CarbonData Partition 功能介绍

与

上汽集团 CarbonData实践分享

上汽集团
SAIC MOTOR

曹　鲁

2017.9.2

关于我

曹　鲁

2010年毕业于武汉大学计算机学院

曾负责某金融行业公司BI、ETL系统开发，某互联网电商行业公司数据仓库的容量管理、性能调优等

2016年加入上汽集团数据业务部
负责大数据平台架构设计与开发

关注大数据技术与开源社区

Mail：caolu@saicmotor.com

上汽集团
SAIC MOTOR

上汽集团数据业务部

主要负责

- 规划和实施数据管理体系
- 建设大数据基础架构和分析平台
- 拓展和提升集团内企业数据业务能力
- 推动人工智能技术在集团业务中的应用

为上汽集团战略转型和创新发展提供支持和服务。

APACHE CarbonData

上汽集团
SAIC MOTOR

议题

CarbonData Partition 功能介绍

- CarbonData背景、文件结构、建索引过程简介
- Partition功能将带来的改变
- PartitionTable的建表语句
- PartitionTable的数据加载过程
- PartitionTable的查询过程
- Partition 的新增(add)、拆分(split)及删除(drop)

上汽集团的CarbonData实践

APACHE CarbonData

CarbonData Partition 功能介绍

上汽集团 SAIC MOTOR

Apache CarbonData

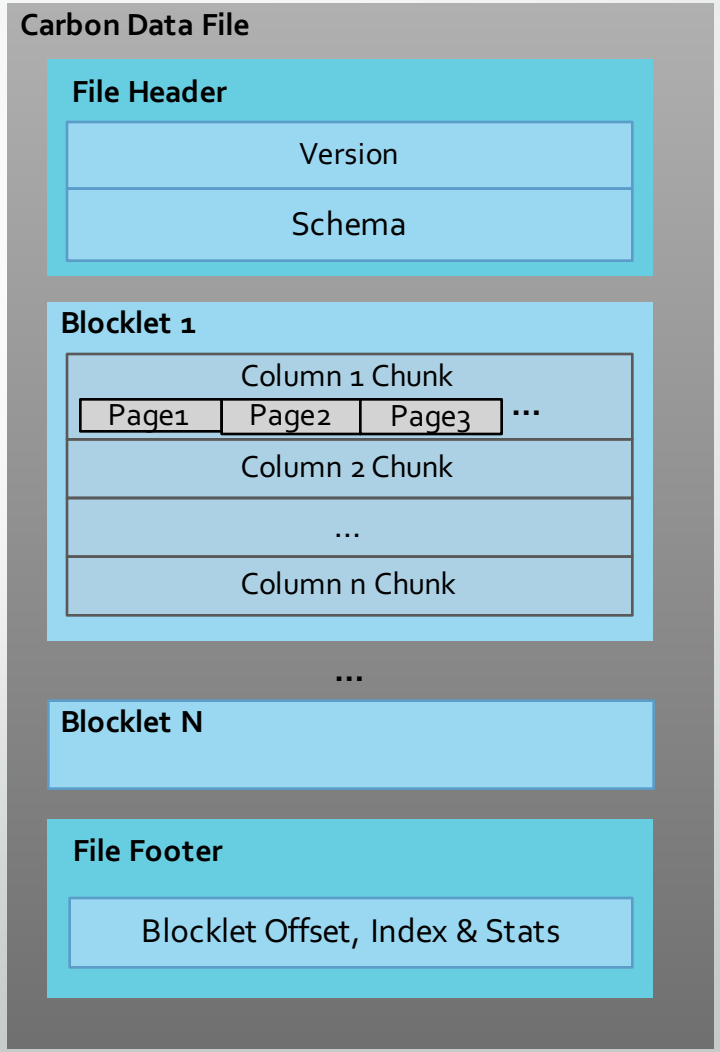
是大数据平台上一种带索引的列式存储数据格式

- 2016年6月由华为公司贡献至apache基金会，成为apache孵化项目
- 2017年4月正式毕业为apache顶级项目

APACHE CarbonData

CarbonData Partition 功能介绍 —— 文件结构简介

- Blocklet: 文件内的数据块
- Column chunk: Blocklet内的列数据
- 元数据和索引信息
  - Header：Version，Schema
  - Footer：Blocklet Offset, Index & statistics



CarbonData Partition 功能介绍 —— 索引生成过程

字典编码 + 排序

MDK => Multi Dimension Key

Years	Quarters	Months	Territory	Country	Quantity	Sales
2003	QTR1	Jan	EMEA	Germany	142	11,432
2003	QTR1	Jan	APAC	China	541	54,702
2003	QTR1	Jan	EMEA	Spain	443	44,622
2003	QTR1	Feb	EMEA	Denmark	545	58,871
2003	QTR1	Feb	EMEA	Italy	675	56,181
2003	QTR1	Mar	APAC	India	52	9,749
2003	QTR1	Mar	EMEA	UK	570	51,018
2003	QTR1	Mar	Japan	Japan	561	55,245
2003	QTR2	Apr	APAC	Australia	525	50,398
2003	QTR2	Apr	EMEA	Germany	144	11,532

Encoding

[1,1,1,1,1] : [142,11432]
[1,1,1,3,2] : [541,54702]
[1,1,1,1,3] : [443,44622]
[1,1,2,1,4] : [545,58871]
[1,1,2,1,5] : [675,56181]
[1,1,3,3,6] : [52,9749]
[1,1,3,1,7] : [570,51018]
[1,1,3,2,8] : [561,55245]
[1,2,4,3,9] : [525,50398]
[1,2,4,1,1] : [144,11532]

Sort
(MDK Index)

