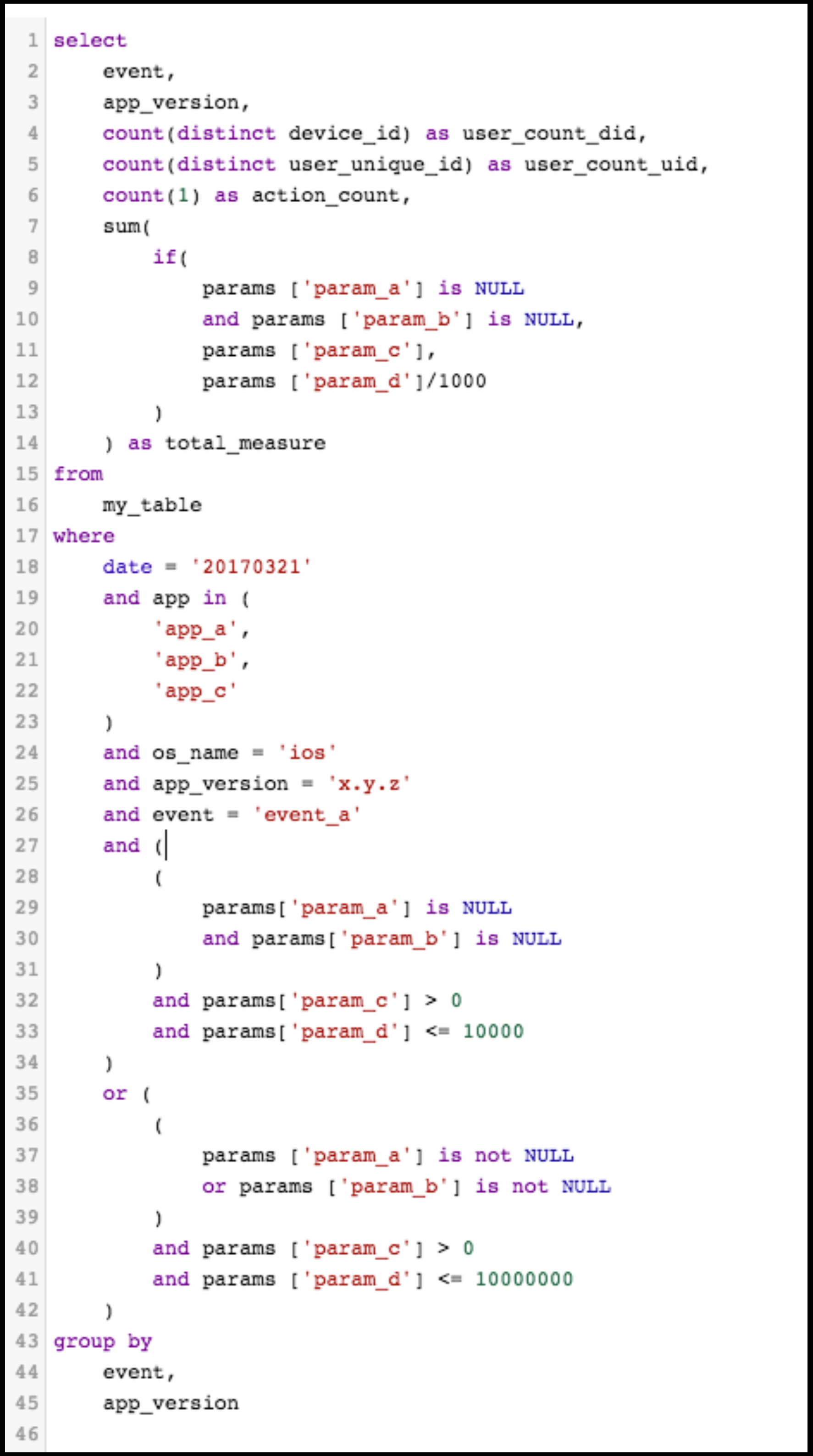
OLAP at Toutiao

yangchaozhong@bytedance.com

Challenge

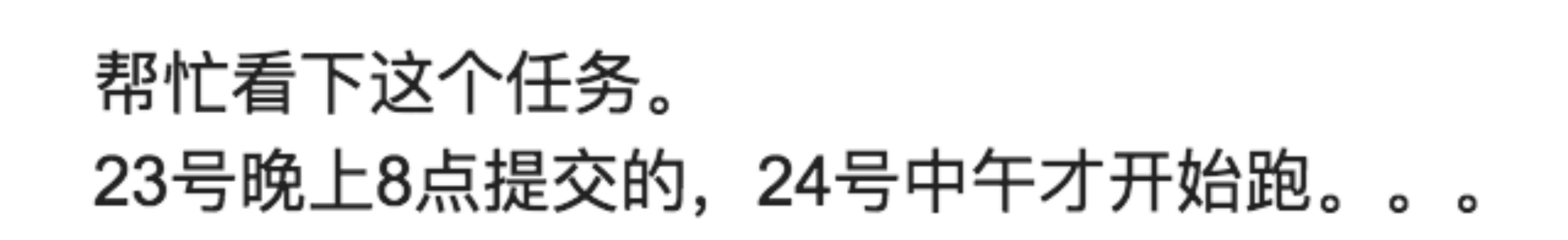
- 14000+ Hive Table
- 20+ billion rows for Top 10 daily partitions (300+ billion rows for Top 1)
- Data Security
- Variety of complex HQLs
- 50+ HS2 instances running on Marathon with Consul (trouble shooting)

