尚硅谷大数据项目之实时项目

ClickHouse

(作者：尚硅谷大数据研发部)

版本：V2.0

# ClickHouse入门

ClickHouse 是俄罗斯的Yandex于2016年开源的列式存储数据库（DBMS），使用C++语言编写，主要用于在线分析处理查询（OLAP），能够使用SQL查询实时生成分析数据报告。

## ClickHouse的特点

### 列式存储

以下面的表为例：

|  |  |  |
| --- | --- | --- |
| **Id** | **Name** | **Age** |
| 1 | 张三 | 18 |
| 2 | 李四 | 22 |
| 3 | 王五 | 34 |

* **采用行式存储时，数据在磁盘上的组织结构为：**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 张三 | 18 | 2 | 李四 | 22 | 3 | 王五 | 34 |

好处是想查某个人所有的属性时，可以通过一次磁盘查找加顺序读取就可以。但是当想查所有人的年龄时，需要不停的查找，或者全表扫描才行，遍历的很多数据都是不需要的。

* **采用列式存储时，数据在磁盘上的组织结构为：**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 张三 | 李四 | 王五 | 18 | 22 | 34 |

这时想查所有人的年龄只需把年龄那一列拿出来就可以了

* **列式储存的好处：**
* 对于列的聚合，计数，求和等统计操作原因优于行式存储。
* 由于某一列的数据类型都是相同的，针对于数据存储更容易进行数据压缩，每一列选择更优的数据压缩算法，大大提高了数据的压缩比重。
* 由于数据压缩比更好，一方面节省了磁盘空间，另一方面对于cache也有了更大的发挥空间。

### DBMS的功能

几乎覆盖了标准SQL的大部分语法，包括 DDL和 DML，以及配套的各种函数，用户管理及权限管理，数据的备份与恢复。

### 多样化引擎

ClickHouse和MySQL类似，把表级的存储引擎插件化，根据表的不同需求可以设定不同的存储引擎。目前包括合并树、日志、接口和其他四大类20多种引擎。

### 高吞吐写入能力

ClickHouse采用类LSM Tree的结构，数据写入后定期在后台Compaction。通过类LSM tree的结构，ClickHouse在数据导入时全部是顺序append写，写入后数据段不可更改，在后台compaction时也是多个段merge sort后顺序写回磁盘。顺序写的特性，充分利用了磁盘的吞吐能力，即便在HDD上也有着优异的写入性能。

官方公开benchmark测试显示能够达到50MB-200MB/s的写入吞吐能力，按照每行100Byte估算，大约相当于50W-200W条/s的写入速度。

### 数据分区与线程级并行

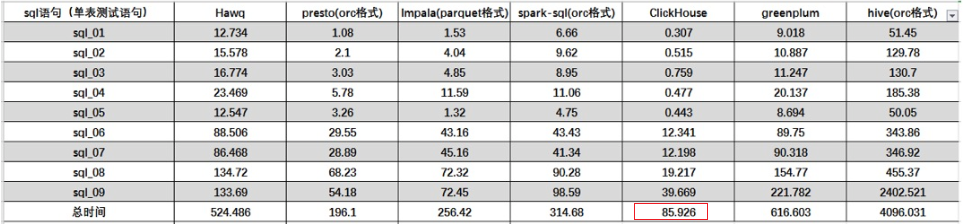
ClickHouse将数据划分为多个partition，每个partition再进一步划分为多个index granularity(索引粒度)，然后通过多个CPU核心分别处理其中的一部分来实现并行数据处理。在这种设计下，单条Query就能利用整机所有CPU。极致的并行处理能力，极大的降低了查询延时。

所以，ClickHouse即使对于大量数据的查询也能够化整为零平行处理。但是有一个弊端就是对于单条查询使用多cpu，就不利于同时并发多条查询。所以对于高qps的查询业务，ClickHouse并不是强项。

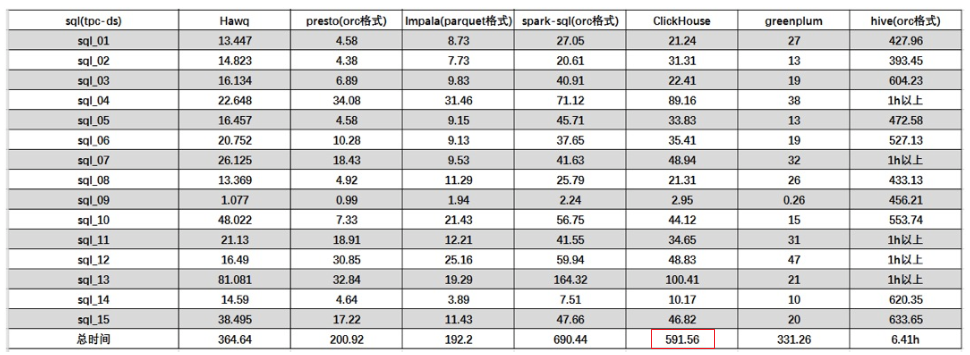
### 性能对比

某网站精华帖，中对几款数据库做了性能对比。

* **单表查询**



* **关联查询**



结论: ClickHouse像很多OLAP数据库一样，单表查询速度由于关联查询，而且ClickHouse的两者差距更为明显。

# ClickHouse的安装

## 准备工作

### 确定防火墙处于关闭状态

### CentOS取消打开文件数限制

* 在hadoop102的 /etc/security/limits.conf文件的末尾加入以下内容

[atguigu@hadoop102 ~]$ sudo vim /etc/security/limits.conf

\* soft nofile 65536

\* hard nofile 65536

\* soft nproc 131072

\* hard nproc 131072

* 在hadoop102的/etc/security/limits.d/20-nproc.conf文件的末尾加入以下内容

[atguigu@hadoop102 ~]$ sudo vim /etc/security/limits.d/20-nproc.conf

\* soft nofile 65536

\* hard nofile 65536

\* soft nproc 131072

\* hard nproc 131072

* 执行同步操作

[atguigu@hadoop102 ~]$ sudo /home/atguigu/bin/xsync /etc/security/limits.conf

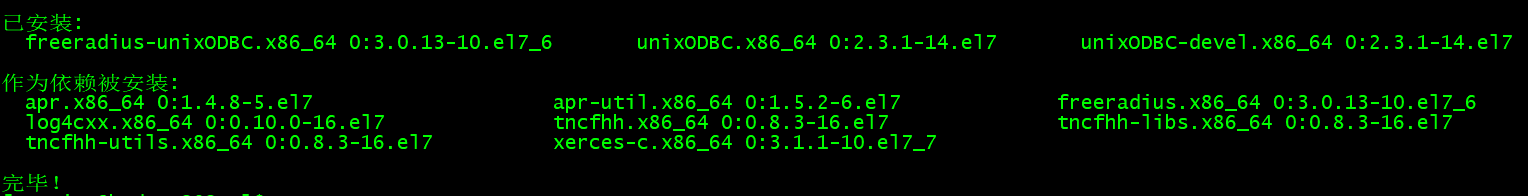
[atguigu@hadoop102 ~]$ sudo /home/atguigu/bin/xsync /etc/security/limits.d/20-nproc.conf

### 安装依赖

[atguigu@hadoop102 ~]$ sudo yum install -y libtool



[atguigu@hadoop102 ~]$ sudo yum install -y \*unixODBC\*



在hadoop103、hadoop104上执行以上操作

### CentOS取消SELINUX

* 修改/etc/selinux/config中的SELINUX=disabled

[atguigu@hadoop102 ~]$ sudo vim /etc/selinux/config

SELINUX=disabled

**注意：别改错了**

* 执行同步操作

[atguigu@hadoop102 ~]$ sudo /home/atguigu/bin/xsync /etc/selinux/config

* 重启三台服务器

## 单机安装

官网：<https://clickhouse.yandex/>

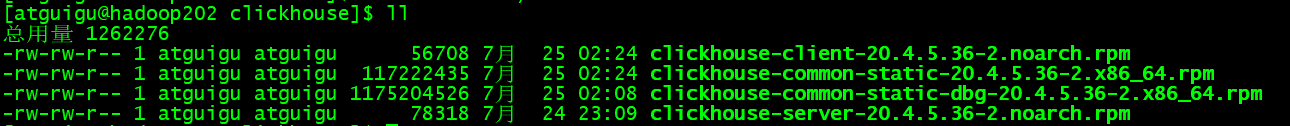
下载地址：<http://repo.red-soft.biz/repos/clickhouse/stable/el6/>