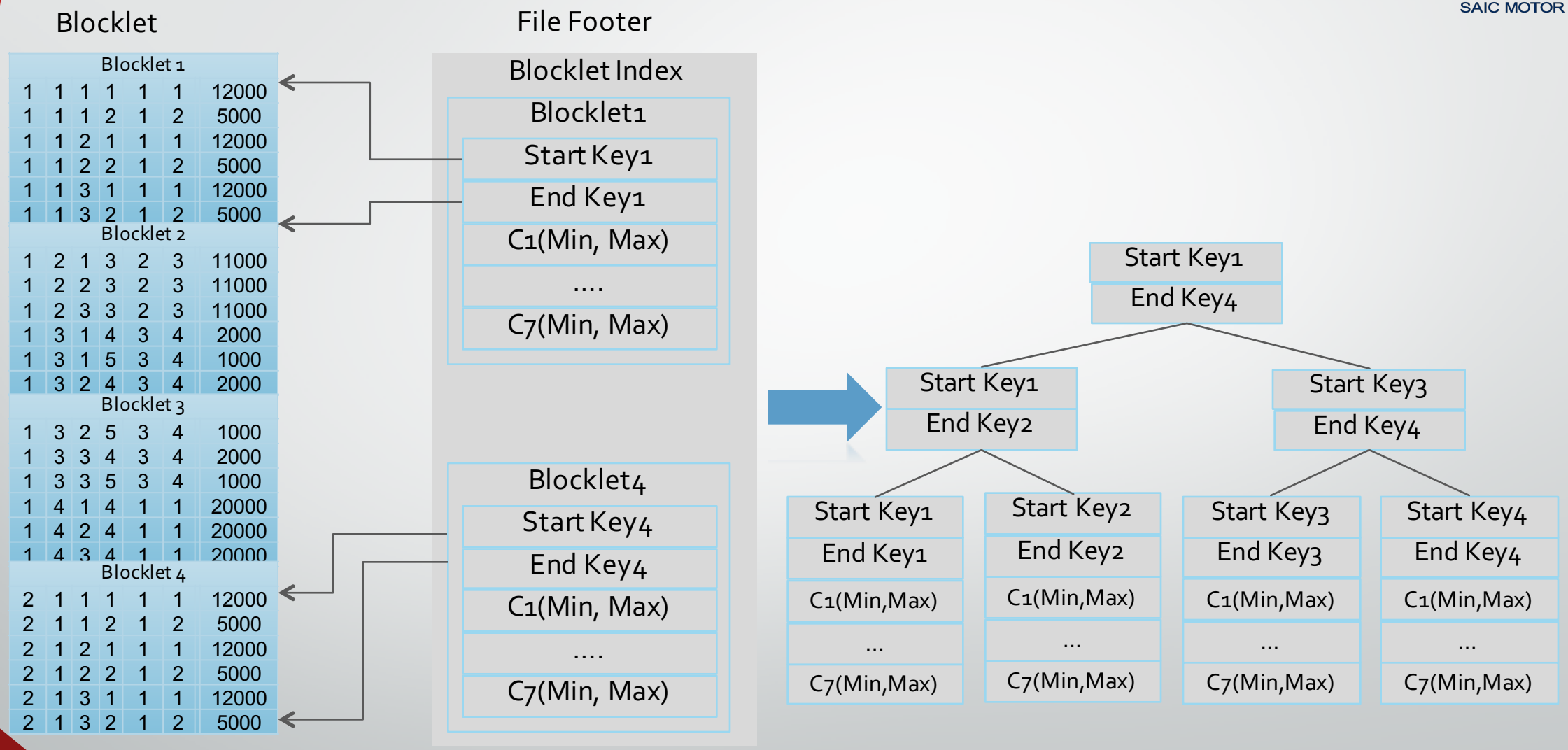
Sorted MDK Index

[1,1,1,1,1] : [142,11432]
[1,1,1,1,3] : [443,44622]
[1,1,1,3,2] : [541,54702]
[1,1,2,1,4] : [545,58871]
[1,1,2,1,5] : [675,56181]
[1,1,3,1,7] : [570,51018]
[1,1,3,2,8] : [561,55245]
[1,2,4,3,9] : [525,50398]
[1,2,4,1,1] : [144,11532]

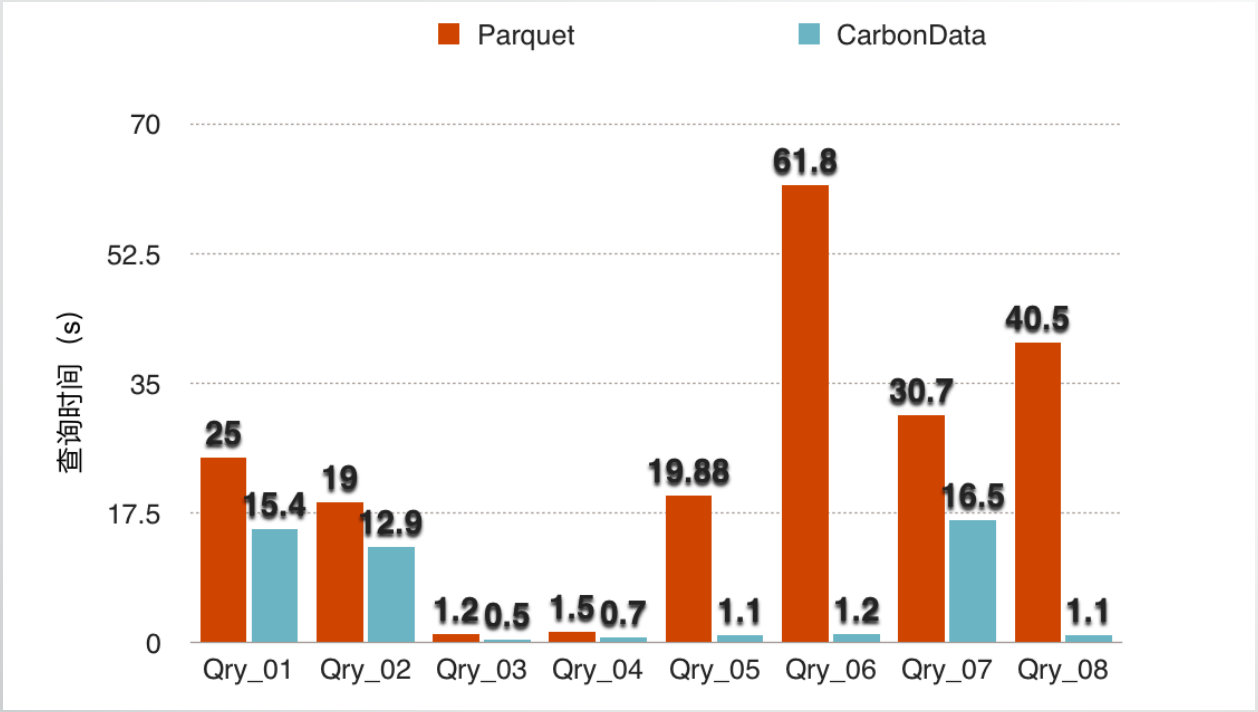
Blocklet (Columnar view)

C1	C2	C3	C4	C5	C6	C7
1	1	1	1	1	142	11432
1	1	1	1	3	443	44622
1	1	1	3	2	541	54702
1	1	2	1	4	545	58871
1	1	2	1	5	675	56181
1	1	3	1	7	570	51018
1	1	3	2	8	561	55245
1	1	3	3	6	52	9749
1	2	4	1	1	144	11532
1	2	4	3	9	525	50398

CarbonData Partition 功能介绍 —— 索引生成过程



CarbonData Partition 功能介绍——初次性能测试



相对较慢的查询场景：

- 没有排序维度过滤条件的聚合查询，即无法使用Btree索引
- 有排序维度列作为过滤条件，但聚合列在MDKey中排序较为靠后的聚合查询

CarbonData对比Parquet查询性能测试（Dec. 2016, CarbonData 0.2.0）

CarbonData Partition 功能介绍

为什么我们要做Partition功能?

or

Partition功能将带来哪些改变?

