CARLSON SCHOOL
OF MANAGEMENT
UNIVERSITY OF MINNESOTA

Introduction To Hive

MSBA 6330 Prof Liu

Slides credits: Cloudera Academic Partners Program

Introduction To Hive

- In this chapter, you will learn
  - What Hive is
  - How Hive differs from a relational database
  - Ways in which organizations use Hive
  - How to invoke and interact with Hive

Introduction to Hive

WHAT IS HIVE?

Overview of Apache Hive

- Apache Hive is a data warehouse facility for reading, writing, and managing large datasets in distributed storage and queried using SQL syntax
  - Initially runs on MapReduce, but now also supports Tez and Spark as execution engines
- With many of the standard SQL functionalities, Hive is designed for:
  - Analysts with SQL expertise
  - BI tools that generate SQL
  - ETL
- History
  - Originally developed by Facebook for data warehousing in 2007
  - Apache Hive first release (v0.3) in 2010.
  - Hive on Tez released on 2013.
  - Hive 2 released on 2016 (with LLAP support) – 25x speed up over Hive 1
  - Hive 3 released on 2018

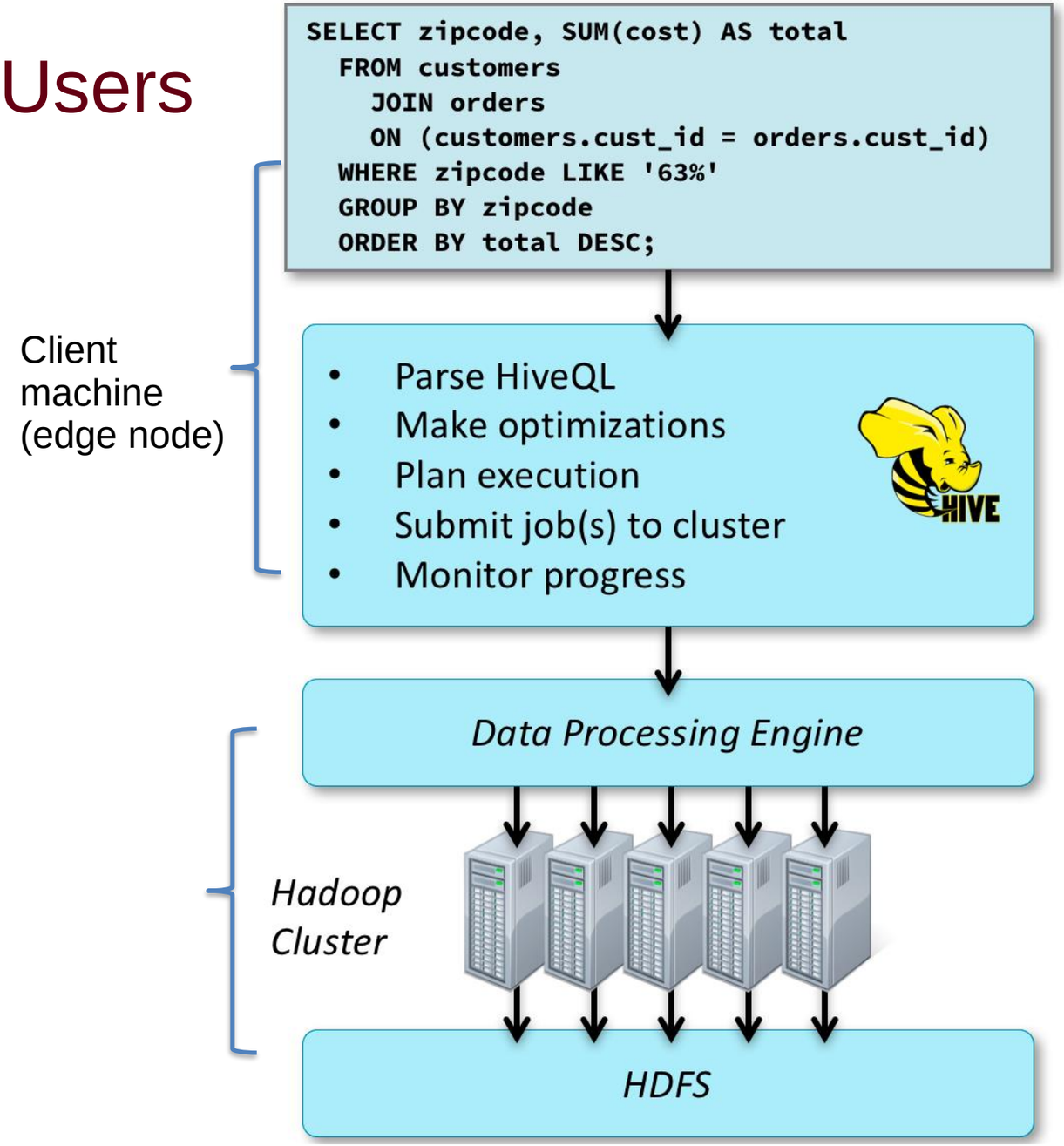
Overview of Apache Hive (2)