Case study



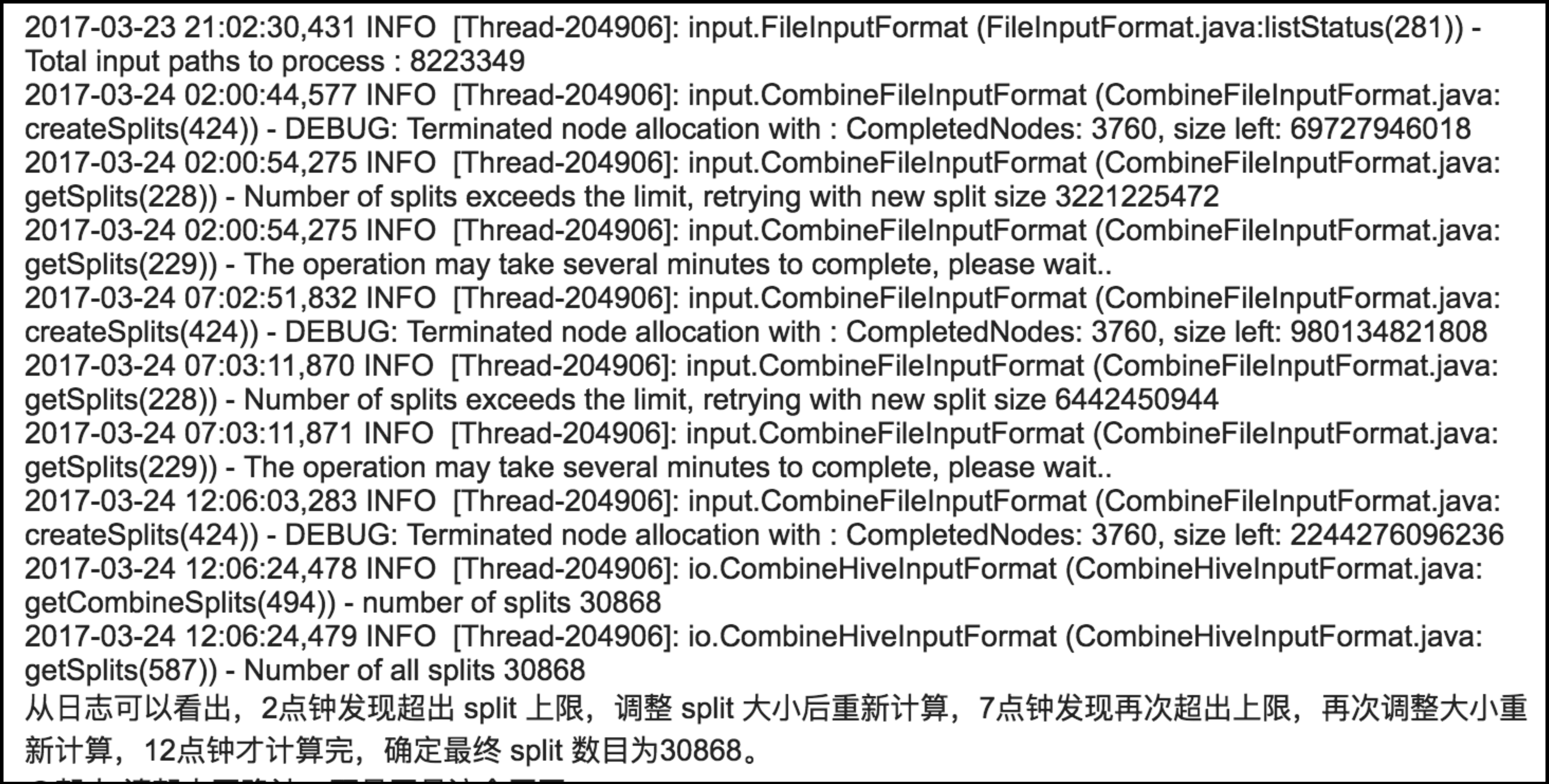
```sql
select
    event,
    app_version,
    count(distinct device_id) as user_count_did,
    count(distinct user_unique_id) as user_count_uid,
    count(1) as action_count,
    sum(
        if(
            params ['param_a'] is NULL
            and params ['param_b'] is NULL,
            params ['param_c'],
            params ['param_d']/1000
        )
    ) as total_measure
from
    my_table
where
    date = '20170321'
    and app in (
        'app_a',
        'app_b',
        'app_c'
    )
    and os_name = 'ios'
    and app_version = 'x.y.z'
    and event = 'event_a'
    and (
        (
            params['param_a'] is NULL
            and params['param_b'] is NULL
        )
        and params['param_c'] > 0
        and params['param_d'] <= 10000
    )
    or (
        (
            params ['param_a'] is not NULL
            or params ['param_b'] is not NULL
        )
        and params ['param_c'] > 0
        and params ['param_d'] <= 10000000
    )
group by
    event,
    app_version

