移动数据收集及处理之迷思

Mystery of Mobile Data Collecting & Processing

黄洋成

Chief Architect@TalkingData

TalkingData

中国最大的独立第三方移动数据服务平台

Analytics

移动应用统计分析

Game Analytics

GAME ANALYTICS
TalkingData 用数据说话
TalkingData Game Analytics跨平台游戏数据分析—
全面监测数据指标、深入分析玩家行为、数据驱动游戏设计
52%
26%
60%
获取 30天 免费试用

游戏统计分析

Campaign

推广追踪

Insight

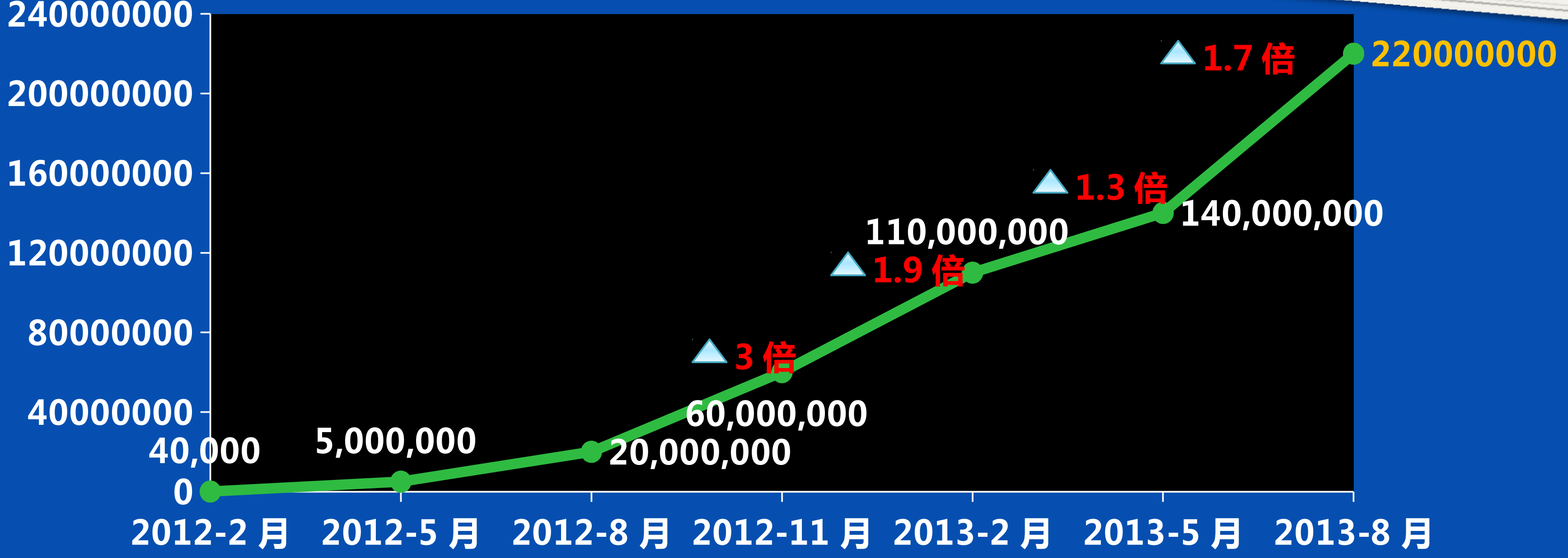
咨询和定制解决方案

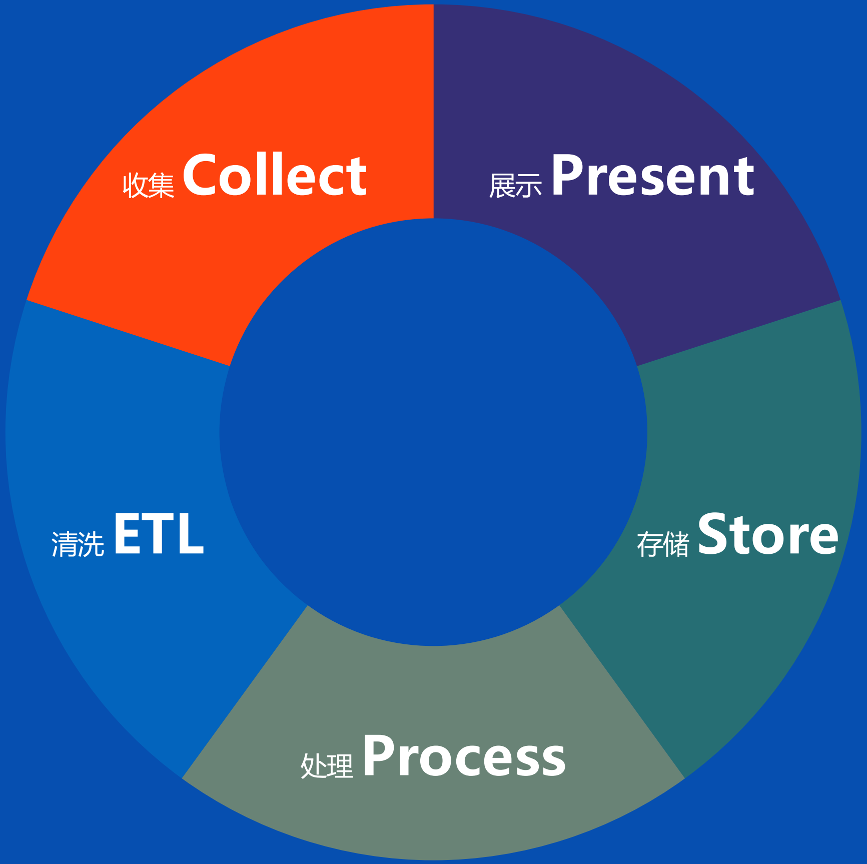
Big Data + Mobile Internet

TalkingData
Mobile·Data·Value

过去的日子

TalkingData