CarbonData Partition 功能介绍——Partition带来的改变

上汽集团 SAIC MOTOR

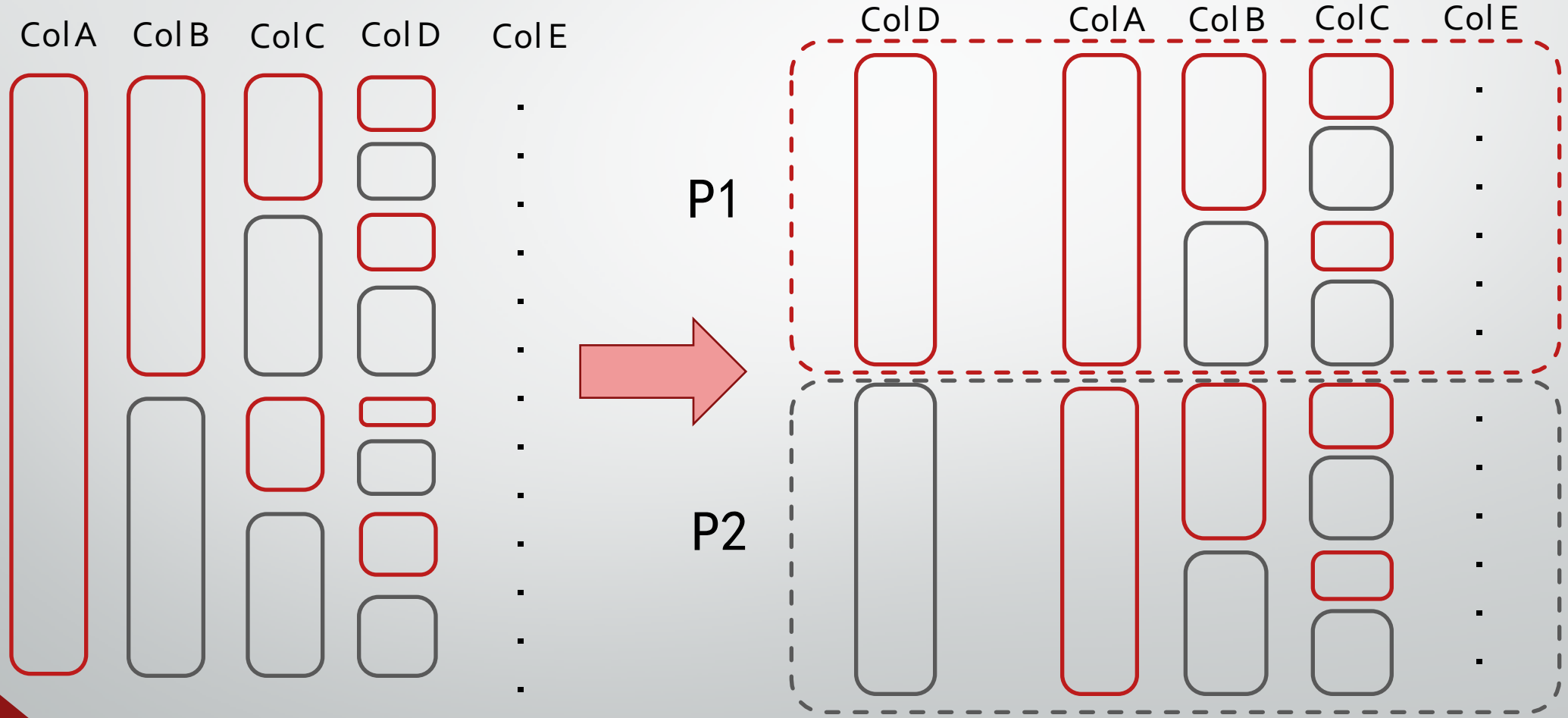
改变1：数据将基于Partition列更为集中存储，查询时可过滤掉大量block，减少spark task数量

APACHE CarbonData

CarbonData Partition 功能介绍——Partition带来的改变

上汽集团 SAIC MOTOR

改变2：可以使其他列在排序中更靠前，提升查询性能



APACHE CarbonData

CarbonData Partition 功能介绍

目前支持三种类型Partition Table：

- Hash Partition
- Range Partition
- List Partition
- Range Interval（Developing）

CarbonData Partition 功能介绍

上汽集团 SAIC MOTOR

Hash Partition DDL Syntax

```
CREATE TABLE [IF NOT EXISTS] [db_name.]table_name
[(col_name data_type , ...)]
PARTITIONED BY (partition_col_name data_type)
STORED BY 'carbondata'
[TBLPROPERTIES ('PARTITION_TYPE'='HASH',
'PARTITION_NUM'='N' ...)]
//N is the number of hash partitions
```

Example

```
create table if not exists hash_partition_table(
col_A String,
col_B Int,
col_C Long,
col_D Decimal(10,2),
col_F Timestamp )
partitioned by (col_E Long)
stored by 'carbondata'
tblproperties('partition_type'='Hash','partition_num'='9')
```

APACHE CarbonData

CarbonData Partition 功能介绍

上汽集团 SAIC MOTOR

Range Partition DDL Syntax

```
CREATE TABLE [IF NOT EXISTS] [db_name.]table_name
[(col_name data_type , ...)]
PARTITIONED BY (partition_col_name data_type)
STORED BY 'carbondata'
[TBLPROPERTIES ('PARTITION_TYPE'='RANGE',
'RANGE_INFO'='2014-01-01, 2015-01-01, 2016-01-01' ...)]
```

Example

```
create table if not exists range_partition_table(
col_A String,
col_B Int,
col_C Long,
col_D Decimal(10,2),
col_E Long )
partitioned by (col_F Timestamp)
stored by 'carbondata'
tblproperties('partition_type'='Range',
'range_info'='2015-01-01, 2016-01-01, 2017-01-01, 2017-02-01')
```

RangeInfo 必须为升序

APACHE CarbonData

CarbonData Partition 功能介绍

上汽集团 SAIC MOTOR

List Partition DDL Syntax

```
CREATE TABLE [IF NOT EXISTS] [db_name.]table_name
[(col_name data_type , ...)]
PARTITIONED BY (partition_col_name data_type)
STORED BY 'carbondata'
[TBLPROPERTIES ('PARTITION_TYPE'='LIST', 'LIST_INFO'='A, B, C' ...)]
```

Example

```
create table if not exists list_partition_table(
col_B Int,
col_C Long,
col_D Decimal(10,2),
col_E Long,
col_F Timestamp )
partitioned by (col_A String)
stored by 'carbondata'
tblproperties('partition_type'='List', 'list_info'='aaaa, bbbb, (cccc, dddd), eeee')
```

ListInfo 支持One Level Group

APACHE CarbonData

CarbonData Partition 功能介绍 —— Show Partition

上汽集团 SAIC MOTOR

Syntax: SHOW PARTITIONS [db_name.]table_name

Hash Partition

```
+--------------------+
|partition           |
+--------------------+
|vin = HASH_NUMBER(5)|
+--------------------+
```

Range Partition

```
+-------------------------------------+
|partition                            |
+-------------------------------------+
|0, logdate = DEFAULT                 |
|1, logdate < 2014/01/01              |
|2, 2014/01/01 <= logdate < 2015/01/01|
|3, 2015/01/01 <= logdate < 2016/01/01|
+-------------------------------------+
```

```
SegmentDir/0_batchno0-0-1502703086921.carbonindex
           ^
SegmentDir/part-0-0_batchno0-0-1502703086921.carbondata
                    ^
```

