CHINA · BEIJING

Apache Kudu在网易实时数仓的实践

何李夫

网易杭研院软件工程师

主办方 Geekbang. 极客邦科技 InfoQ

关注 QCon 公众号

收获国内外一线大厂实践与技术大咖同行成长

演讲视频 干货整理 大咖采访 行业趋势

SPEAKER
INTRODUCE

何李夫

- Apache Kudu Committer & PMC member
- 2014年加入网易杭研院，曾负责分布式缓存系统开发，现负责实时数仓存储引擎开发
- 网易之前，设计和开发了新三板交易所核心交易系统、短信增值业务平台等

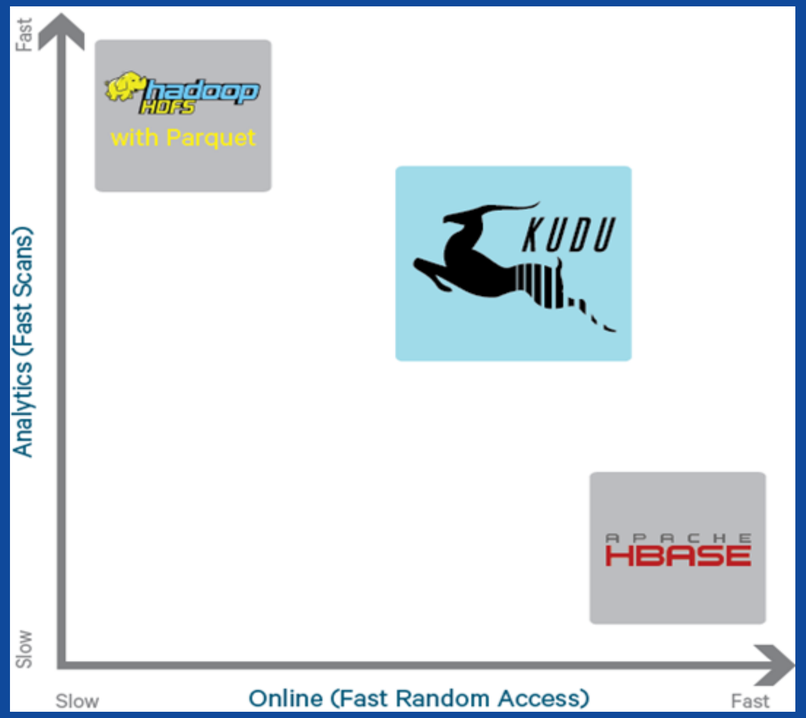
TABLE OF

CONTENTS 大纲

- 系统概述
- 生产实践
- 运维交流

系统概述

- Kudu定位



Fast Analytics on **Fast Data**

快速读写：低延时、高吞吐

实时数据：状态数据、时序日志

系统概述

- Kudu优势
  - 列式存储引擎
  - 支持实时写入
  - 支持更新和删除

像传统的数据库

系统概述

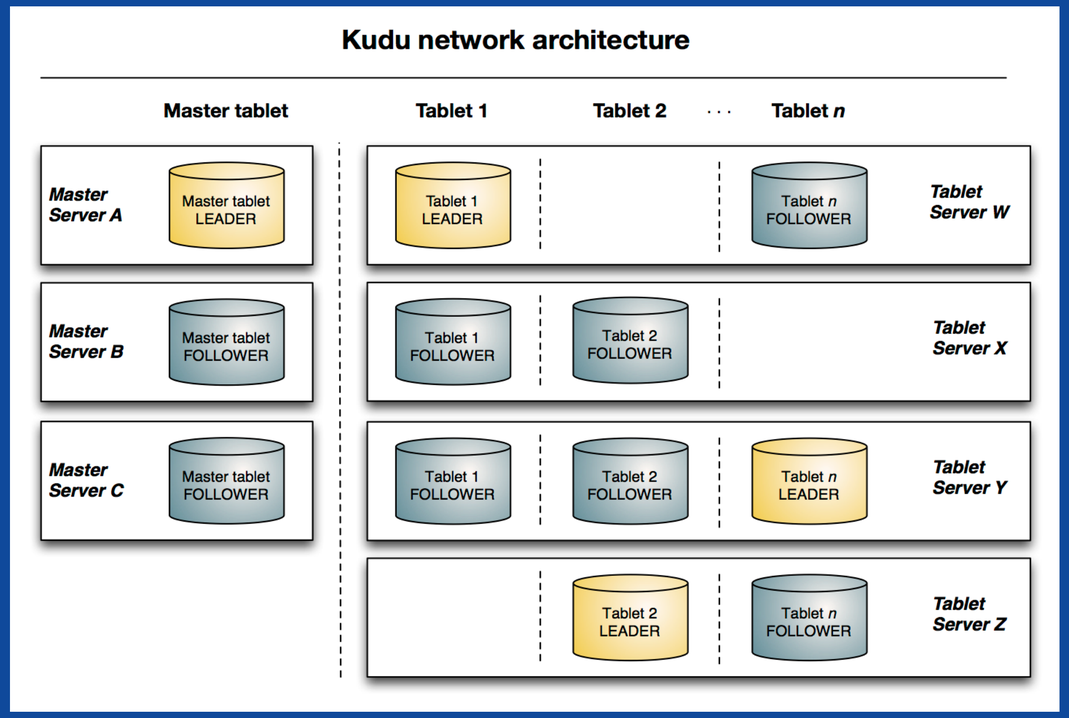
- Kudu设计

```
CREATE TABLE table1 (
  day STRING COMMENT "日期",
  test1 BIGINT COMMENT 'test1',
  test2 STRING COMMENT 'test2',
  test3 DECIMAL(10, 2) COMMENT "test3",
  primary key(day, test1)
)
PARTITION BY HASH(test1) PARTITIONS 3,
RANGE (day)
(
    PARTITION VALUE = '2019-12-05',
    PARTITION VALUE = '2019-12-06'
)
stored as kudu;
```