自发布以来，我们的 SDK 已覆盖了
超过 **3 亿**台设备。
为超过 12000 款 App 提供数据服务。




收集 Collect
展示 Present
存储 Store
处理 Process
清洗 ETL

WHAT to collect

DEVICE
IDentification

- IMEI?
- MAC?
- IDFA?

SESSION
RECognition

CONTEXTUAL
DATA

- Network
- Location
- Sensors
- Voice/Image

InfoQ

TalkingData
Mobile·Data·Value

HOW to collect

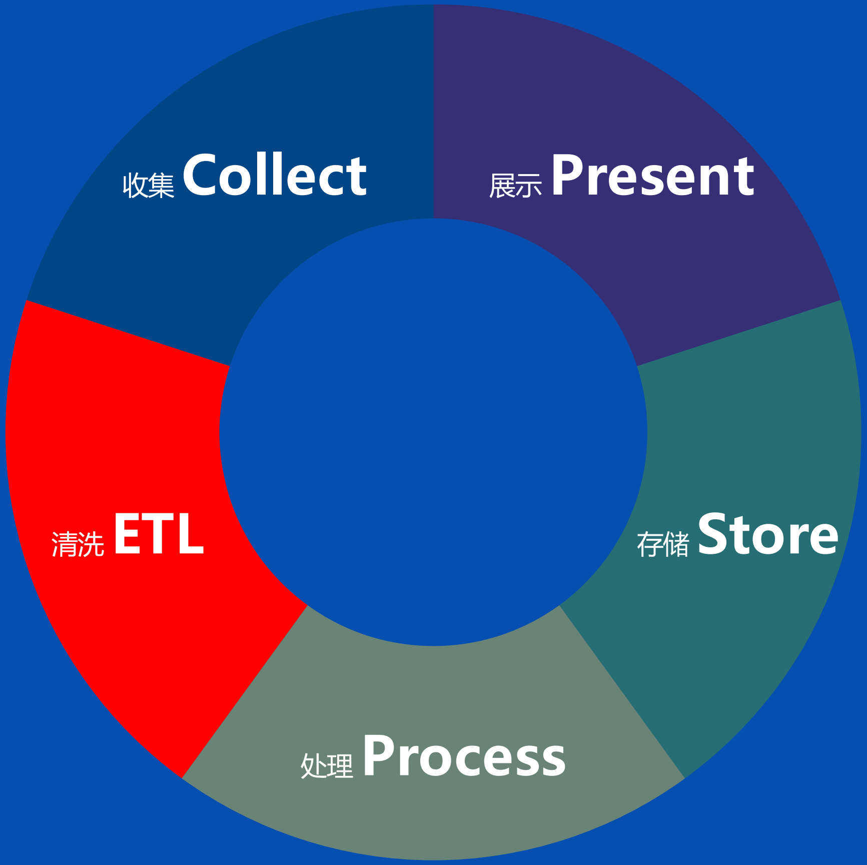
STORE and Forward

- Unreliable connection
- Unreliable execution

CLOCK Issues

POWER Consumption

- Log vs structured
- Compress
- Keep-alive?