- Uses a SQL-like language called **HiveQL**
  - A subset ANSI SQL-92 standard, plus a few extensions found in MySQL and Oracle SQL dialects


```
  SELECT    zipcode, SUM(cost) AS total
    FROM    customers
    JOIN    orders
      ON    customers.cust_id = orders.cust_id
   WHERE    zipcode LIKE '63%'
GROUP BY    zipcode
ORDER BY    total DESC;
```

High-Level Overview For Hive Users

- Hive runs on the client machine
  - Turns HiveQL queries into a directed acyclic graph of MapReduce, Tez, or Spark jobs.
  - Submits those jobs to Hadoop cluster for execution
    - In the form of an execution plan



Why Use Apache Hive?

- More productive than writing MapReduce directly
  - Five lines of HiveQL might be equivalent to 100 lines or more of Java
- Brings large-scale data analysis to a broader audience
  - No software development experience required
  - Leverage existing knowledge of SQL
- Offers interoperability with other systems
  - Extensible through UDFs, JDBC/ODBC, and external scripts
  - Many business intelligence (BI) tools support Hive
    - Tableau, Datameter, Microstrategy, Pentaho, Qlikview
- Caveat Emptor
  - Remember that Hive generates Hadoop jobs, making it ultimately a batch processing platform, not a real-time / interactive platform*