### 在hadoop102的/opt/software下创建clickhouse目录

[atguigu@hadoop102 software]$ mkdir clickhouse



### 将/2.资料/ClickHouse下4个文件上传到hadoop102的software/clickhouse目录下

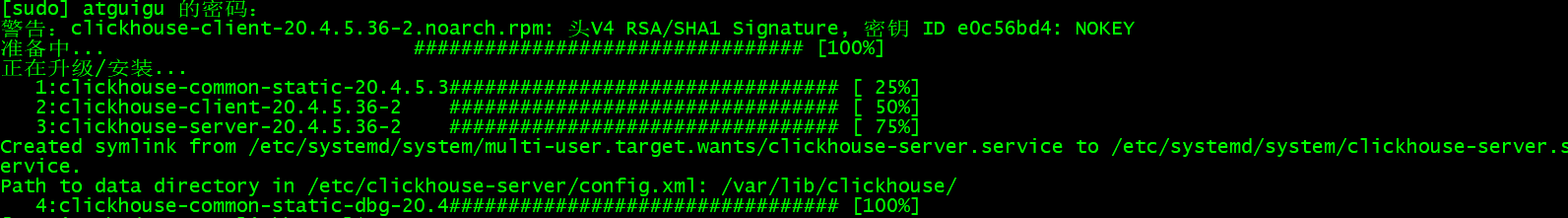


### 将安装文件同步到hadoop103、hadoop104

[atguigu@hadoop102 software]$ xsync clickhouse

### 分别在三台机子上安装这4个rpm文件

[atguigu@hadoop102 clickhouse]$ sudo rpm -ivh \*.rpm

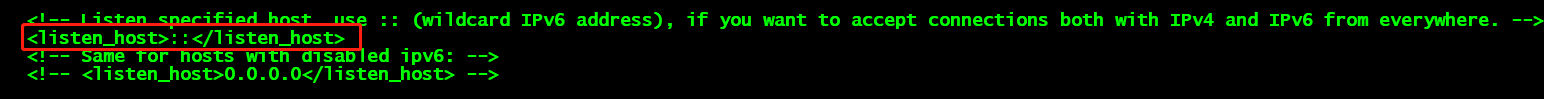


sudo rpm -qa|grep clickhouse查看安装情况

### 修改配置文件

[atguigu@hadoop102 clickhouse]$ sudo vim /etc/clickhouse-server/config.xml

把 **<listen\_host>::</listen\_host>** 的注释打开，这样的话才能让ClickHouse被除本机以外的服务器访问



分发配置文件

sudo /home/atguigu/bin/xsync /etc/clickhouse-server/config.xml

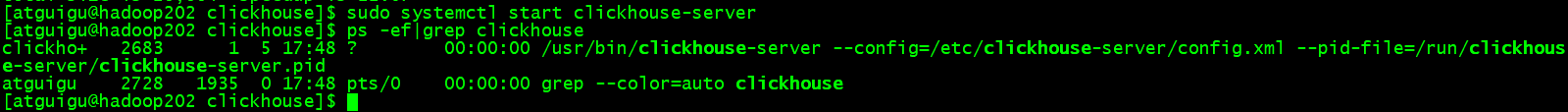
在这个文件中，有ClickHouse的一些默认路径配置，比较重要的

数据文件路径：<path>/var/lib/clickhouse/</path>

日志文件路径：<log>/var/log/clickhouse-server/clickhouse-server.log</log>

### 启动Server

[atguigu@hadoop102 clickhouse]$ sudo systemctl start clickhouse-server

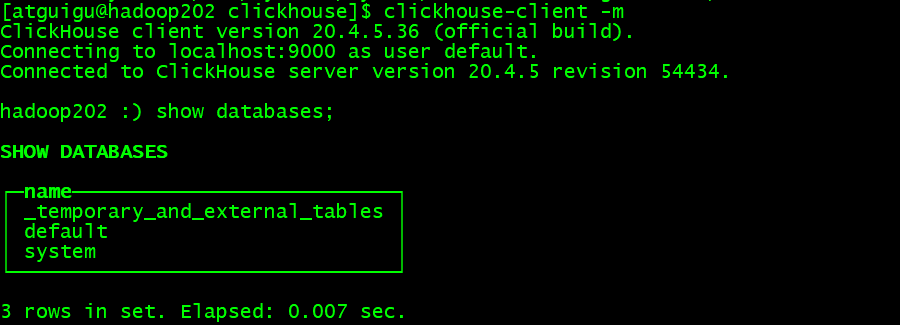


### 三台机器上关闭开机自启

[atguigu@hadoop102 clickhouse]$sudo systemctl disable clickhouse-server

### 使用client连接server

[atguigu@hadoop102 clickhouse]$ clickhouse-client -m



-m :可以在命令窗口输入多行命令

# 数据类型

## 整型

固定长度的整型，包括有符号整型或无符号整型。

整型范围（-2n-1~2n-1-1）：

Int8 - [-128 : 127]

Int16 - [-32768 : 32767]

Int32 - [-2147483648 : 2147483647]

Int64 - [-9223372036854775808 : 9223372036854775807]

无符号整型范围（0~2n-1）：

UInt8 - [0 : 255]

UInt16 - [0 : 65535]

UInt32 - [0 : 4294967295]

UInt64 - [0 : 18446744073709551615]

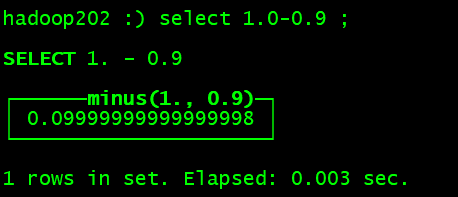
**使用场景： 个数、数量、也可以存储型id。**

## 浮点型

Float32 - float

Float64 – double

建议尽可能以整数形式存储数据。例如，将固定精度的数字转换为整数值，如时间用毫秒为单位表示，因为浮点型进行计算时可能引起四舍五入的误差。



**使用场景：一般数据值比较小，不涉及大量的统计计算，精度要求不高的时候。比如保存商品的重量。**

## 布尔型

没有单独的类型来存储布尔值。可以使用 UInt8 类型，取值限制为 0 或 1。

## Decimal 型

有符号的浮点数，可在加、减和乘法运算过程中保持精度。对于除法，最低有效数字会被丢弃（不舍入）。

有三种声明：

* Decimal32(s)，相当于Decimal(9-s,s)，有效位数为1~9
* Decimal64(s)，相当于Decimal(18-s,s)，有效位数为1~18
* Decimal128(s)，相当于Decimal(38-s,s)，有效位数为1~38

s标识小数位

**使用场景： 一般金额字段、汇率、利率等字段为了保证小数点精度，都使用Decimal进行存储。**

## 字符串

* String

字符串可以任意长度的。它可以包含任意的字节集，包含空字节。

* FixedString(N)

固定长度 N 的字符串，N 必须是严格的正自然数。当服务端读取长度小于 N 的字符串时候，通过在字符串末尾添加空字节来达到 N 字节长度。 当服务端读取长度大于 N 的字符串时候，将返回错误消息。

与String相比，极少会使用FixedString，因为使用起来不是很方便。

**使用场景：名称、文字描述、字符型编码。 固定长度的可以保存一些定长的内容，比如一些编码，性别等但是考虑到一定的变化风险，带来收益不够明显，所以定长字符串使用意义有限。**

## 枚举类型

包括 Enum8 和 Enum16 类型。Enum 保存 'string'= integer 的对应关系。

Enum8 用 'String'= Int8 对描述。

Enum16 用 'String'= Int16 对描述。

* **用法演示**

创建一个带有一个枚举 Enum8('hello' = 1, 'world' = 2) 类型的列

CREATE TABLE t\_enum

(

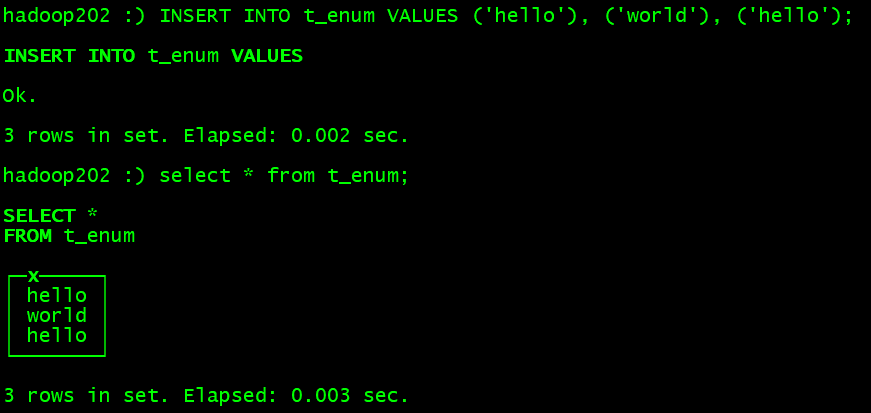
x Enum8('hello' = 1, 'world' = 2)

)

ENGINE = TinyLog;

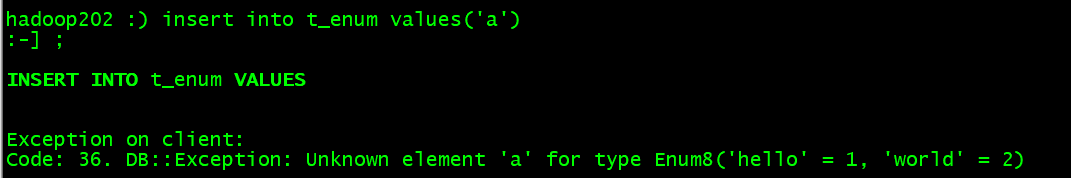
* 这个 x 列只能存储类型定义中列出的值：'hello'或'world'

hadoop102 :) INSERT INTO t\_enum VALUES ('hello'), ('world'), ('hello');



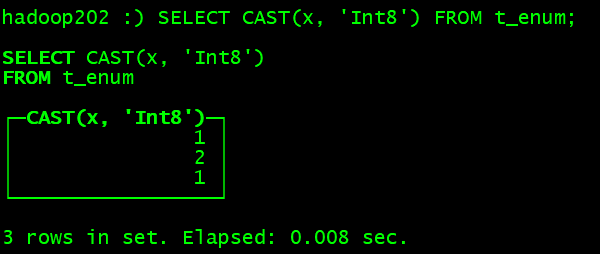
* 如果尝试保存任何其他值，ClickHouse 抛出异常

hadoop102 :) insert into t\_enum values('a')