```

Case study



帮忙看下这个任务。
23号晚上8点提交的，24号中午才开始跑。。。

Case study



```
2017-03-23 21:02:30,431 INFO  [Thread-204906]: input.FileInputFormat (FileInputFormat.java:listStatus(281)) -
Total input paths to process : 8223349
2017-03-24 02:00:44,577 INFO  [Thread-204906]: input.CombineFileInputFormat (CombineFileInputFormat.java:
createSplits(424)) - DEBUG: Terminated node allocation with : CompletedNodes: 3760, size left: 69727946018
2017-03-24 02:00:54,275 INFO  [Thread-204906]: input.CombineFileInputFormat (CombineFileInputFormat.java:
getSplits(228)) - Number of splits exceeds the limit, retrying with new split size 3221225472
2017-03-24 02:00:54,275 INFO  [Thread-204906]: input.CombineFileInputFormat (CombineFileInputFormat.java:
getSplits(229)) - The operation may take several minutes to complete, please wait..
2017-03-24 07:02:51,832 INFO  [Thread-204906]: input.CombineFileInputFormat (CombineFileInputFormat.java:
createSplits(424)) - DEBUG: Terminated node allocation with : CompletedNodes: 3760, size left: 980134821808
2017-03-24 07:03:11,870 INFO  [Thread-204906]: input.CombineFileInputFormat (CombineFileInputFormat.java:
getSplits(228)) - Number of splits exceeds the limit, retrying with new split size 6442450944
2017-03-24 07:03:11,871 INFO  [Thread-204906]: input.CombineFileInputFormat (CombineFileInputFormat.java:
getSplits(229)) - The operation may take several minutes to complete, please wait..
2017-03-24 12:06:03,283 INFO  [Thread-204906]: input.CombineFileInputFormat (CombineFileInputFormat.java:
createSplits(424)) - DEBUG: Terminated node allocation with : CompletedNodes: 3760, size left: 2244276096236
2017-03-24 12:06:24,478 INFO  [Thread-204906]: io.CombineHiveInputFormat (CombineHiveInputFormat.java:
getCombineSplits(494)) - number of splits 30868
2017-03-24 12:06:24,479 INFO  [Thread-204906]: io.CombineHiveInputFormat (CombineHiveInputFormat.java:
getSplits(587)) - Number of all splits 30868
```

从日志可以看出，2点钟发现超出 split 上限，调整 split 大小后重新计算，7点钟发现再次超出上限，再次调整大小重新计算，12点钟才计算完，确定最终 split 数目为30868。