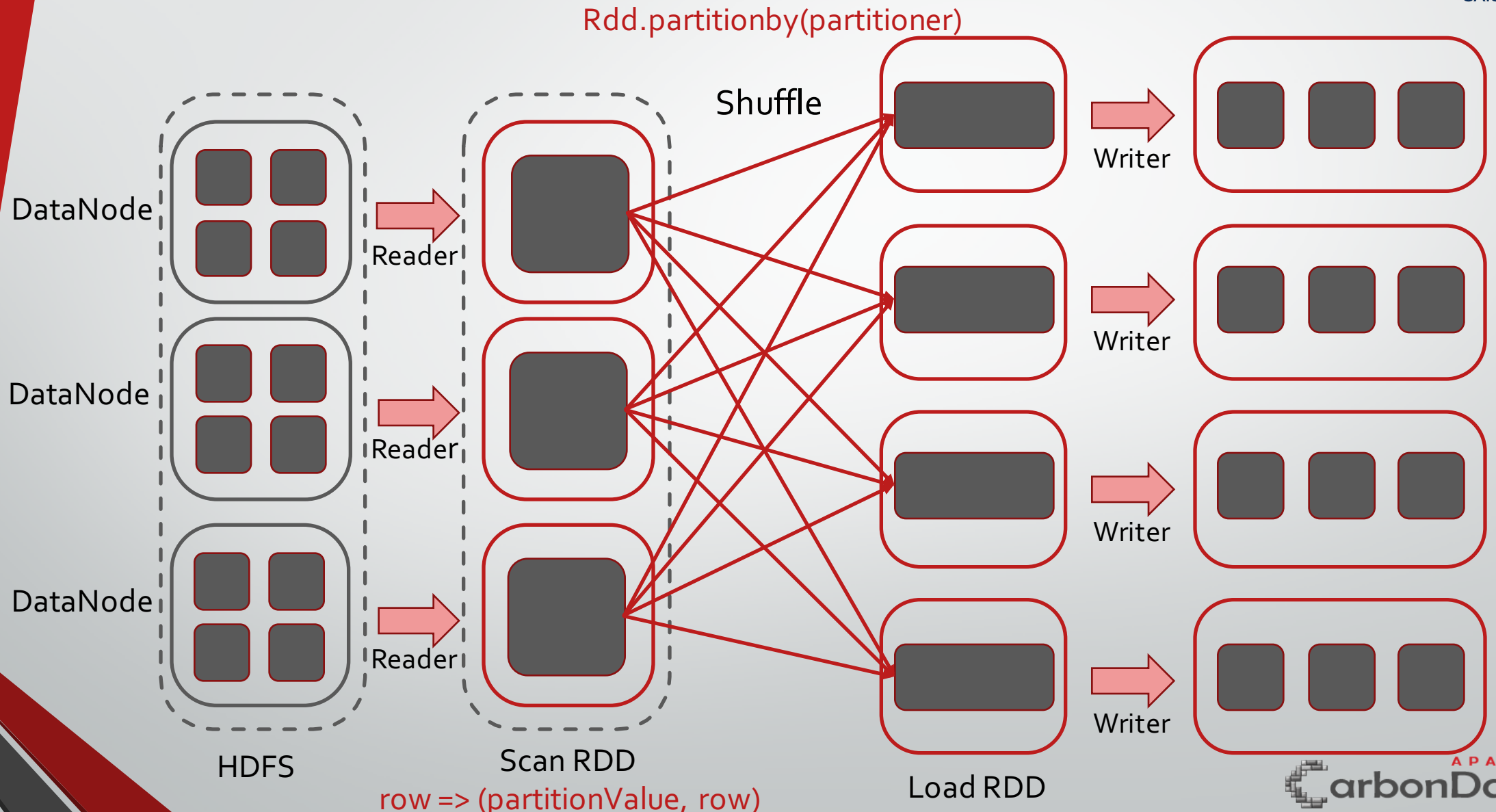
Partition Id

List Partition

```
+--------------------------+
|partition                 |
+--------------------------+
|0, country = DEFAULT      |
|1, country = China        |
|2, country = UK, US       |
|3, country = Japan        |
|7, country = Canada       |
|8, country = Russia       |
|9, country = Fiji         |
|5, country = Korea        |
|6, country = India        |
+--------------------------+
```

APACHE CarbonData

CarbonData Partition 功能介绍——数据加载过程



CarbonData Partition 功能介绍——查询过程

1. 根据SQL中的过滤条件=, <=, <, >, >=, in, not in以及表达式右值确定命中的partitionId

Example:
某表A以年龄为分区字段，创建range partition table
分区信息为（'range_info' = '10, 20, 30, 40, 50, 60, 70, 80'）
SQL过滤条件: age >= 20 and age < 40
得到命中的分区ID为3， 4

2. 如果有其他在排过序的维度列有过滤条件，则在driver端根据Btree索引获取blocklet 所在的文件名，如没有则获取全部，再根据文件名中的partitionId，筛选得到需要读取的文件，最后再下发spark task进行读取

```
SegmentDir/0_batchno0-0-1502703086921.carbonindex
           ^
SegmentDir/part-0-0_batchno0-0-1502703086921.carbondata
                     ^
```

Partition Id

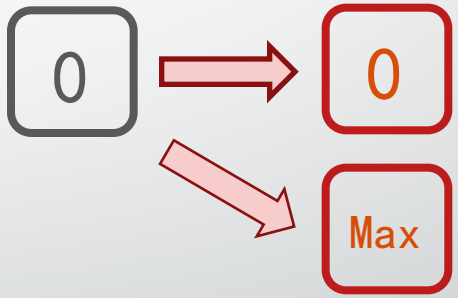
CarbonData Partition 功能介绍 —— Add Partition

Syntax： ALTER TABLE [db_name].table_name ADD PARTITION('new_partition')

Process：

1. 修改PartitionInfo

2. 读取默认分区（PartitionId为0）数据并根据新的PartitionInfo重新导入数据



3. 删除旧的默认分区数据文件

CarbonData Partition 功能介绍 —— Split Partition

上汽集团 SAIC MOTOR

Syntax：

```
ALTER TABLE [db_name].table_name  SPLIT PARTITION(partition_id) INTO('new_partition1', 'new_partition2'...)
```

Process：

1. 修改PartitionInfo

Note： 不能Split 0分区（默认分区）

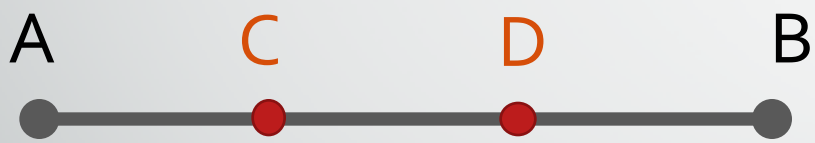
APACHE CarbonData

CarbonData Partition 功能介绍 —— Split Partition

Process：

2.读取目标分区数据并根据新的PartitionInfo重新导入数据

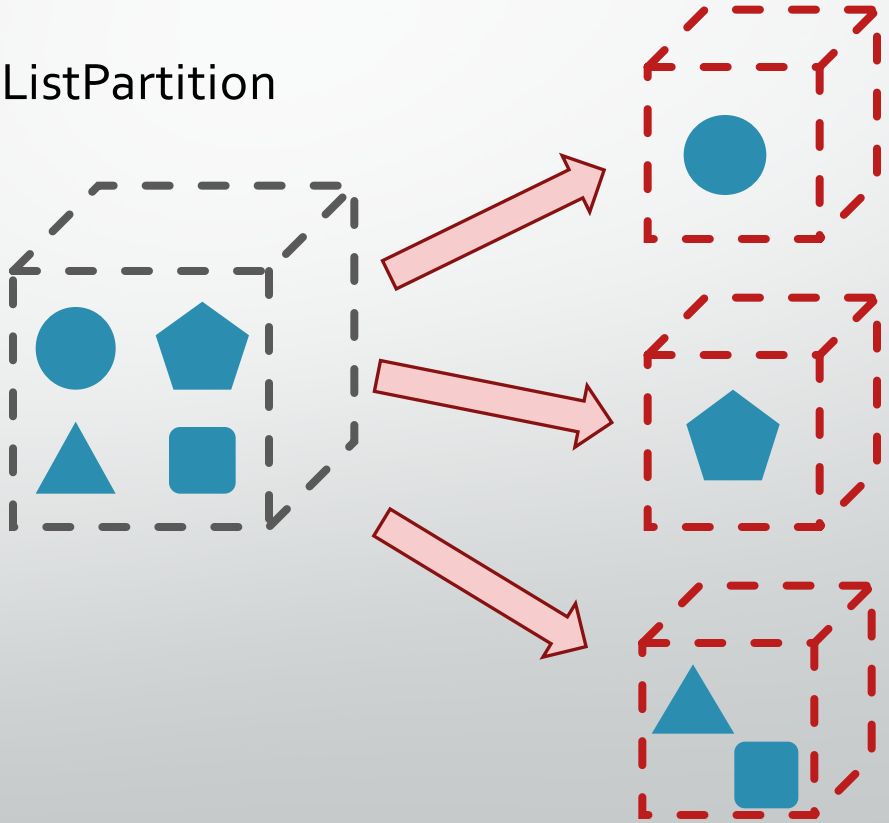
RangePartition



[A, B) =>

[A, C), [C, D), [D, B)

