DTCC

2017第八届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2017

小数据到大数据之路

范学鑫

个人微信号

本来生活

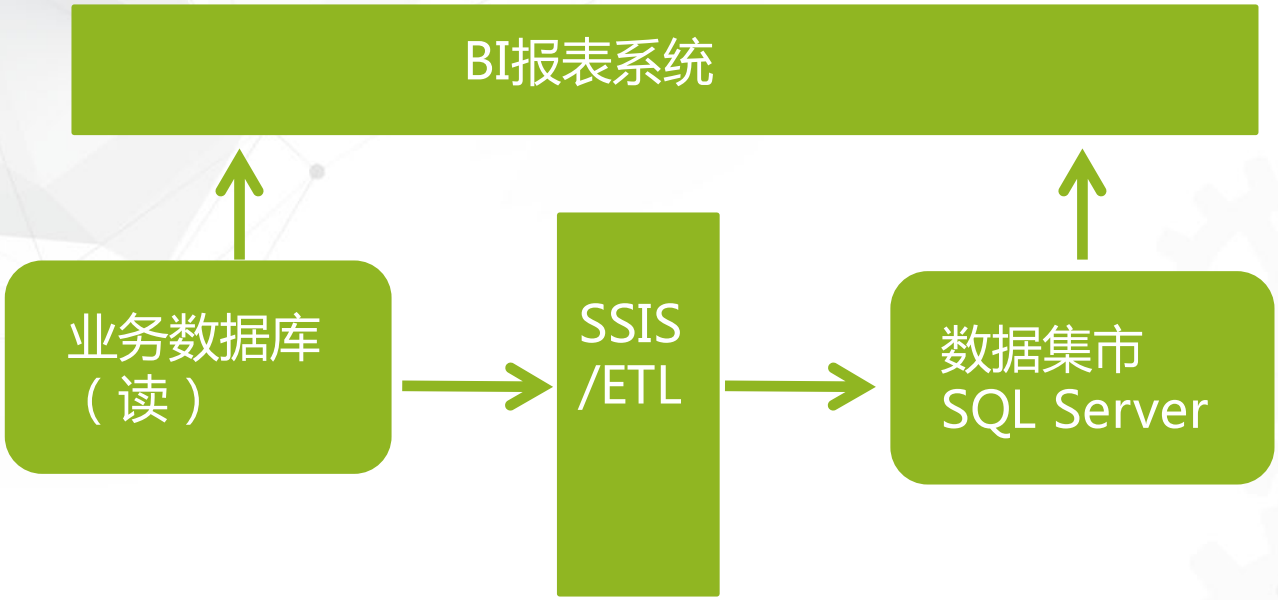
介绍本来生活网数据体系的演变：SQL Server到Hadoop生态

目录

Part I 本来BI从这里开始

本来生活网从褚橙开始走上生鲜电商的舞台

BI报表系统1.0



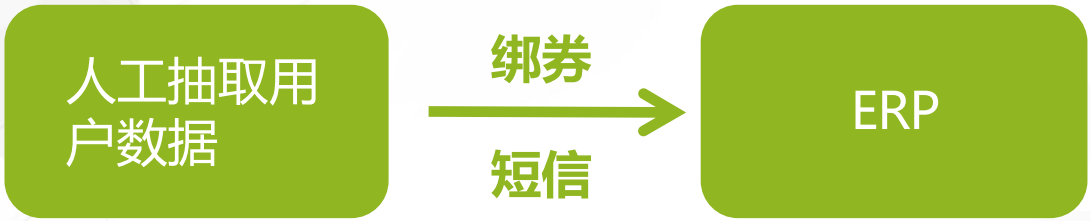
遇到的问题

① 集市数据粒度粗，导致报表的重用性差

② 同样的统计时间段，在不同时间点跑出来的数据不一样

③ 数据口径的统一很困难

CRM ？



遇到的问题

