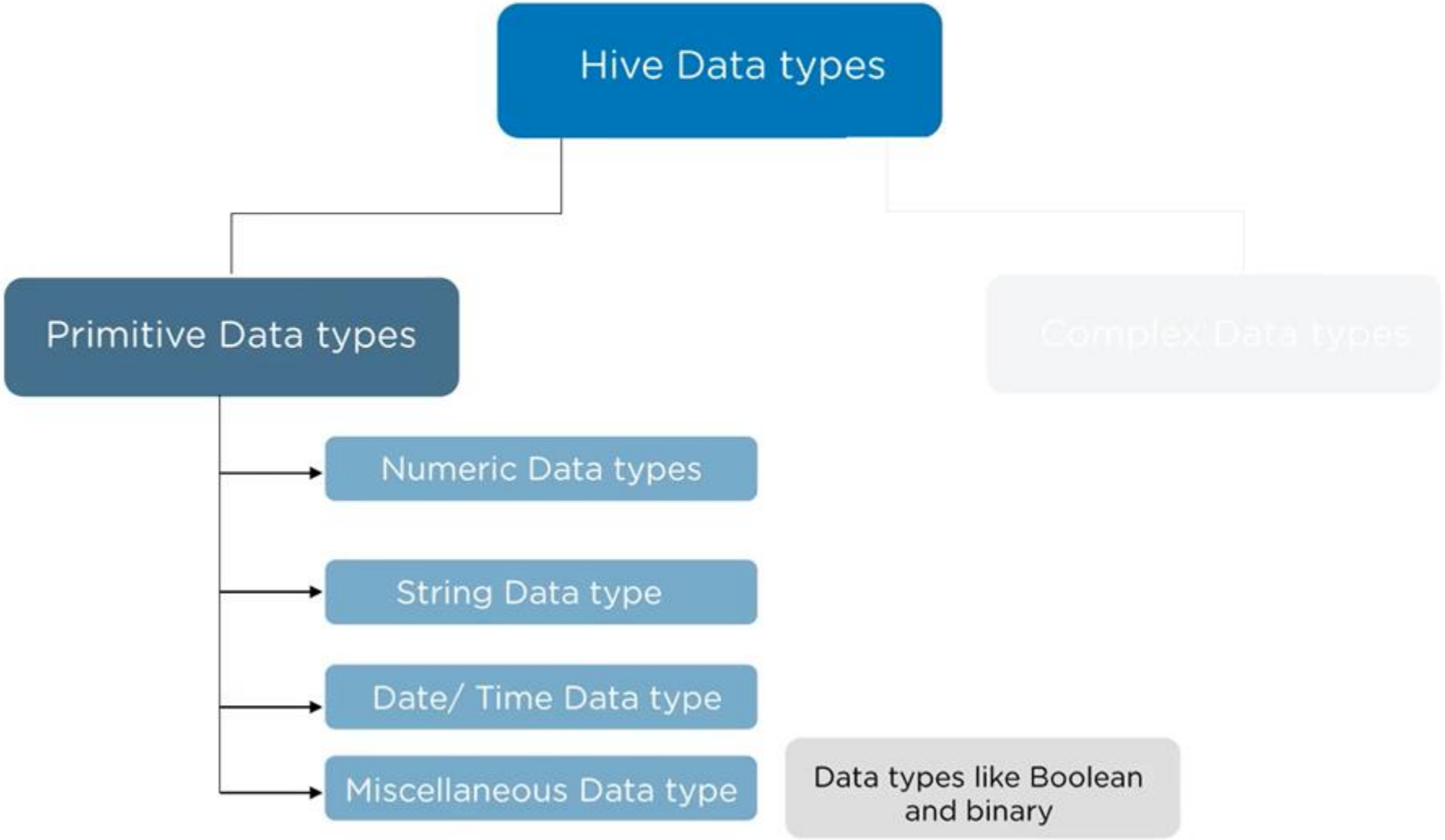
Hive Data types



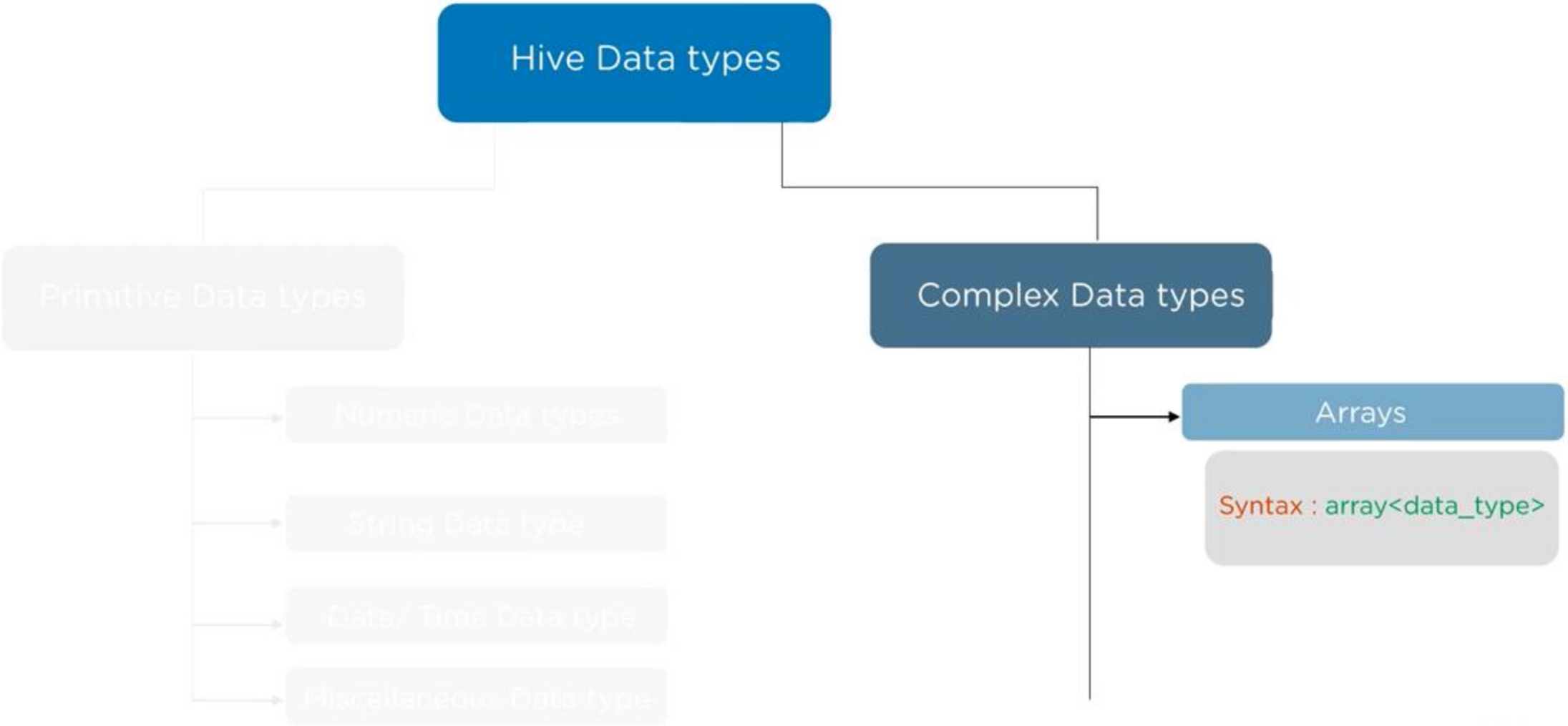
Hive Data types



