GOOAGOO购阿购

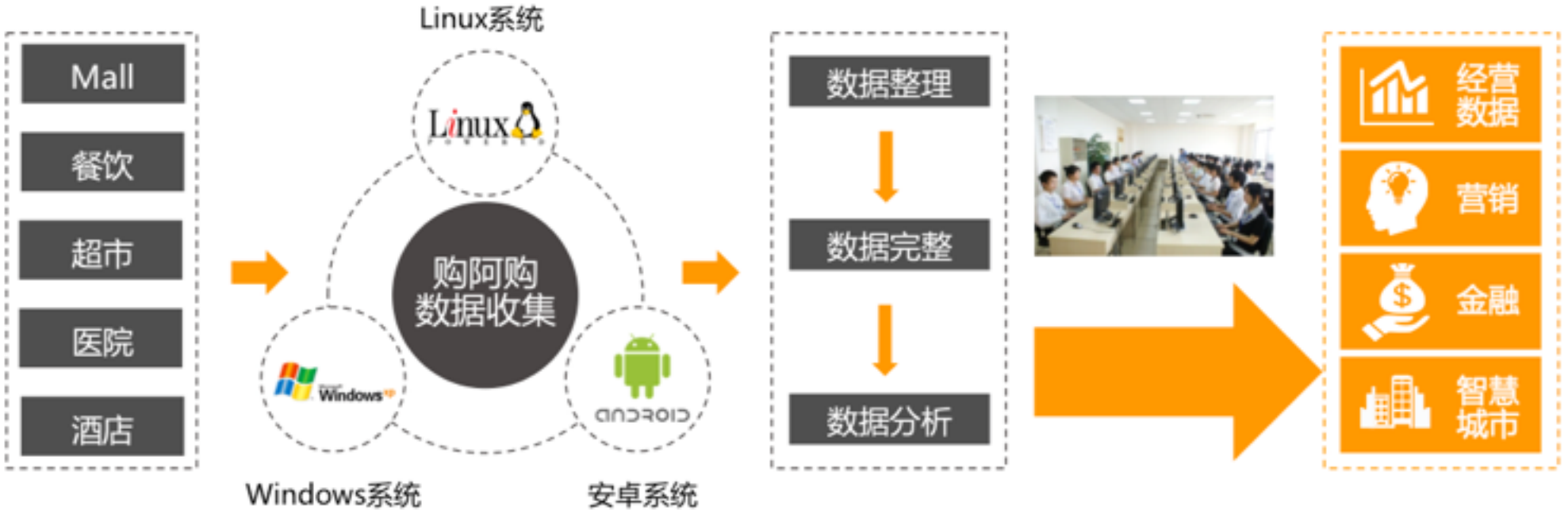
购阿购数据分析平台 ElasticSearch实践

by 万斌

Agenda

- 使用场景
- 需求
- 如何做的技术选型
- ElasticSearch的优势
- 我们怎么做的？
- ElasticSearch不适合做哪些？
- 线上运维

我们的使用场景



多场景收集 → 多系统收集 → 全方位整理 → 数据服务

面临的需求

数据分析的维度和指标多样

数据量的增长迅速

数据是对客户提供服务，对稳定性要求高

小公司、时间紧、人少

技术选型

使用传统数据库技术方案，如MySQL

使用Hadoop生态里的离线处理方案，如使用Hive/Spark等

使用数据分析引擎，如阿里云开放搜索/ElasticSearch等

ElasticSearch的优势

基于分布式、扩展性好

友好的Restful接口

Agg的功能强大

生态丰富

我们怎么做的？

固定mapping字段

Aggreation 功能(terms agg/histogram agg)

Kibana（Marvel/Sense插件)

利用alias切换全量索引

监控JVM的内存变化

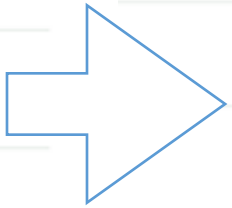
使用terms Aggregation时要注意的问题

	Shard A	Shard B	Shard C
1	Product A (25)	Product A (30)	Product A (45)
2	Product B (18)	Product B (25)	Product C (44)
3	Product C (6)	Product F (17)	Product Z (36)
4	Product D (3)	Product Z (16)	Product G (30)
5	Product E (2)	Product G (15)	Product E (29)
6	Product F (2)	Product H (14)	Product H (28)
7	Product G (2)	Product I (10)	Product Q (2)
8	Product H (2)	Product Q (6)	Product D (1)
9	Product I (1)	Product J (8)	
10	Product J (1)	Product C (4)	

使用terms Aggregation时要注意的问题

The shards will return their top 5 terms so the results from the shards will be:

	Shard A	Shard B	Shard C
1	Product A (25)	Product A (30)	Product A (45)
2	Product B (18)	Product B (25)	Product C (44)
3	Product C (6)	Product F (17)	Product Z (36)
4	Product D (3)	Product Z (16)	Product G (30)
5	Product E (2)	Product G (15)	Product E (29)



Product A (100)
Product Z (52)
Product C (50)
Product G (45)
Product B (43)