① 大量的人力花在了抽数、绑券这种价值感很低的事情上。

② 出错率居高不下、业务抱怨不断。

③ 人员流动率高。

流量分析1.0？（第三方统计）

① 只能看到概览性质的数据（PV、UV）。

② 很难把流量数据和业务系统内的数据关联。

③ 关键路径的转化率很难统计。

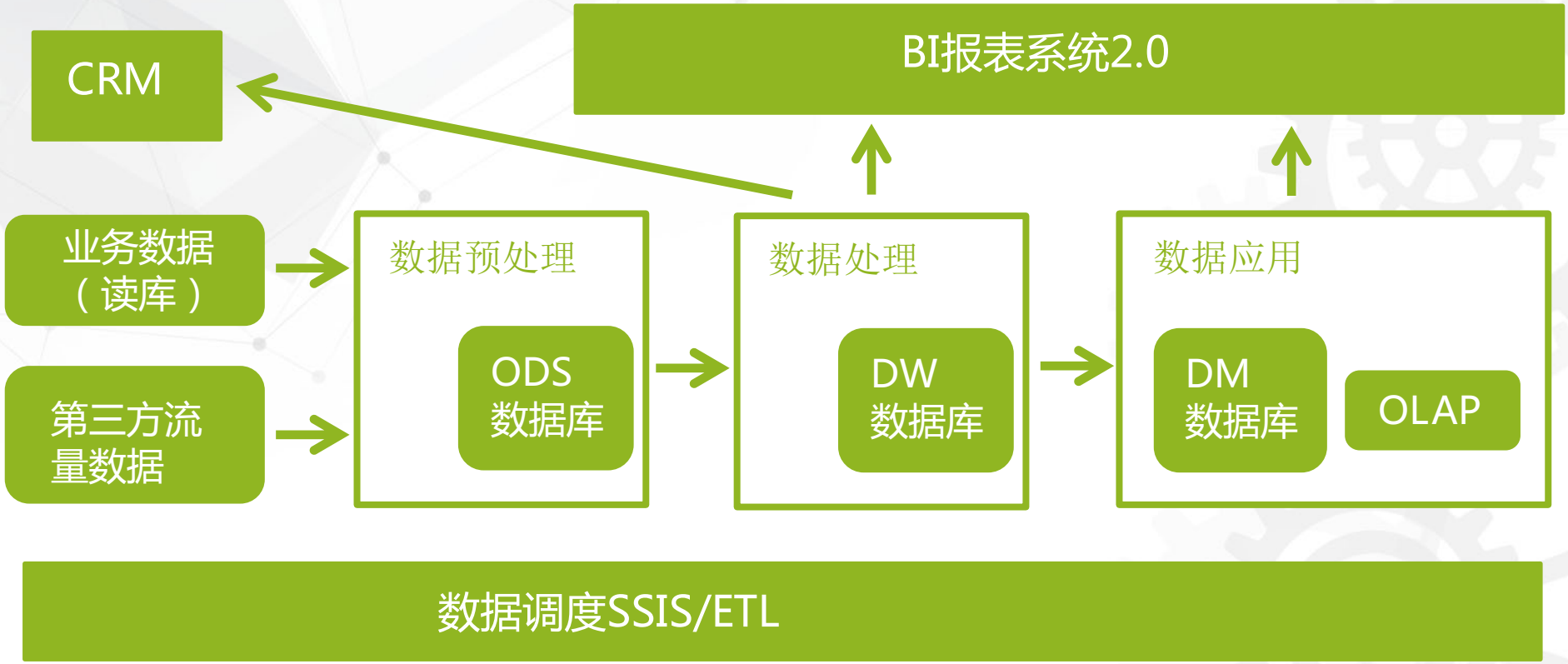
④ 微信的细分流量很难统计。

⑤ ……

Part 2 数据仓库时代的BI

越来越多的水果大品在本来生活网获得了非常好的销量和口碑

BI报表系统2.0(基于数据仓库）



优势和解决的问题

① 在一定数据量下Kimball 的维度建模表现优异。

② 按不同层次的粒度对数据建模，拥有良好的重用性。

③ 数据仓库层统一报表数据源，解决数据一致性问题。