    **With LLAP, Hive queries can be executed in sub-second response time.*

Introduction to Hive

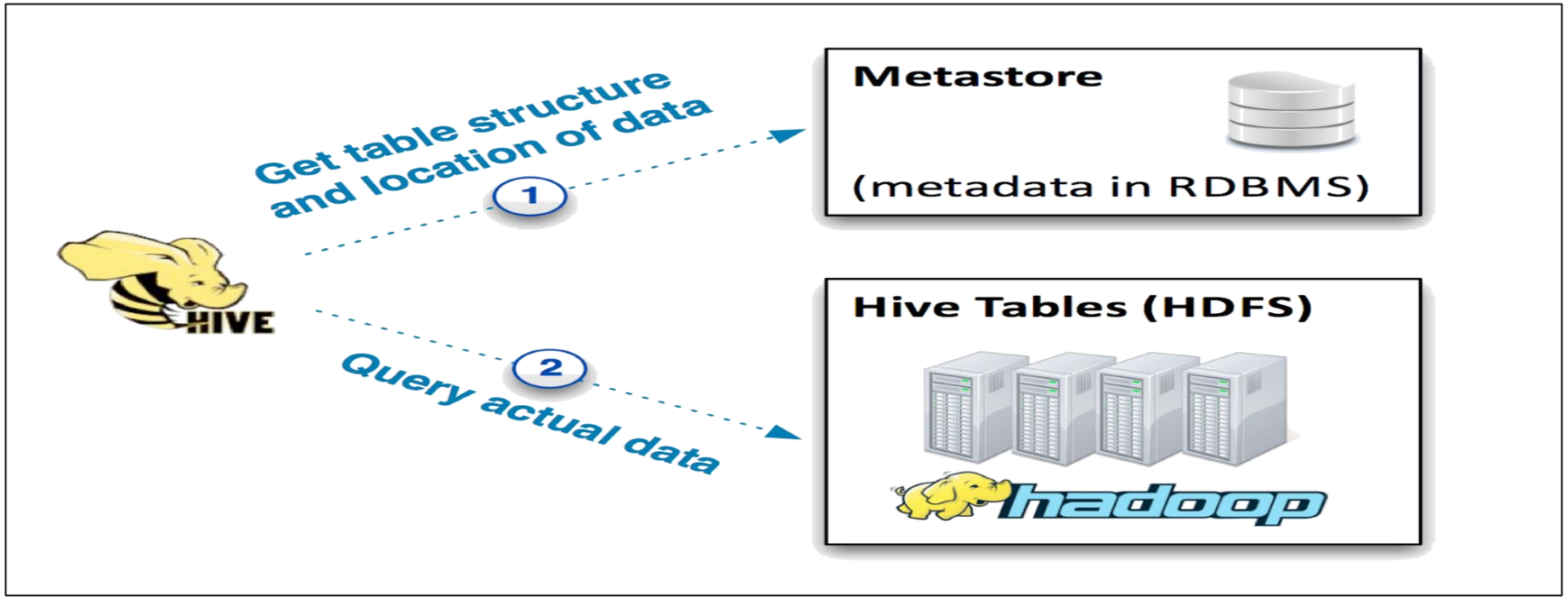
HIVE SCHEMA AND DATA STORAGE

How Hive Loads and Stores Data (1)

- Hive's queries operate on tables, just like in an RDBMS
- A table has two components
  - Meta data
    - Specify the structure and location of data
    - Defined when table is created
    - Stored in the **metastore**, contained in an RDBMS (typically Derby or MySQL)
  - Data
    - Typically in an **HDFS directory** containing one or more files
    - Default path: `/user/hive/warehouse/<table_name>`
    - Can be in any of several formats for storage and retrieval

How Hive Loads and Stores Data (2)

- Hive consults the metastore to determine data format and location
- The query itself operates on data stored on a filesystem (typically HDFS)



Hive Managed versus External Tables

- (In Hive 3) **Managed** (internal) tables:
  - Fully under Hive control
  - ACID on by default
  - Default storage format: ORC
- (In Hive 3) External tables:
  - Outside control and management of data
  - No ACID
  - Default storage format: Text

Atomicity, Consistency, Isolation, Durability

- *Before Hive 3, the default format is text for both managed and external tables.*

Hive vs. RDBMS

- Hive is often considered data warehousing for Hadoop
- Hive shares many similarities with an RDBMS but there is at least one important difference
  - In an RDBMS, you create the table with rigid structure that must be specified before any data is added to the table (called "**schema on write**").
  - But with Hive you can store the data in HDFS without knowing its format at all. You only need to specify the format (fields, types, etc) of the data when you need to read it (called "**schema on read**").
- Pros and Cons of **Schema on Read**:
  - **Pro**: This provides far more flexibility and speed on write
  - **Con**: a conflict between the expected and actual data formats won't be detected at the time records are added to a table.

Hive vs. RDBMS

- Client-server RDBMS has many strengths
  - Very fast response time (milliseconds)
  - Support for transactions, with ACID (Atomicity, Consistency, Isolation, Durability) guarantees.
  - Allow frequent modification of small number of records
  - Can serve thousands of simultaneous clients
- Hive does not turn your Hadoop cluster into an RDBMS
  - Initially has no support insert/update/deletion (IUD)*
  - Initially has no support for ACID transactions*
  - No referential integrity
  - Batch job – latency is high compared to RDBMS

* Later versions (2.0+) of Hive support ACID transactions for certain types of tables.

Hive vs RDBMS

Feature	RDBMS	Hive
Query Language	SQL	HiveQL (subset of SQL)
Update Individual Records	Yes	Managed tables only*
Delete Individual Records	Yes	Managed tables only*
Transactions	Yes	Managed tables only*
Index Support	Extensive	Limited
Latency	Very Low	High**
Data Size	Terabytes	Petabytes
Storage cost	Very high	Very Low

*Starting from Hive version **0.14.0+ (2014)**: INSERT/UPDATE/DELETE with ACID supports are available for certain types of tables.
** Starting from Hive 2, Hive can use LLAP to achieve sub-second response time for some small, frequent queries.

