Hive

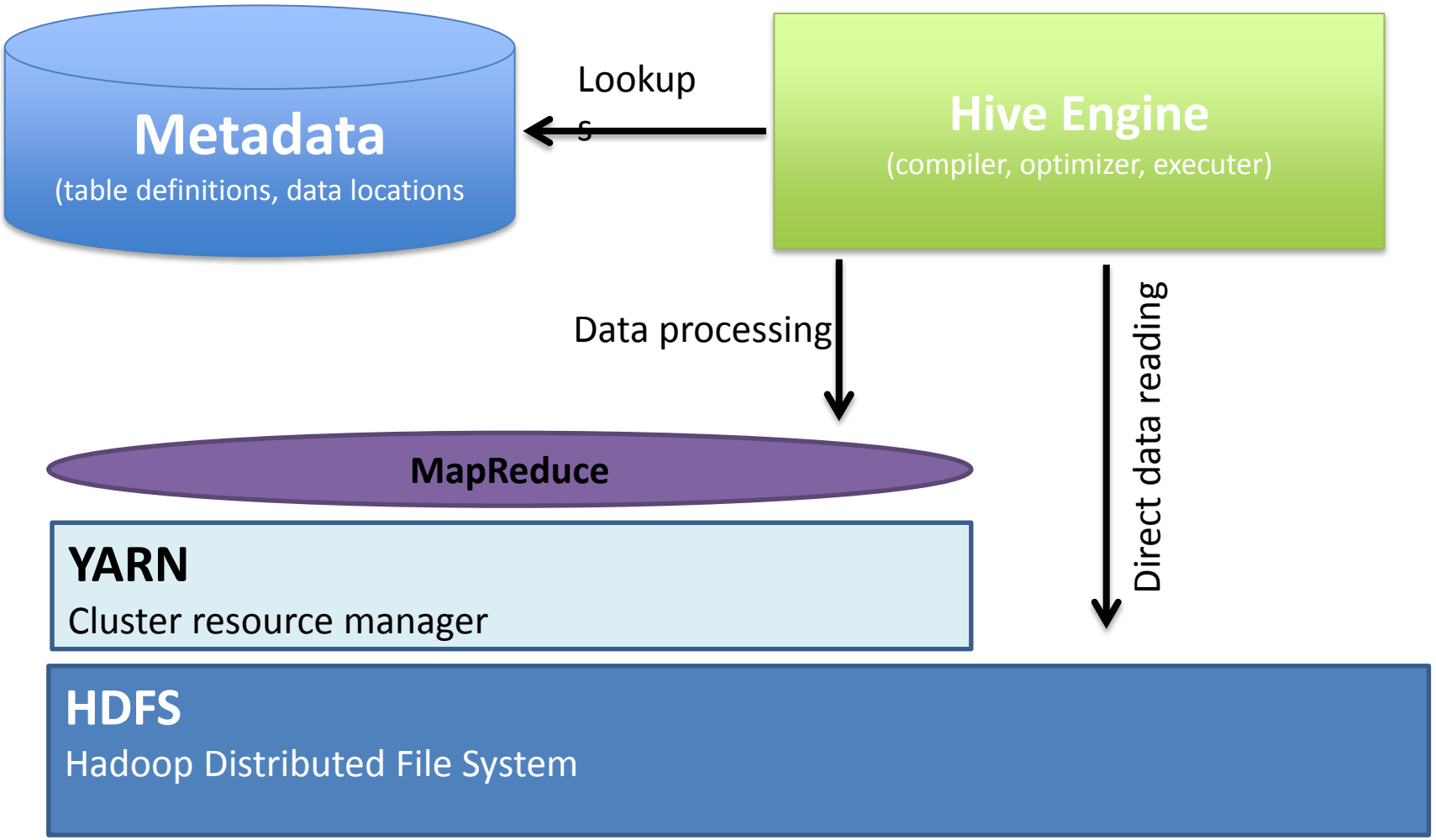
What is Hive?