ElasticSearch不适合做什么？

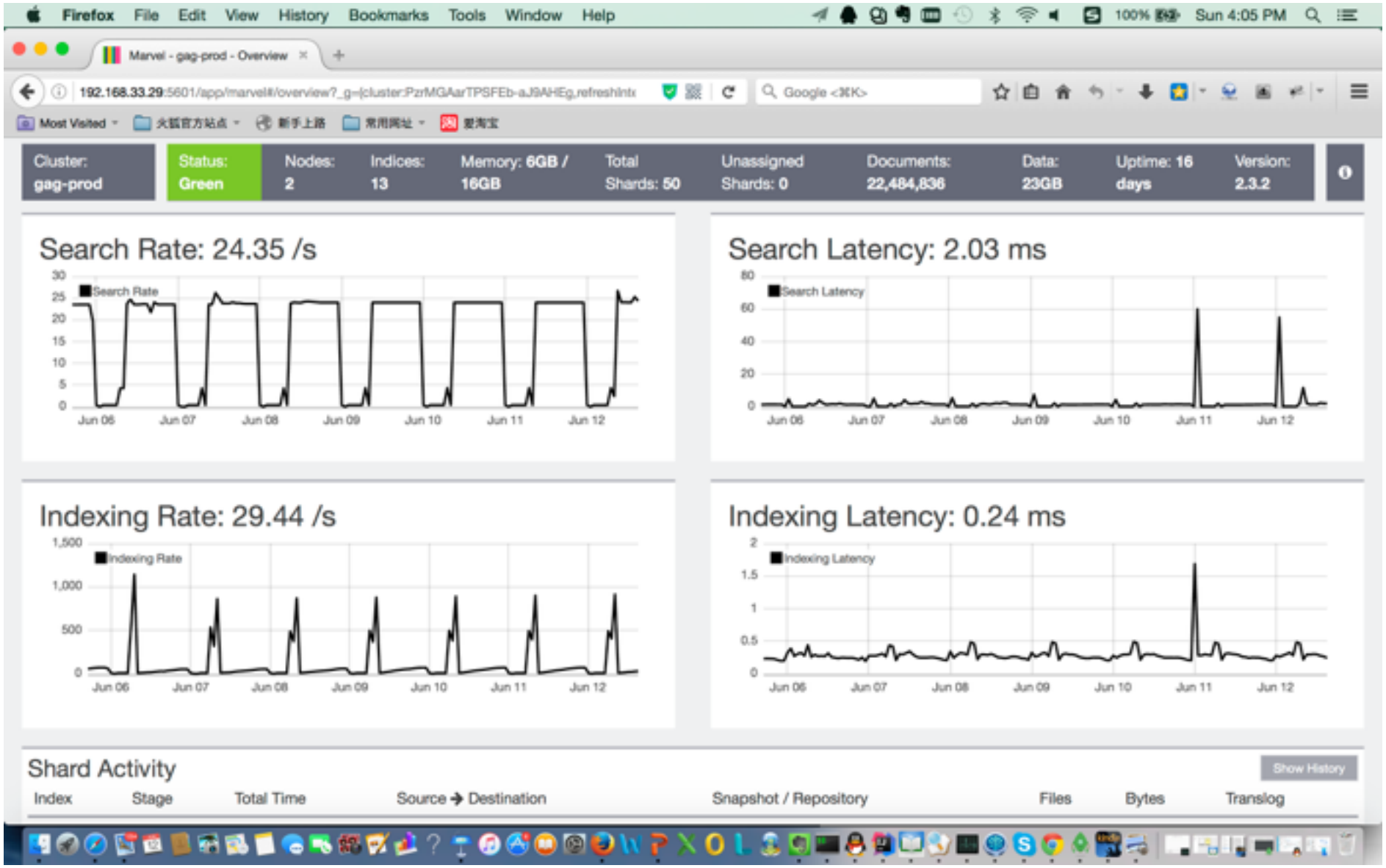
当数据量很大时，要求100%的准确的Agg统计

文档的嵌套层次不宜太多，如果太多，需要做适度的ETL工作

查询接口不如SQL那么友好

对有数据安全要求的需求

我们的线上状况


Firefox File Edit View History Bookmarks Tools Window Help
Marvel - gag-prod - Overview
Most Visited
Cluster: gag-prod
Status: Green
Nodes: 2
Indices: 13
Memory: 6GB / 16GB
Total Shards: 50
Unassigned Shards: 0
Documents: 22,484,836
Data: 23GB
Uptime: 16 days
Version: 2.3.2
Search Rate: 24.35 /s
Search Rate
Search Latency: 2.03 ms
Search Latency
Indexing Rate: 29.44 /s
Indexing Rate
Indexing Latency: 0.24 ms
Indexing Latency
Jun 06
Jun 07
Jun 08
Jun 09
Jun 10
Jun 11
Jun 12
Shard Activity
Show History
Index
Stage
Total Time
Source → Destination
Snapshot / Repository
Files
Bytes
Translog

Q&A