Introduction to Hive

HIVE USE CASES

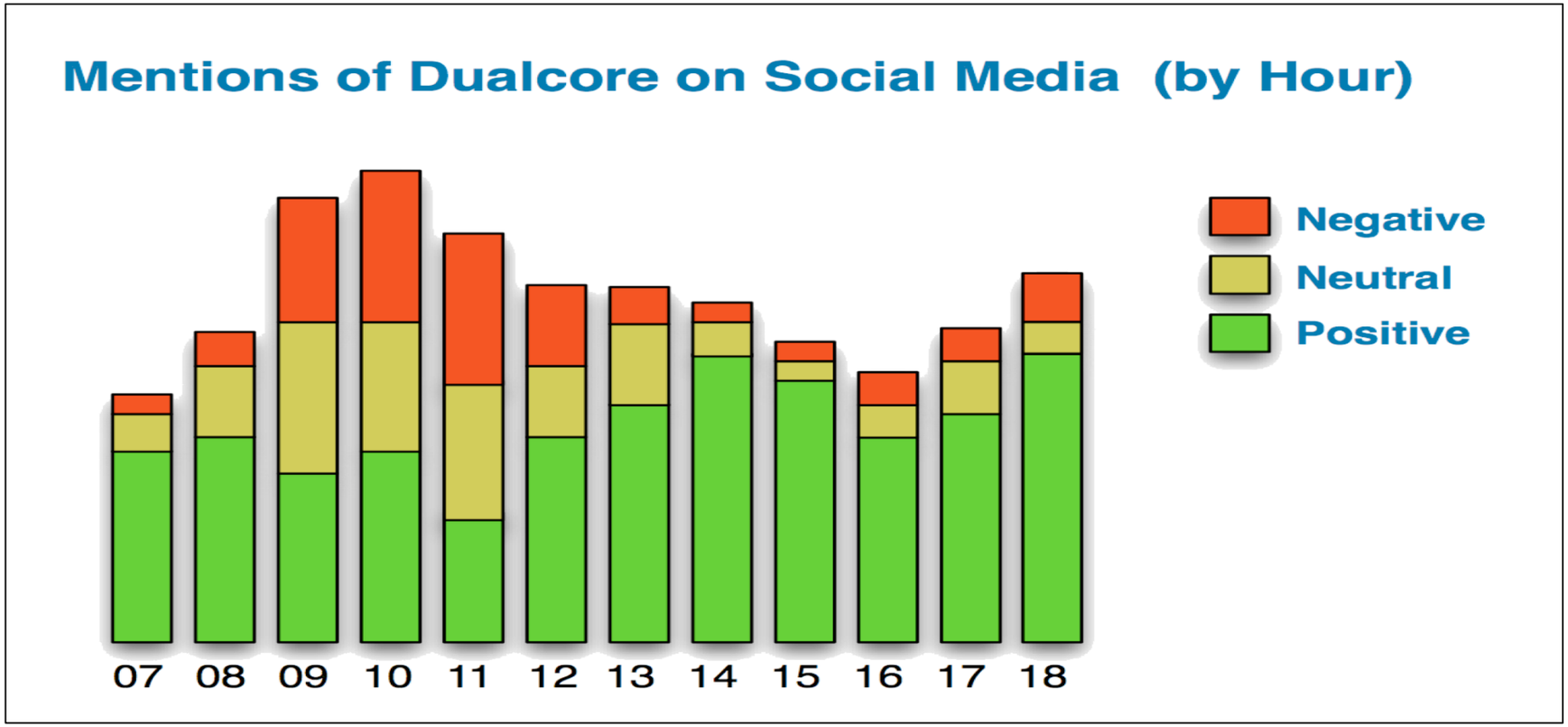
Hive Use Cases

- With its familiar interface, Hive is the tool-of-choice for a variety of batch processing workloads, including:
  - Data preparation
  - ETL
  - Data mining
  - Ad optimization

Source: https://www.cloudera.com/products/apache-hadoop/apache-hive.html

Use Case: Sentiment Analysis

- Many organizations use Hive to analyze social media



Example: https://xebia.com/blog/sentiment-analysis-using-apache-hive/

Use Case: Log File Analytics

- Because Hive is flexible in its data format, it can be used to store non traditional tables e.g. web log files.
- Hive allows you to treat a directory of log files like a table
  - Allows SQL-like queries against semi-structured data