* 如果需要看到对应行的数值，则必须将 Enum 值转换为整数类型

hadoop102 :) SELECT CAST(x, 'Int8') FROM t\_enum;



**使用场景：对一些状态、类型的字段算是一种空间优化，也算是一种数据约束。但是实际使用中往往因为一些数据内容的变化增加一定的维护成本，甚至是数据丢失问题。所以谨慎使用。**

## 时间类型

目前ClickHouse 有三种时间类型

* + Date接受**年-月-日**的字符串比如 ‘2019-12-16’
* Datetime接受**年-月-日 时:分:秒**的字符串比如 ‘2019-12-16 20:50:10’
* Datetime64接受**年-月-日 时:分:秒.亚秒**的字符串比如‘2019-12-16 20:50:10.66’

日期类型，用两个字节存储，表示从 1970-01-01 (无符号) 到当前的日期值。

还有很多数据结构，可以参考官方文档：<https://clickhouse.yandex/docs/zh/data_types/>

## 数组

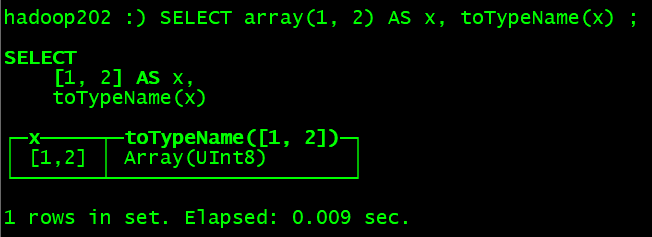
**Array(T)：**由 T 类型元素组成的数组。

T 可以是任意类型，包含数组类型。 但不推荐使用多维数组，ClickHouse 对多维数组的支持有限。例如，不能在MergeTree表中存储多维数组。

* 创建数组方式1，使用array函数

array(T)

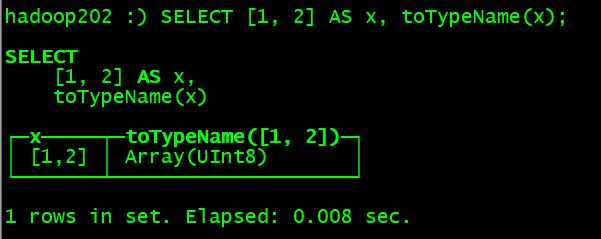
hadoop102 :) SELECT array(1, 2) AS x, toTypeName(x) ;



* 创建数组方式2：使用方括号

[]

hadoop102 :) SELECT [1, 2] AS x, toTypeName(x);



# 表引擎

## 表引擎的使用

表引擎是ClickHouse的一大特色。可以说， 表引擎决定了如何存储表的数据。包括：

* 数据的存储方式和位置，写到哪里以及从哪里读取数据。
* 支持哪些查询以及如何支持。
* 并发数据访问。
* 索引的使用（如果存在）。
* 是否可以执行多线程请求。
* 数据复制参数。

表引擎的使用方式就是必须显式在创建表时定义该表使用的引擎，以及引擎使用的相关参数。

**特别注意：引擎的名称大小写敏感**

## TinyLog

以列文件的形式保存在磁盘上，不支持索引，没有并发控制。一般保存少量数据的小表，生产环境上作用有限。可以用于平时练习测试用。

如：

create table t\_tinylog ( id String, name String) **engine=TinyLog**;

## Memory

内存引擎，数据以未压缩的原始形式直接保存在内存当中，服务器重启数据就会消失。读写操作不会相互阻塞，不支持索引。简单查询下有非常非常高的性能表现（超过10G/s）。

一般用到它的地方不多，除了用来测试，就是在需要非常高的性能，同时数据量又不太大（上限大概 1 亿行）的场景。

## MergeTree

ClickHouse中最强大的表引擎当属MergeTree（合并树）引擎及该系列（\*MergeTree）中的其他引擎，支持索引和分区，地位可以相当于innodb之于Mysql。 而且基于MergeTree，还衍生除了很多小弟，也是非常有特色的引擎。

* **建表语句**

create table t\_order\_mt(

id UInt32,

sku\_id String,

total\_amount Decimal(16,2),

create\_time Datetime

) engine =**MergeTree**

**partition by** toYYYYMMDD(create\_time)

**primary key** (id)

**order** **by** (id,sku\_id);

* **插入数据**

insert into t\_order\_mt values

(101,'sku\_001',1000.00,'2020-06-01 12:00:00') ,

(102,'sku\_002',2000.00,'2020-06-01 11:00:00'),

(102,'sku\_004',2500.00,'2020-06-01 12:00:00'),

(102,'sku\_002',2000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',12000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',600.00,'2020-06-02 12:00:00');

MergeTree其实还有很多参数(绝大多数用默认值即可)，但是三个参数是更加重要的，也涉及了关于MergeTree的很多概念。

### partition by 分区(可选)

* **作用**

学过hive的应该都不陌生，分区的目的主要是降低扫描的范围，优化查询速度

* **如果不填**

只会使用一个分区。

* **分区目录**

MergeTree 是以列文件+索引文件+表定义文件组成的，但是如果设定了分区那么这些文件就会保存到不同的分区目录中。

* **并行**

分区后，面对涉及跨分区的查询统计，ClickHouse会以分区为单位并行处理。

* **数据写入与分区合并**

任何一个批次的数据写入都会产生一个临时分区，不会纳入任何一个已有的分区。写入后的某个时刻（大概10-15分钟后），ClickHouse会自动执行合并操作（等不及也可以手动通过optimize执行），把临时分区的数据，合并到已有分区中。

optimize table xxxx final;

* 例如

再次执行上面的插入操作

insert into t\_order\_mt values

(101,'sku\_001',1000.00,'2020-06-01 12:00:00') ,

(102,'sku\_002',2000.00,'2020-06-01 11:00:00'),

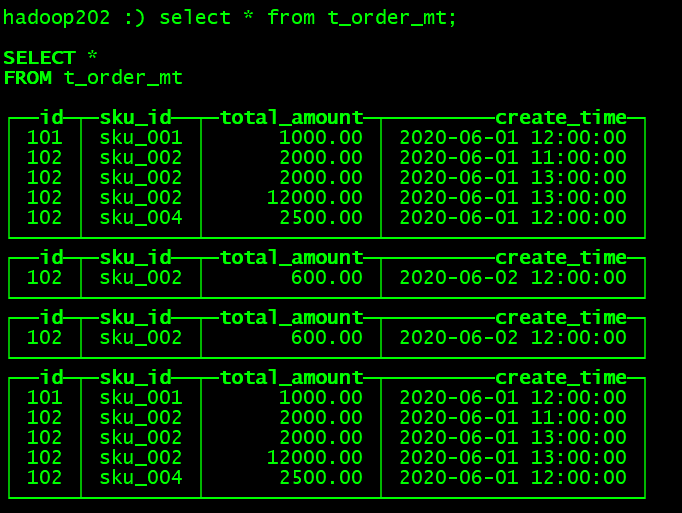
(102,'sku\_004',2500.00,'2020-06-01 12:00:00'),

(102,'sku\_002',2000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',12000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',600.00,'2020-06-02 12:00:00');

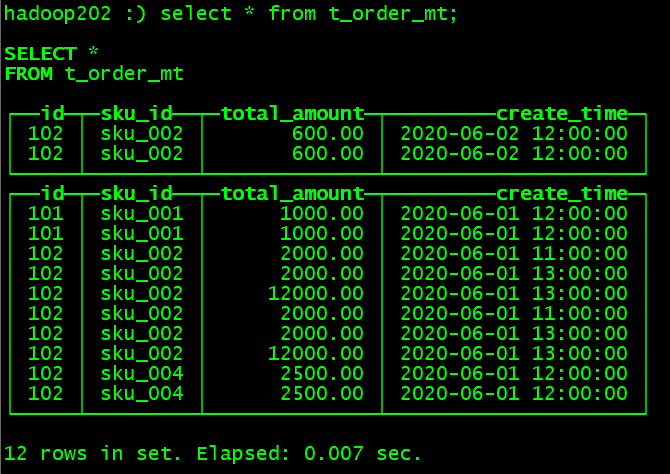
查看数据并没有纳入任何分区



手动optimize之后

hadoop102 :) optimize table t\_order\_mt final;

再次查询



### primary key主键(可选)

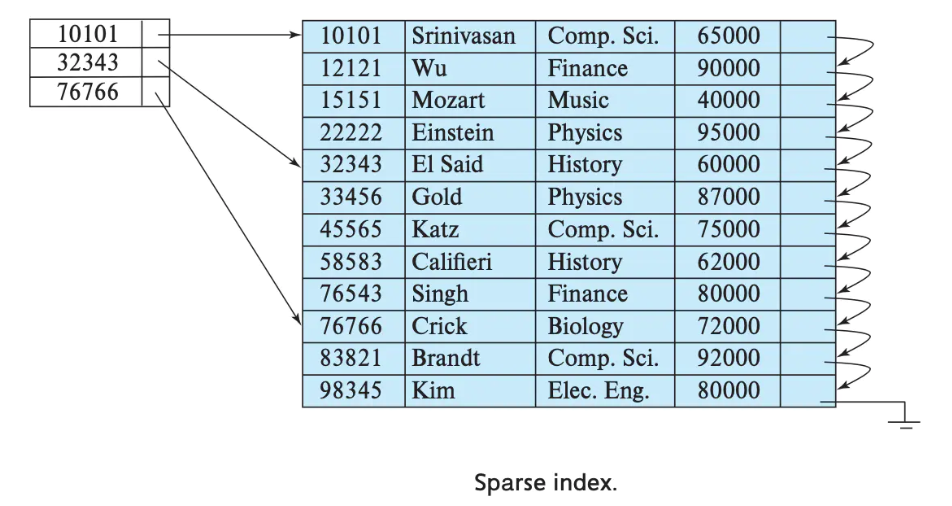
ClickHouse中的主键，和其他数据库不太一样，**它只提供了数据的一级索引，但是却不是唯一约束**。这就意味着是可以存在相同primary key的数据的。