ListPartition



3.删除旧分区数据文件

CarbonData Partition 功能介绍 —— Drop Partition

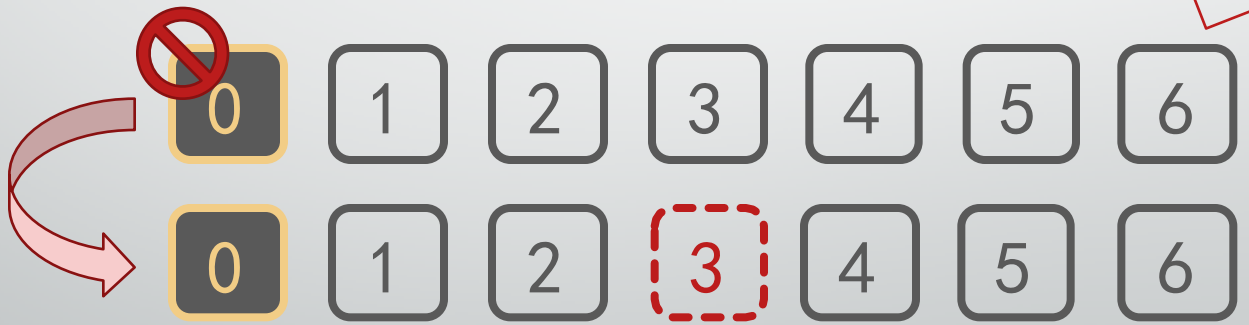
上汽集团 SAIC MOTOR

Syntax：

//Drop partition definition only and keep data
ALTER TABLE [db_name].table_name DROP PARTITION(partition_id)
//Drop both partition definition and data
ALTER TABLE [db_name].table_name DROP PARTITION(partition_id) WITH DATA

Process：

1. 修改PartitionInfo



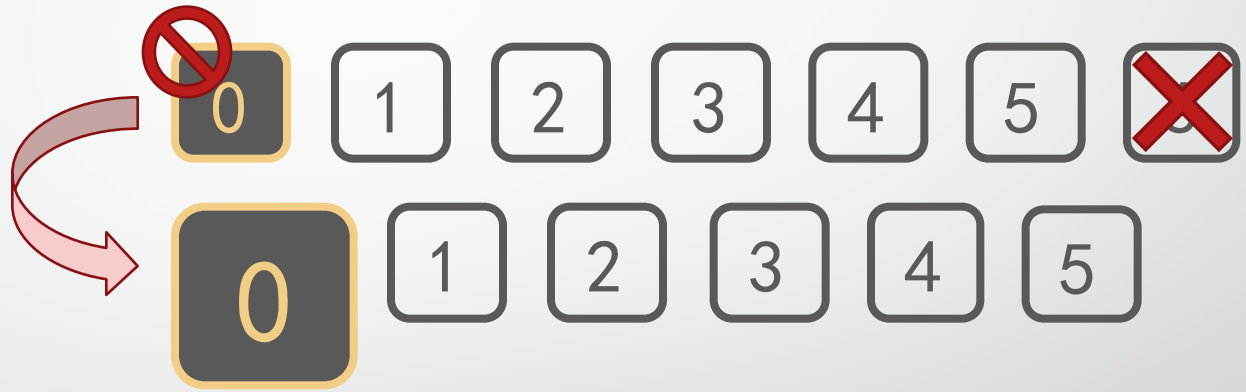
APACHE CarbonData

CarbonData Partition 功能介绍 —— Drop Partition

Process：

2a. Drop Partition 但保留数据 —— RangePartition

Drop最后一个partition
数据merge到default partiton



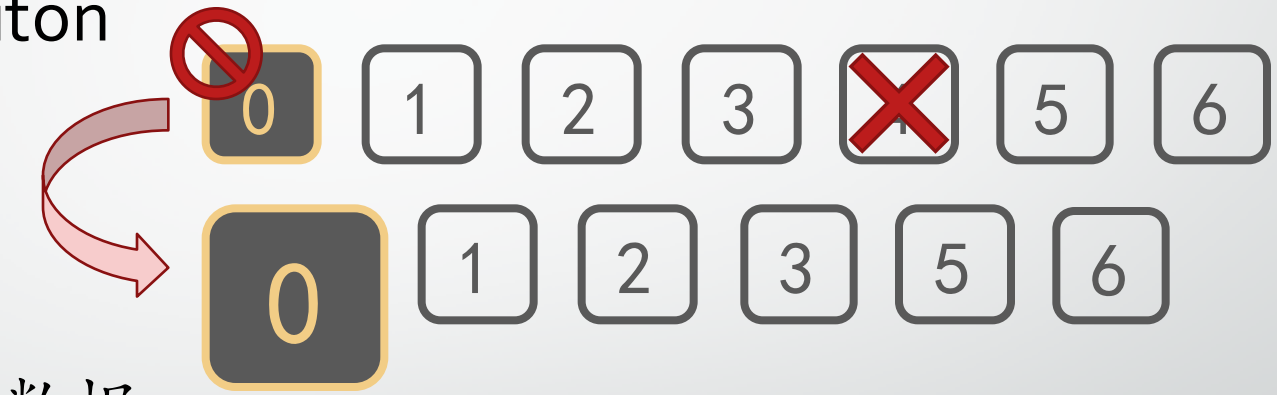
Drop非最后partition
数据merge到 next partition

CarbonData Partition 功能介绍 —— Drop Partition

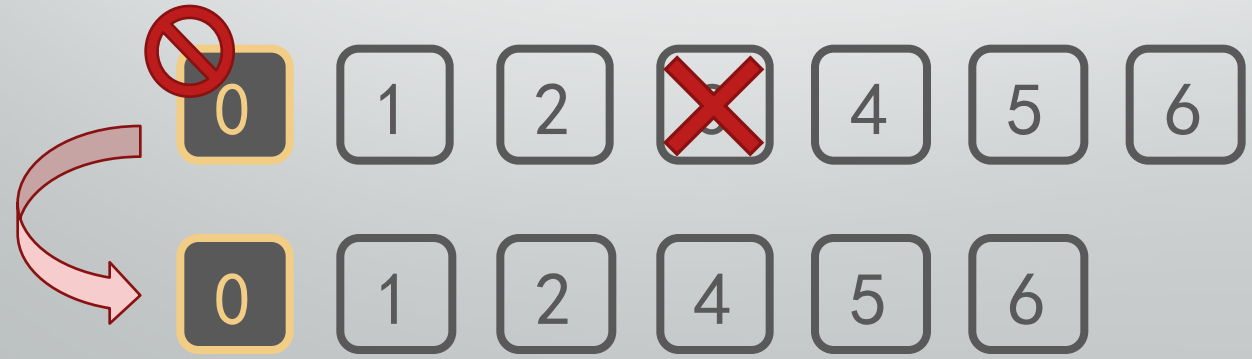
Process：

2a. Drop Partition 但保留数据 —— ListPartition

数据merge到default partiton



2b. Drop Partition 不保留数据



上汽集团CarbonData实践分享

上汽集团 SAIC MOTOR

集群环境：1台Spark client server　6台DataNode server

启用资源：

```
spark-shell
--master yarn
--deploy-mode client
--num-executors 6
--driver-memory 10g
--executor-memory 50g
--executor-cores 5
```