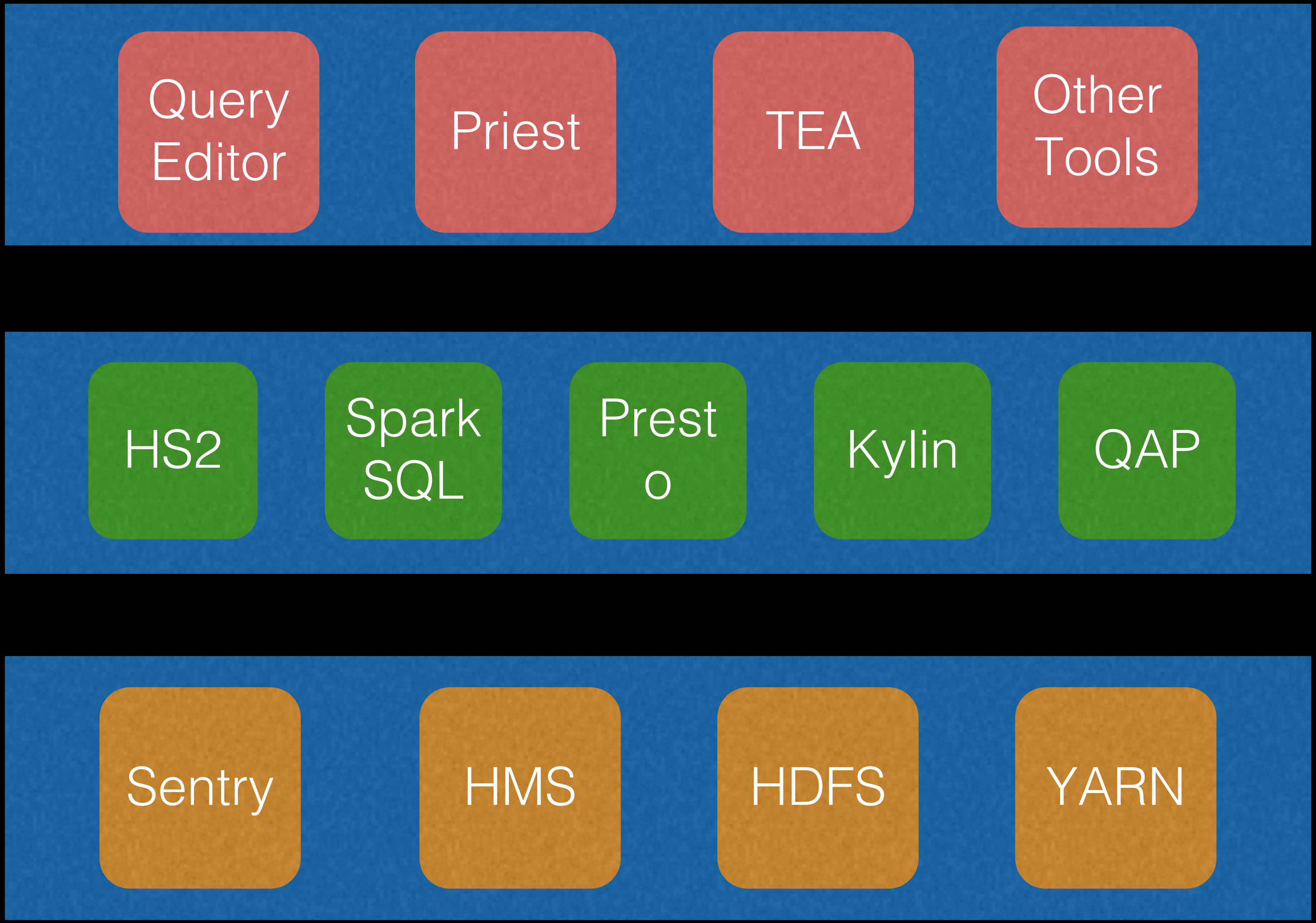
Case study

- Operator Priority: and > or
  - A and B or C <> A and (B or C)
- PartitionPruner#compactExpr

Open Source Projects

- Apache Hive (HMS + HS2)
- Apache Spark (SQL)
- Presto
- Apache Kylin (Data Cube)
- Apache Sentry (Authorization)

Architecture Overview



Query Editor

Priest

TEA

Other Tools

HS2

Spark SQL

Prest
o

Kylin

QAP

Sentry

HMS

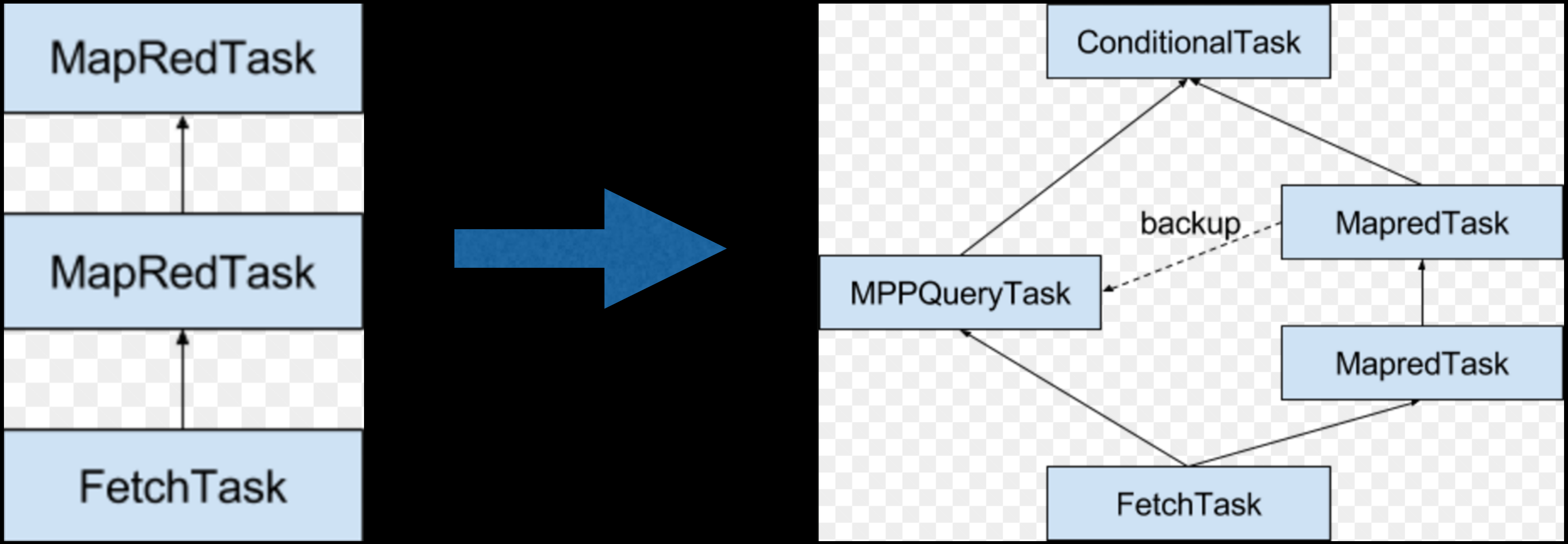
HDFS

YARN

What we did?

- Introduce Presto
  - rewrite HQL to ANSI SQL
  - deployed on YARN by slider
  - HiveMPP

What we did?