主键的设定主要依据是查询语句中的where 条件。

根据条件通过对主键进行某种形式的二分查找，能够定位到对应的index granularity,避免了全表扫描。·

index granularity： 直接翻译的话就是索引粒度，指在稀疏索引中两个相邻索引对应数据的间隔。ClickHouse中的MergeTree默认是8192。官方不建议修改这个值，除非该列存在大量重复值，比如在一个分区中几万行才有一个不同数据。

稀疏索引：



稀疏索引的好处就是可以用很少的索引数据，定位更多的数据，代价就是只能定位到索引粒度的第一行，然后再进行进行一点扫描。

### order by(必选)

order by 设定了**分区内**的数据按照哪些字段顺序进行有序保存。

order by是MergeTree中唯一一个必填项，甚至比primary key 还重要，因为当用户不设置主键的情况，很多处理会依照order by的字段进行处理（比如后面会讲的去重和汇总）。

**要求：主键必须是order by字段的前缀字段。**

比如order by 字段是 (id,sku\_id) 那么主键必须是id 或者(id,sku\_id)

### 二级索引

目前在ClickHouse的官网上二级索引的功能是被标注为**实验性**的。

#### 使用二级索引前需要增加设置

是否允许使用实验性的二级索引

set allow\_experimental\_data\_skipping\_indices=1;

#### 创建测试表

create table t\_order\_mt2(

id UInt32,

sku\_id String,

total\_amount Decimal(16,2),

create\_time Datetime,

**INDEX a total\_amount TYPE minmax GRANULARITY 5**

) engine =MergeTree

partition by toYYYYMMDD(create\_time)

primary key (id)

order by (id, sku\_id);

其中**GRANULARITY N** 是设定二级索引对于一级索引粒度的粒度。

#### 插入数据

insert into t\_order\_mt2 values

(101,'sku\_001',1000.00,'2020-06-01 12:00:00') ,

(102,'sku\_002',2000.00,'2020-06-01 11:00:00'),

(102,'sku\_004',2500.00,'2020-06-01 12:00:00'),

(102,'sku\_002',2000.00,'2020-06-01 13:00:00'),

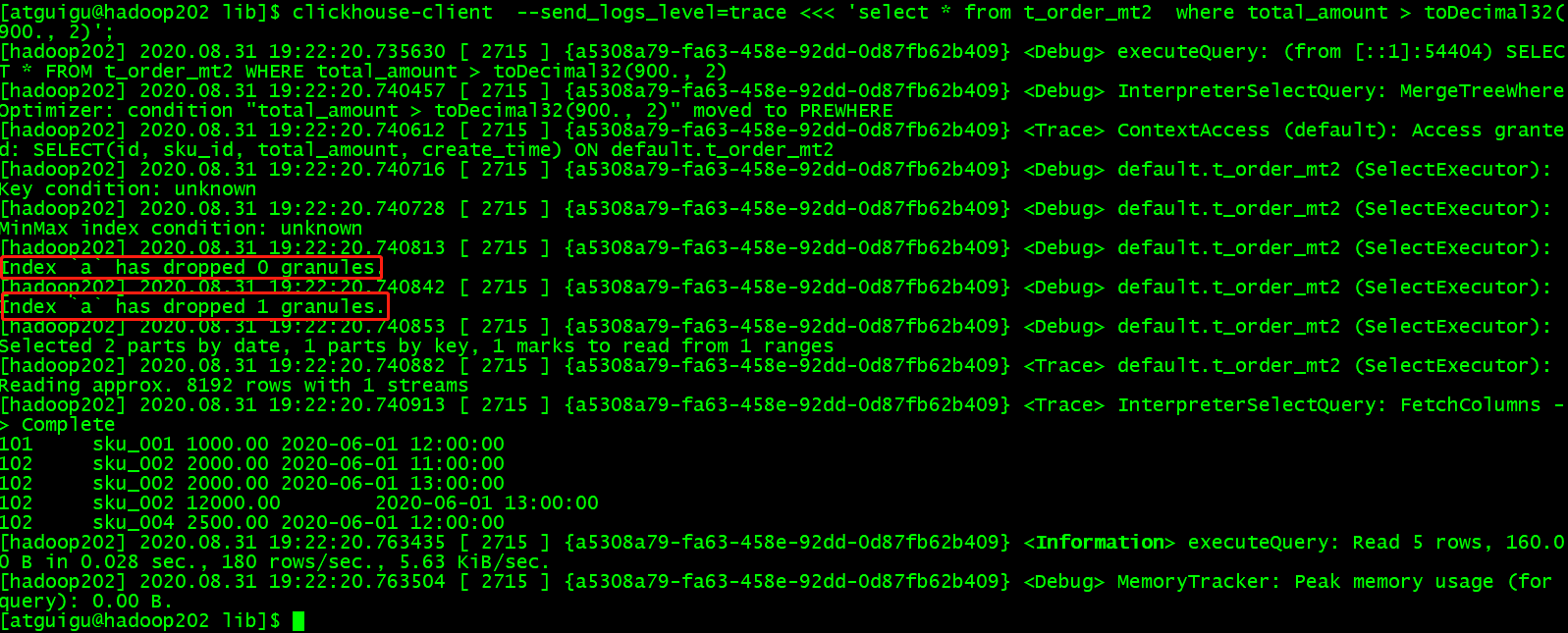
(102,'sku\_002',12000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',600.00,'2020-06-02 12:00:00');

#### 对比效果

那么在使用下面语句进行测试，可以看出二级索引能够为非主键字段的查询发挥作用。

[atguigu@hadoop102 lib]$ clickhouse-client --send\_logs\_level=trace <<< 'select \* from t\_order\_mt2 where total\_amount > toDecimal32(900., 2)';



### 数据TTL

TTL即Time To Live，MergeTree提供了可以管理数据或者列的生命周期的功能。

#### 列级别TTL

* **创建测试表**

create table t\_order\_mt2(

id UInt32,

sku\_id String,

total\_amount Decimal(16,2) TTL create\_time+interval 10 SECOND,

create\_time Datetime

) engine =MergeTree

partition by toYYYYMMDD(create\_time)

primary key (id)

order by (id, sku\_id);

* **插入数据（注意：根据实际时间改变）**

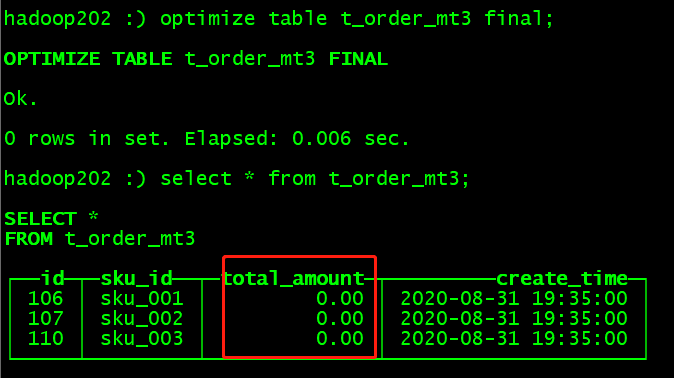
insert into t\_order\_mt2 values

(106,'sku\_001',1000.00,'2020-06-12 22:52:30'),

(107,'sku\_002',2000.00,'2020-06-12 22:52:30'),

(110,'sku\_003',600.00,'2020-06-13 12:00:00');

* 手动合并，查看效果 到期后，指定的字段数据归0



#### 表级TTL

下面的这条语句是数据会在create\_time 之后10秒丢失

alter table t\_order\_mt3 MODIFY TTL create\_time + INTERVAL 10 SECOND;

涉及判断的字段必须是Date或者Datetime类型，推荐使用分区的日期字段。

能够使用的时间周期：

- SECOND

- MINUTE

- HOUR

- DAY

- WEEK

- MONTH

- QUARTER

- YEAR

## ReplacingMergeTree

ReplacingMergeTree是MergeTree的一个变种，它存储特性完全继承MergeTree，只是多了一个去重的功能。 尽管MergeTree可以设置主键，但是primary key其实没有唯一约束的功能。如果你想处理掉重复的数据，可以借助这个ReplacingMergeTree。

* **去重时机**

数据的去重只会在合并的过程中出现。合并会在未知的时间在后台进行，所以你无法预先作出计划。有一些数据可能仍未被处理。

* **去重范围**

如果表经过了分区，去重只会在分区内部进行去重，不能执行跨分区的去重。

所以ReplacingMergeTree能力有限， ReplacingMergeTree 适用于在后台清除重复的数据以节省空间，但是它不保证没有重复的数据出现。

* 案例演示
* **创建表**

create table t\_order\_rmt(

id UInt32,

sku\_id String,

total\_amount Decimal(16,2) ,

create\_time Datetime

) engine = ReplacingMergeTree(create\_time)

partition by toYYYYMMDD(create\_time)

primary key (id)

order by (id, sku\_id);