收集 Collect
展示 Present
清洗 ETL
存储 Store
处理 Process

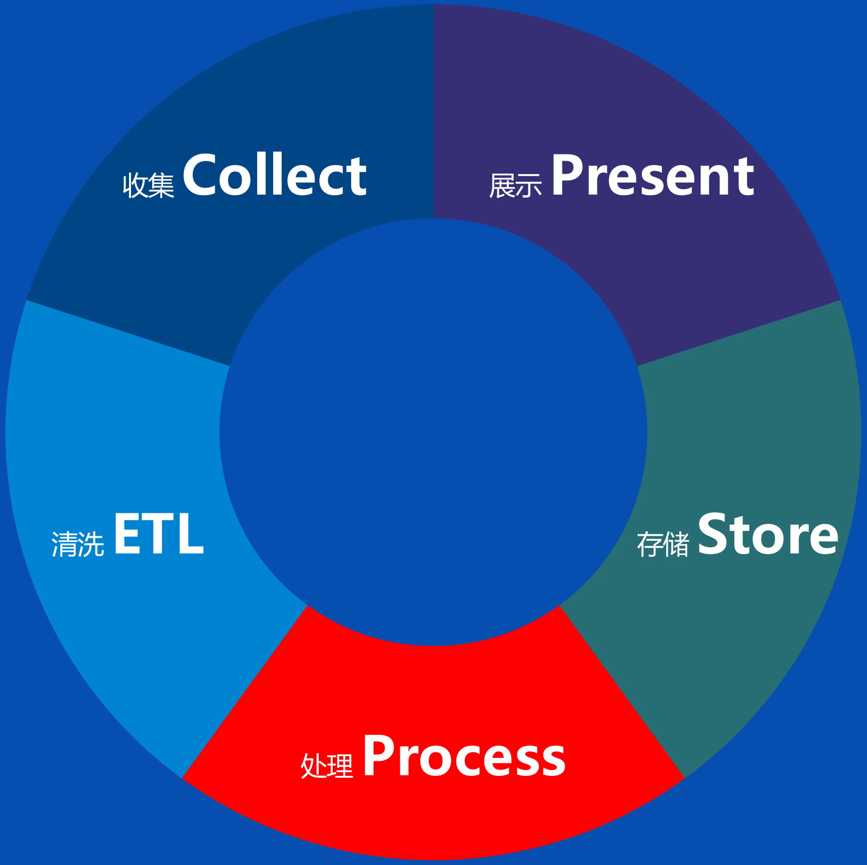
ETL 数据清洗

De-Duplication

Matching

Normalization

- Time
- Location
- Mobile Device
- Carrier


收集 Collect
展示 Present
清洗 ETL
存储 Store
处理 Process

REPORT REQuirement

LATENCY

Action → Report in minutes

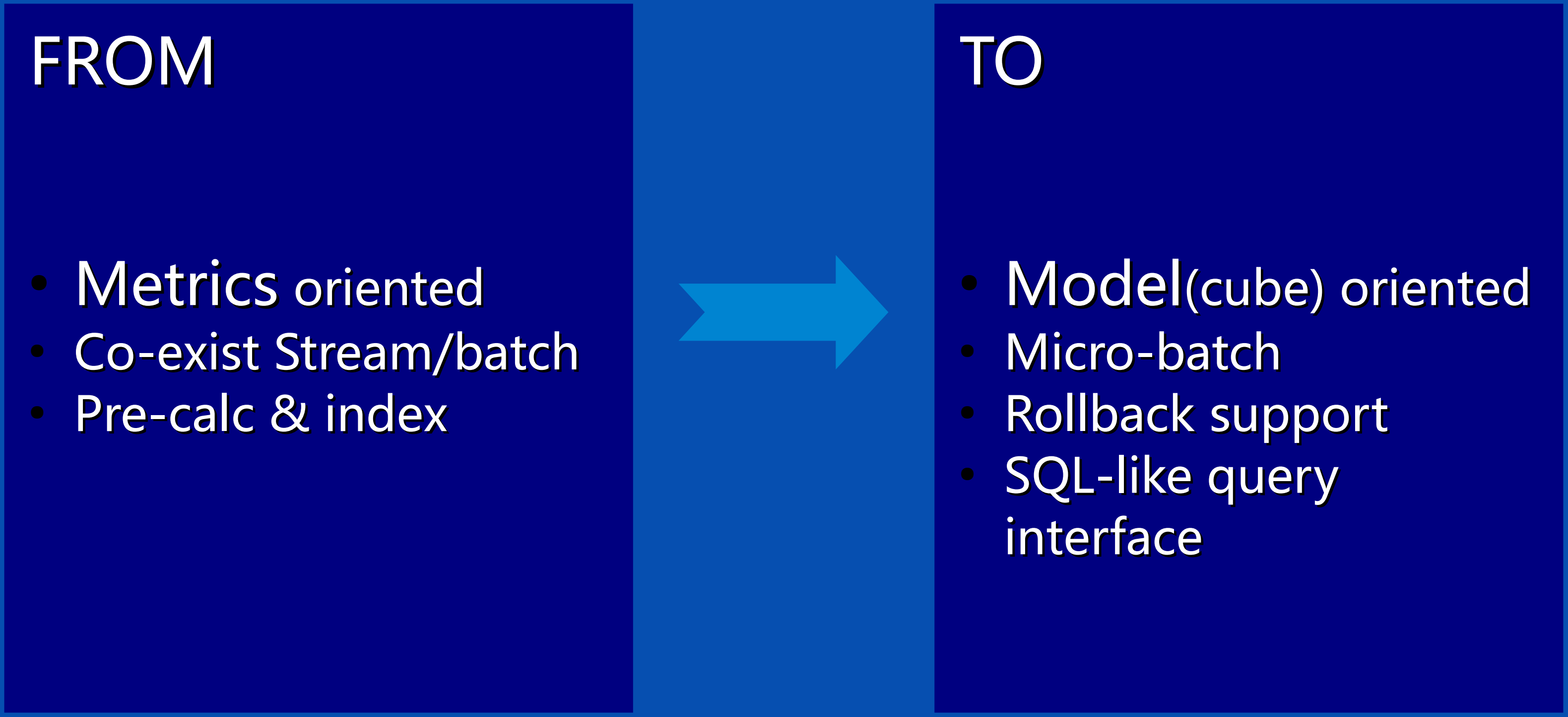
Comprehensive

Report

Quick

Response

REPORT Processing



FROM

- Metrics oriented
- Co-exist Stream/batch
- Pre-calc & index

TO

- Model(cube) oriented
- Micro-batch
- Rollback support
- SQL-like query interface

