

# Tutorial A

Note: The commands in the image are visible when zoomed in.

## Commands 1-14

```
hadoop@ip-172-31-40-133:~
E:::E EEEE M:::M:::M::: M:::M:::M::: R:::R R:::R
E:::E M:::M M:::M M:::M M:::M M:::M R:::R R:::R
E:::EEEEEEEE M:::M M:::M M:::M M:::M R:::RRRRRR:::R
E:::M:::M M:::M M:::M M:::M M:::M R:::R R:::R
E:::EEEEEEEE M:::M M:::M M:::M R:::RRRRRR:::R
E:::E M:::M M:::M M:::M R:::R R:::R
E:::E EEEE M:::M M:::M M:::M R:::R R:::R
E:::EEEEEEEE:::E M:::M M:::M M:::R R:::R
E:::M:::M M:::M M:::M R:::R R:::R
EEEEEEEEEEEEEEEE M:::M M:::M R:::R R:::R

[hadoop@ip-172-31-40-133 ~]$ hadoop fs -mkdir HV
mkdir: 'HV': File exists
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -mkdir HV/input
mkdir: 'HV/input': File exists
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -copyFromLocal CustomerDetails.txt HV/input
copyFromLocal: HV/input/customerDetails.txt: File exists
[hadoop@ip-172-31-40-133 ~]$ hadoop fs -copyFromLocal payments.txt HV/input
[hadoop@ip-172-31-40-133 ~]$ hive
Hive Session ID = 47ad2a7a-5504-a63c-e4c6idd15970

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> CREATE DATABASE bank_db COMMENT 'This database is created for analyzing customer late payments';
OK
Time taken: 1.368 seconds
hive> show databases;
OK
bank_db
default
Time taken: 0.242 seconds, Fetched: 2 row(s)
hive> use bank_db;
OK
Time taken: 0.052 seconds
hive> CREATE TABLE IF NOT EXISTS bank_db.payments (c_id int, transaction_id string, transaction_date string, late string) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
Time taken: 0.394 seconds
hive> CREATE TABLE IF NOT EXISTS bank_db.payments_ptn (c_id int, transaction_id string, transaction_date string) PARTITIONED BY (late string) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
OK
Time taken: 0.18 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS bank_db.customers_ext (c_id int, c_firstname string, c_lastname string, c_street string, c_city string, c_state string, c_zip string, c_yob int, c_gender string, credit_card string, internet string, mobile string) COMMENT 'This is a table stored externally in HV/bank subdirectory of the HDFS.' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '/user/hive/warehouse/bank_db/customers_ext';
OK
Time taken: 0.062 seconds
hive> show tables;
OK
customers_ext
payments
payments_ptn
Time taken: 0.042 seconds, Fetched: 3 row(s)
hive> drop table payments_ptn;
OK
Time taken: 0.412 seconds
hive> alter table customers_ext RENAME to customers;
OK
Time taken: 0.144 seconds
hive> describe payments;
OK
c_id int
transaction_id string
transaction_date string
late string
Time taken: 0.067 seconds, Fetched: 4 row(s)
hive> describe customers;
```