测试数据样本：

荣威RX5 2017年1月1日~1月30日数据

每天3~5亿条，19~30G

30天共109亿条，约667G

APACHE CarbonData

上汽集团CarbonData实践分享

上汽集团 SAIC MOTOR

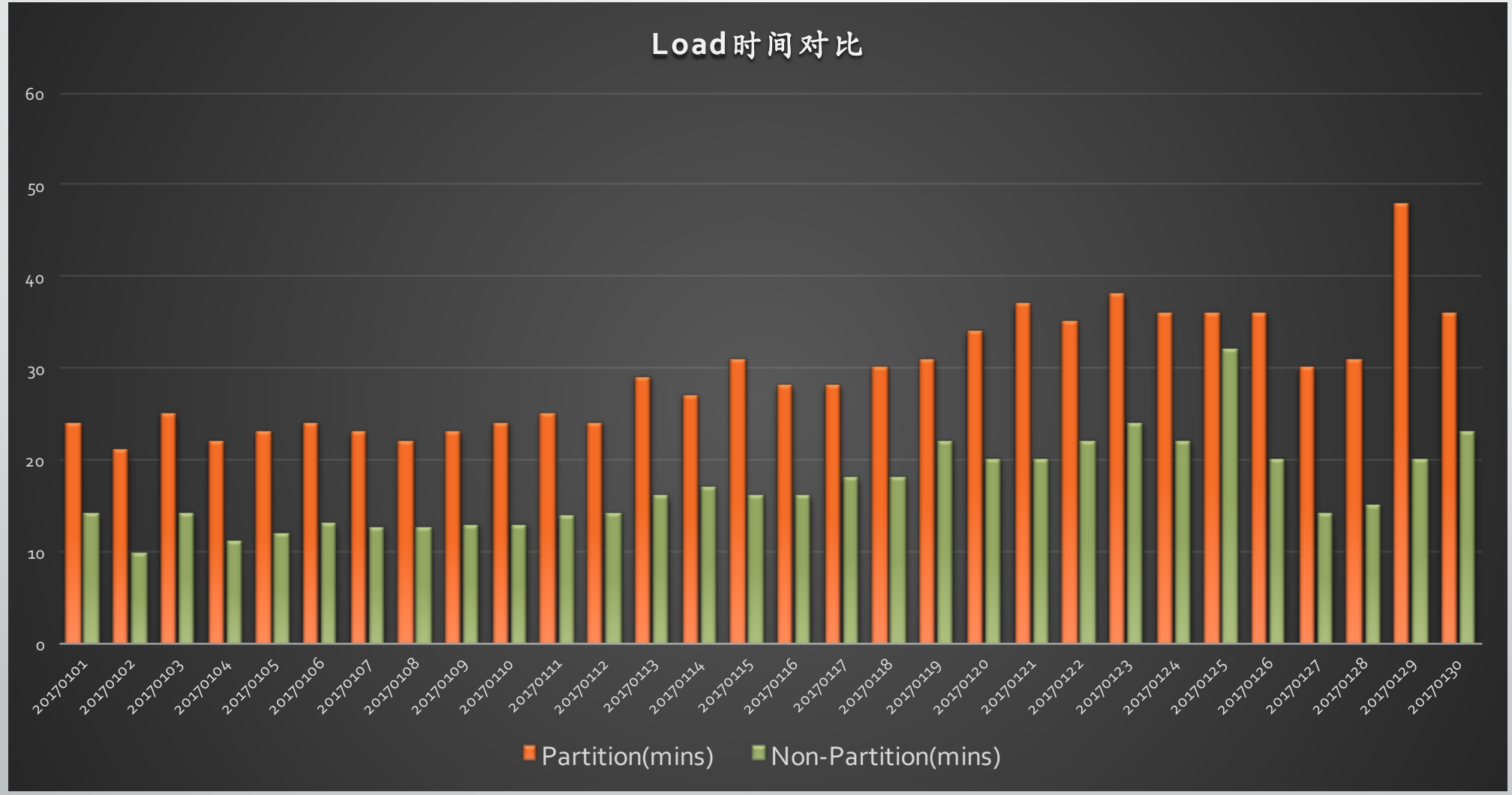
建表语句：

```
CREATE TABLE IF NOT EXISTS rx5_carbon_partition_test(
Id string,
create_time timestamp,
Field_C  Int,
Field_D  String,
...
...
Field_XX
)
PARTITIONED BY (create_time_hour int)
STORED BY 'carbondata'
TBLPROPERTIES('PARTITION_TYPE' = 'RANGE',
'RANGE_INFO' = '1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24',
'SORT_COLUMNS' = 'Id, Field_C, Field_D ')
```

