

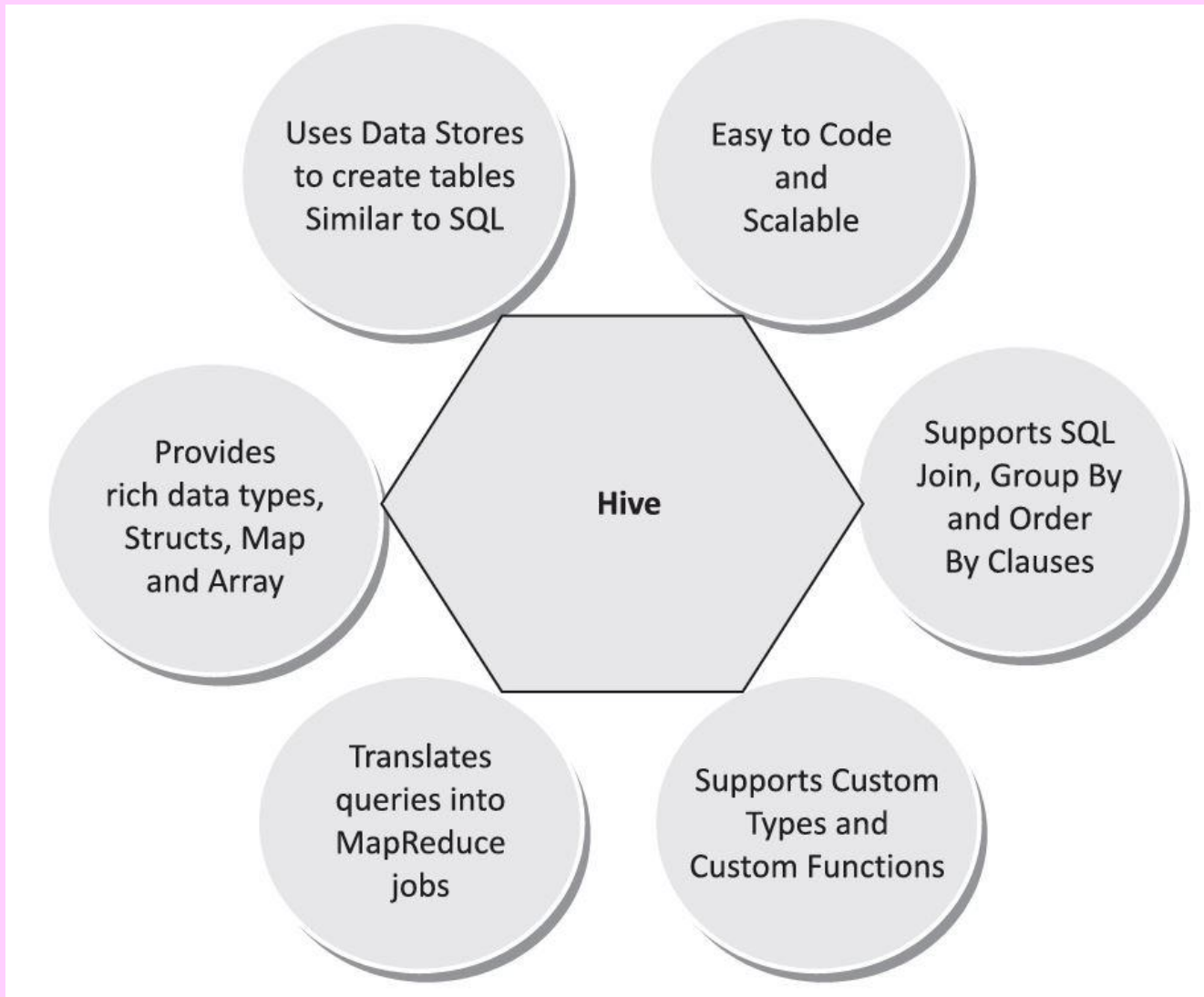
Lesson 5

Hive Characteristics and Functions

Hive

- A data-store, data warehouse infrastructure
- Provides data summarization
- Creates Tables, Files, Databases
- Ad hoc querying for data
- Also a processing tool on the top of the Hadoop

Figure 4.9 Main features of Hive



Hive Characteristics

- Capability to translate the queries into MapReduce jobs, making Hive scalable

Hive Characteristics

- Handles data warehouse applications, therefore, suitable for the analysis of static data of extremely large size, where fast response time is not a criterion

Hive Characteristics

- Supports web interface as well, which means application API as well as web browser client that can access the Hive DB serve

Hive Characteristics

- Provides an SQL dialect, Hive Query Language, (abbreviated HiveQL or HQL)

Hive Characteristics

- Results of HiveQL Query and the data load in the tables which store at Hadoop cluster