- Data warehousing layer on top of Hadoop
  - table abstractions
- SQL-like language (HiveQL) for “batch” data processing
- SQL is translated into one or series of MapReduce executions
- Good for ad-hoc reporting queries on HDFS data
  - however generated MR executions can be sub optimal

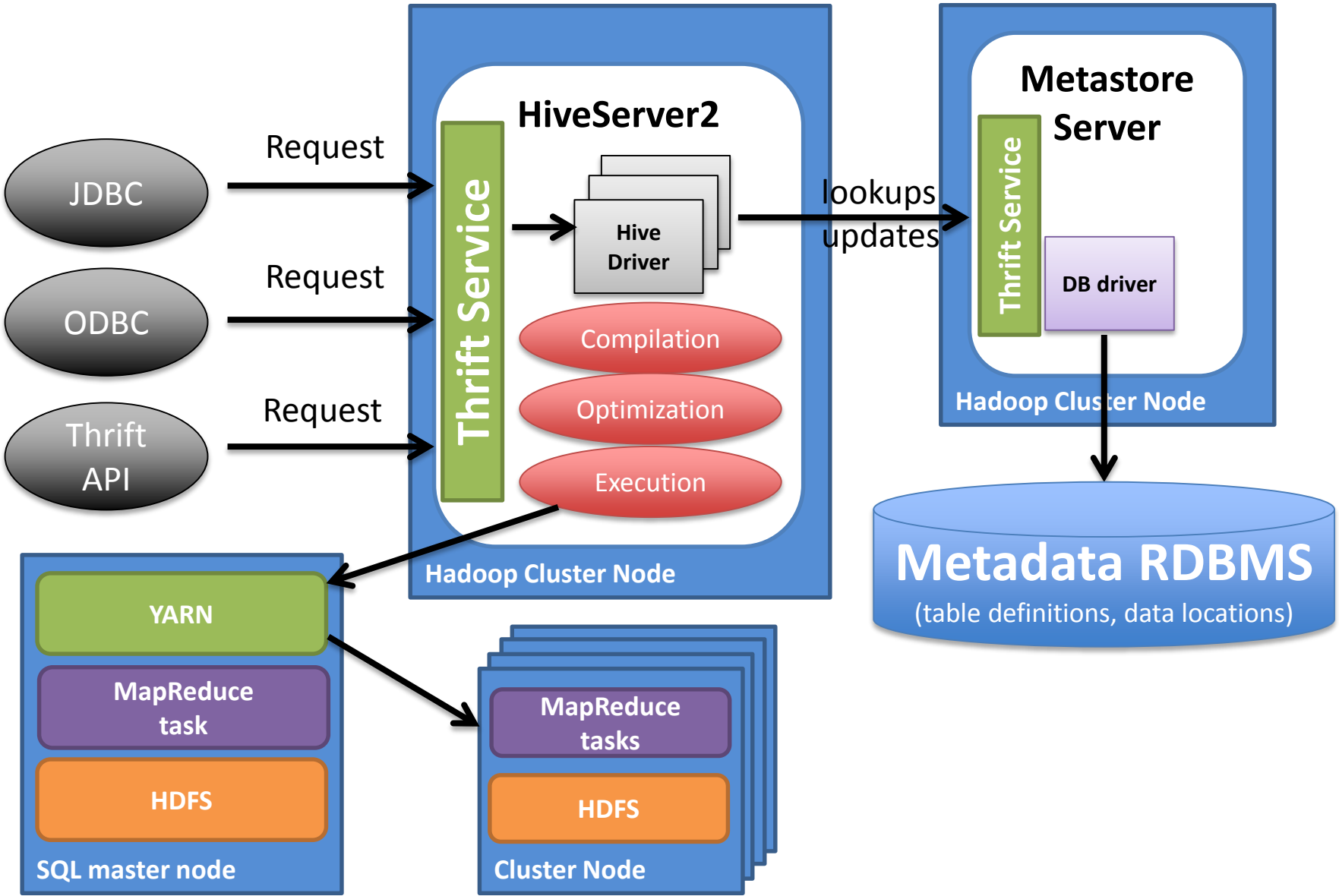
What is not…

- Not a relational database
  - not transactions
  - no row updates
  - no indexes
  - no constraints
- No interactive querying
- No DBMS