MapRedTask

MapRedTask

FetchTask

ConditionalTask

backup

MapredTask

MPPQueryTask

MapredTask

FetchTask

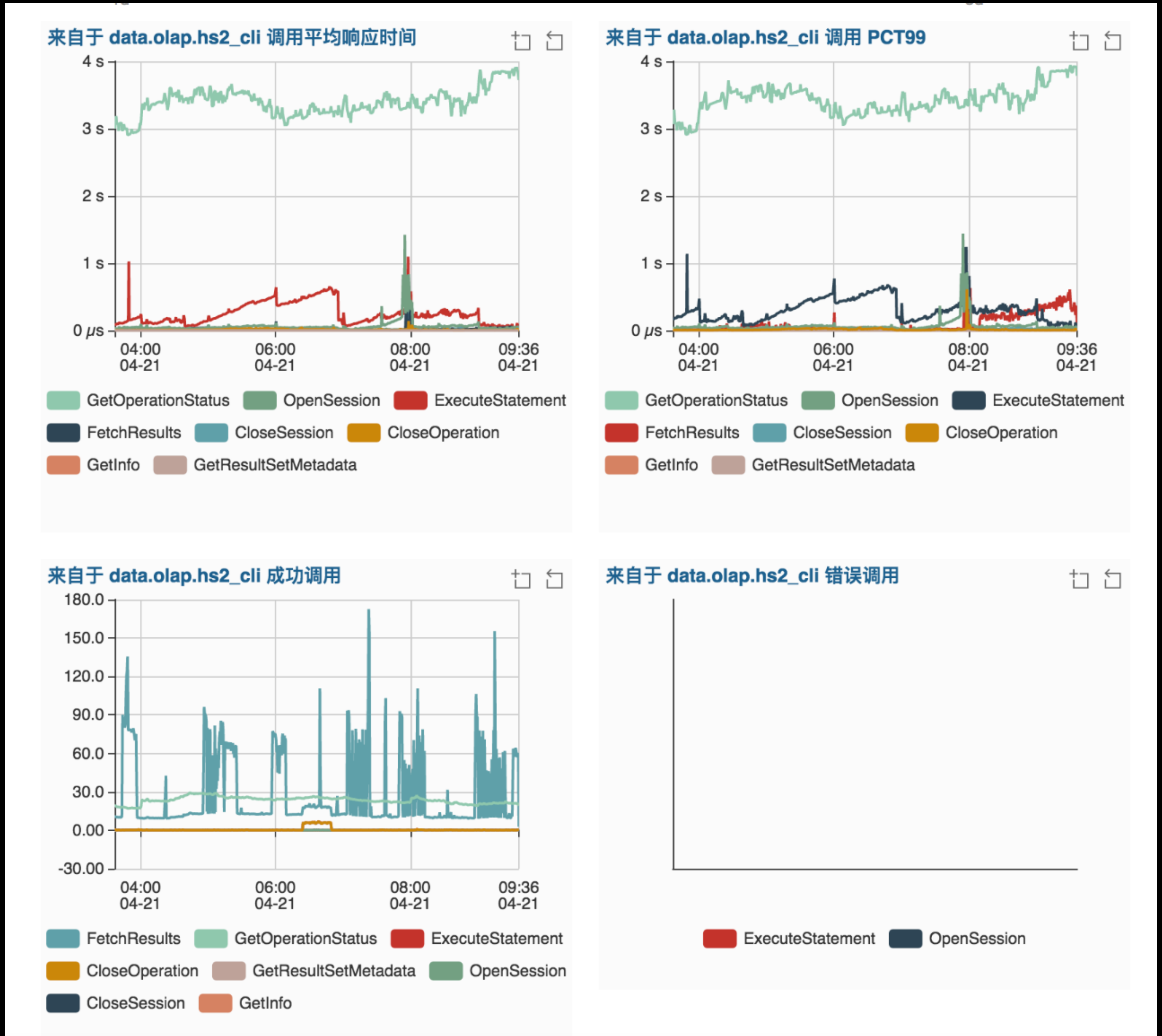
What we did?

- Query Analysis Platform (QAP)
    - Caching HMS & HDFS RPC to speed up HQL semantic analysis.
    - Extract Query Cost Features. (Cardinality Estimation)
    - Predict elapsed time for every HQL query. (decision tree regressor)

What we did?

- HMS & HS2 as a Service
  - We have 50+ HS2 instances
  - Emit metrics for Every HS2 RPC call.

What we did?