- SQL like Schema
- Composite Primary Key
- Range & Hash Partitions

系统概述

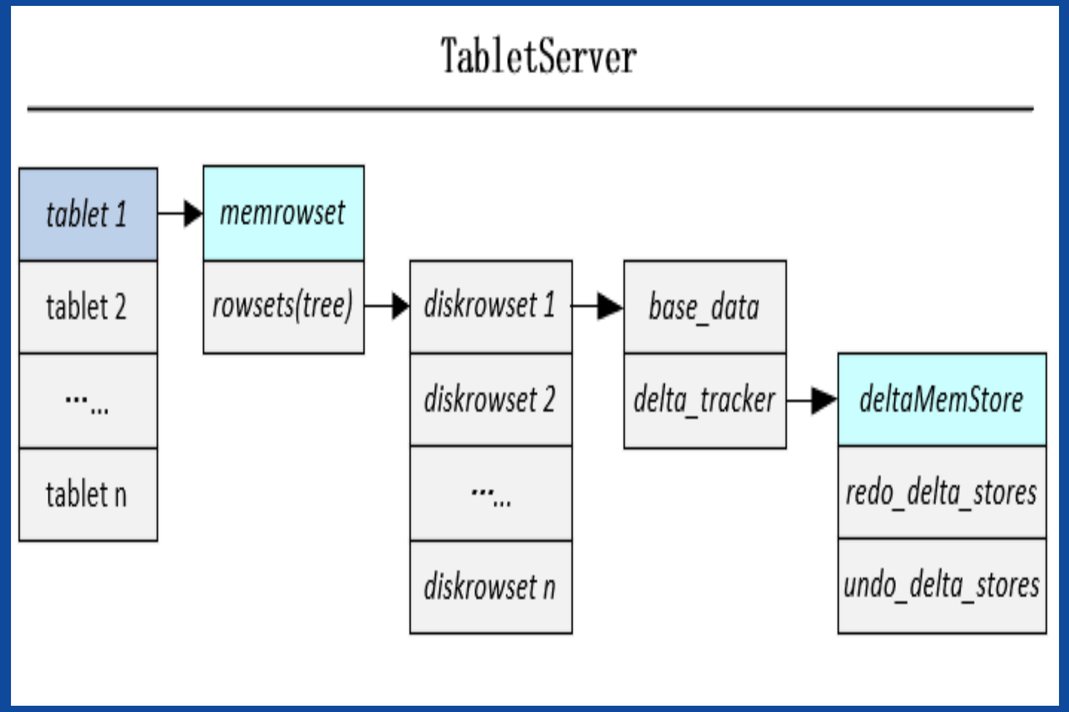
- Kudu设计



- Master & TServer
- Flexible Partitioning
- Raft Consensus
- All metadata in RAM

系统概述

- Kudu设计



- LSM
- Interval Tree
- Primary key sequence

系统概述

- Kudu设计

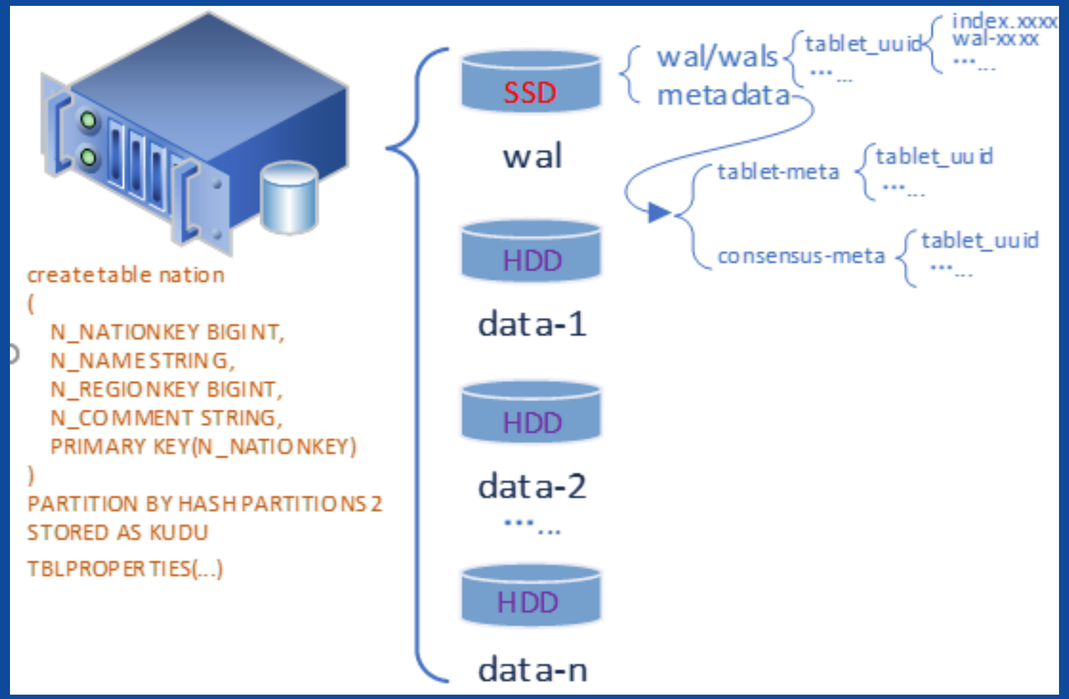
```
                MemRowSet Row
+-----------------------------------------------------+
| insertion timestamp  | mutation head | row data... |
+-------------------------|-------------------------+
                          |
                          v          First mutation
                  +-----------------------------------------------+
                  | mutation timestamp | next_mut | change record |
                  +--------------------|--------------------------+
                            __________/
                           /
                           |         Second mutation
                  +--------v--------------------------------------+
                  | mutation timestamp | next_mut | change record |
                  +--------------------|--------------------------+
                            __________/
                           /
                           ...
```

```
+--------------+       +-----------+      +--------------+
| UNDO records | <---  | base data | ---> | REDO records |
+--------------+       +-----------+      +--------------+
```