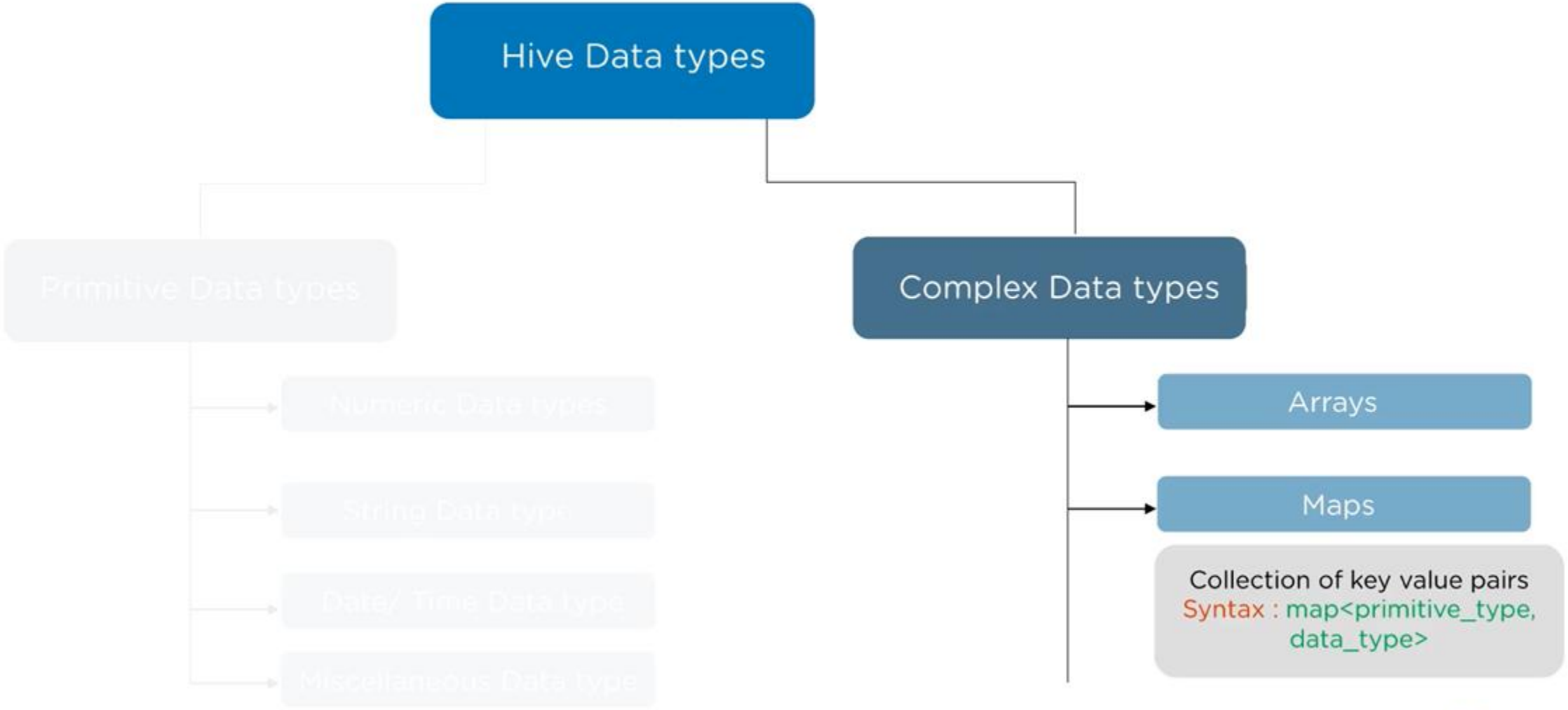
Hive Data types



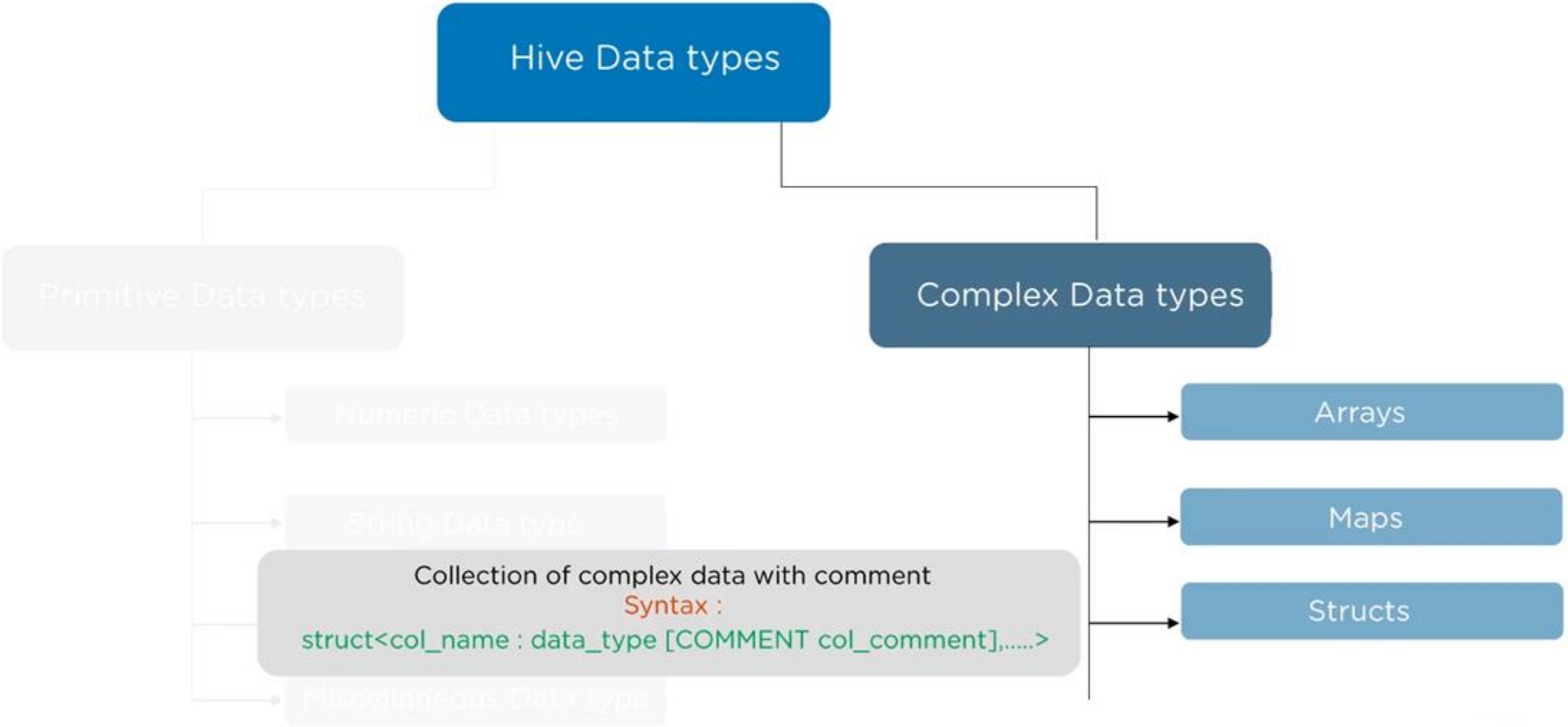
Hive Data types