REPORT Achievements

MODEL based
Processing

QL/Script
Query Front

Report
Tools

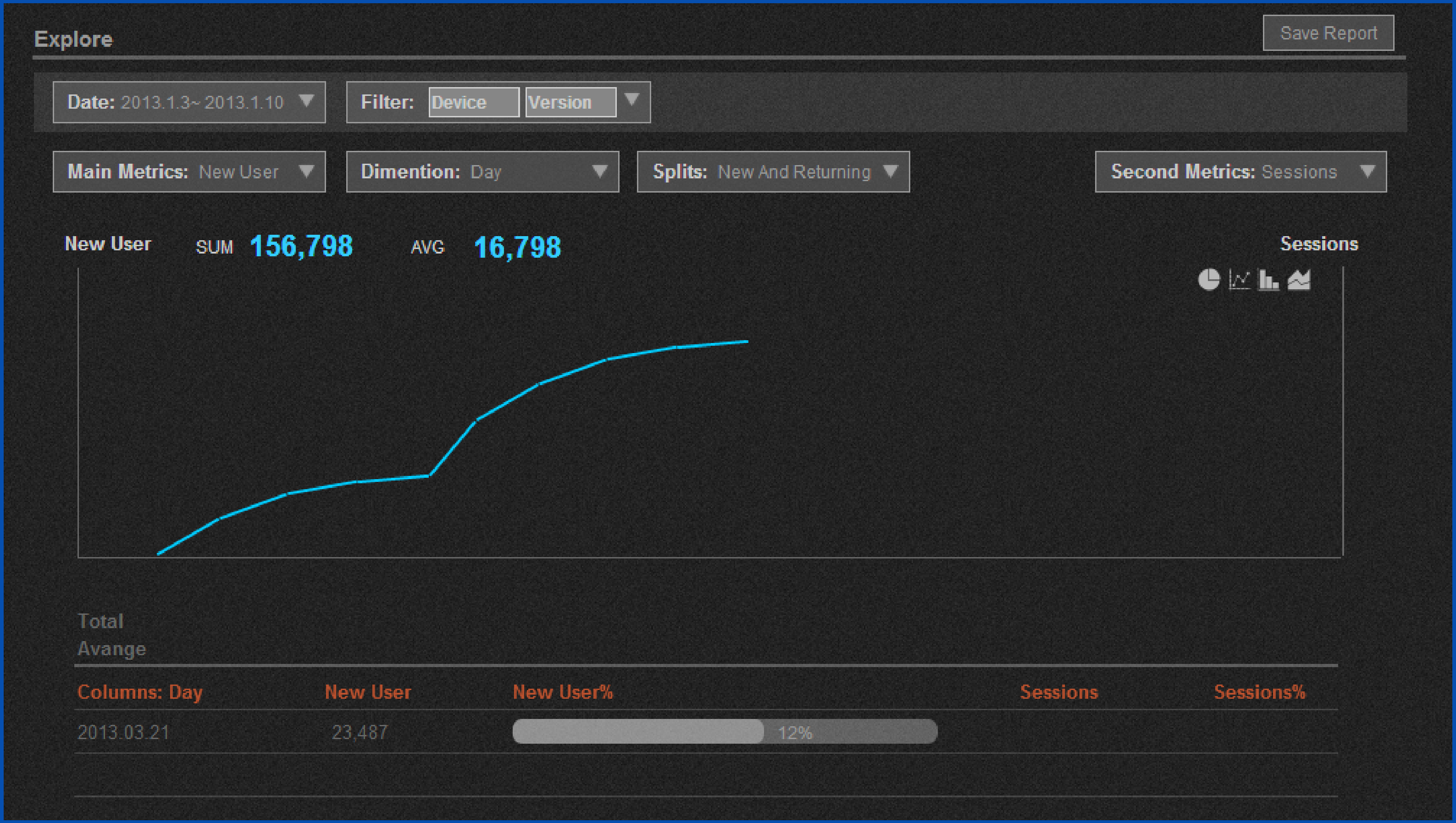
- ongoing

Sample SCRIPT

留存 (Retention) 计算：

```
days = [20131023, 20131024, 20131025];
result = new HashMap<Integer, Integer>();
for(day in days){
    new = select user_id from new_user where product_id =
'1313930' and _day_of_rec_time = day;
    active = select user_id from active_user where product_id =
'1313930' and _day_of_rec_time = (day + 1);
    retention = intersection(new, active);
    result.put(day, retention.cardinality());
}
return result;
```

Target Report



Explore