**ReplacingMergeTree() 填入的参数为版本字段，重复数据保留版本字段值最大的。**

**如果不填版本字段，默认按照插入顺序保留最后一条。**

* **向表中插入数据**

insert into t\_order\_rmt values

(101,'sku\_001',1000.00,'2020-06-01 12:00:00') ,

(102,'sku\_002',2000.00,'2020-06-01 11:00:00'),

(102,'sku\_004',2500.00,'2020-06-01 12:00:00'),

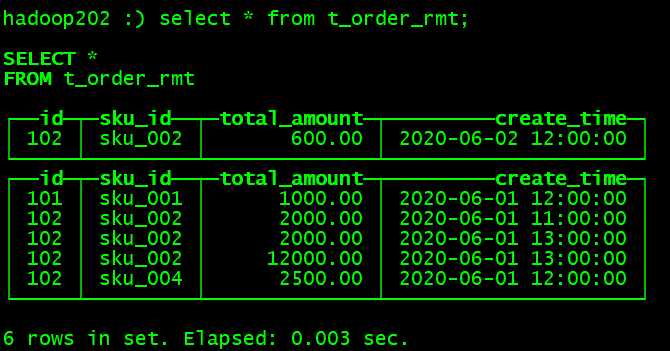
(102,'sku\_002',2000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',12000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',600.00,'2020-06-02 12:00:00');

* **执行第一次查询**

hadoop102 :) select \* from t\_order\_rmt;

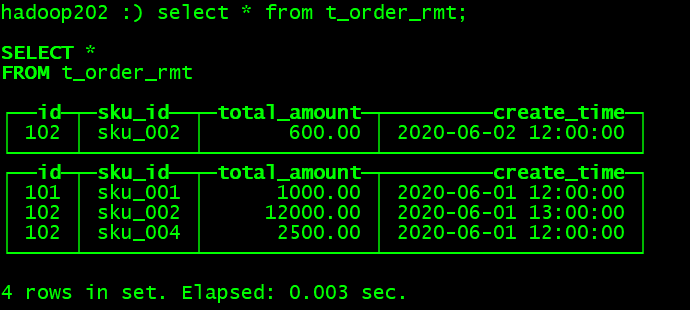


* **手动合并**

OPTIMIZE TABLE t\_order\_rmt FINAL;

* **再执行一次查询**

hadoop102 :) select \* from t\_order\_rmt;



* **通过测试得到结论**
  + 实际上是使用order by 字段作为唯一键
  + 去重不能跨分区
  + 只有合并才会进行去重
  + 认定重复的数据保留，版本字段值最大的
  + 如果版本字段相同则按插入顺序保留最后一笔

## SummingMergeTree

对于不查询明细，只关心以维度进行汇总聚合结果的场景。如果只使用普通的MergeTree的话，无论是存储空间的开销，还是查询时临时聚合的开销都比较大。

ClickHouse 为了这种场景，提供了一种能够“预聚合”的引擎SummingMergeTree

* 案例演示
* **创建表**

create table t\_order\_smt(

id UInt32,

sku\_id String,

total\_amount Decimal(16,2) ,

create\_time Datetime

) engine =SummingMergeTree(total\_amount)

partition by toYYYYMMDD(create\_time)

primary key (id)

order by (id,sku\_id );

* **插入数据**

insert into t\_order\_smt values

(101,'sku\_001',1000.00,'2020-06-01 12:00:00'),

(102,'sku\_002',2000.00,'2020-06-01 11:00:00'),

(102,'sku\_004',2500.00,'2020-06-01 12:00:00'),

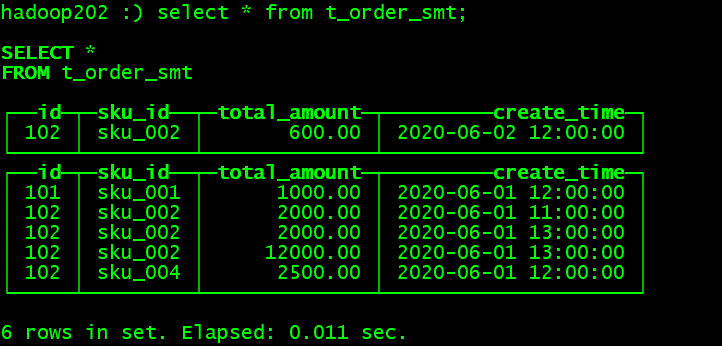
(102,'sku\_002',2000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',12000.00,'2020-06-01 13:00:00'),

(102,'sku\_002',600.00,'2020-06-02 12:00:00');

* **执行第一次查询**

hadoop102 :) select \* from t\_order\_smt;

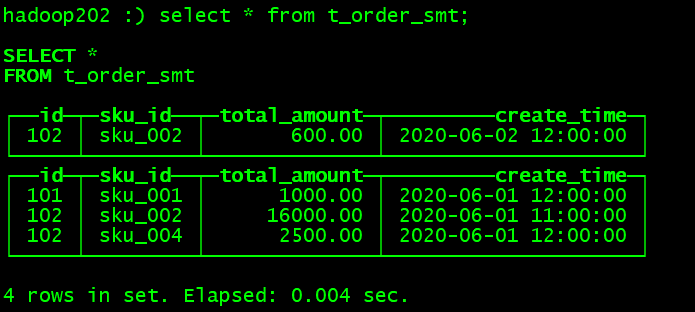


* **手动合并**

OPTIMIZE TABLE t\_order\_smt FINAL;

* **再执行一次查询**

hadoop102 :) select \* from t\_order\_smt;



* **通过结果可以得到以下结论**
  + 以SummingMergeTree（）中指定的列作为汇总数据列
  + 可以填写多列必须数字列，如果不填，以所有非维度列且为数字列的字段为汇总数据列
  + 以order by 的列为准，作为维度列
  + 其他的列按插入顺序保留第一行
  + 不在一个分区的数据不会被聚合
* **开发建议**

设计聚合表的话，唯一键值、流水号可以去掉，所有字段全部是维度、度量或者时间戳。

* **问题**

能不能直接执行以下SQL得到汇总值

select total\_amount from XXX where province\_name=’’ and create\_date=’xxx’

不行，可能会包含一些还没来得及聚合的临时明细

如果要是获取汇总值，还是需要使用sum进行聚合，这样效率会有一定的提高，但本身ClickHouse是列式存储的，效率提升有限，不会特别明显。

select sum(total\_amount) from province\_name=’’ and create\_date=‘xxx’

# SQL操作

基本上来说传统关系型数据库（以MySQL为例）的SQL语句，ClickHouse基本都支持，这里不会从头讲解SQL语法只介绍ClickHouse与标准SQL（MySQL）不一致的地方。

## Insert

基本与标准SQL（MySQL）基本一致

* 标准

insert into [table\_name] values(…),(….)

* 从表到表的插入

insert into [table\_name] select a,b,c from [table\_name\_2]

## Update 和 Delete

ClickHouse提供了Delete和Update的能力，这类操作被称为Mutation查询，它可以看做Alter 的一种。

虽然可以实现修改和删除，但是和一般的OLTP数据库不一样，**Mutation语句是一种很“重”的操作，而且不支持事务**。

“重”的原因主要是每次修改或者删除都会导致放弃目标数据的原有分区，重建新分区。所以尽量做批量的变更，不要进行频繁小数据的操作。

* **删除操作**

alter table t\_order\_smt delete where sku\_id ='sku\_001';

* **修改操作**

alter table t\_order\_smt update total\_amount=toDecimal32(2000.00,2) where id =102;

由于操作比较“重”，所以 Mutation语句分两步执行，同步执行的部分其实只是进行新增数据新增分区和并把旧分区打上逻辑上的失效标记。直到触发分区合并的时候，才会删除旧数据释放磁盘空间，一般不会开放这样的功能给用户，由管理员完成。

## 查询操作

ClickHouse基本上与标准SQL 差别不大

* 支持子查询
* 支持CTE(Common Table Expression 公用表表达式 with 子句)
* 支持各种JOIN， 但是JOIN操作无法使用缓存，所以即使是两次相同的JOIN语句，ClickHouse也会视为两条新SQL
* 窗口函数(官方正在测试中...)
* 不支持自定义函数
* GROUP BY 操作增加了 with rollup\with cube\with total 用来计算小计和总计。
* 插入数据

hadoop102 :) alter table t\_order\_mt delete where 1=1;