- MVCC
- Snapshot scanners
- Time-travel scanners
- Change-history queries

系统概述

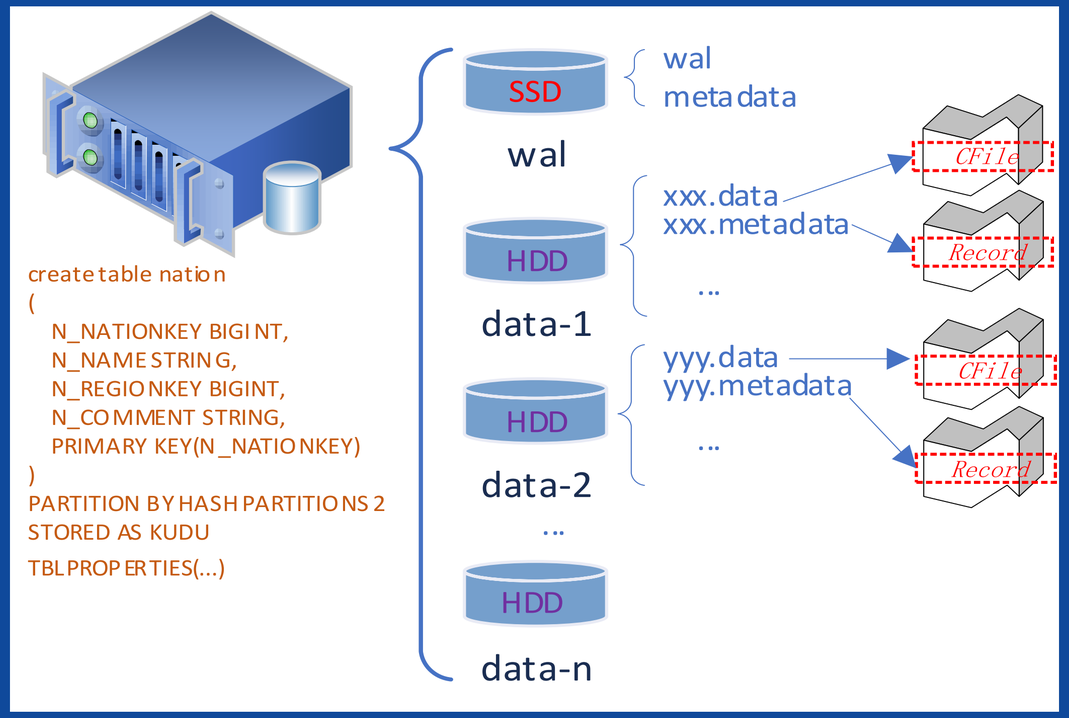
- Kudu设计



- WAL
- Tablet metadata
- Consensus metadata

系统概述

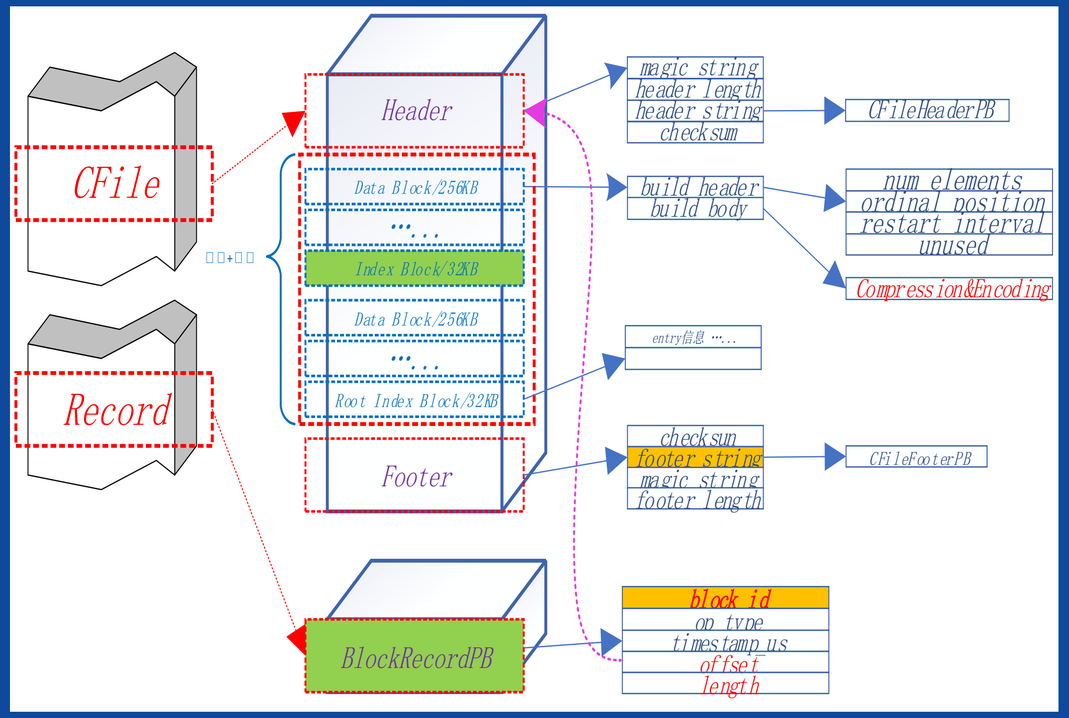
- Kudu设计



- Columnar Storage
- Append Only File
- Bloom File