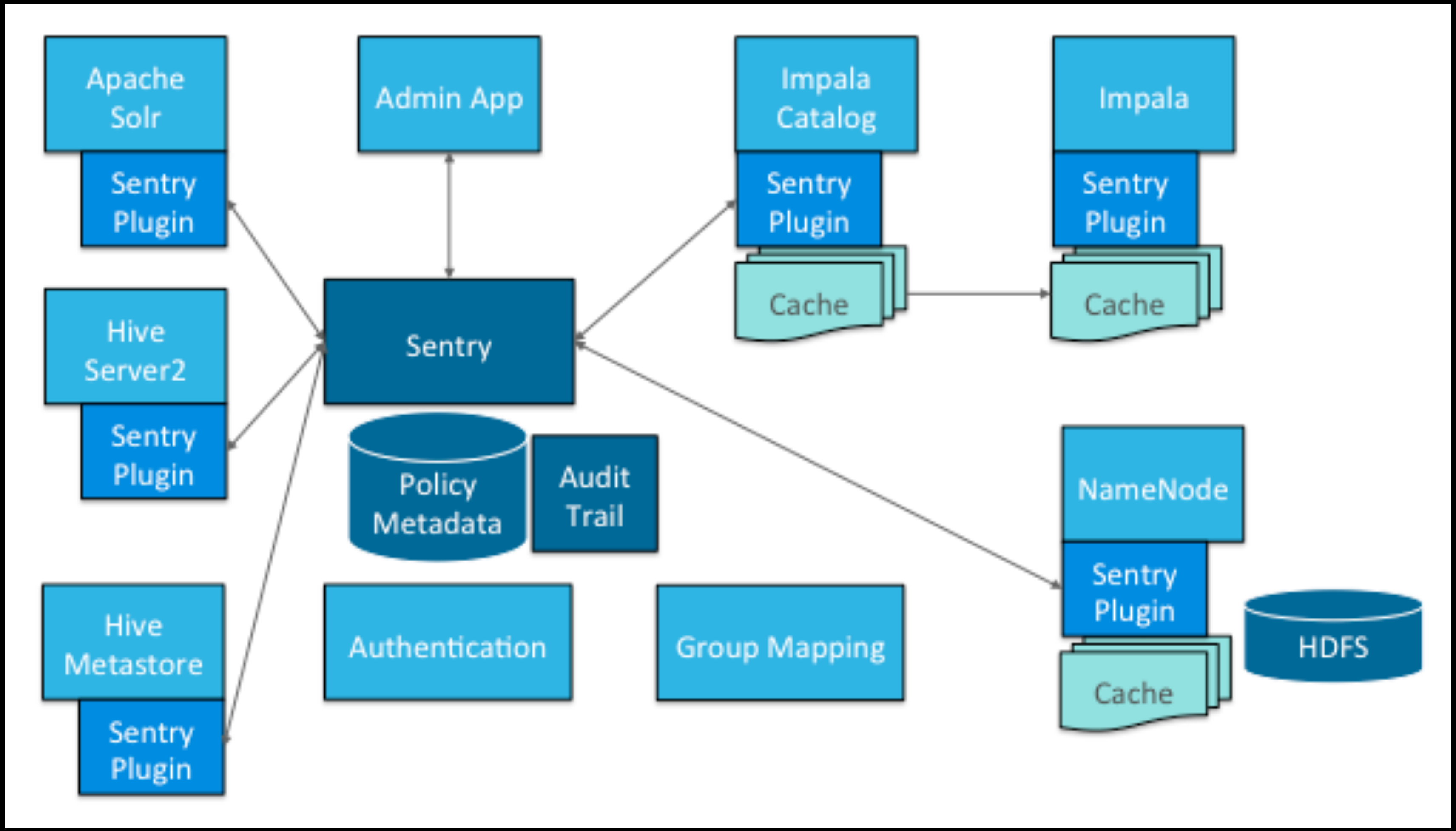


Metrics for Every HS2 RPC call

What we did?

- Introduce Apache Sentry
    - Integrated into our people system.
    - Work at HS2/HMS/CLI as an authorization plugin.
    - Hook: preAnalyze && postAnalyze