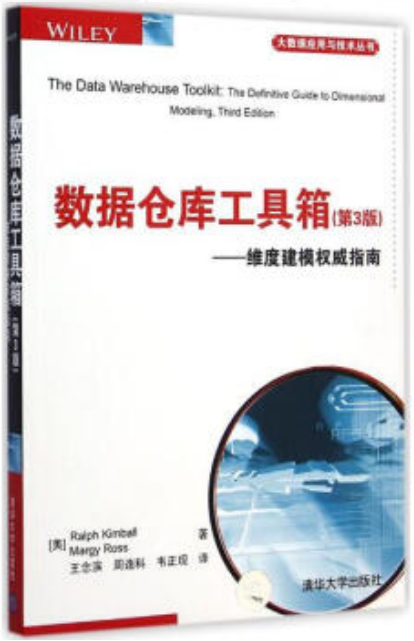
新的问题

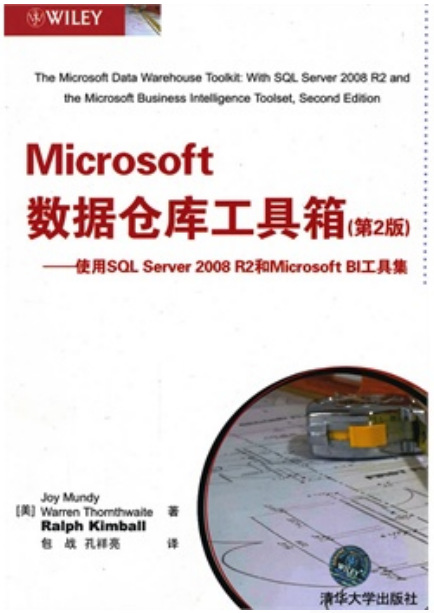
① 流量数据接入后，SQL Server 数据仓库的硬盘空间不够用。

② T+1的方式招到的业务的挑战越来越多。

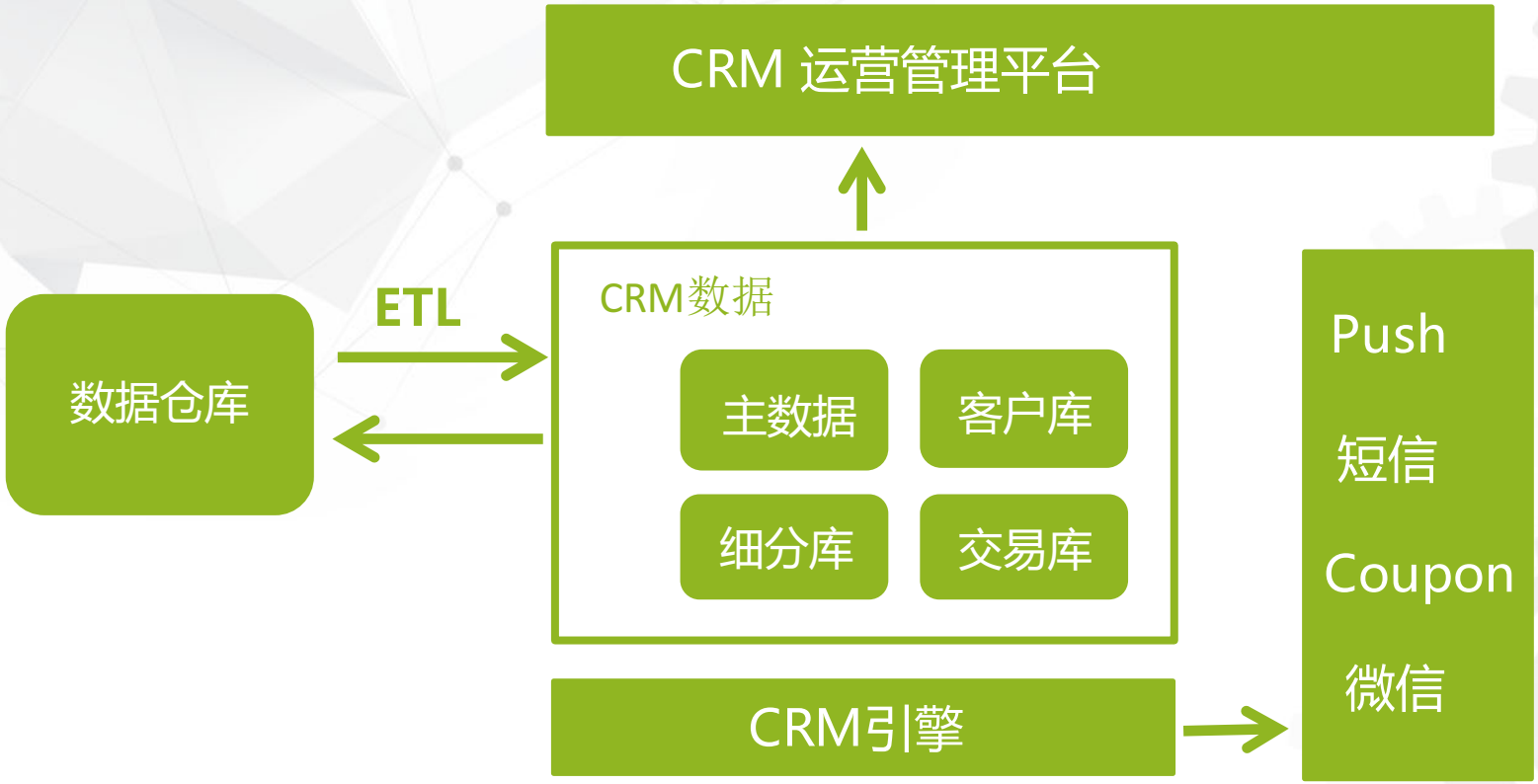
③ ETL的依赖越来越复杂，ETL的执行时间越来越久。

一些建议和推荐


WILEY
The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition
数据仓库工具箱(第3版)
——维度建模权威指南
[美] Ralph Kimball
Margy Ross
王志海 周连科 韦正现 译
清华大学出版社