- Punch Hole

系统概述

- Kudu设计



- Project pushdown
- Predicate pushdown
- Encoding & compression

系统概述

- Kudu with Hadoop ecosystem

- Impala、Presto
- Spark、 Hive、 MapReduce
- Flink、Spark Streaming、Flume

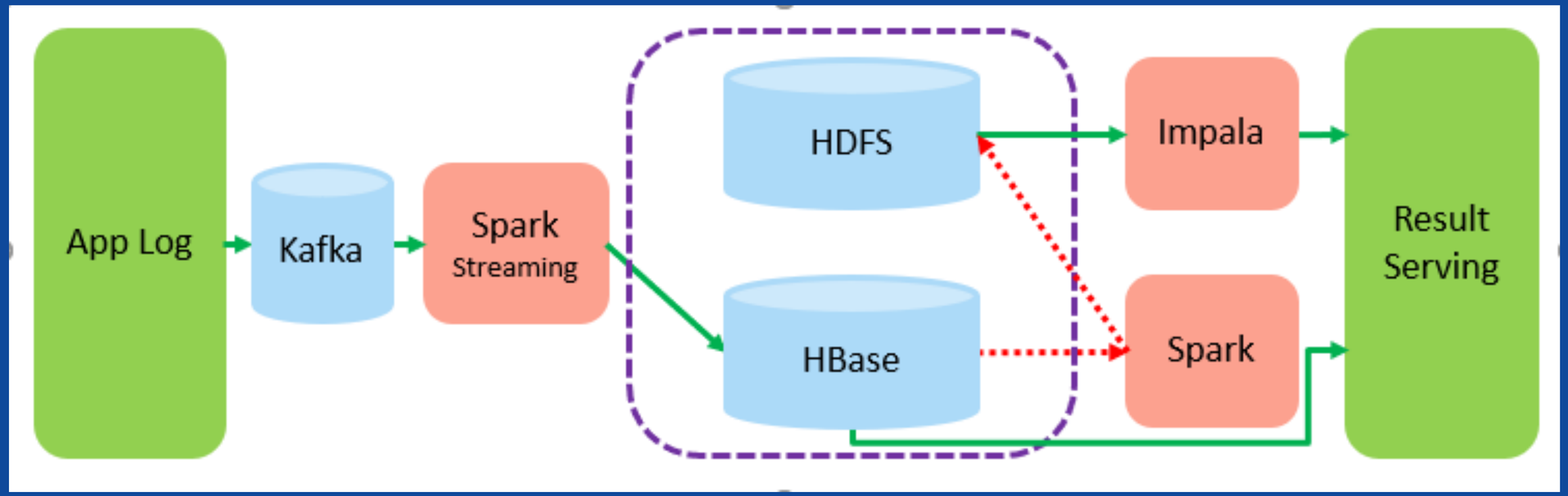
TABLE OF

CONTENTS 大纲

- 系统概述
- 生产实践
- 运维交流