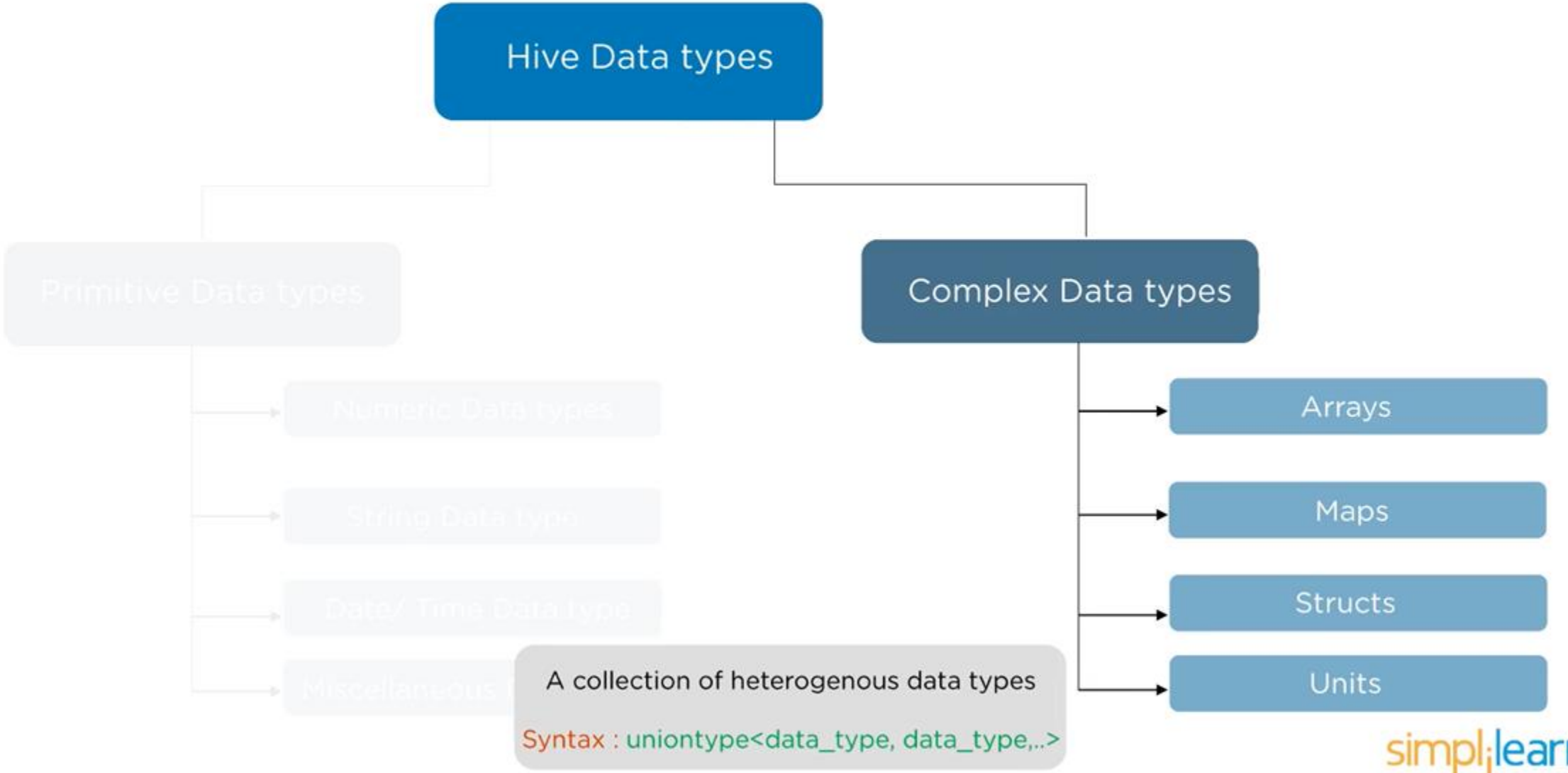
Hive Data types



Hive Data types



Hive Data types



Different modes of Hive

Hive operates in two modes depending on the number and size of data nodes

- Is used when Hadoop is having one data node and the data is small
- Processing will be very fast on smaller datasets which are present in local machine