WILEY
The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset, Second Edition
Microsoft
数据仓库工具箱(第2版)
——使用SQL Server 2008 R2和Microsoft BI工具集
Joy Mundy
[美] Warren Thornthwaite 著
Ralph Kimball
包 战 孔祥亮 译
清华大学出版社

CRM 2.0（基于SQL Server）



优势和解决的问题

① BI人员从常规的抽数中得到了解放。

② 客户选取灵活(交易行为、客户属性、标签等维度进行选取)。

③ 客户触达（短信、优惠券）系统化以及自动化。

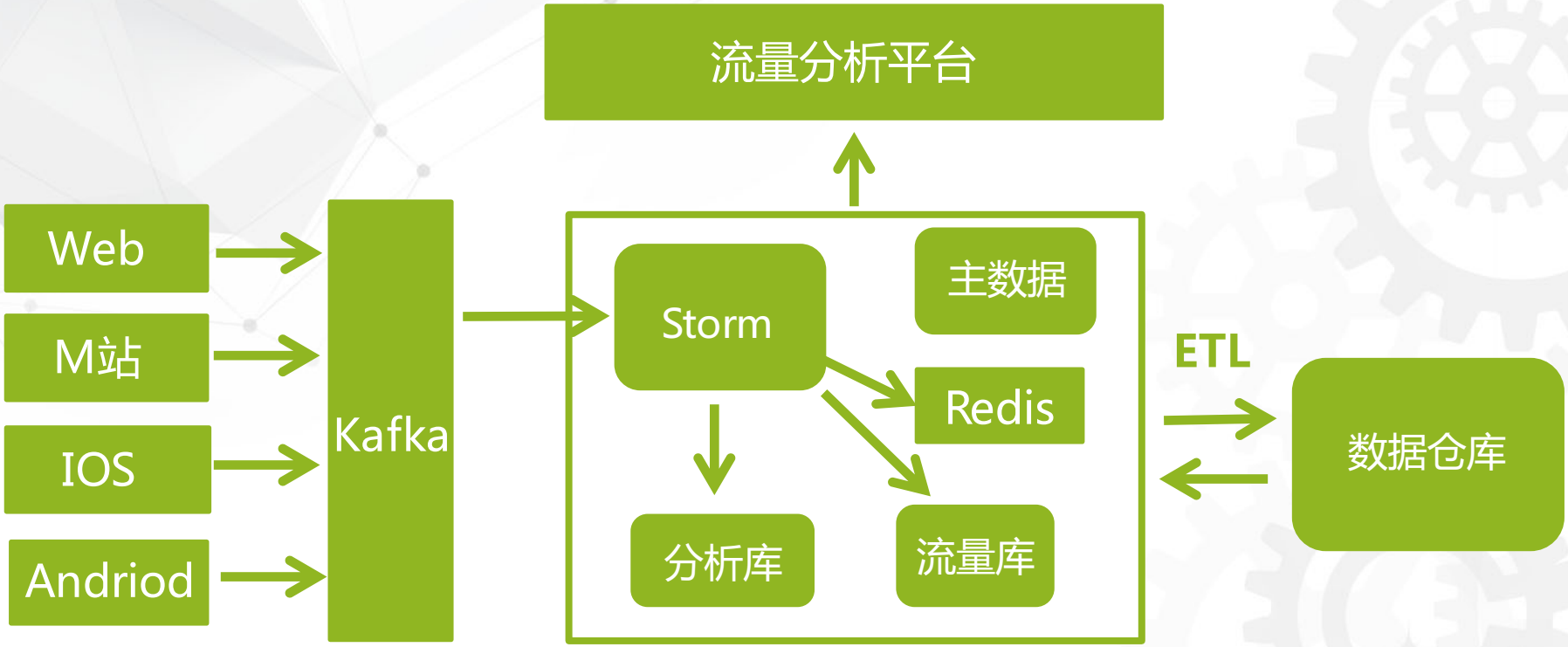
新的问题

① 还不支持对用户访问行为进行营销。

② 隔天数据无法满营销需求。

③ 通过SQL拼装、临时表的方式创建客户细分，性能越来越差。