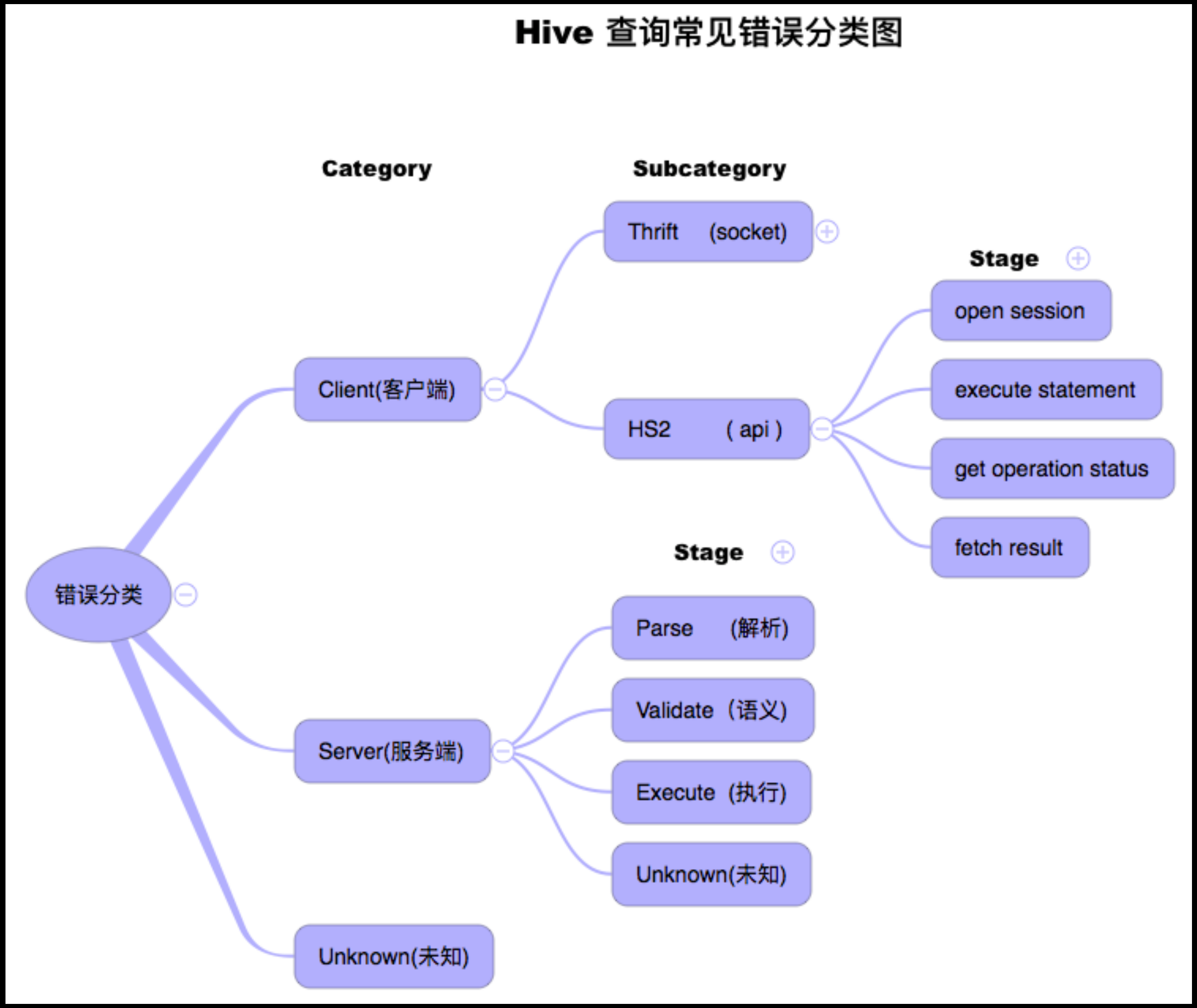
What we did?


Apache Solr
Sentry Plugin
Admin App
Impala Catalog
Sentry Plugin
Cache
Impala
Sentry Plugin
Cache
Hive Server2
Sentry Plugin
Sentry
Policy Metadata
Audit Trail
NameNode
Sentry Plugin
Cache
HDFS
Hive Metastore
Sentry Plugin
Authentication
Group Mapping

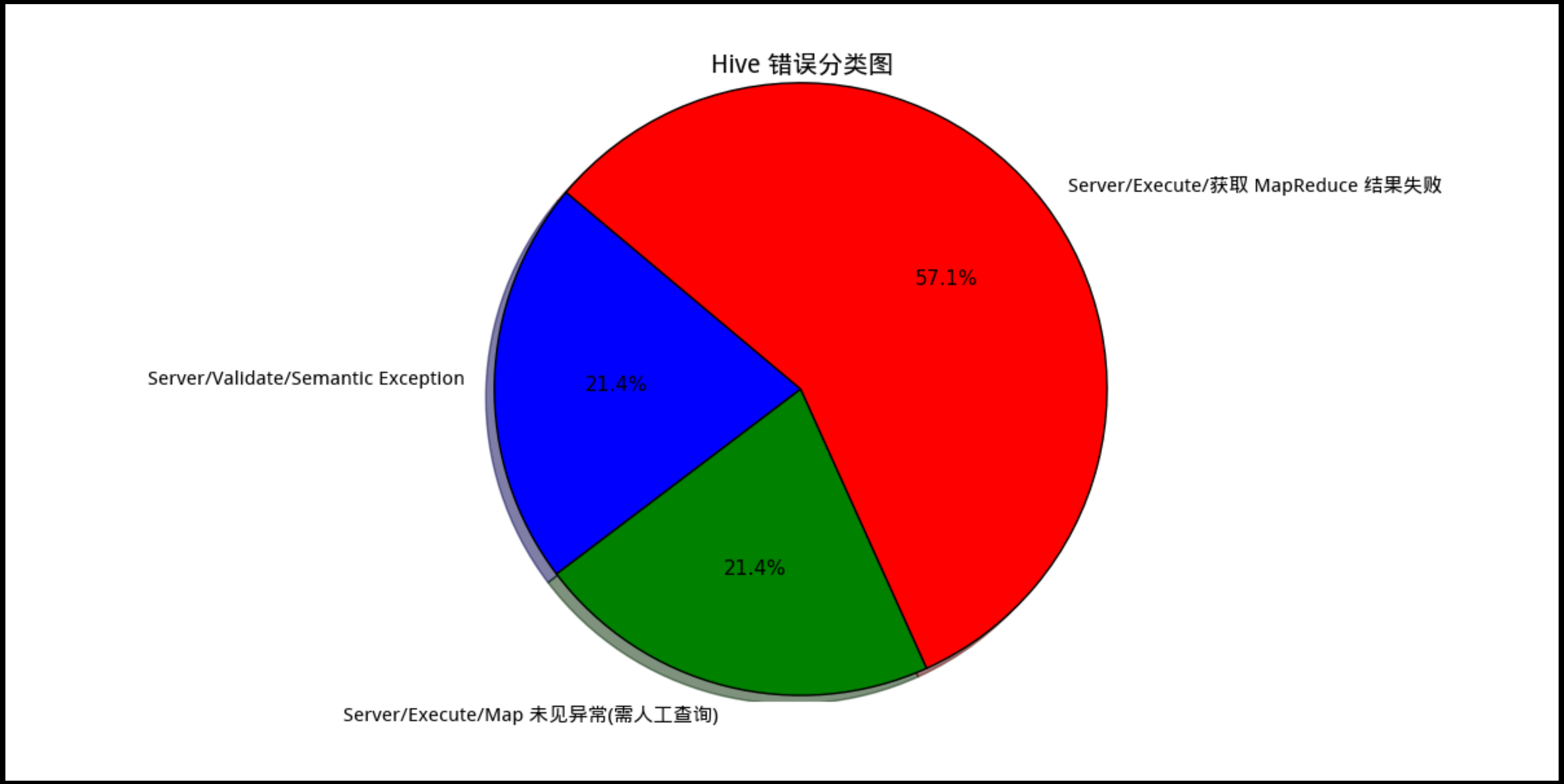
What we did?