Local Mode

MapReduce Mode

Different modes of Hive

Hive operates in two modes depending on the number and size of data nodes

Local Mode

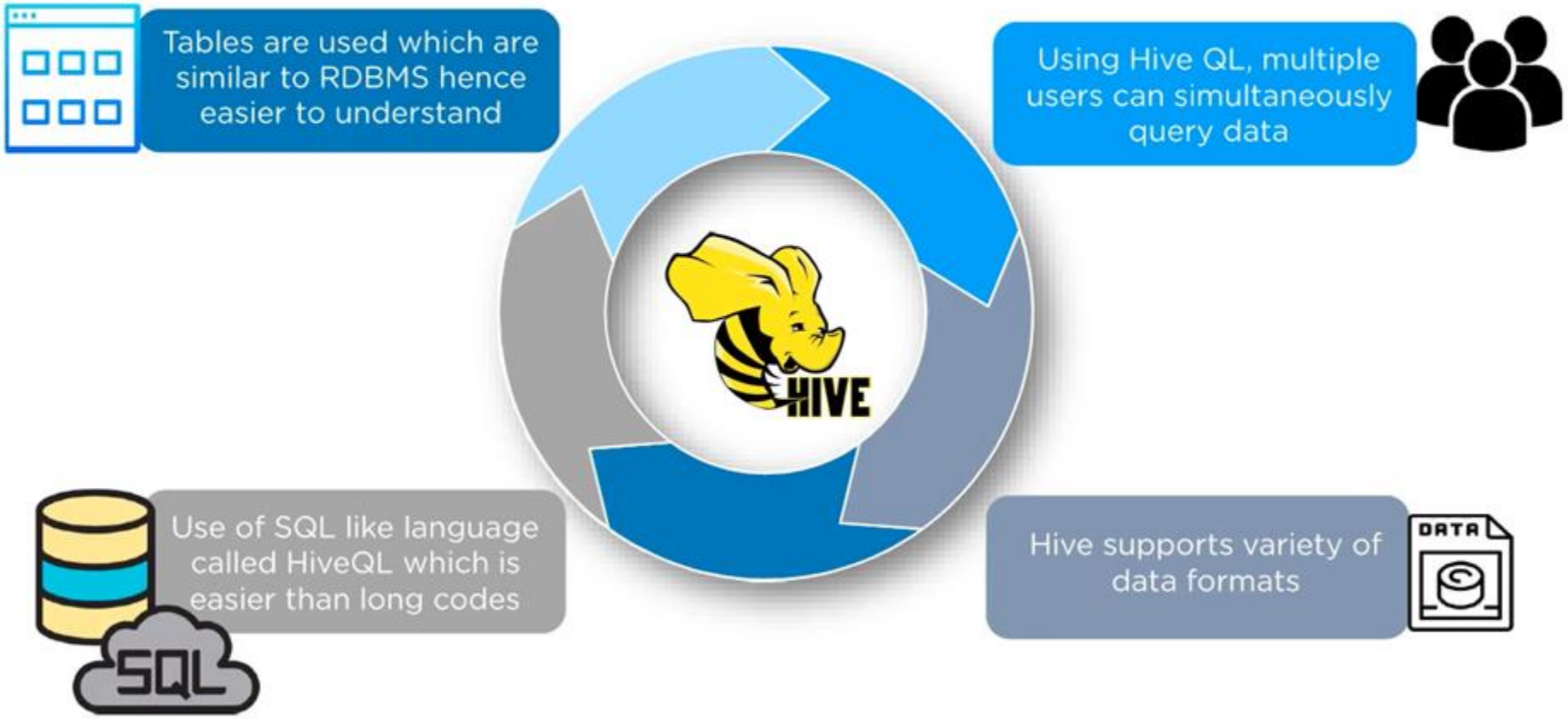
MapReduce Mode

- Is used when Hadoop is having multiple data nodes and the data is spread across various data notes
- Processing large datasets can be more efficient using this mode

Difference between Hive and RDBMS

Hive	RDBMS
Hive enforces schema on read	RDBMS enforces schema on write
Hive data size is in petabytes	Data size is in terabytes
Hive is based on the notion of write once and read many times	RDBMS is based on the notion of read and write many times
Hive resembles a traditional database by supporting SQL but it is not a database. It is a data warehouse	RDBMS is a type of database management system which is based on the relational model of data
Easily scalable at low cost	Not scalable at low cost