流量分析 2.0（基于SQL Server）



优势和解决的问题

① 通过UUID把流量和业务库的客户串联了起来。

② 可以定制专属于生鲜电商的转化漏斗。

③ 拥有明细数据可以进行多维度的分析和挖掘。

新的问题

① SQL Server 硬盘不够用，水平扩展很困难。

② 做了N多优化，数据库的写入依然是瓶颈。

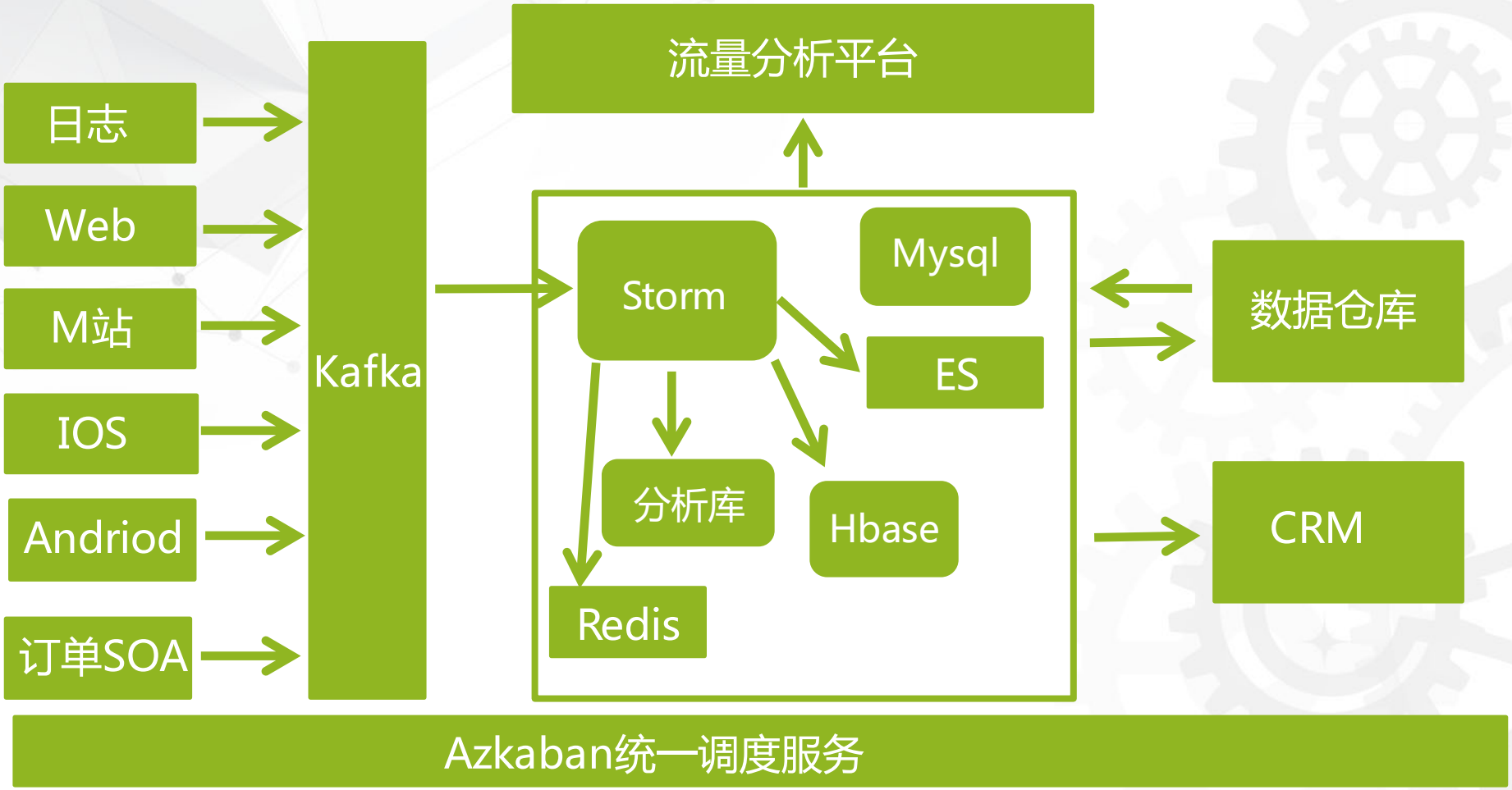
一些建议和推荐


网站分析实战
如何以数据驱动决策，提升网站价值
Web Analytics in Action
网站分析实战
如何以数据驱动决策，提升网站价值

Part 3 Hadoop体系下的BI

在本来生活网上越来越多的商品开始走红，比如像这块牛肉

流量分析 3.0（基于Hadoop）



技术工具介绍

① 使用了Flume采集日志，备选方案：LogStash。

② 采用JS+Nginx+Tomcat+Kafka采集用户行为。