insert into t\_order\_mt values

(101,'sku\_001',1000.00,'2020-06-01 12:00:00'),

(101,'sku\_002',2000.00,'2020-06-01 12:00:00'),

(103,'sku\_004',2500.00,'2020-06-01 12:00:00'),

(104,'sku\_002',2000.00,'2020-06-01 12:00:00'),

(105,'sku\_003',600.00,'2020-06-02 12:00:00'),

(106,'sku\_001',1000.00,'2020-06-04 12:00:00'),

(107,'sku\_002',2000.00,'2020-06-04 12:00:00'),

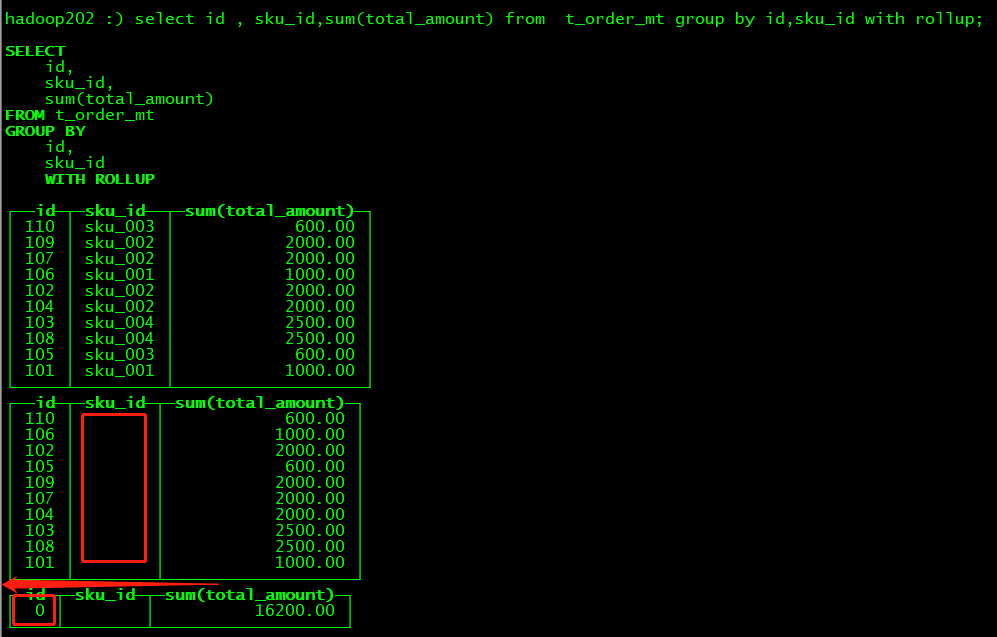
(108,'sku\_004',2500.00,'2020-06-04 12:00:00'),

(109,'sku\_002',2000.00,'2020-06-04 12:00:00'),

(110,'sku\_003',600.00,'2020-06-01 12:00:00');

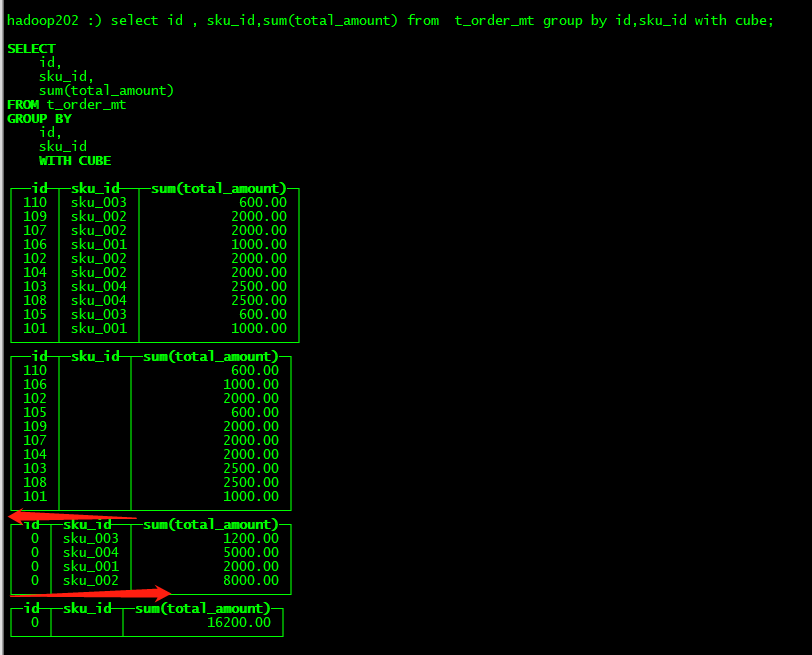
* with rollup：**从右至左**去掉维度进行小计

hadoop102 :) select id,sku\_id,sum(total\_amount) from t\_order\_mt group by id,sku\_id with rollup;



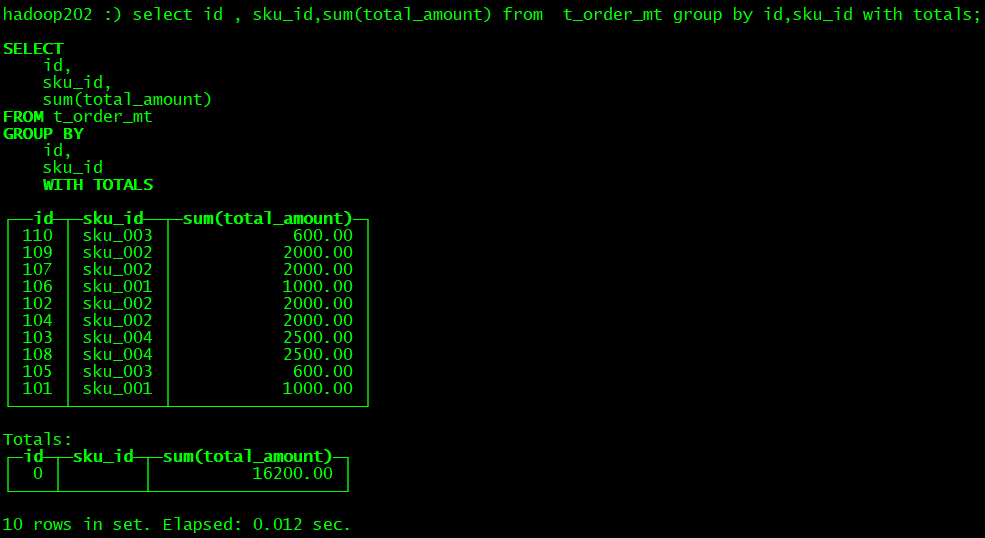
* with cube : 从右至左去掉维度进行小计，再从左至右去掉维度进行小计

hadoop102 :) select id , sku\_id,sum(total\_amount) from t\_order\_mt group by id,sku\_id with cube;



* with totals: 只计算合计

hadoop102 :) select id,sku\_id,sum(total\_amount) from t\_order\_mt group by id,sku\_id with totals;



## alter操作

同MySQL的修改字段基本一致

* **新增字段**

alter table **tableName** add column **newcolname** **String** after **col1;**

* **修改字段类型**

alter table tableName modify column newcolname String;

* **删除字段**

alter table tableName drop column newcolname;

## 导出数据

clickhouse-client --query "select \* from t\_order\_mt where create\_time='2020-06-01 12:00:00'" --format CSVWithNames> /opt/module/data/rs1.csv1

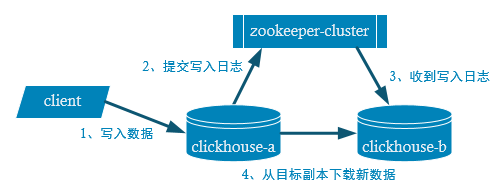
更多支持格式参照：

<https://clickhouse.tech/docs/v19.14/en/interfaces/formats/#csvwithnames>

# 副本

副本的目的主要是保障数据的高可用性，即使一台ClickHouse节点宕机，那么也可以从其他服务器获得相同的数据。

## 副本写入流程



## 配置步骤

* 启动zookeeper集群
* 在hadoop102的/etc/clickhouse-server/config.d目录下创建一个名为metrika.xml的配置文件,内容如下：

<?xml version="1.0"?>

<yandex>

<zookeeper-servers>

<node index="1">

<host>hadoop102</host>

<port>2181</port>

</node>

<node index="2">

<host>hadoop103</host>

<port>2181</port>

</node>

<node index="3">

<host>hadoop104</host>

<port>2181</port>

</node>

</zookeeper-servers>

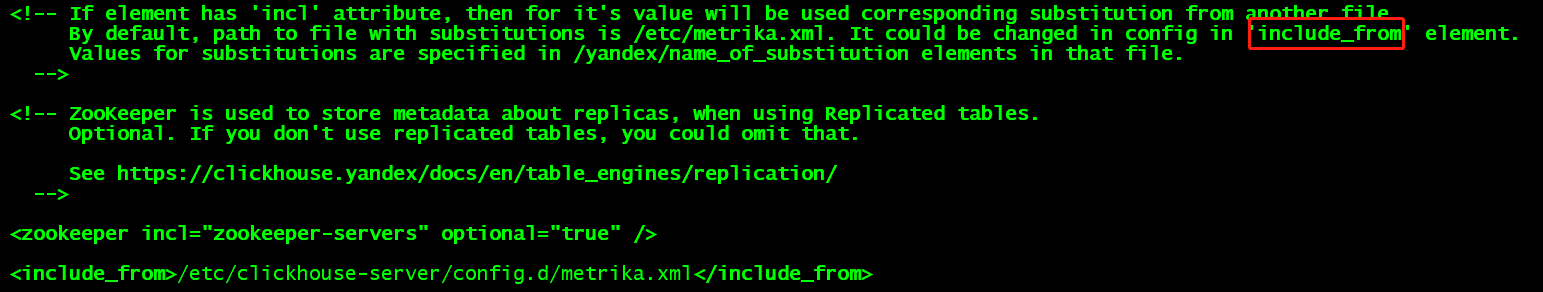
</yandex>

* 同步到hadoop103和hadoop104上

sudo /home/atguigu/bin/xsync /etc/clickhouse-server/config.d/metrika.xml

* 在 hadoop102的/etc/clickhouse-server/config.xml中增加

<include\_from>/etc/clickhouse-server/config.d/metrika.xml</include\_from>



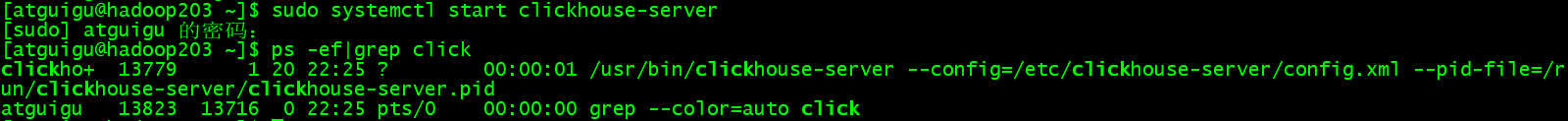
* 同步到hadoop103和hadoop104上

sudo /home/atguigu/bin/xsync /etc/clickhouse-server/config.xml

* 分别在hadoop102和hadoop103上启动ClickHouse服务

注意：因为修改了配置文件，如果以前启动了服务需要重启

[atguigu@hadoop102|3 ~]$ sudo systemctl start clickhouse-server



注意：我们演示副本操作只需要在hadoop102和hadoop103两台服务器即可，上面的操作，我们hadoop104可以你不用同步，我们这里为了保证集群中资源的一致性，做了同步。