Features of Hive

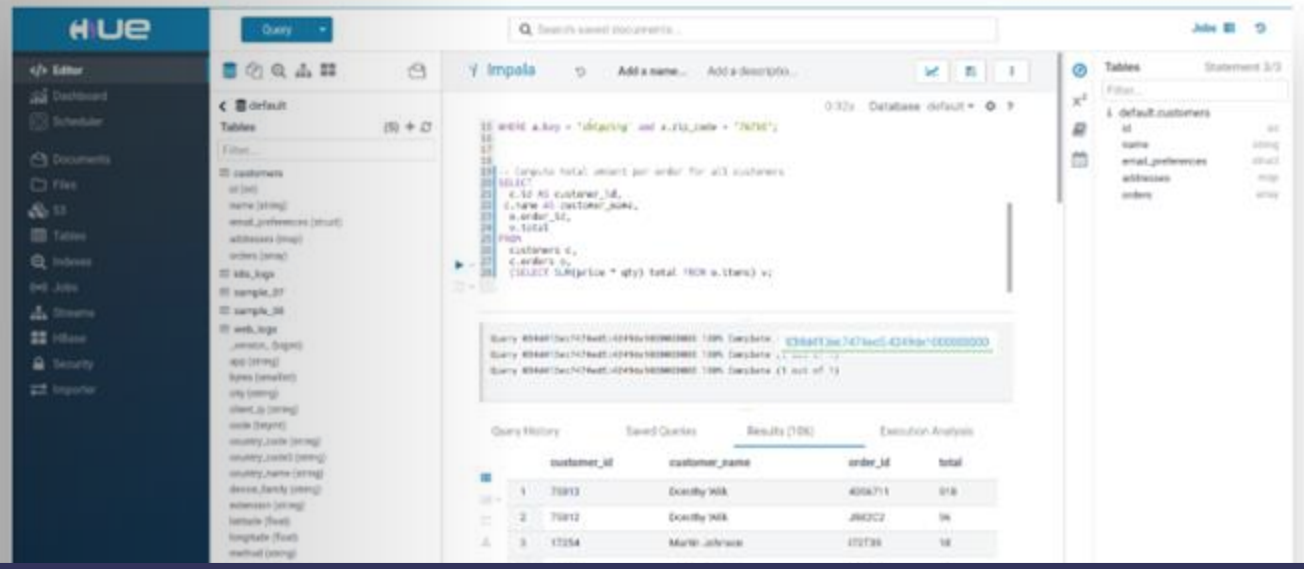


Go to https://gethue.com/

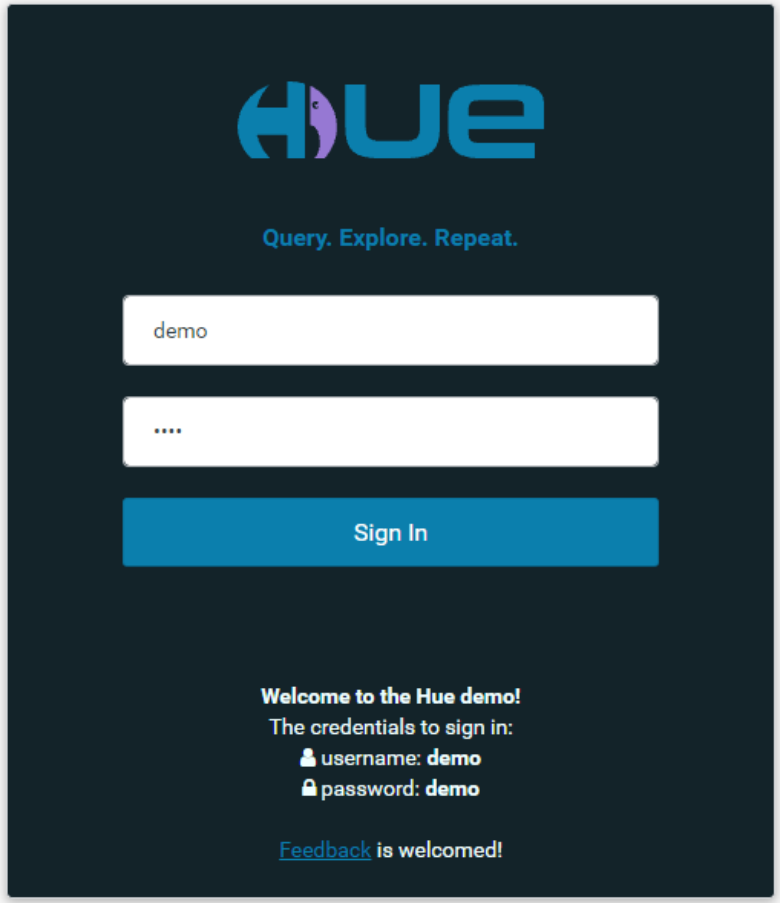
Query. Explore. Share.

Hue is an open source SQL Assistant for Databases & Data Warehouses

TRY HUE NOW



username: **demo**
password: **demo**



Search saved documents...

Editor

- MySQL
- Phoenix SQL
- Trino (Presto SQL)
- Flink SQL
- Hive
- ksqlDB
- Dask Sql
- SparkSQL
- Notebook

MySQL Add a name... Add a description...