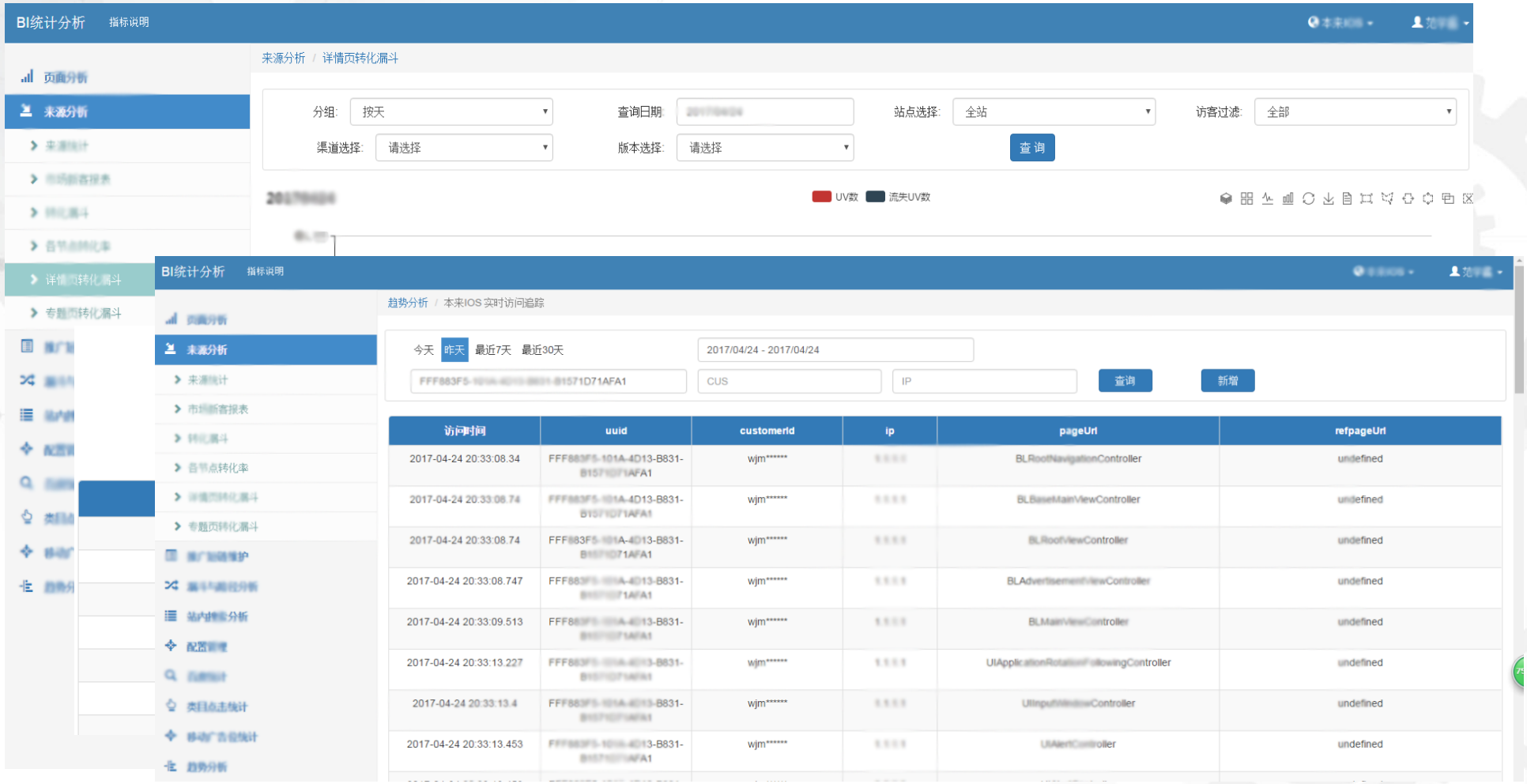
③ Storm实时处理，根据规则分别写入Mysql、Hbase、ES。

④ 使用Redis作为Storm实时处理的缓存。

⑤ ElasticSearch 作为快速查询的工具。

⑥ MySql 作为配置库以及报表集市库

举个栗子

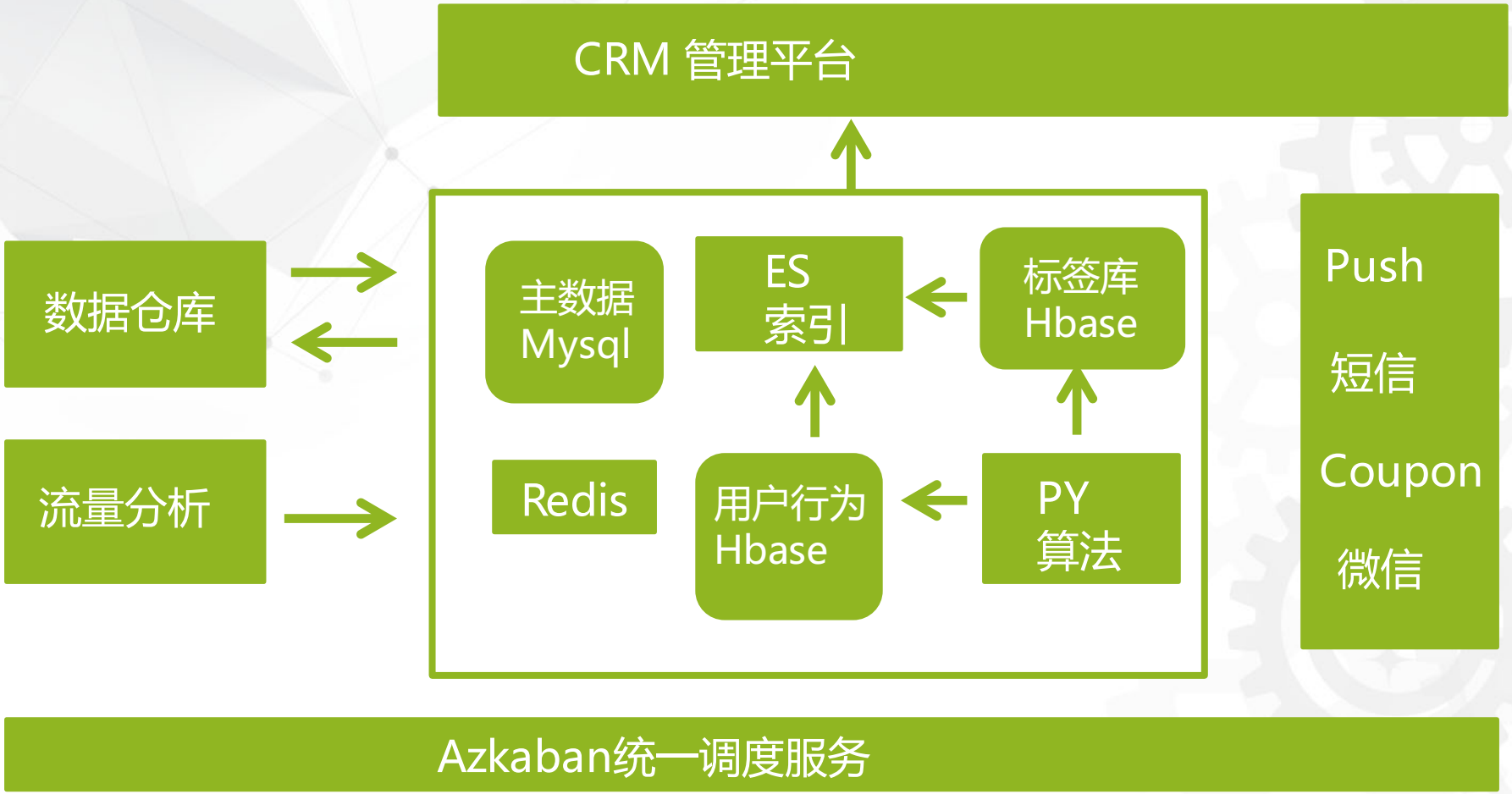

BI统计分析
指标说明
来源分析 / 详情页转化漏斗
分组: 按天
查询日期:
站点选择: 全站
访客过滤: 全部
渠道选择: 请选择
版本选择: 请选择
查 询
UV数
流失UV数
BI统计分析
指标说明
趋势分析 / 本来IOS 实时访问追踪
今天 昨天 最近7天 最近30天
2017/04/24 - 2017/04/24
CUS
IP
查询
新增
访问时间 uuid customerId ip pageUrl refpageUrl
2017-04-24 20:33:08.34 wjm****** BLRootNavigationController undefined
2017-04-24 20:33:08.74 wjm****** BLBaseMainViewController undefined
2017-04-24 20:33:08.74 wjm****** BLRootViewController undefined
2017-04-24 20:33:08.747 wjm****** undefined
2017-04-24 20:33:09.513 wjm****** undefined
2017-04-24 20:33:13.227 wjm****** undefined
2017-04-24 20:33:13.4 wjm****** undefined
2017-04-24 20:33:13.453 wjm****** undefined