Hive overview



Hive Architecture



Metastore

- Contains tables definitions and data location on HDFS
- Stored in additional RDBMS
  - very often on one of a cluster machines
  - MySQL, PostgreSQL, Oracle, Derby…
- Used by many other Hadoop components
  - via HCatalog service

Hive Table

- Table
  - Definitions stored in a Hive metastore (RDBMS)
  - Data are stored in files on HDFS
- Data files can be in various formats
  - but a type has to by unique within a single table
- Table partitioning and bucketing is supported
- EXTERNAL tables (DMLs are not possible)
- Table statistics can be gathered – important for getting optimal execution plans

Interacting with Hive

- Remotely
  - JDBC and ODBC drivers
  - Thrift API
  - beeline (via JDBC, HiveServer2 support)
- Locally
  - hive-shell (deprecated)
  - beeline

Operations

- Based on SQL-92 specification
- DDL
  - `CREATE TABLE, ALTER TABLE, DROP TABLE….`
- DML
  - `INSERT, INSERT OVERWRITE…`
- SQL
  - `SELECT…`
    - `DISTINCT…JOIN WHERE…GROUP BY…HAVING…ORDER BY…LIMIT`
  - REGEXP supported
  - Subqueries only in the FROM cluase

Data Types

- TINYINT – 1 byte
- BOOLEAN
- SMALLINT – 2 bytes
- INT – 4 bytes
- BIGINT – 8 bytes
- DOUBLE
- STRING
- STRUCT – named fields
- MAP – key-value pairs collection
- ARRAY – order collection of records in the same type

Other features

- Views
- Build in functions
  - floor, rand, cast, case, if, concat, substr etc
  - ‘show functions’
- User defined functions
  - have to be delivered in jar

Using Hive CLIs

- Starting hive shell (deprecated)

```
> hive
```

- Use beeline instead (supports new HiveServer2)

```
>beeline
```

  - Connection in remote mode

```
!connect jdbc:hive2://localhost:10000/default
```

  - Connection in embedded mode

```
!connect jdbc:hive2://
```

Useful Hive commands

- Get all databases

```
show databases
```

- Set a default database

```
use <db_name>
```

- Show tables in a database

```
show tables
```

- Show table definition

```
desc <table_name>
```

- Explaining plan

```
EXPLAIN [EXTENDED|DEPENDENCY|AUTHORIZATION]  <query>
```

Hands on Hive (1)

- All scripts are available with:

```
mkdir hive; cd hive
wget https://cern.ch/zbaranow/hive.zip
unzip hive.zip
hdfs dfs –put ~/tutorials/data data
```

- **To execute a script in beeline**

```
!run <script_name>
```

- Creation of an external table from existing data (name=geneva)
  - external.sql
- Creation of a external table without “ (name=geneva_clean)
  - external_clean.sql
- Creation of a local table ‘as select’ from external (name=weather)
  - standard.sql
- Querying the data
  - queries.sql

Hands-on Hive (2)

- Explain plan
  - explain.sql
- Table statistics
  - stats.sql
- Creation of a partitioned table (name=weather_part)
  - partitioned.sql
- Creation of partitioned and bucketed table (name=weather_part_buck)
  - bucketing.sql
- Creation of a table stored in a parquet format (name=weather_parquet)
  - parquet.sql
- Creation of a compressed table (name=weather_compr)
  - compressed.sql

Hands on Hive (JDBC)

- Compile the code

```
javac HiveJdbcClient.java
```

- Run
  - set classpath

```
source ./setHiveEnv.sh
```

  - Execute

```
java -cp $CLASSPATH HiveJdbcClient
```

This talk does not cover…

- Views
  - Object representing a sql statement
- SerDe
  - Serializer and Deserializer of data
  - There are predefined for common data formats
  - Custom Ser/De can be written by a user
- Writing UDF
- Querying Hbase

Summary

- Provides table abstraction layer on HDFS data
- SQL on Hadoop translated to MapReduce jobs
- The Hive Query Language has some limitations
- For batch processing, not interactive
- Can append data
  - But not row updates or deletions