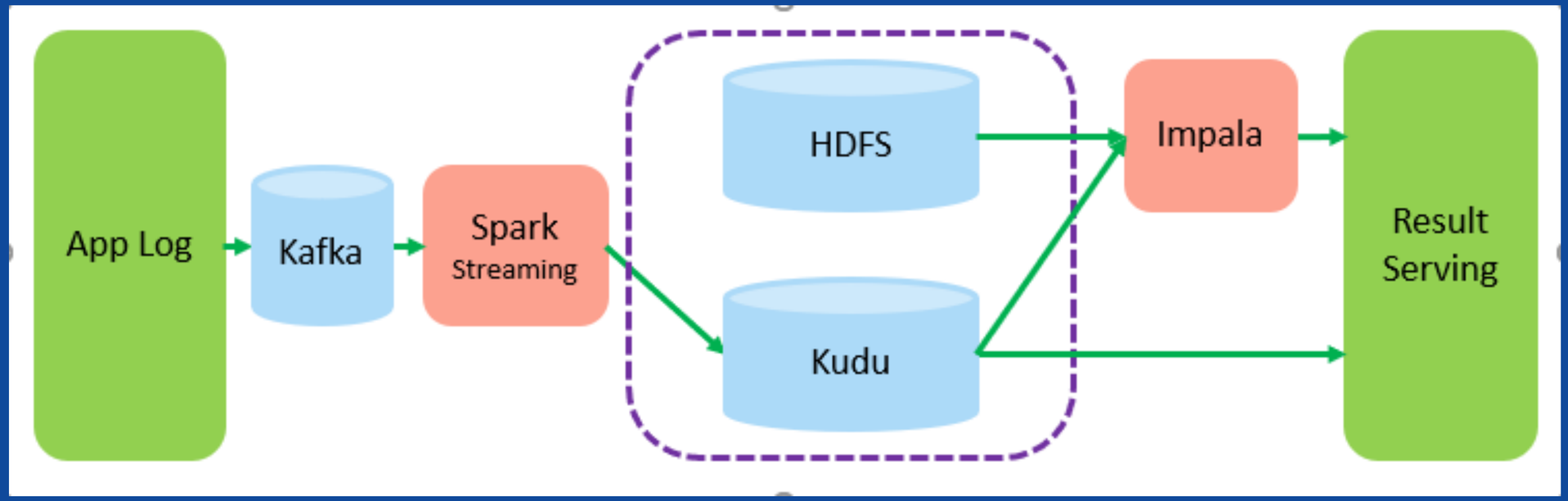
生产实践

- Case1：实时数据采集



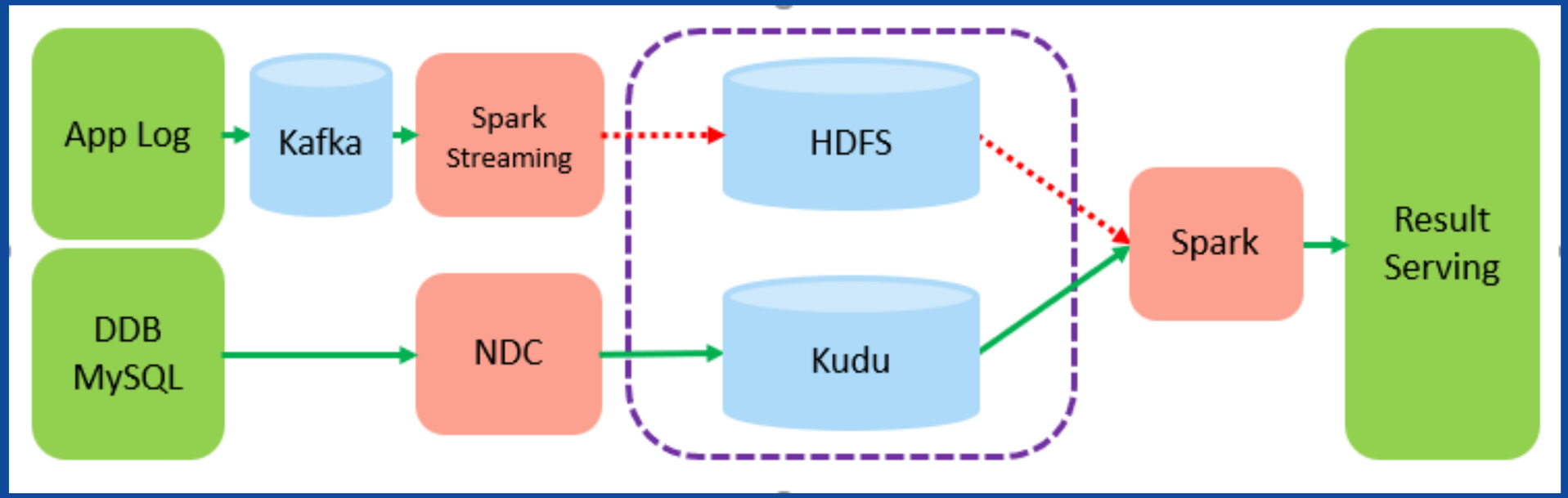
生产实践

- Case1：实时数据采集



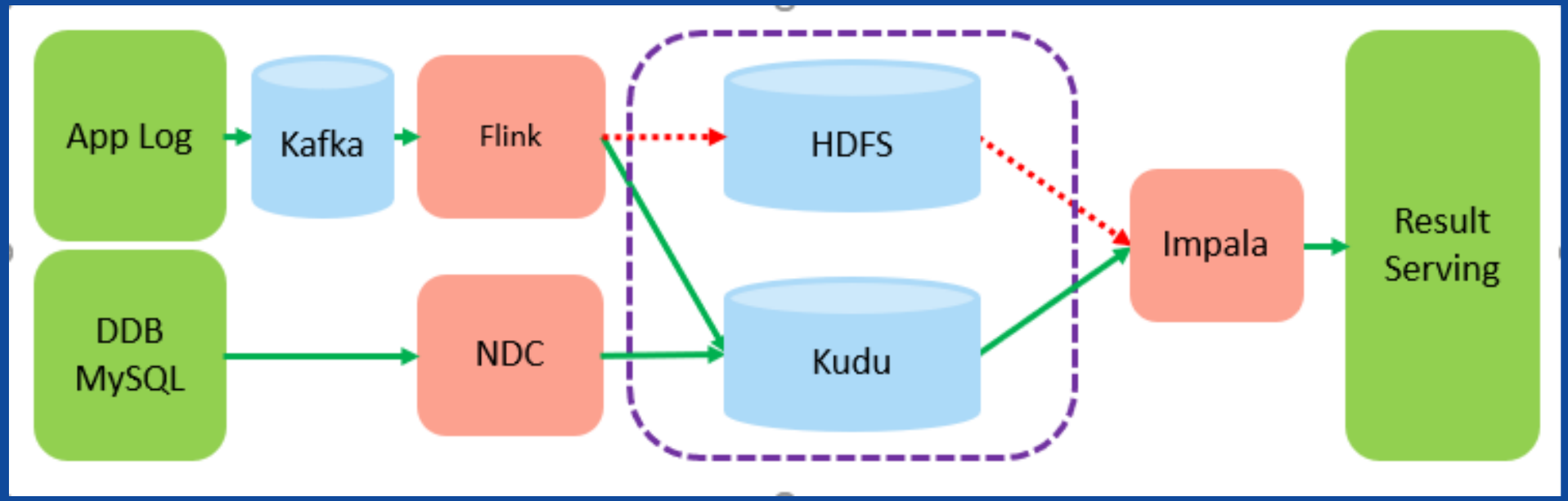
生产实践

- Case2：维表数据关联应用



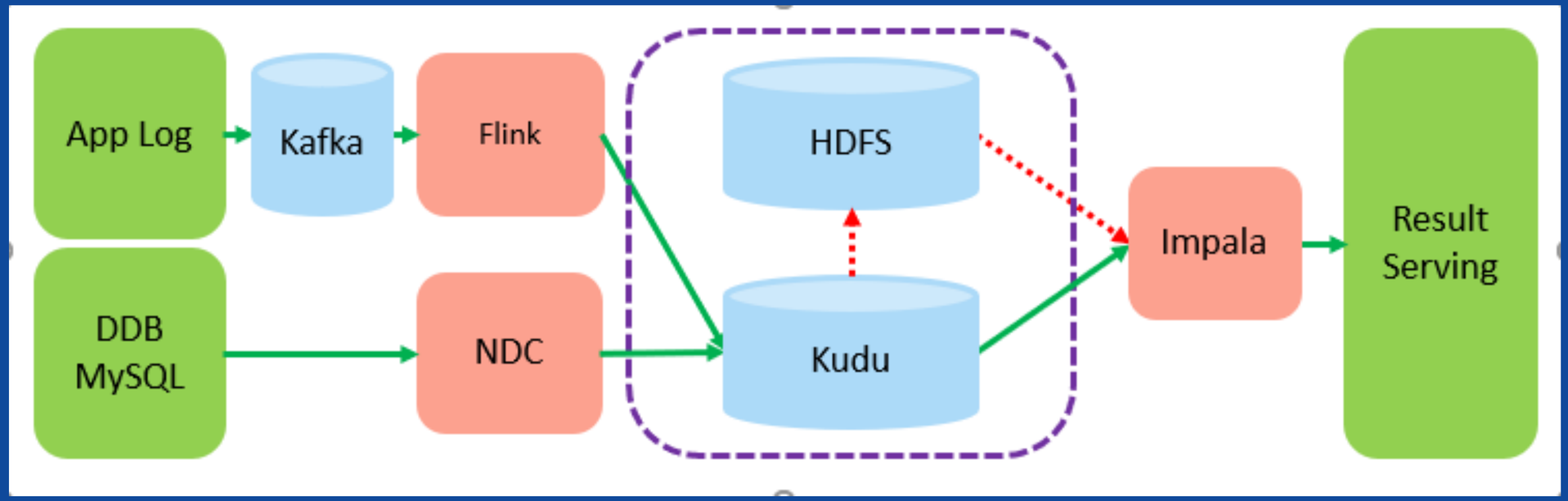
生产实践

- Case3：实时流量分析



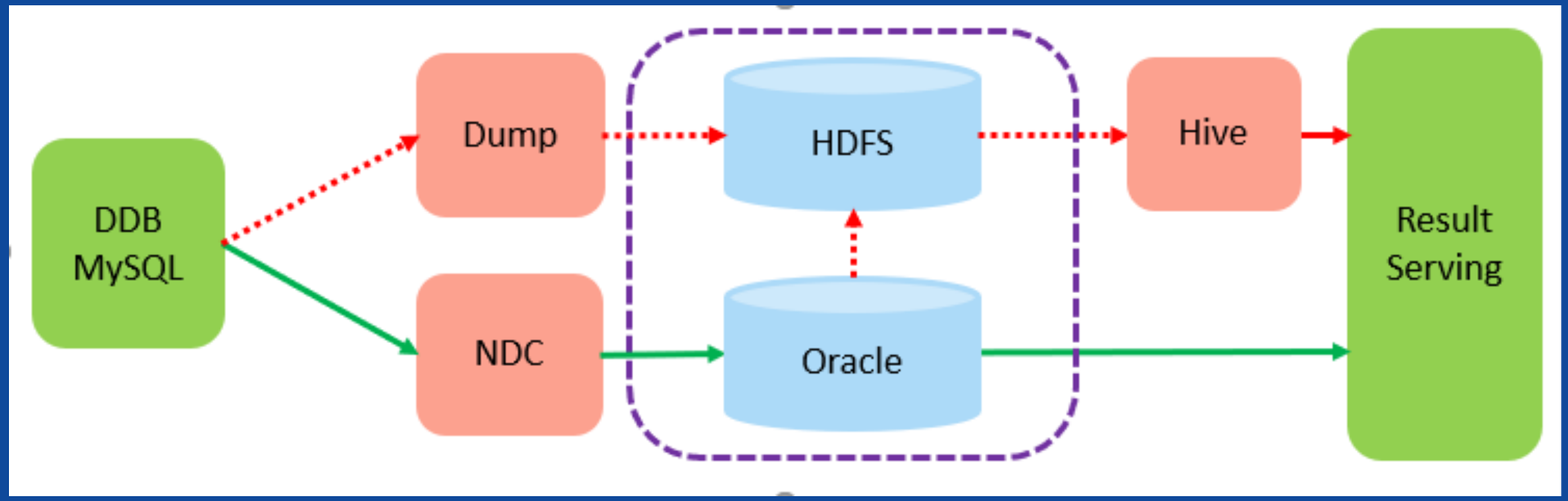