We will see an example of this later

Dualcore Inc. Public Web Site (June 1 - 8)					
Product	**Unique Visitors**	**Page Views**	**Average Time on Page**	**Bounce Rate**	**Conversion Rate**
Tablet	5,278	5,894	17 seconds	23%	65%
Notebook	4,139	4,375	23 seconds	47%	31%
Stereo	2,873	2,981	42 seconds	61%	12%
Monitor	1,749	1,862	26 seconds	74%	19%
Router	987	1,139	37 seconds	56%	17%
Server	314	504	53 seconds	48%	28%
Printer	86	97	34 seconds	27%	64%

Examples:
http://cuddletech.com/?p=795 (Flume + Pig + Hive, step by step)
https://hortonworks.com/blog/hadoop_tutorial_visualizing_server_logs/ (HortonWorks Sandbox Tutorial, Flume + Hive)
http://www.irdindia.in/journal_ijraet/pdf/vol4_iss2/27.pdf (Twitter, Flume, Hive)

Hive at Facebook

- ~200 people/month run jobs on Hadoop Hive
- Analyst (non-engineers) use Hadoop through Hive
- 95% of jobs are Hive Jobs
- Types of applications
  - Reporting – measures of user engagement, daily aggregations of impression/click counts
  - Ad hoc analysis
  - Machine Learning (assemble training data)
    - Ad optimization, user engagement as a function of user attributes

Source:Hive: A Petabyte Scale Data Warehouse System on Hadoop (PDF, 2010)

Introduction to Hive

INTERACTING WITH HIVE

Using Beeline (The Hive Shell)

- The following examples uses the dualcore database. To install it, run:
  - `"./scripts/analyst/advance_labs.sh lab1"` in the bash shell.
- You can execute HiveQL statements in the Beeline
  - This interactive tool is similar to the MySQL shell
- Start Beeline by specifying a URL for a hive server

Add `-n username -p password` if needed.

```
> beeline -u jdbc:hive2://
```

```
0: jdbc:hive2://> use dualcore;
0: jdbc:hive2://> SELECT lname, fname FROM customers
WHERE state = 'CA' LIMIT 50;
+-------------+-------------+
| lname       | fname       |
+-------------+-------------+
| Ham         | Marilyn     |
| Franks      | Gerard      |
...
| Falgoust    | Jennifer    |
+-------------+-------------+
50 rows selected (15.829 seconds)
0: jdbc:hive2://>
```

Using Beeline

- Beeline commands start with "!"
  - No terminator character (a SQL query ends with ;)
- Some commands
  - `!connect url` : connects to a different hive server
  - `!exit` : exits the shell
  - `!help` : shows the full list of commands
  - `!verbose`: shows additional details of queries

```
0: jdbc:hive2://localhost:10000> !exit
```

- Press Enter to execute a query or command

https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients for documentation on beeline

Accessing Hive From The Command Line

- You can also execute a file containing HiveQL code using the `-f` (file) option

The drawback is that this spins up Hive jars each time whereas Hive CLI does it only once.

```
$ beeline –u ... -f myquery.hql
```

- Or use HiveQL directly from the command line using the `-e` (execute) option

```
$ beeline –u ... -e 'SELECT * FROM users'
```

- Use the `--silent=true` option to suppress informational messages

```
$ beeline --silent=true –u ...
```

Hive Configuration

- Many aspects of Hive's behavior are configured through properties
  - Use set -v in Hive to see current values

```
0: ...> set -v;
```

- You can also use set to specify property values
  - The following enables columns headers in query results

```
0: ...> set hive.cli.print.header=true;
0: ...> set hive.execution.engine;   #to read the current value
0: ...> set hive.execution.engine=spark;  #set new value, does not work on our VM
```

- Hive runs the `.hiverc` file in your home directory at startup
  - Your can edit it to specify per-user defaults (e.g. setting execution engine, setting printer header)