Date: 2013.1.3~ 2013.1.10

Filter: Device Version

Main Metrics: New User

Dimention: Day

Splits: New And Returning

Second Metrics: Sessions

New User SUM 156,798 AVG 16,798

Sessions

Total
Avange

Columns: Day	New User	New User%	Sessions	Sessions%
2013.03.21	23,487	12%		

InfoQ

TalkingData
Mobile·Data·Value

简单直接最有效

Do it straightforward

关注视角

DATA Mining WORKS

Process

- Azkaban based workflow
- ETL based on Pig
- Machine Learning
  - Most on single machine
  - Some on Spark

Input

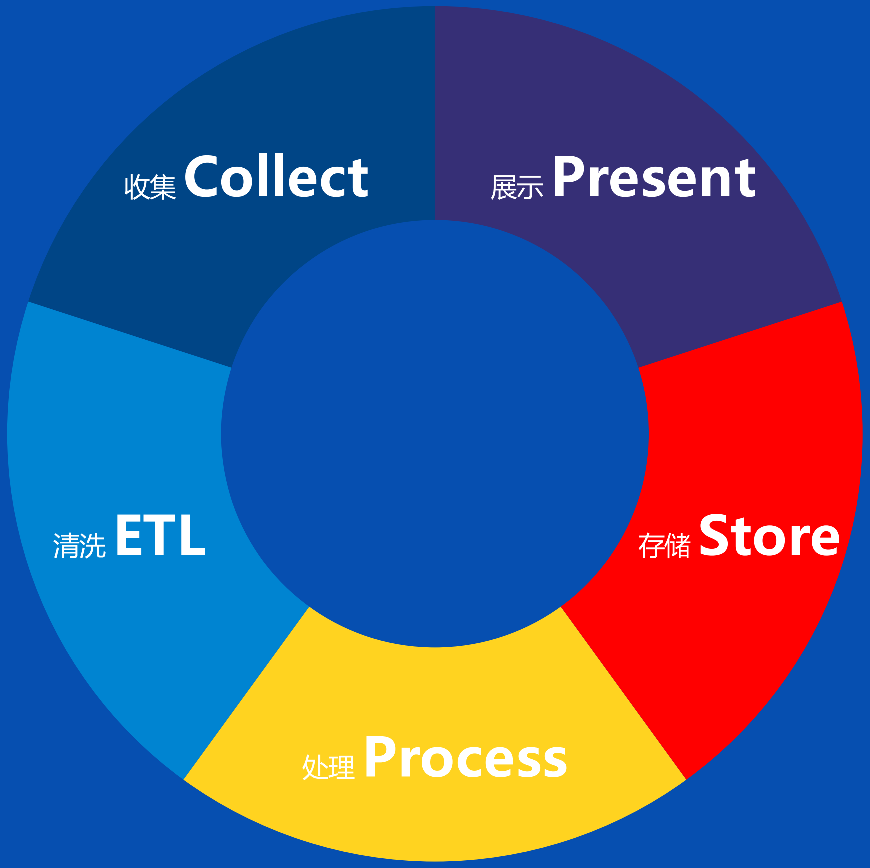
- Apps desc
- Device → Apps
  - Explore to download
  - Use
  - Pay