生产实践

- Case3：实时流量分析



生产实践

- Case4：实时数仓ETL



生产实践

- Case4：实时数仓ETL

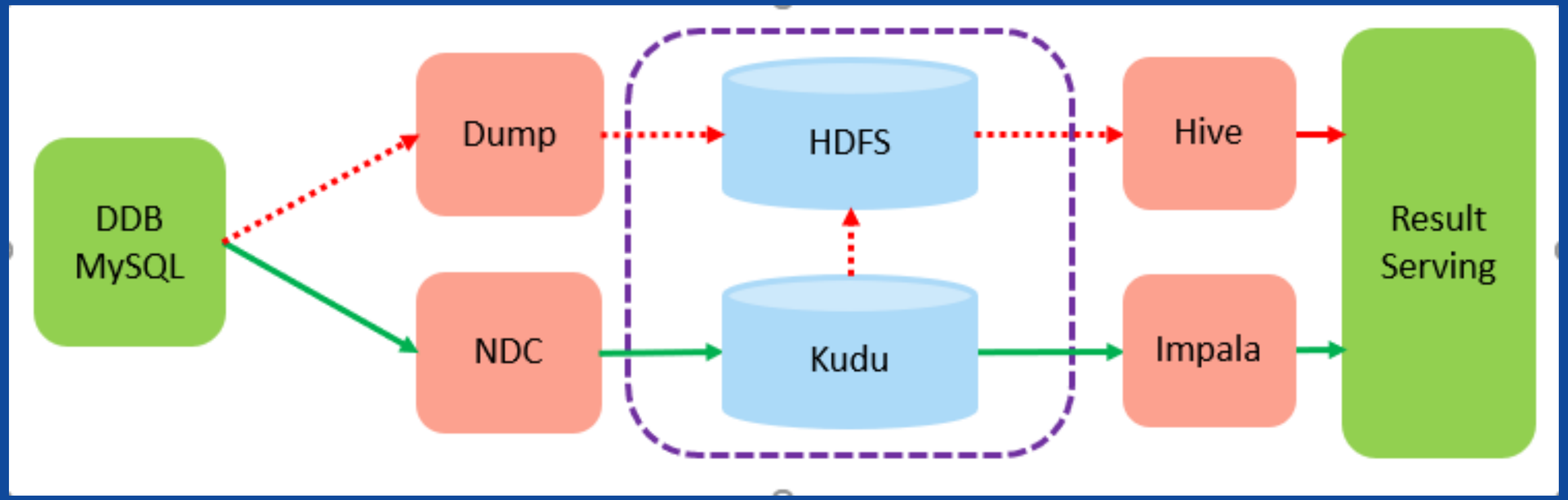


TABLE OF

CONTENTS 大纲

- 系统概述
- 生产实践
- 运维交流

运维交流

- 生产部署
  - 每个集群4个节点起，上规模后1个master独立
  - 每个节点WAL独占1块SSD、其他多块SAS或SSD
  - 启动ntp、nscd（/etc/hosts）、关闭thp
  - 万兆网络