* 在hadoop102和hadoop103上分别建表

**副本只能同步数据，不能同步表结构，所以我们需要在每台机器上自己手动建表**

* hadoop102

create table t\_order\_rep2 (

id UInt32,

sku\_id String,

total\_amount Decimal(16,2),

create\_time Datetime

) engine =ReplicatedMergeTree('/clickhouse0225/table/01/t\_order\_rep','rep\_102')

partition by toYYYYMMDD(create\_time)

primary key (id)

order by (id,sku\_id);

* hadoop103

create table t\_order\_rep2 (

id UInt32,

sku\_id String,

total\_amount Decimal(16,2),

create\_time Datetime

) engine =ReplicatedMergeTree('/clickhouse0225/table/01/t\_order\_rep','rep\_103')

partition by toYYYYMMDD(create\_time)

primary key (id)

order by (id,sku\_id);

* 参数解释

ReplicatedMergeTree 中，

**第一个参数是**分片的zk\_path一般按照： /clickhouse/table/{shard}/{table\_name} 的格式写，如果只有一个分片就写01即可。

**第二个参数是**副本名称，相同的分片副本名称不能相同。

* 在hadoop102上执行insert语句

insert into t\_order\_rep2 values

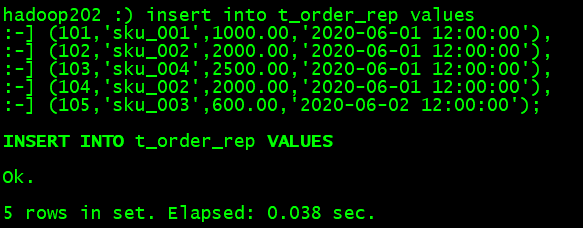
(101,'sku\_001',1000.00,'2020-06-01 12:00:00'),

(102,'sku\_002',2000.00,'2020-06-01 12:00:00'),

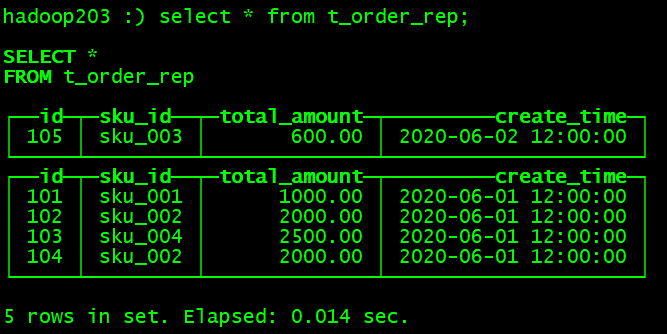
(103,'sku\_004',2500.00,'2020-06-01 12:00:00'),

(104,'sku\_002',2000.00,'2020-06-01 12:00:00'),

(105,'sku\_003',600.00,'2020-06-02 12:00:00');



* 在hadoop103上执行select，可以查询出结果，说明副本配置正确



# 分片集群

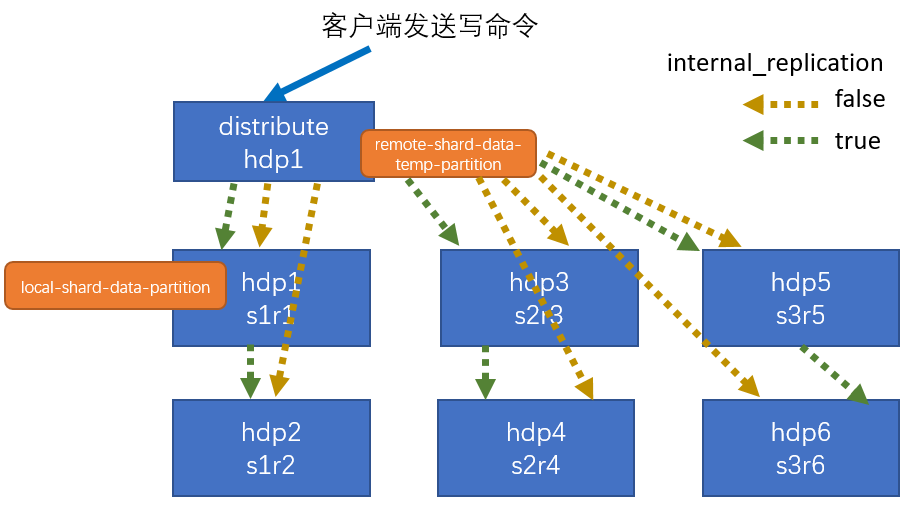
副本虽然能够提高数据的可用性，降低丢失风险，但是每台服务器实际上必须容纳全量数据，对数据的横向扩容没有解决。

要解决数据水平切分的问题，需要引入分片的概念。通过分片把一份完整的数据进行切分，不同的分片分布到不同的节点上，再通过Distributed表引擎把数据拼接起来一同使用。

Distributed表引擎本身不存储数据，有点类似于MyCat之于MySql，成为一种中间件，通过分布式逻辑表来写入、分发、路由来操作多台节点不同分片的分布式数据。

注意：ClickHouse的集群是表级别的，实际企业中，大部分做了高可用，但是没有用分片，避免降低查询性能以及操作集群的复杂性。

## 集群写入流程（3分片2副本共6个节点）



## 集群读取流程（3分片2副本共6个节点）

## 3分片2副本共6个节点集群配置（供参考）

配置的位置还是在之前的/etc/clickhouse-server/config.d/metrika-shard.xml，内容如下

<yandex>

<clickhouse\_remote\_servers>

<gmall\_cluster> <!-- 集群名称-->

<shard> <!--集群的第一个分片-->

<internal\_replication>true</internal\_replication>

<!--该分片的第一个副本-->

<replica>

<host>hadoop101</host>

<port>9000</port>

</replica>

<!--该分片的第二个副本-->

<replica>

<host>hadoop102</host>

<port>9000</port>

</replica>

</shard>

<shard> <!--集群的第二个分片-->

<internal\_replication>true</internal\_replication>

<replica> <!--该分片的第一个副本-->

<host>hadoop103</host>

<port>9000</port>

</replica>

<replica> <!--该分片的第二个副本-->

<host>hadoop104</host>

<port>9000</port>

</replica>

</shard>

<shard> <!--集群的第三个分片-->

<internal\_replication>true</internal\_replication>

<replica> <!--该分片的第一个副本-->

<host>hadoop105</host>

<port>9000</port>

</replica>

<replica> <!--该分片的第二个副本-->

<host>hadoop106</host>

<port>9000</port>

</replica>

</shard>

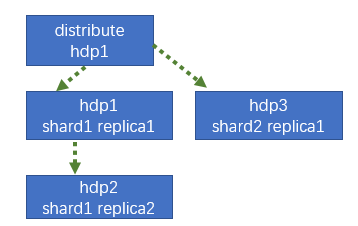
</gmall\_cluster>

</clickhouse\_remote\_servers>

</yandex>

## 配置三节点版本集群及副本

### 集群及副本规划（2个分片，只有第一个分片有副本）



|  |  |  |
| --- | --- | --- |
| **hadoop102** | **hadoop103** | **hadoop104** |
| <macros>  <shard>01</shard>  <replica>rep\_1\_1</replica>  </macros> | <macros>  <shard>01</shard>  <replica>rep\_1\_2</replica>  </macros> | <macros>  <shard>02</shard>  <replica>rep\_2\_1</replica>  </macros> |

### 配置步骤

#### 在hadoop102的/etc/clickhouse-server/config.d目录下创建metrika-shard.xml文件

<?xml version="1.0"?>

<yandex>

<clickhouse\_remote\_servers>

<gmall\_cluster> <!-- 集群名称-->

<shard> <!--集群的第一个分片-->

<internal\_replication>true</internal\_replication>

<replica> <!--该分片的第一个副本-->

<host>hadoop102</host>

<port>9000</port>

</replica>

<replica> <!--该分片的第二个副本-->

<host>hadoop103</host>

<port>9000</port>

</replica>

</shard>

<shard> <!--集群的第二个分片-->

<internal\_replication>true</internal\_replication>

<replica> <!--该分片的第一个副本-->

<host>hadoop104</host>

<port>9000</port>

</replica>

</shard>

</gmall\_cluster>

</clickhouse\_remote\_servers>

<zookeeper-servers>

<node index="1">

<host>hadoop102</host>

<port>2181</port>

</node>

<node index="2">

<host>hadoop103</host>

<port>2181</port>

</node>

<node index="3">

<host>hadoop104</host>

<port>2181</port>

</node>

</zookeeper-servers>

<macros>

<shard>01</shard> <!--不同机器放的分片数不一样-->

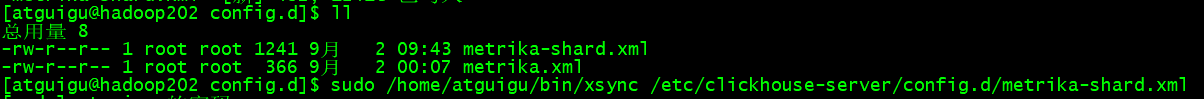
<replica>rep\_1\_1</replica> <!--不同机器放的副本数不一样-->