共51个字段，其中绝大部分为维度

APACHE CarbonData

上汽集团CarbonData实践分享



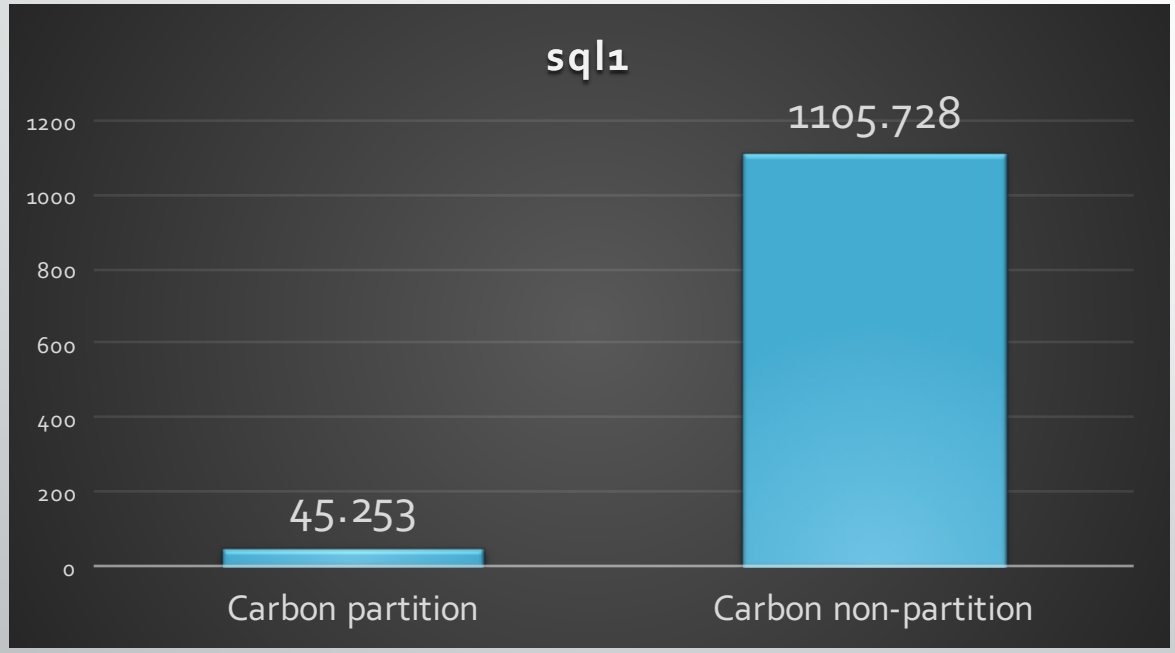
上汽集团CarbonData实践分享

上汽集团 SAIC MOTOR

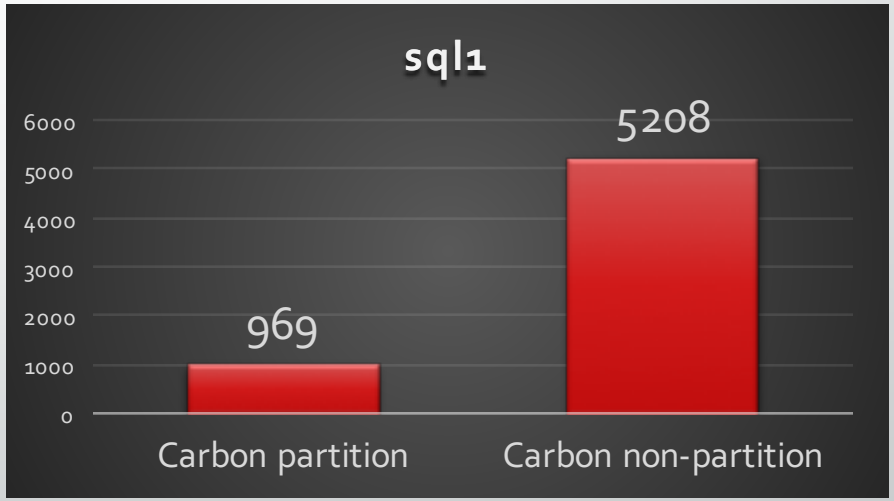
性能测试Case1：无排序维度列作为过滤条件，有partition列上的范围过滤条件的聚合查询

```
select Date(create_time), count(distinct ID)
from default.rx5_carbon_partition_test
where create_time_hour >= 17 and create_time_hour < 20
group by Date(create_time)
```

无法应用索引
或在索引中排序位置较为靠后



查询时间对比



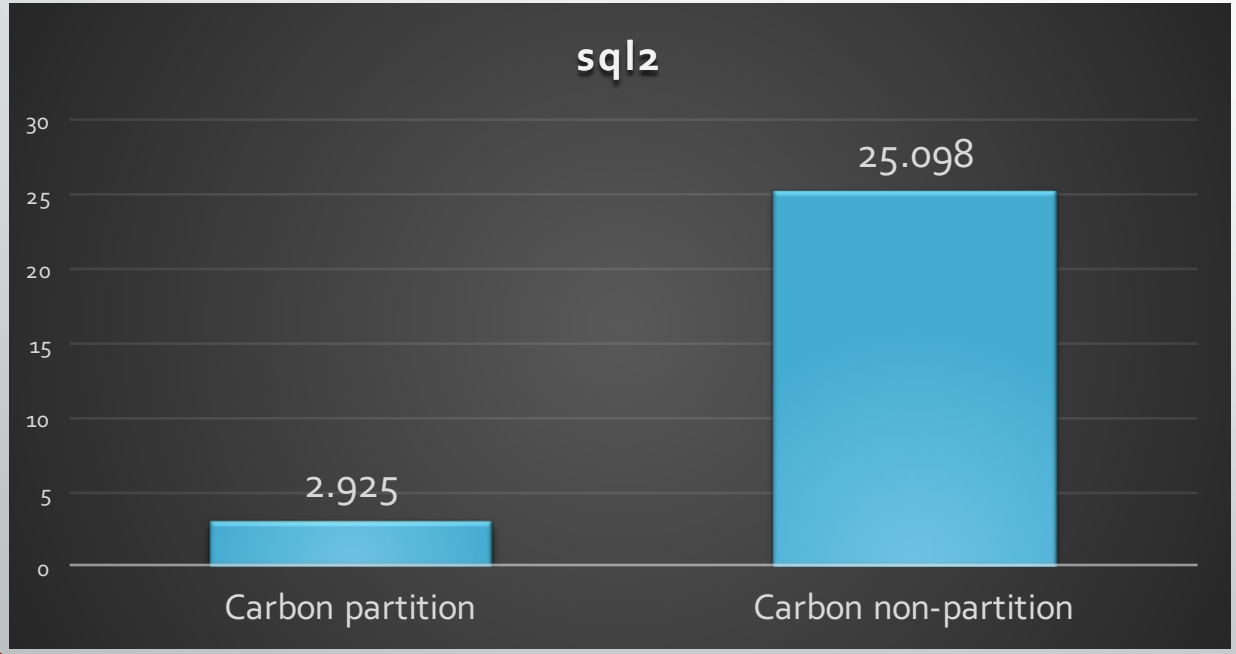
Spark Task数对比

APACHE CarbonData

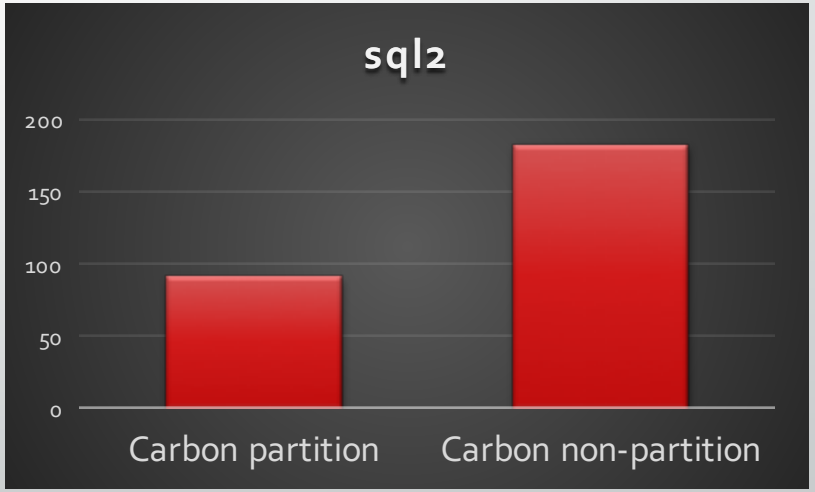
上汽集团CarbonData实践分享

性能测试Case2：有维度列作为过滤条件，有Partition列过滤条件的聚合查询

```
select Date(create_time), count(*)
from default.rx5_carbon_partition_test
where create_time_hour >= 17 and create_time_hour < 20
and id = 'XXXX0000000001'
group by Date(create_time)
```



查询时间对比



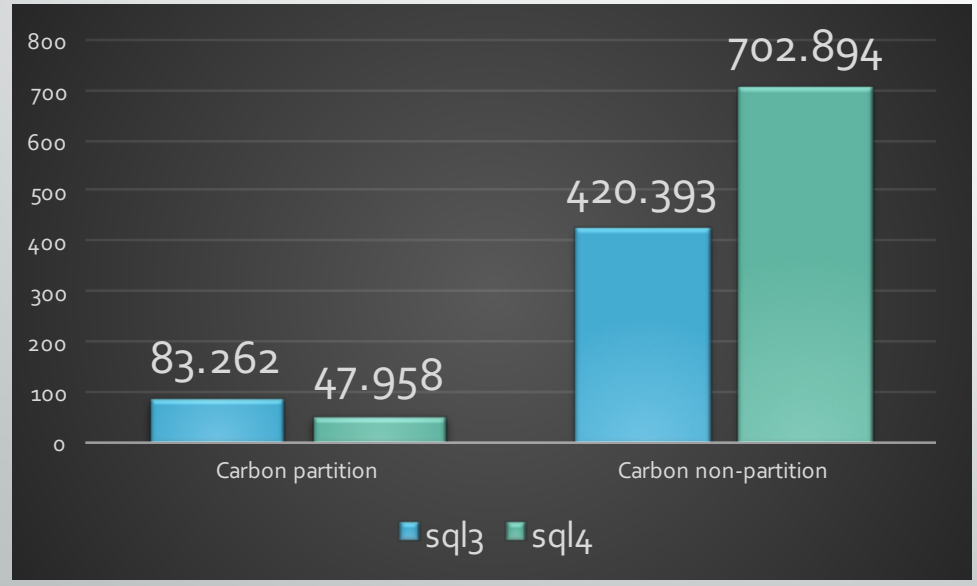
Spark Task数对比

上汽集团CarbonData实践分享

性能测试Case3：没有过滤条件的聚合查询
性能测试Case4：没有过滤条件的聚合查询，聚合列为partition列

```
select Date(create_time), count(*)
from default.rx5_carbon_partition_test
group by Date(create_time)
```