解决的问题

① Hbase的快速写入解决了SQL Server的写入瓶颈。

② Hdfs的分布式存储，解决了SQL Server存储问题。

CRM 3.0（基于Hadoop）



优势和解决的问题

① 近实时的数据更新，满足业务时效性需求。

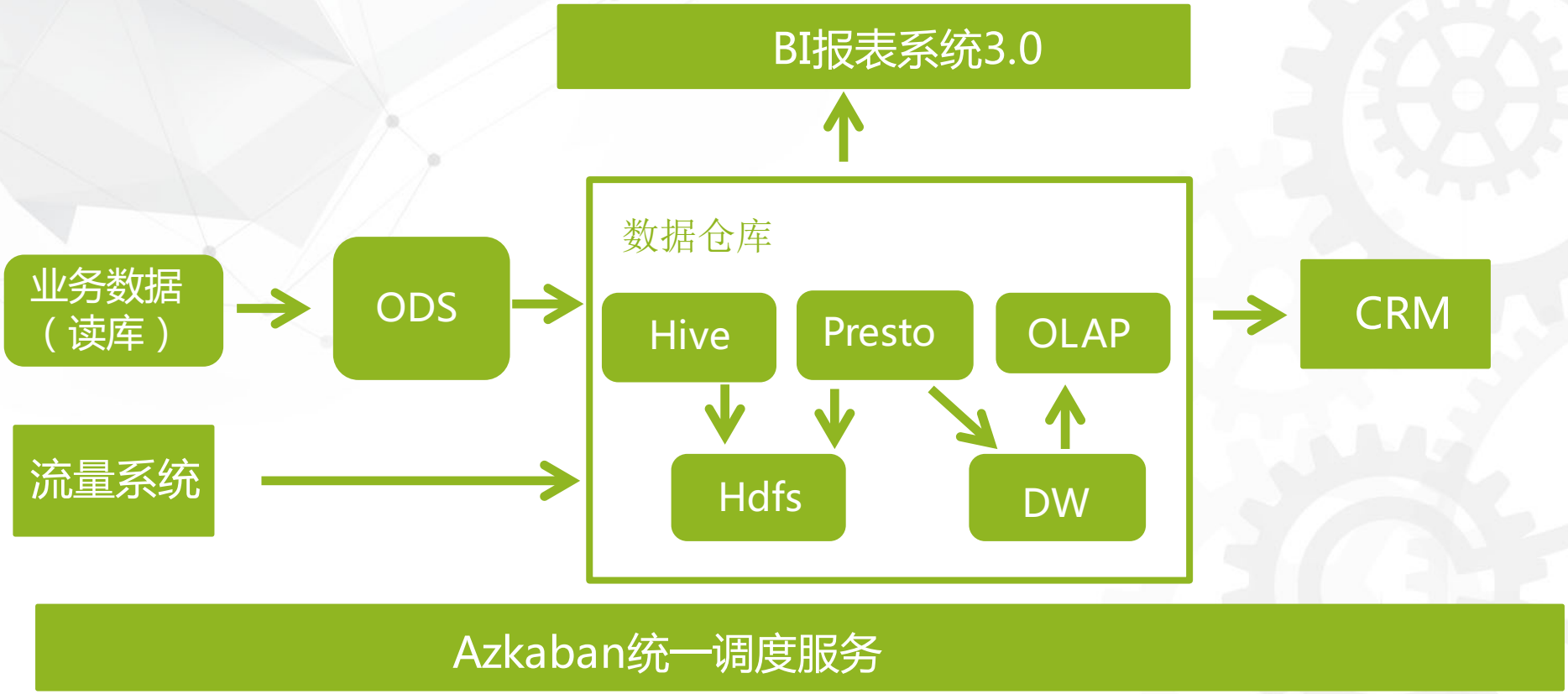
② 更加丰富的标签体系、更精准的营销。

③ 可以设定规则根据特定用户行为进行触发式营销。

④ 基于ElasticSearch的用户细分，性能有了数量级的提升。

⑤ 利用Python丰富的算法包进行离线分析和打标签

BI报表系统3.0(Hadoop)



优势和解决的问题

① 引入Hive解决数据存储问题。

② Presto提升查询性能以及多数据源问题。

③ 用Azkaban统一调度工具管理所有离线分析/ETL的Job和Job间的依赖。

Last

按需选择数据架构

数据需求多了 ➡ 有了报表系统和CRM

报表多了 ➡ 有了多维度分析系统

感觉数据粒度 粗了 ➡ 研发流量分析系统

数据分析多了 ➡ 开始沉淀规则（标签）

本来生活网BI一直再探索...

本来生活

个人微信号

THANKS

SequeMedia 盛拓传媒 IT168.com ITPUB ChinaUnix