## Commands 15-20

```
hadoop@ip-172-31-40-133:~$
c_id int
transaction_id string
transaction_date string
last string
Time taken: 0.067 seconds, Fetched: 4 row(s)
hive> describe customers;
OK
c_id int
c_firstname string
c_lastname string
c_street string
c_city string
c_state string
c_zip string
c_yob int
c_gender string
credit_card string
internet string
mobile string
Time taken: 0.067 seconds, Fetched: 12 row(s)
hive> LOAD DATA INPATH 'HV/input/LatePayments.txt' OVERWRITE INTO TABLE payments;
Loading data to table bank_db.payments
OK
Time taken: 0.734 seconds
hive> LOAD DATA INPATH 'HV/input/CustomerDetails.txt' OVERWRITE INTO TABLE customers;
Loading data to table bank_db.customers
OK
Time taken: 0.229 seconds
hive> SELECT * FROM payments;
OK
181 146268743-8 23-9-2020 TRUE
172 396589804-7 28-9-2020 FALSE
183 553752031-8 25-9-2020 FALSE
183 559786553-4 13-12-2018 TRUE
175 108659198-9 13-9-2016 TRUE
176 360007735-5 27-9-2016 FALSE
177 238309554-X 28-11-2014 FALSE
178 644696918-4 31-5-2014 TRUE
175 318241713-5 15-7-2016 TRUE
180 84360172-2 6-7-2016 FALSE
181 80763386-8 6-7-2016 FALSE
182 845886260-4 23-4-2014 FALSE
183 270161074-X 19-7-2016 TRUE
171 887426135-0 4-7-2020 FALSE
172 401125380-4 23-8-2015 TRUE
173 277486266-7 27-12-2019 FALSE
174 82118455-2 4-12-2019 TRUE
Time taken: 1.682 seconds, Fetched: 17 row(s)
hive> SELECT * FROM customers;
OK
171 Madelyn Hensley 10075 Thierer Plaza New York New York 81377 1976 M TRUE FALSE FALSE
172 Lanny Foster 23501 Park Meadow Dr Austin Texas 13490 1981 F TRUE FALSE FALSE
173 Karina Livingston 95 Anderson Park Chattanooga Tennessee 04519 1977 F TRUE TRUE TRUE TRUE
174 Avery McCormick 09982 Sunfield Parkway Chicago Illinois 38350 1992 F TRUE FALSE FALSE
175 Peter King 0486 Dryden Road Chicago Illinois 21012 1991 M TRUE TRUE TRUE TRUE
176 Bret Ibarra 14 Transport Plaza San Diego California 12865 1984 M TRUE FALSE FALSE
177 Leonardo Wheeler 806 Cory Crossing New York New York 44464 1972 F TRUE TRUE TRUE TRUE
178 Kenneth Noble 8 South Terrace Hixon Tennessee 52890 1989 F FALSE TRUE TRUE TRUE
179 Marcia Matthews 0380 Knutson Road Dallas Texas 80477 1967 F TRUE FALSE TRUE TRUE
180 Avis Kramer 49624 Hancock Junction New York New York 62542 1965 W TRUE FALSE TRUE TRUE
181 Lynnette Tate 30741 Paget Court New York New York 34886 1987 M TRUE FALSE TRUE TRUE
182 Lakisha Estrada 50 Dahlia Crossing Dallas Texas 16042 1991 F FALSE TRUE FALSE TRUE
183 Bill Silva 4 McBride Crossing Detroit Michigan 15871 1968 M TRUE TRUE TRUE
Time taken: 0.172 seconds, Fetched: 13 row(s)
hive>
```

## Tutorial B

### Command 1

```
hadoop@ip-172-31-40-133:~$
183 Bill Silva 4 McBride Crossing Detroit
Time taken: 0.172 seconds, Fetched: 13 row(s)
hive> select c_id, c_firstname, c_lastname
> from customers
> where c_state = 'New York';
OK
171 Madelyn Hensley
177 Leonardo Wheeler
180 Avis Kramer
181 Lynnette Tate
Time taken: 0.579 seconds, Fetched: 4 row(s)
```

## Command 2

```
hadoop@ip-172-31-40-133:~  
Time taken: 0.579 seconds, Fetched: 4 row(s)  
hive> select c_state, count(c_id) as total  
      > from customers  
      > group by c_state;  
Query ID = hadoop_20241117072339_29b05345-faf2-4735-814f-58911b889152  
Total jobs = 1  
Launching Job 1 out of 1  
Tez session was closed. Reopening...  
Session re-established.  
Session re-established.  
Status: Running (Executing on YARN cluster with App id application_1731826520887_0002)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1 .....	container	SUCCEEDED	1	1	0	0	0	0	
Reducer 2 .....	container	SUCCEEDED	2	2	0	0	0	0	

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 6.82 s

```
OK  
Illinois      2  
New York      4  
California    1  
Michigan      1  
Tennessee    2  
Texas        3  
Time taken: 18.105 seconds, Fetched: 6 row(s)
```

## Command 3

```
hadoop@ip-172-31-40-133:~  
Time taken: 18.105 seconds, Fetched: 6 row(s)  
hive> select c.c_id, c.c_firstname, c.c_lastname, p.transaction_id, p.transaction_date  
      > from customers c join payments p on c.c_id = p.c_id  
      > where p.late = 'TRUE';  
Query ID = hadoop_20241117072418_ea99b505-9697-4ce7-92de-a98a23bc56b7  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1731826520887_0002)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 2 .....	container	SUCCEEDED	1	1	0	0	0	0	
Map 1 .....	container	SUCCEEDED	1	1	0	0	0	0	

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 10.70 s

```
OK  
172  Lonny  Foster  401129380-4    23-8-2015  
174  Avery  McCormick  851112155-2    4-12-2019  
175  Peter  King    108659198-9    13-9-2016  
178  Bennett Noble  694690715-8    31-5-2014  
179  Marcia  Matthews  318241713-5    15-7-2016  
181  Lynnette Tate    146268743-8    23-8-2020  
183  Bill    Silva    270161074-X    19-7-2016  
183  Bill    Silva    559786593-4    13-12-2018  
Time taken: 11.517 seconds, Fetched: 8 row(s)
```