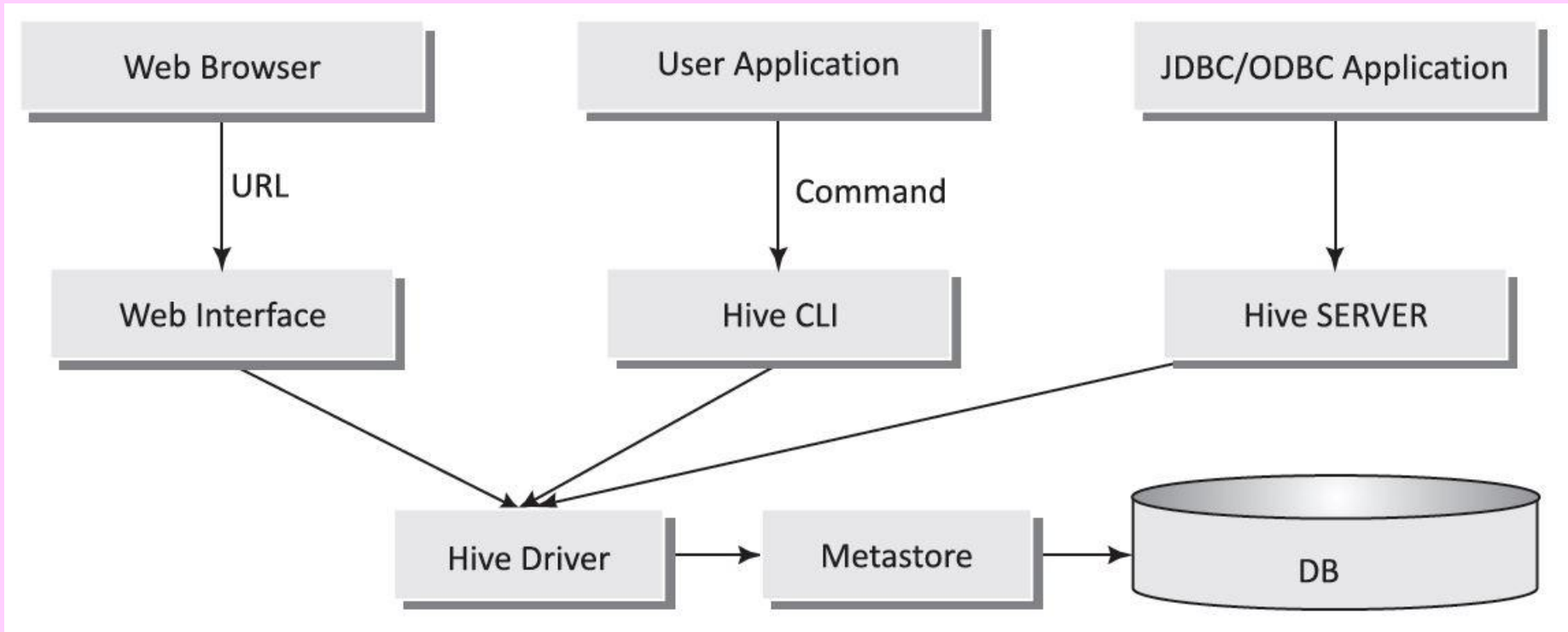
Hive Limitations

- Not a full database
- Does not provide update, alter and deletion of records in the database
- Not developed for unstructured data,
- Not designed for real time queries, and

Hive Limitations

- Performs the partition, always from last column

Figure 4.10 Hive architecture



Hive Server (Thrift)

- An optional service
- Remote client submits requests to Hive
- Retrieves results
- Thrift Server exposes a very simple client API to execute HiveQL statements.

Client Requests to Hive Server (Thrift)

- Requests can be in a variety of programming languages

Hive CLI (Command Line Interface)

- A popular interface
- Interact with Hive
- Hive runs in local mode that uses local storage when running the CLI on a Hadoop cluster instead of HDFS

Hive Web Interface (HWI)

- Hive can be accessed using a web browser as well.
- A HWI Server running on some designated code

Metastore

- The system catalog
- Stores the schema or metadata of tables, databases, columns in a table, their data types, and HDFS mapping

Hive Driver

- Manages the lifecycle of a HiveQL statement during compilation, optimization and execution

Hive Installation

- Windows 10, Ubuntu 16.04 and MySQL

Hive Installation

- Compatible version of Hive with Java— Hive 1.2 onward supports Java 1.7 or newer
- JDK for Java compiler(Javac) and interpreter
- Hadoop