Accessing Hive With Hue

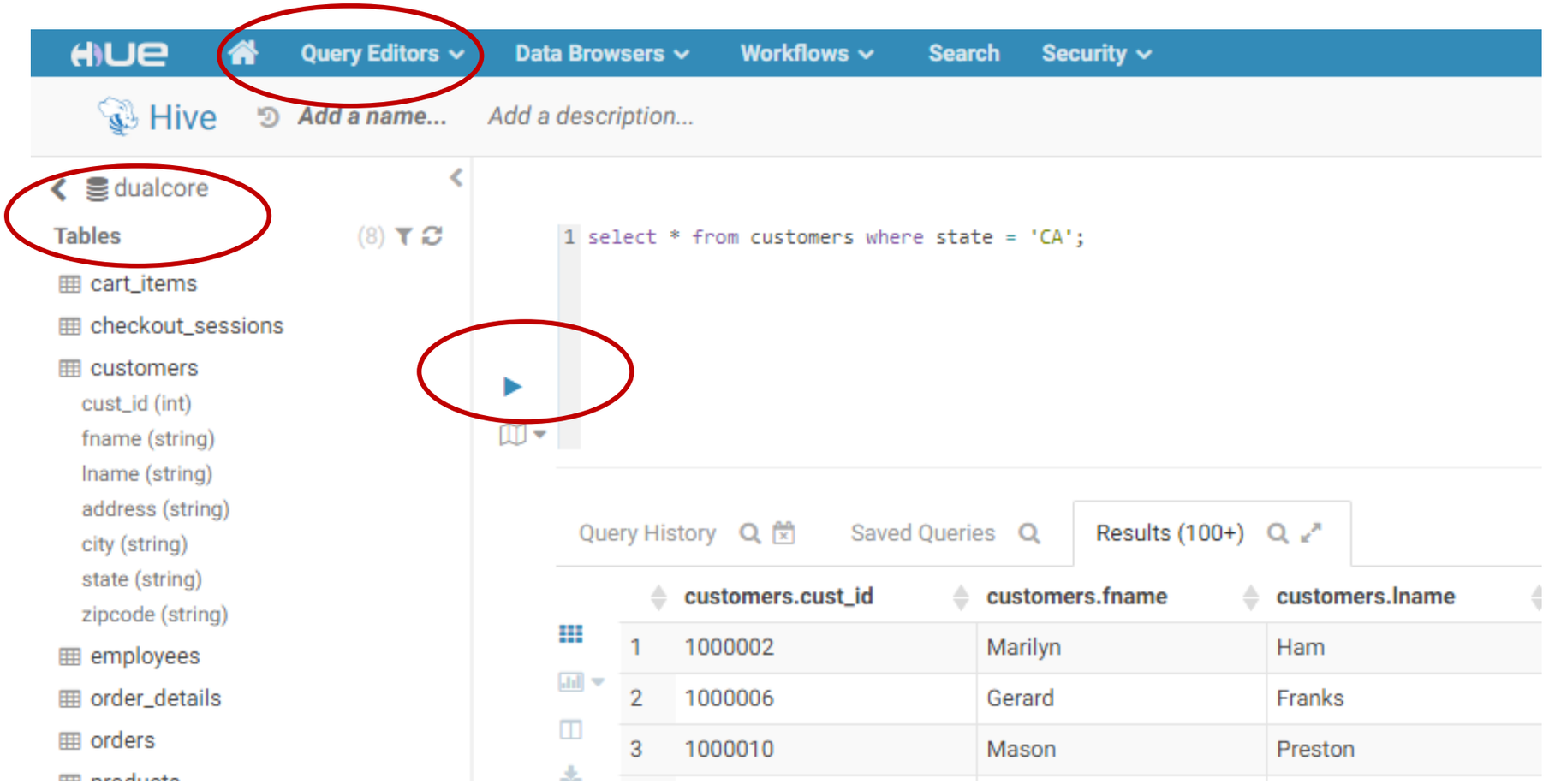
- Alternatively, you can access Hive through Hue
  - Web based UI for many Hadoop related services, including Hive
- To use Hue, browse to `http://localhost:8888/`
  - May need to start Hue service first

```
$ sudo service hue start
```