## Command 4

```
hadoop@ip-172-31-40-133:~  
Time taken: 11.517 seconds, Fetched: 8 row(s)  
hive> select c_state, count(c_id)  
  > from customers  
  > where credit_card = 'FALSE'  
  > group by c_state;  
Query ID = hadoop_20241117072447_63025eee-f516-4c90-8296-f581b4f42274  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1731826520887_0002)  
  
-----  
VERTICES    MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
-----  
Map 1 ..... container  SUCCEEDED    1         1         0         0         0         0  
Reducer 2 ..... container  SUCCEEDED    2         2         0         0         0         0  
-----  
VERTICES: 02/02  [=====>>] 100%  ELAPSED TIME: 6.55 s  
-----  
OK  
New York      1  
Tennessee    1  
Texas        1  
Time taken: 7.002 seconds, Fetched: 3 row(s)
```

## Command 5

```
hadoop@ip-172-31-40-133:~  
Time taken: 7.002 seconds, Fetched: 3 row(s)  
hive> select c.c_state, count(c.c_id) as total  
  > from customers c join payments p on c.c_id = p.c_id  
  > where p.late = 'TRUE'  
  > group BY c.c_state;  
Query ID = hadoop_20241117072507_f2885d2a-2aa7-4e0b-8474-43cb7dbb449a  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1731826520887_0002)  
  
-----  
VERTICES    MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
-----  
Map 3 ..... container  SUCCEEDED    1         1         0         0         0         0  
Map 1 ..... container  SUCCEEDED    1         1         0         0         0         0  
Reducer 2 ..... container  SUCCEEDED    2         2         0         0         0         0  
-----  
VERTICES: 03/03  [=====>>] 100%  ELAPSED TIME: 13.92 s  
-----  
OK  
Illinois      2  
New York      1  
Michigan      2  
Tennessee    1  
Texas        2  
Time taken: 14.591 seconds, Fetched: 5 row(s)
```

## Command 6

```
hadoop@ip-172-31-40-133:~
Time taken: 14.591 seconds, Fetched: 5 row(s)
hive> select *
> from customers
> where mobile = 'FALSE';
OK
171 Madelyn Hensley 10075 Thierer Plaza New York New York 81377 1976 M TRUE FALSE FALSE
172 Lonny Foster 23901 Park Meadow Dr Austin Texas 13498 1981 F TRUE FALSE FALSE
174 Avery McCormick 09992 Sunfield Parkway Chicago Illinois 38300 1992 F TRUE FALSE FALSE
176 Bret Ibarra 14 Transport Place San Diego California 12865 1984 M TRUE FALSE FALSE
182 Lakisha Estrada 50 Dahle Crossing Dallas Texas 16042 1991 F FALSE TRUE FALSE
Time taken: 0.224 seconds, Fetched: 5 row(s)
```

## Command 7

```
hadoop@ip-172-31-40-133:~
Time taken: 0.224 seconds, Fetched: 5 row(s)
hive> select c.c_id, c.c_firstname, c.c_lastname, c.c_street, c.c_city, c.c_state, c.c_zip, c.c_yob, c.c_gender, c.credit_card, c.internet, c.mobile
> from customers c join payments p on c.c_id = p.c_id
> where p.plate = 'TRUE';
Query ID = hadoop_20241117072546_15926144-0bfb-4b33-9585-b9aa3fe6da42
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1731826520887_0002)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 2 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 10.98 s
-----
OK
172 Lonny Foster 23901 Park Meadow Dr Austin Texas 13498 1981 F TRUE FALSE FALSE
174 Avery McCormick 09992 Sunfield Parkway Chicago Illinois 21012 1991 F TRUE FALSE FALSE
175 Peter King 0486 Dryden Road Chicago Illinois 21012 1991 M TRUE FALSE TRUE
178 Bennett Noble 8 South Terrace Hixson Tennessee 52890 1989 F FALSE TRUE TRUE
179 Marcia Matthews 0380 Knutson Road Dallas Texas 80477 1967 F TRUE FALSE TRUE
181 Lynnette Tate 30741 Paget Court New York New York 34886 1987 M TRUE TRUE FALSE TRUE
183 Bill Silva 4 McBride Crossing Detroit Michigan 15871 1968 M TRUE TRUE TRUE
193 Bill Silva 4 McBride Crossing Detroit Michigan 15871 1968 M TRUE TRUE TRUE
Time taken: 11.635 seconds, Fetched: 8 row(s)
```

## Command 8

```
hadoop@ip-172-31-40-133:~
Time taken: 11.635 seconds, Fetched: 8 row(s)
hive> select c.c_id, c.c_firstname, c.c_lastname, c.c_city, c.c_state
> from customers c join payments p on c.c_id = p.c_id
> where p.plate = 'FALSE'
> group by c.c_id, c.c_firstname, c.c_lastname, c.c_city, c.c_state;
Query ID = hadoop_20241117072613_5eac1ddd-526d-49cc-a583-901cfa49e89b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1731826520887_0002)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 3 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  2      2      0      0      0      0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 12.28 s
-----
OK
171 Madelyn Hensley New York New York
172 Lonny Foster Austin Texas
173 Karina Livingston Chattanooga Tennessee
176 Bret Ibarra San Diego California
177 Leonardo Wheeler New York New York
180 Avis Kramer New York New York
181 Lynnette Tate New York New York
182 Lakisha Estrada Dallas Texas
183 Bill Silva Detroit Michigan
Time taken: 13.068 seconds, Fetched: 9 row(s)
```