运维交流

- 生产部署
  - Maintenance Thread

`--maintenance_manager_num_threads`

Size of the maintenance manager thread pool. For spinning disks, the number of threads should not be above the number of devices

Type	int32
Default	1
Tags	stable

机械盘数量1/3
固态盘适当提高

  - Across all data dir

`--fs_target_data_dirs_per_tablet`

Indicates the target number of data dirs to spread each tablet's data across. If greater than the number of data dirs available, data will be striped across those available. A value of 0 indicates striping should occur across all healthy data dirs. Using fewer data dirs per tablet means a single drive failure will be less likely to affect a given tablet.

Type	int32
Default	3
Tags	evolving,advanced

运维交流

- Schema Partitions

```
create table nation (
    N_NATIONKEY INT,
    N_NAME STRING,
    N_REGIONKEY INT,
    N_COMMENT STRING,
    PRIMARY KEY(N_NATIONKEY)
)
STORED AS KUDU
TBLPROPERTIES ('kudu.master_addresses'='kudu1,kudu2,kudu3');
```

单Range分区：小数据量维表

```
create table nation (
    N_NATIONKEY INT,
    N_NAME STRING,
    N_REGIONKEY INT,
    N_COMMENT STRING,
    PRIMARY KEY(N_NATIONKEY)
)
PARTITION BY HASH (N_NATIONKEY) PARTITIONS 4
STORED AS KUDU
TBLPROPERTIES ('kudu.master_addresses'='kudu1,kudu2,kudu3');
```

Hash分区不能修改

运维交流

- Schema Partitions

```
CREATE TABLE app_log (
    day STRING,
    servertime BIGINT,
    action STRING,
    page_type STRING,
    os STRING,
    deviceudid STRING,
    logid STRING,
    ip STRING,
    sessionid STRING,
    ... ...
    PRIMARY KEY (day, servertime, action, page_type, os, deviceudid, logid)
)
PARTITION BY HASH (servertime) PARTITIONS 300,
RANGE (day)
(
  PARTITION VALUE = '2019-12-06'
)
STORED AS KUDU
TBLPROPERTIES ('kudu.master_addresses'='kudu1,kudu2,kudu3')
```

尽量保持主键的单调性

Range分区需要定时任务支持

运维交流

- 性能分析

```
CREATE TABLE `test_table3` (
    `col1` BIGINT,
    `col2` BIGINT,
    `col3` BIGINT,
    PRIMARY KEY (`col1`, `col2`)
)
PARTITION BY HASH (col1, col2) Partitions 3
stored as kudu
```

```
select * from test_table3 where col1 = 6 and col2= 6;
select * from test_table3 where col2 = 6;
select * from test_table3 where col1 = 6;
```

Query

```
SELECT col1,
       col2,
       col3
  FROM impala::music_kudu_internal.test_table3
 WHERE PRIMARY KEY >= (6, 6)
   AND PRIMARY KEY < (6, 7)
```

分区裁剪+主键索引

```
SELECT col1,
       col2,
       col3
  FROM impala::music_kudu_internal.test_table3
 WHERE col2 = 6
```

全表扫描

```
SELECT col1,
       col2,
       col3
  FROM impala::music_kudu_internal.test_table3
 WHERE PRIMARY KEY >= (6)
   AND PRIMARY KEY < (7)
```

全表扫描+主键索引

运维交流

- 性能分析

扫描请求

每列数据读取详细信息

Scanner id	State	Query	Requestor	Duration	Time since start	Timing
2673e8c14c0d4ff087e46b7810afa813	Complete	SELECT dt, agency_id, contract_id, company_id, income_1m, income_tm, income_1y, income_yoy, income_mom, income_1p, income_tp, income_type, data_type FROM impala::music_kudu_internal.test_copyright_agency_income_report_m WHERE company_id = 0 AND contract_id = 2010 AND income_type = 0	impala	146 ms	1.31 min	real: 146 ms user: 136 ms sys: 8 ms

持续时间=RTT+Server耗时

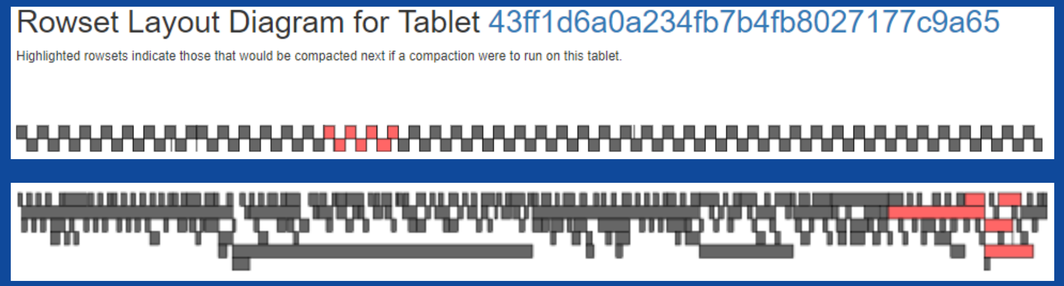
Server耗时

Column Stats			
column	cells read	bytes read	blocks read
dt	327.68k	2.8K	5
agency_id	196.61k	145.5K	6
contract_id	3.52M	2.29M	111
company_id	196.61k	249.2K	6
income_type	3.52M	1.64M	57
income_1m	196.61k	1.30M	6
income_tm	196.61k	1.31M	6
income_1y	2.39M	60B	3
income_yoy	2.39M	65.4K	3
income_mom	471.61k	1.11M	5
income_1p	220.20k	1.28M	6
income_tp	203.46k	1.31M	6
data_type	327.68k	9.3K	5
total	14.16M	10.71M	225