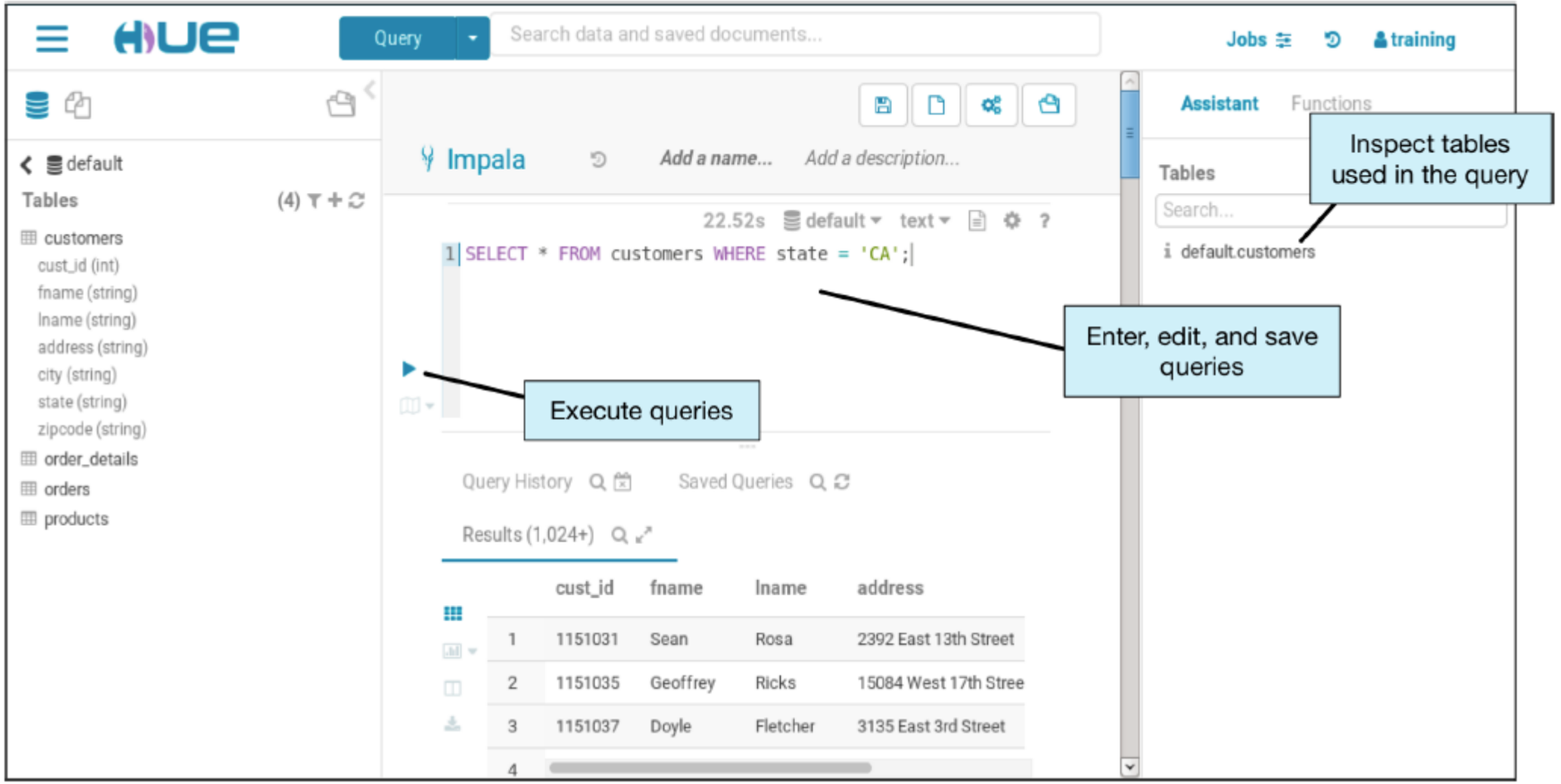
On our VM, you need to enter "cloudera' as username and password

- Hue's Hive interface is called Beeswax
  - Launch by clicking its icon
- Beeswax features include:
  - Creating tables
  - Running queries
  - Browsing tables
  - Saving queries for later execution