No databases found

```
Example: SELECT * FROM tablename, or press CTRL + space
```

Execute 5000 More

Query History Saved Queries Results Chart Execution Analysis

Search... Clear Export import

4 hours ago	!	SHOW databases
23 hours ago	!	SELECT * from t
1 day ago	!	select * from test
1 day ago	!	select -- 维度类字段 account_id,-2 as anchor_uid, client_source,app_channel, platform_type,platform,terminal,os_name,client_type,live_mode, --'{}' as entrance, -- entrance修复后切换回来 '-2' as entrance, -- 回滚数据解决entrance问题 使用entrance_key '-2' as template,-2 as game_type,-2 as game_type_source,'-2' as live_type, '-2' as category_attr,-2 as room_id,-2 as channel_id,'-2' as channel_type, is_login, if(is_first_login = 1 or is_first_login_nofilter = 1, 1, 0) as is_first_login, is_first_login_pc, is_first_login_mob, is_first_login_web, -- last类字段 account_uid,account_ccid,account_name, -- 使用account_id维度聚合取last account_sex,account_level,vip_level,noble_level, app_ver,sdk_ver,os_ver, '-2' as anchor_id,-2 as anchor_ccid,'-2' as anchor_name, -- 使用anchor_uid维度聚合取last -2 as anchor_sex,-2 as anchor_level,'-2' as anchor_grade,'-2' as anchor_belong,'-2' as anchor_talent_tag, -2 as is_effective_sign,'-2' as fir...
1 day ago	!	CREATE DATABASE mynameiskhan;
1 day ago	!	select * from aaa

HiveQL

he Hive Query Language (HiveQL) is a query language for Hive to process and analyze structured data in a Metastore.

Hive Create Database Syntax

In Hive, CREATE DATABASE statement is used to create a Database, this takes an optional clause IF NOT EXISTS, using this option, it creates only when database not already exists.

```
CREATE DATABASE [IF NOT EXISTS] <database_name>
```

Note: Creating a database with already existing name in a database returns an error.

Hive Create Table Syntax

By using CREATE TABLE statement you can create a table in Hive, It is similar to SQL and CREATE TABLE statement takes multiple optional clauses,

```
CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.] table_name
[(col_name data_type [column_constraint] [COMMENT col_comment], ...)]
[PARTITIONED BY (col_name data_type [COMMENT 'col_comment'], ...)]
[CLUSTERED BY (col_name, col_name,.......]
[COMMENT table_comment]
[ROW FORMAT row_format]
[FIELDS TERMINATED BY char]
[LINES TERMINATED BY char]
[LOCATION 'hdfs_path']
[STORED AS file_format]
```

In Hive, table can be created with or without the database, If you wanted to create in a database, specify database name qualifier.

DROP TABLE Syntax

```
DROP TABLE [IF EXISTS] table_name [PURGE];
```

DATABSE and SCHEMA can be used interchangeably in Hive as both refer to the same.

DROP DATABASE Syntax

```
DROP DATABASE [IF EXISTS] database_name [RESTRICT|CASCADE];
```

Hive DROP DATABASE consists of several optional clauses, using these we can change the behavior of the Hive statements.

- IF EXISTS – Use *IF EXISTS* to check if the database exists before running a drop database statement.
- RESTRICT – The default behavior is *RESTRICT*, where *DROP DATABASE* will fail if the database is not empty
- CASCADE – Use *CASCADE* option, if you wanted to drop all tables before dropping the database.

INSERT INTO Syntax

The **Hive INSERT INTO** syntax will be as follows.

```
INSERT INTO TABLE tablename1
[PARTITION (partcol1=val1, partcol2=val2 ...)]
select_statement1 FROM from_statement;
```

INSERT OVERWRITE Syntax

The **Hive INSERT OVERWRITE** syntax will be as follows.

```
INSERT OVERWRITE TABLE tablename1
[PARTITION (partcol1=val1, partcol2=val2 ...)
[IF NOT EXISTS]]
select_statement1 FROM from_statement;
```