## Command 9

```
Time taken: 13.068 seconds, Fetched: 9 row(s)
hive> select c.c_firstname, c.c_firstname, c.c_city, c.c_state
> from customers c join payments p on c.c_id = p.c_id
> where p.late = 'TRUE' and c.c_yob > 1985;
Query ID = hadoop_20241117072642_1cbd1556-0026-4367-9be8-ae42f372e542
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1731826520887_0002)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 2	.....	container	SUCCEEDED	1	1	0	0	0	0
Map 1	.....	container	SUCCEEDED	1	1	0	0	0	0

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 10.48 s

```
OK
Avery Avery Chicago Illinois
Peter Peter Chicago Illinois
Bennett Bennett Hixson Tennessee
Lynnette Lynnette New York New York
Time taken: 11.304 seconds, Fetched: 4 row(s)
```

## Command 10

```
Time taken: 11.304 seconds, Fetched: 4 row(s)
hive> select c.c_id, c.c_firstname, c.c_lastname, c.c_city, c.c_state
> from customers c join payments p on c.c_id = p.c_id
> where p.late = 'FALSE' and c.internet = 'TRUE'
> group by c.c_id, c.c_firstname, c.c_lastname, c.c_city, c.c_state;
Query ID = hadoop_20241117072710_30e4db71-53f8-4f19-8753-37b7f8468beb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1731826520887_0002)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 3	.....	container	SUCCEEDED	1	1	0	0	0	0
Map 1	.....	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	.....	container	SUCCEEDED	2	2	0	0	0	0

VERTICES: 03/03 [=====>>] 100% ELAPSED TIME: 11.45 s

```
OK
173 Karina Livingston Chattanooga Tennessee
182 Lakisha Estrada Dallas Texas
183 Bill Silva Detroit Michigan
Time taken: 12.02 seconds, Fetched: 3 row(s)
hive>
```



## Q. What are the similarities and differences of Hive versus traditional RDBMSs?

### Similarities:

1. **SQL-Like Query Language:** Both Hive and traditional RDBMSs use SQL or SQL-like languages for database operations. In Hive, you used HiveQL commands, which are like SQL commands in RDBMSs.
  - Example: CREATE DATABASE, CREATE TABLE, SELECT statements.
2. **Data Organization:** Both systems organize data in tables, rows, and columns.
  - Example: Creating tables like payments, payments\_ptn, and customers.
3. **Schema Enforcement:** Both enforce schemas to define the structure of data.
  - Example: Defining column data types and table structures in Hive.

### Differences:

1. **Data Storage:**
  - **RDBMS:** Typically store data in proprietary formats optimized for quick read and write operations.
  - **Hive:** Designed to work with large datasets stored in Hadoop Distributed File System (HDFS) or other distributed storage systems, handling semi-structured and unstructured data.
  - **Example:** External table stored in HDFS (customers\_ext).
2. **Processing Model:**
  - **RDBMS:** Executes queries directly using their own storage engines. Optimized for real-time transactional processing.
  - **Hive:** Translates queries into MapReduce jobs or other Hadoop-based processing tasks, optimized for batch processing and large-scale data analysis.
  - **Example:** Loading data into Hive tables and performing batch queries.
3. **Data Volume:**
  - **RDBMS:** Best suited for handling gigabytes to terabytes of data.
  - **Hive:** Designed to handle petabytes of data distributed across a cluster of machines.
  - **Example:** Partitioned table payments\_ptn for efficient querying of large datasets.

#### 4. Performance:

- **RDBMS:** Generally faster for transactional queries due to optimized storage engines and indexing techniques.
- **Hive:** Slower for real-time queries since it translates queries into MapReduce or other batch processing jobs.
- **Example:** Executing joins and aggregations in Hive, which may be slower compared to RDBMS.

#### 5. ACID Transactions:

- **RDBMS:** Supports ACID (Atomicity, Consistency, Isolation, Durability) properties for transactions.
- **Hive:** Initially did not support full ACID properties; newer versions have transactional capabilities but are still more limited compared to RDBMSs.
- **Example:** Loading data into Hive tables without full ACID support.

#### 6. Schema-on-Write vs. Schema-on-Read:

- **RDBMS:** Follows schema-on-write, where the schema is enforced when data is written to the database.
- **Hive:** Uses schema-on-read, where the schema is applied when data is read.
- **Example:** Defining table schemas in Hive but applying them during query execution.