</macros>

</yandex>

#### 将hadoop102的metrika-shard.xml同步到103和104

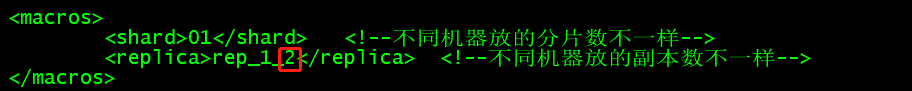
sudo /home/atguigu/bin/xsync /etc/clickhouse-server/config.d/metrika-shard.xml



#### 修改103和104中metrika-shard.xml宏的配置

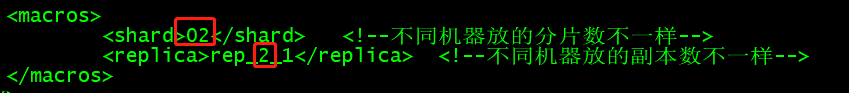
* **103**

[atguigu@hadoop103 ~]$ sudo vim /etc/clickhouse-server/config.d/metrika-shard.xml

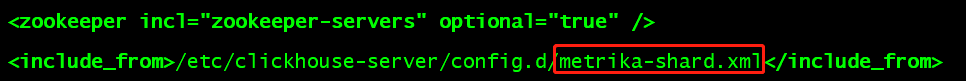


* **104**

[atguigu@hadoop104 ~]$ sudo vim /etc/clickhouse-server/config.d/metrika-shard.xml



#### 在hadoop102上修改/etc/clickhouse-server/config.xml



#### 同步/etc/clickhouse-server/config.xml到103和104

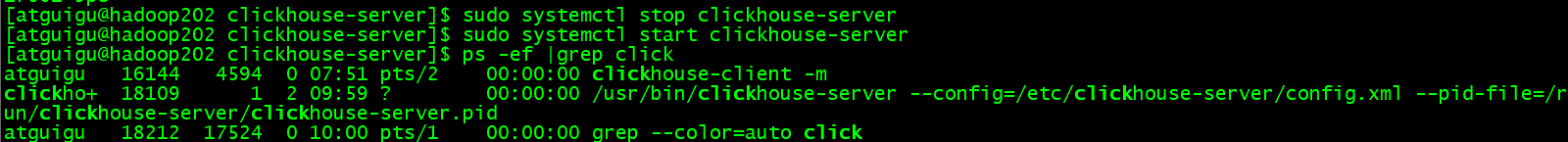
[atguigu@hadoop102 ~]$ sudo /home/atguigu/bin/xsync /etc/clickhouse-server/config.xml

#### 重启三台服务器上的ClickHouse服务

[atguigu@hadoop102 clickhouse-server]$ sudo systemctl stop clickhouse-server

[atguigu@hadoop102 clickhouse-server]$ sudo systemctl start clickhouse-server

[atguigu@hadoop102 clickhouse-server]$ ps -ef |grep click



#### 在hadoop102上执行建表语句

* 会自动同步到hadoop103和hadoop104上
* 集群名字要和配置文件中的一致
* 分片和副本名称从配置文件的宏定义中获取

create table st\_order\_mt on cluster gmall\_cluster (

id UInt32,

sku\_id String,

total\_amount Decimal(16,2),

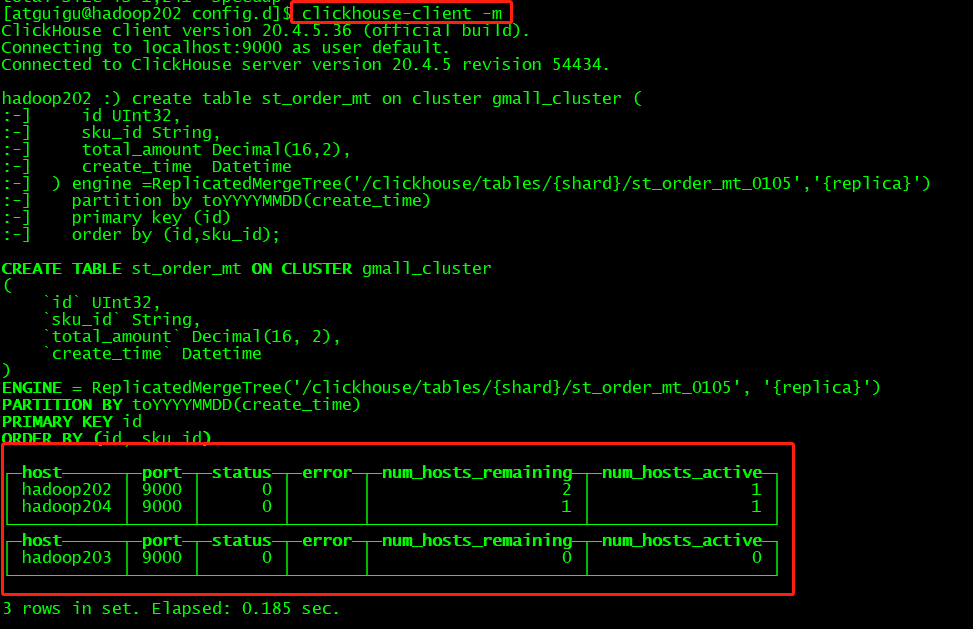
create\_time Datetime

) engine =ReplicatedMergeTree('/clickhouse/tables/{shard}/st\_order\_mt\_0225','{replica}')

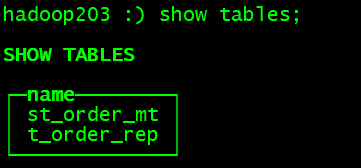
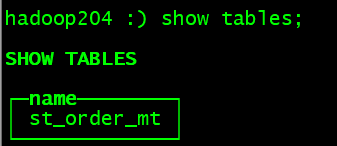
partition by toYYYYMMDD(create\_time)

primary key (id)

order by (id,sku\_id);



可以到hadoop103和hadoop104上查看表是否创建成功

#### 在hadoop102上创建Distribute 分布式表

create table st\_order\_mt\_all on cluster gmall\_cluster

(

id UInt32,

sku\_id String,

total\_amount Decimal(16,2),

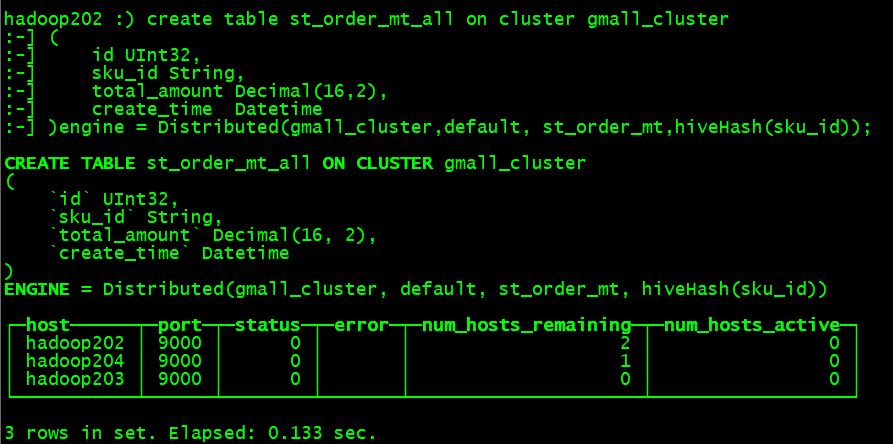
create\_time Datetime

)engine = Distributed(gmall\_cluster,default, st\_order\_mt,hiveHash(sku\_id));

**参数含义**

Distributed(集群名称，库名，本地表名，分片键)

分片键必须是整型数字，所以用hiveHash函数转换，也可以rand()



#### 在hadoop102上插入测试数据

insert into st\_order\_mt\_all values

(201,'sku\_001',1000.00,'2020-06-01 12:00:00') ,

(202,'sku\_002',2000.00,'2020-06-01 12:00:00'),

(203,'sku\_004',2500.00,'2020-06-01 12:00:00'),

(204,'sku\_002',2000.00,'2020-06-01 12:00:00'),

(205,'sku\_003',600.00,'2020-06-02 12:00:00');

#### 通过查询分布式表和本地表观察输出结果

* **分布式表**

SELECT \* FROM st\_order\_mt\_all;

* **本地表**

select \* from st\_order\_mt;

* 观察数据的分布

|  |  |
| --- | --- |
| st\_order\_mt\_all |  |
| hadoop102:  st\_order\_mt |  |
| hadoop103:  st\_order\_mt |  |
| hadoop104:  st\_order\_mt |  |

## 项目为了节省资源，就使用单节点，不用集群

不需要求改文件引用，因为已经使用集群建表了，如果改为引用metrika-shard.xml的话，启动会报错。我们以后用的时候只启动102即可。