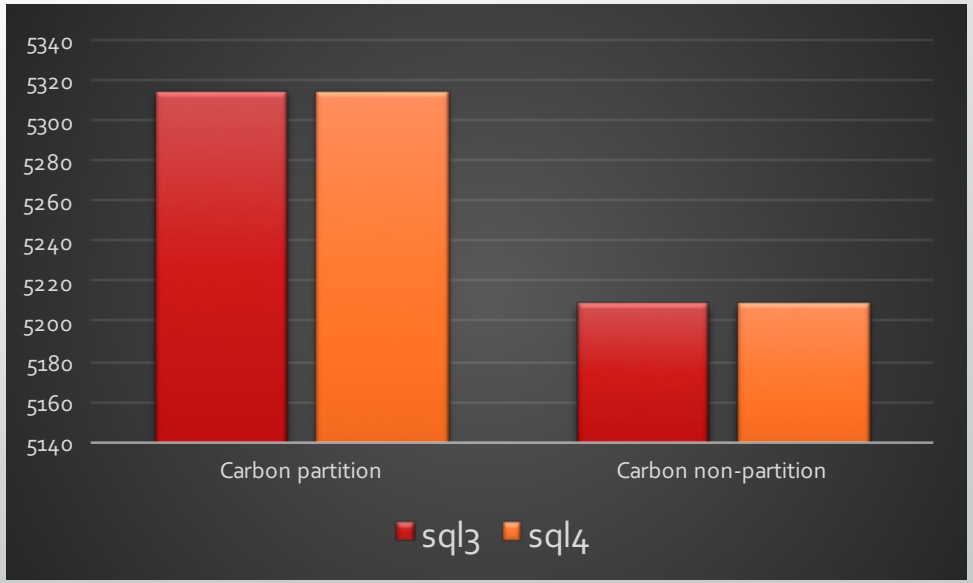
```
select Date(create_time), count(*)
from default.rx5_carbon_nonpartition_test
group by Date(create_time)
```



查询时间对比

```
select create_time_hour, count(*)
from default.rx5_carbon_partition_test
group by create_time_hour
```

```
select hour(create_time), count(*)
from default.rx5_carbon_nonpartition_test
group by hour(create_time)
```

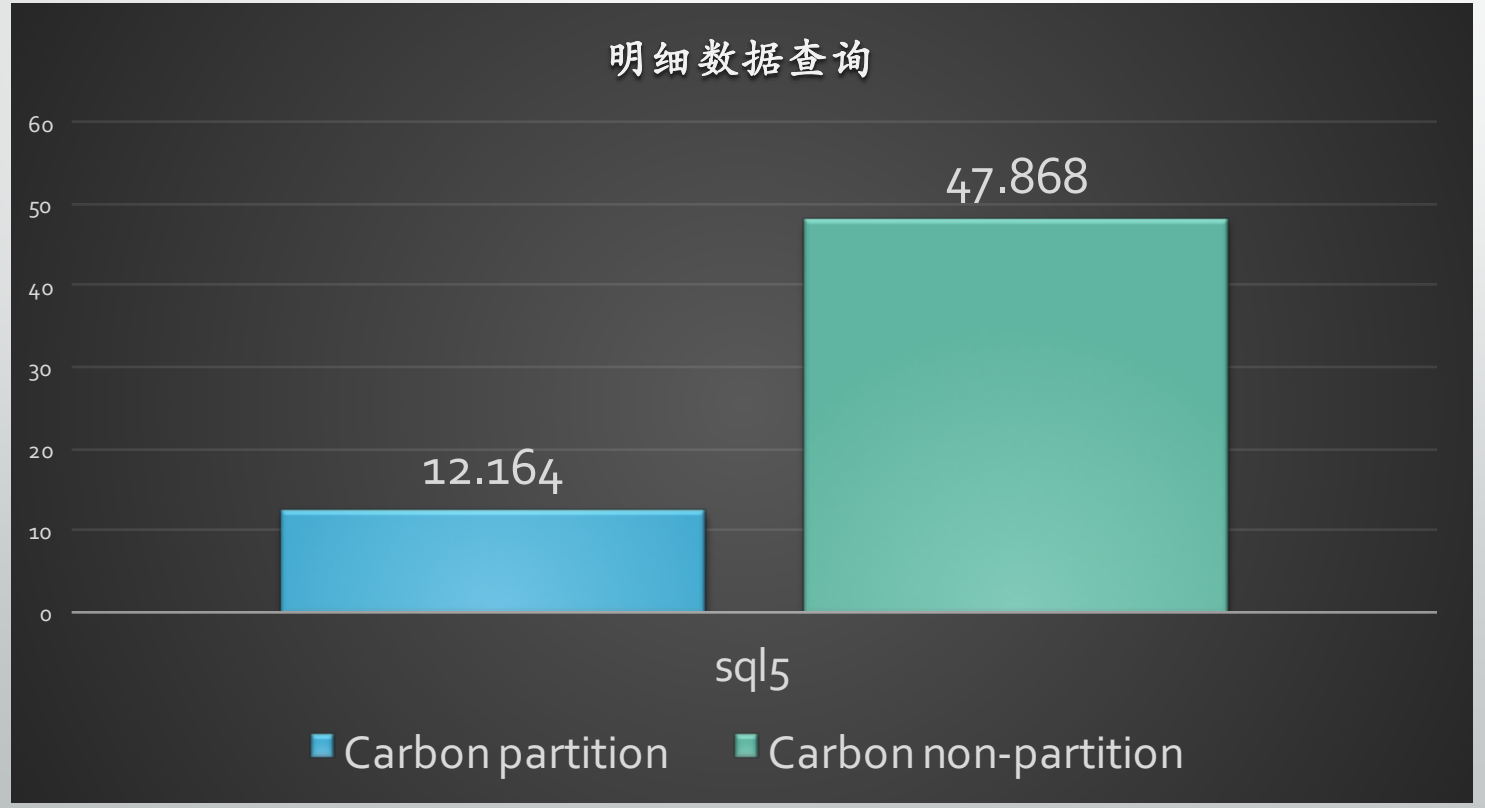


Spark Task数对比

上汽集团CarbonData实践分享

性能测试5：明细数据查询

```
select * from default.rx5_carbon_partition_test
where create_time_hour>= 17 and create_time_hour < 20
and id= 'XXX127893791231023'
and date(create_time) = '2017-01-04'
```



CarbonData Partition的性能调优

上汽集团 SAIC MOTOR

1. 选择最合适的Partition列

- 结合业务，了解哪些列会作为常用的过滤条件
- 结合数据，可粗略统计一下数据分布情况，避免选择数据倾斜严重的列
- 如有多种选择，可再结合数据加载效率考虑

2. 尽可能的使用Partition列作为过滤条件

例如Partition列为A，你根据业务需求在Column B上有筛选条件，但你注意到A与B列之间存在某种固定的mapping关系，这时就可以根据B列的过滤条件再新增一个partition列的过滤条件，以提高查询效率。

APACHE CarbonData

CarbonData Partition TODO List

1. Range Interval Partition
2. Value-Based Partition
3. Alter Table Merge Partition
4. Optimization for join between partition tables

……

谢谢

扫描二维码获取
上汽集团数据业务部开放职位

欢迎加入Apache CarbonData社区:
发邮件至 dev-subscribe@carbondata.apache.org加入邮件组

关于使用CarbonData过程中遇到任何问题均可发邮件至
dev@carbondata.apache.org

官网： http://carbondata.apache.org/