谓词列

运维交流

- 阈值监控
  - 控制单个节点分片数量
  - 控制单个分片数据大小
  - 注意Rowset Layout Diagram



运维交流

- 指标监控

不安全 | [redacted]:8050/metrics?types=tablet&attributes=table_name,kudu_rest_test_lhm_33&metrics=rows_inserted,rows_upserted,on_disk_size


```
[
    {
        "type": "tablet",
        "id": "cd08a60a9d5e42b68dc88cdf3b214619",
        "attributes": {
            "partition": "RANGE (day) PARTITION \"2019-09-29\" <= VALUES < \"2019-09-29\\000\"",
            "table_name": "impala::music_kudu_internal.kudu_rest_test_lhm_33",
            "table_id": "a6e486e1c35f4886a383474a88280eb3"
        },
        "metrics": [
            {
                "name": "on_disk_size",
                "value": 8412044
            },
            {
                "name": "rows_inserted",
                "value": 0
            },
            {
                "name": "rows_upserted",
                "value": 0
            }
        ]
    },
    {
        "type": "tablet",
        "id": "f14afff37f4d42df85517fa1c33a5c74",
        "attributes": {
            "partition": "RANGE (day) PARTITION \"2019-09-27\" <= VALUES < \"2019-09-27\\000\"",
            "table_name": "impala::music_kudu_internal.kudu_rest_test_lhm_33",
            "table_id": "a6e486e1c35f4886a383474a88280eb3"
        },
        "metrics": [
            {
                "name": "on_disk_size",
                "value": 8412044
            },
            {
                "name": "rows_inserted",
                "value": 0
            },
            {
                "name": "rows_upserted",
                "value": 0
            }
        ]
    },
```


务必使用过滤否则
JSON解析很耗时

采集并汇聚统计数据

当前集群统计信息 [最新] 2019-10-10 11:27:10

table	全部空间MB	每分钟删除	每分钟插入	每分钟更新	每分钟upsert	每分钟upsert更新
impala::ka[redacted]	[redacted]	[redacted]	[redacted]	0	[redacted]	8
impala::k[redacted]	[redacted]	[redacted]	[redacted]	0	[redacted]	0
impala::[redacted]rt	[redacted]	[redacted]	[redacted]	0	[redacted]	0
impala::k[redacted]rt	[redacted]	[redacted]	[redacted]	0	[redacted]	0
impala::k[redacted]internal.[redacted]rt	[redacted]	[redacted]	[redacted]	0	[redacted]	0
impala::[redacted]al.[redacted]	[redacted]	[redacted]	[redacted]	0	[redacted]	0

运维交流

- 节点重启

`--follower_unavailable_considered_failed_sec`

Seconds that a leader is unable to successfully heartbeat to a follower after which the follower is considered to be failed and evicted from the config.

Type	int32
Default	`300`
Tags	runtime,advanced

重启完，不要忘记设置回原值

```
线上保护，设置2小时
./bin/kudu tserver set_flag kudu[redacted] follower_unavailable_considered_failed_sec 7200 --force
./bin/kudu tserver set_flag kudu[redacted] follower_unavailable_considered_failed_sec 7200 --force
./bin/kudu tserver set_flag kudu[redacted] follower_unavailable_considered_failed_sec 7200 --force
./bin/kudu tserver set_flag kudu[redacted] follower_unavailable_considered_failed_sec 7200 --force
```

运维交流

- 滚动升级

GitHub, Inc. [US] | https://github.com/apache/kudu/blob/master/docs/installation.adoc#upgrade

Upgrade from a Previous Version of Kudu

Before upgrading, you should read the Release Notes for the version of Kudu that you are about to install. Pay close attention to the incompatibilities, upgrade, and downgrade notes that are documented there.

Warning	The following upgrade process is only relevant when you have binaries available.

支持滚动升级
升级前先重启一轮

1. Prepare the software.
   - Place the new `kudu-tserver`, `kudu-master`, and `kudu` binaries into the appropriate Kudu binary directory.
2. Upgrade the tablet servers.
   - Set the `follower_unavailable_considered_failed_sec` configuration to a high value (conservatively, twice the expected restart time) to prevent tablet replicas hosted on restarting tablet servers from being evicted and re-replicated.

```
$ ./kudu tserver set_flag <tserver> follower_unavailable_considered_failed_sec 7200
```

   - Restart one tablet server.
   - Wait for all tablet replicas on the tablet server to finish bootstrapping by viewing `/tablets` page in the tablet server web UI.
   - Restarting the tablet server will have reset the `follower_unavailable_considered_failed_sec` configuration. Raise it again as needed.
   - Repeat the previous 3 steps for the remaining tablet servers.
   - Restore the original gflag value of every tablet server (the default is 5 minutes)

运维交流

- 在线换盘

安全 | https://kudu.apache.org/docs/command_line_tools_reference.html#fs-update_dirs

`update_dirs`: Updates the set of data directories in an existing Kudu filesystem

If a data directory is in use by a tablet and is removed, the operation will fail unless --force is also used

Usage:

```
kudu fs update_dirs [-force] [-fs_data_dirs=<dirs>] [-fs_metadata_dir=<dir>] [-fs_wal_dir=<dir>]
```

设置维护窗口 → 关闭坏盘节点 → 剔除异常目录 → 更新配置启动节点 → 取消维护窗口 → 机房换盘

回顾总结

- 适合作为事务库的分析库
- 适合存储实时的日志数据
- 提供快速的数据分析能力

Geekbang. 极客邦科技 | InfoQ 极客传媒

InfoQ官网 全新改版上线

促进软件开发领域知识与创新的传播

关注InfoQ网站
第一时间浏览原创IT新闻资讯

免费下载迷你书
阅读一线开发者的技术干货

THANKS

Global
Architect Summit