Steps for installation

- Refer text Section 4.4.2 for Hive in a Linux based OS

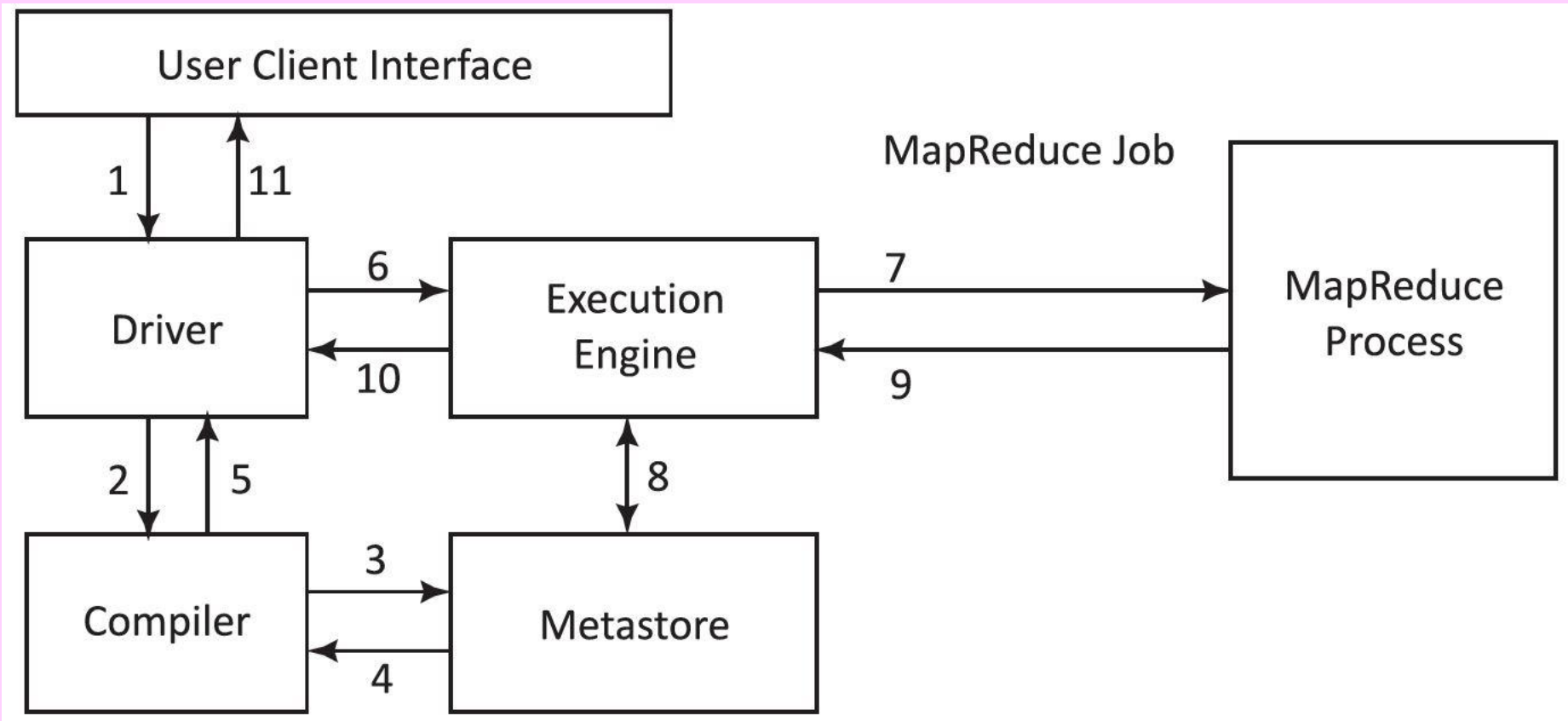
Comparison with RDBMS (Traditional Database)

- No Update and Delete
- No support to transactions
- Larger latency in results
- Supports Peta Bytes Data
- HiveQL
- Limited JDBC/ODBC support

Hive Data and Files

- Hive Data Types and File Formats (Section 4.4.4 Tables 4.5 and 4.6)
- Data Model (Section 4.4.5 Table 4.7)

Fig. 4.11 Hive Integration and Workflow steps



Hive Functions

- **Hive Built-in-Functions (Section 4.4.7 Table 4.8)**

Summary

We learnt :

- Hive– A data warehouse infrastructure
- Creates Tables, Files, Databases
Provides data summarization
- Ad hoc querying
- HWI
- Hive Sever (Thrift)

Summary

We learnt:

- Hive characteristics
- Hive Metadata
- Compression to RDBMS
- Hive Data Types, File Formats, Data Model, Built in Functions

Summary

We learnt:

- Steps in Integration and Workflow
- Integration with the MapReduce and HDFS, thus scalable architecture for Big Data

End of Lesson 5 on Hive Characteristics and Functions