- Category query error log
    - Client or Server ?
    - parse / validate / execute ?

What we did?


Hive 查询常见错误分类图

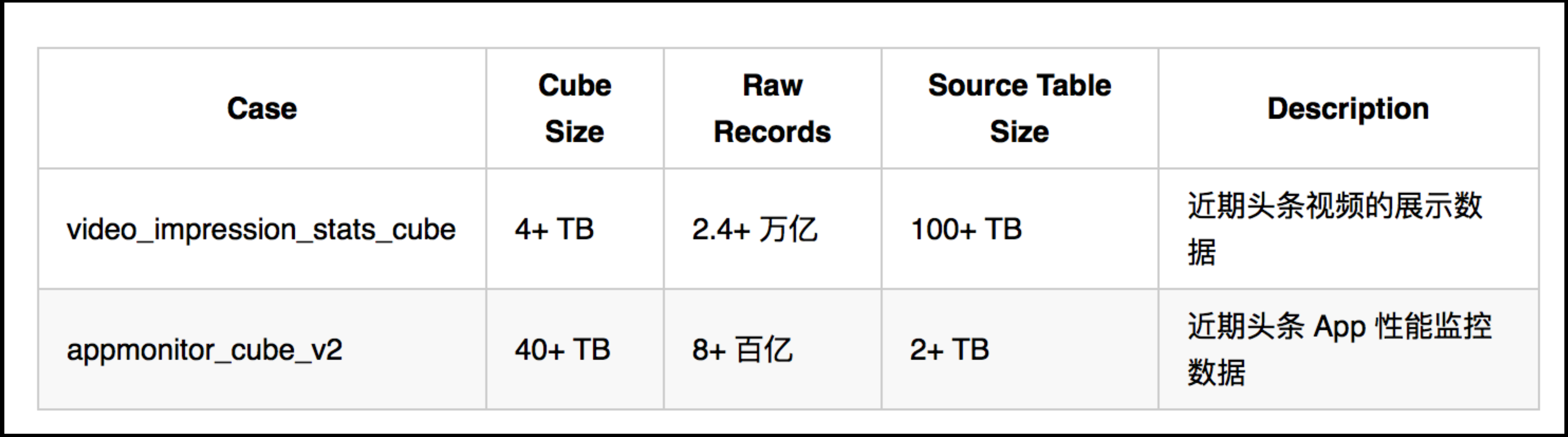
What we did?


Hive 错误分类图
Server/Execute/获取 MapReduce 结果失败
57.1%
Server/Validate/Semantic Exception
21.4%
21.4%
Server/Execute/Map 未见异常(需人工查询)

What we did?

- Introduce Apache Kylin to speed up multi-dimensional analytics.
    - improve cuboid spanning algorithm
    - CuboidJob is triggered by chronos task
    - auto resume CuboidJob
    - popcnt & lzcnt

What we did?



Case	Cube Size	Raw Records	Source Table Size	Description
video_impression_stats_cube	4+ TB	2.4+ 万亿	100+ TB	近期头条视频的展示数据
appmonitor_cube_v2	40+ TB	8+ 百亿	2+ TB	近期头条 App 性能监控数据

Future work

- Integrate Hive & Spark SQL
- Identify and refuse bad query
- Auto suggestion for HQL
- DevOps improvement

We are hiring!

https://job.toutiao.com