Hive数据库通过分区和分桶将大表拆分为小单元，提高查询效率，同时采用 ORC 格式存储和 Snappy 压缩减少存储占用和 I/O 成本；动态分区和分桶规则确保数据加载高效；通过索引、自动 MapJoin、向量化执行等技术加速查询

Movie表

```sql
CREATE TABLE Movie (
  Movie_ID INT,
  Movie_Title STRING,
  Average_Score FLOAT,
  Comment_Num INT,
  Month INT,
  Day INT
)
PARTITIONED BY (Year INT)
CLUSTERED BY (Movie_ID) INTO 10 BUCKETS
STORED AS ORC;


CREATE TABLE Movie_tmp (
  Movie_ID INT,
  Movie_Title STRING,
  Average_Score FLOAT,
  Comment_Num INT,
  Year INT,
  Month INT,
  Day INT
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

INSERT OVERWRITE TABLE Movie PARTITION (year)
SELECT
  movie_id,
  movie_title,
  average_score,
  comment_num,
  month,
  day,
  year
FROM Movie_tmp;

SET hive.exec.dynamic.partition = true;
SET hive.exec.dynamic.partition.mode = nonstrict;
SET hive.enforce.bucketing = true;
SET hive.exec.max.dynamic.partitions = 1000;
SET hive.exec.max.dynamic.partitions.pernode = 100;
SET hive.execution.engine=tez;
```

Actor表

```
CREATE TABLE Actor (
    Actor_ID INT,
    Actor_Name STRING
)
CLUSTERED BY (Actor_ID) INTO 10 BUCKETS;


CREATE TABLE Actor_tmp (
    Actor_ID INT,
    Actor_Name STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

INSERT INTO TABLE Actor
SELECT * FROM Actor_tmp;



CREATE TABLE All_Actor (
    Actor_ID INT,
    Movie_ID INT
)
CLUSTERED BY (Movie_ID, Actor_ID) INTO 10 BUCKETS
STORED AS ORC;
```

```
CREATE TABLE Starring_Actor (
    Actor_ID INT,
    Movie_ID INT
)
CLUSTERED BY (Movie_ID, Actor_ID) INTO 10 BUCKETS
STORED AS ORC;

CREATE TABLE All_Actor_tmp (
    Actor_ID INT,
    Movie_ID INT
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

CREATE TABLE Starring_Actor_tmp (
    Actor_ID INT,
    Movie_ID INT
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

INSERT INTO TABLE All_Actor
SELECT Actor_ID, Movie_ID
FROM All_Actor_tmp;

INSERT INTO TABLE Starring_Actor
SELECT Actor_ID, Movie_ID
FROM Starring_Actor_tmp;

SET hive.auto.convert.join=true;
```

Director表

```
CREATE TABLE Director (
    Director_Name STRING,
    Director_ID INT
)
CLUSTERED BY (Director_ID) INTO 10 BUCKETS
STORED AS ORC;


CREATE TABLE Direct_Movie (
    Director_ID INT,
    Movie_ID INT
)
CLUSTERED BY (Director_ID, Movie_ID) INTO 10 BUCKETS
STORED AS ORC;

CREATE TABLE Dirctor_tmp (
    Dirctor_Name STRING,
    Dirctor_ID INT
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");


CREATE TABLE Direct_Movie_tmp (
    Director_ID INT,
    Movie_ID INT
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

```
INSERT INTO TABLE Director
SELECT Dirctor_Name, Dirctor_ID
FROM Dirctor_tmp;

INSERT INTO TABLE Direct_Movie
SELECT Director_ID, Movie_ID
FROM Direct_Movie_tmp;
```

Genre表

```
CREATE TABLE Genre (
    Genre_ID STRING,
    Genre STRING
)
STORED AS ORC;

CREATE TABLE Movie_Genre (
    Movie_ID INT,
    Genre_ID STRING
)
CLUSTERED BY (Movie_ID,Genre_ID) INTO 10 BUCKETS
STORED AS ORC;


CREATE TABLE Genre_tmp (
    Genre_ID STRING,
    Genre STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");


CREATE TABLE Movie_Genre_tmp (
    Movie_ID INT,
    Genre_ID STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

INSERT INTO TABLE Genre
SELECT Genre_ID, Genre
FROM Genre_tmp;

INSERT INTO TABLE Movie_Genre
SELECT Movie_ID,Genre_ID
FROM Movie_Genre_tmp;
```

Review表

```
CREATE TABLE Review (
    Movie_ID INT,
    Review_ID INT
)
PARTITIONED BY (Score INT)
CLUSTERED BY (Movie_ID) INTO 10 BUCKETS
STORED AS ORC;

CREATE TABLE Review_tmp (
    Movie_ID INT,
    Review_ID INT,
    Score INT
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

INSERT OVERWRITE TABLE Review PARTITION (Score)
SELECT
  Movie_ID,
  Review_ID,
  Score
```

```
FROM Review_tmp;
```

Format表

```
CREATE TABLE Format (
    Format_ID STRING,
    Format STRING
)
STORED AS ORC;

CREATE TABLE Movie_Format (
    Movie_ID INT,
    Format_ID STRING
)
CLUSTERED BY (Movie_ID,Format_ID) INTO 10 BUCKETS
STORED AS ORC;


CREATE TABLE Format_tmp (
    Format_ID STRING,
    Format STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");


CREATE TABLE Movie_Format_tmp (
    Movie_ID INT,
    Format_ID STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar"     = "\"",
  "escapeChar"    = "\\"
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

```sql
INSERT INTO TABLE Format
SELECT Format_ID, Format
FROM Format_tmp;

INSERT INTO TABLE Movie_Format
SELECT Movie_ID,Format_ID
FROM Movie_Format_tmp;
```

优化参数

```sql
SET hive.exec.compress.output=true;
SET mapreduce.output.fileoutputformat.compress=true;
SET mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.SnappyCodec;

SET hive.vectorized.execution.enabled=true;
SET hive.vectorized.execution.reduce.enabled=true;

SET hive.optimize.ppd=true;
SET hive.ppd.remove.duplicatefilters=true;
SET hive.optimize.index.filter=true;

SET hive.cbo.enable=true;
SET hive.compute.query.using.stats=true;
SET hive.stats.fetch.column.stats=true;
SET hive.stats.fetch.partition.stats=true;
```