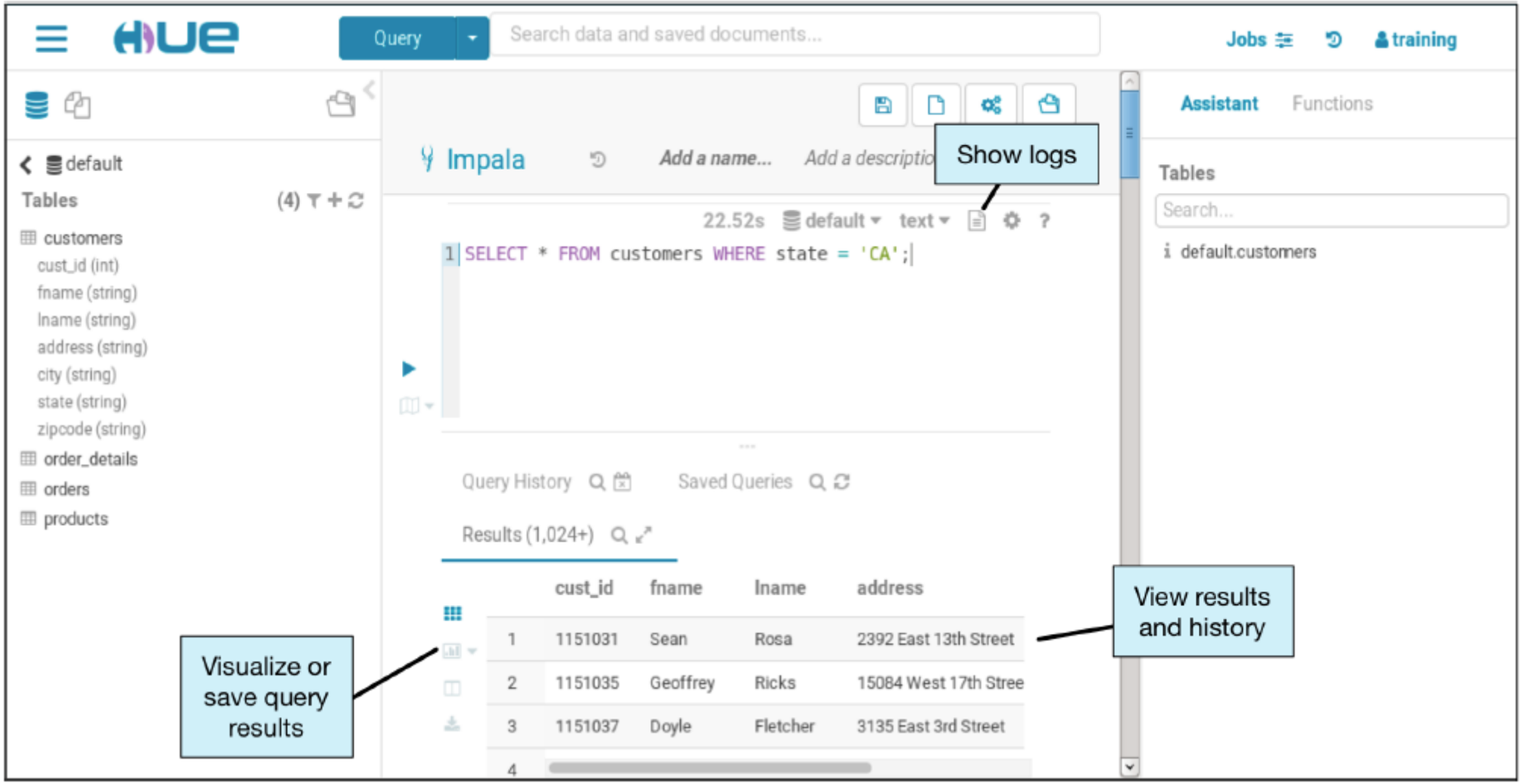
The Hue Query Editor: Databases



The Hue Query Editor: Queries



Query Results In Beeswax



Essential Points

- Hive is a high level abstraction on top of Hadoop
  - Runs jobs on Hadoop based on HiveQL statements
- HiveQL is very similar to SQL
  - Easy to learn for those with relational database experience
  - Hive is not a typical RDBMS, nor replaces one.
- Hive tables are really directories of files in HDFS
  - Information about those tables is kept in Hive's metastore