Output

- ID mapping
- Tags for apps
- Tags for devices
- IP location

Experiments

- App recommend
- Audience targeting Ads



收集 Collect

清洗 ETL

处理 Process

存储 Store

展示 Present

Key Components

Logs

- HDFS
- RCFile/Snappy
- Archived as bzip

META DATA

- Redis

Report DATA

- MySQL/MyISAM
- Toku

NEAR Cache

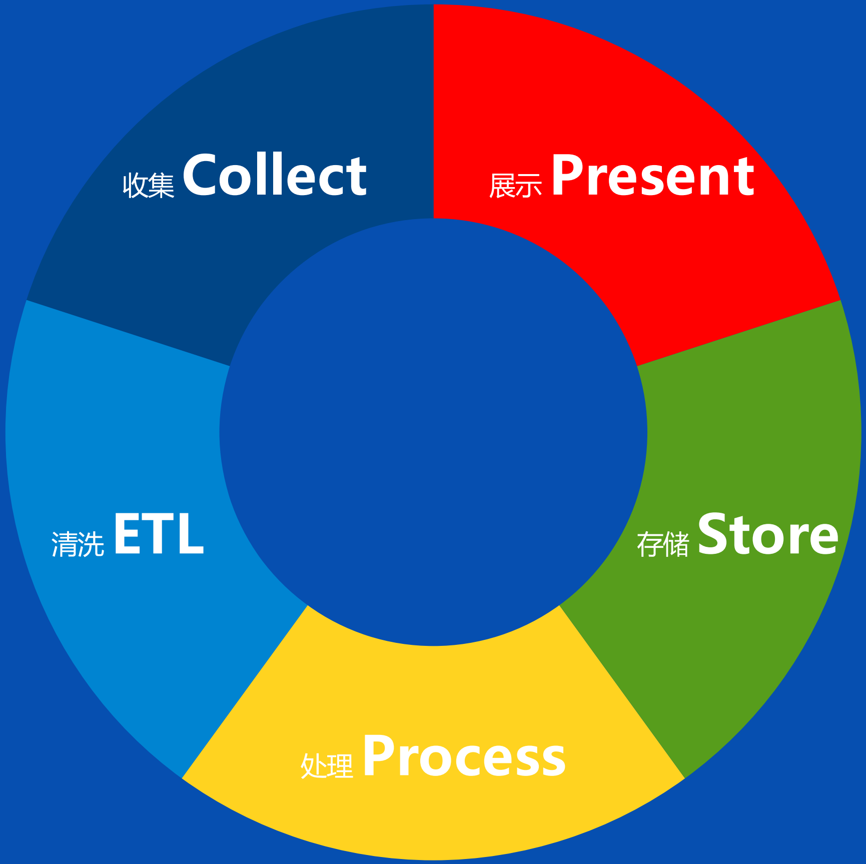
- Off-heap cache
- Off-heap serialization lib
- Off-heap collections

Challenges

NETWORK
DATA Flow Design

Capacity
Plan

Partition
Management


收集 Collect
展示 Present
存储 Store
处理 Process
清洗 ETL

PRESENTATION Challenges

Security

- https
- Authorization check

Complex Pages

- Components
- Rich Webapp
- Service Oriented

Performance

- Parallel
  - Fork-join
  - Intelligent cache
  - Grid computing
  - GPU

谢谢！

Contact:
黄洋成 YC Huang
yc.huang@tendcloud.com
http://weibo.com/yc08
https://www.talkingdata.net/