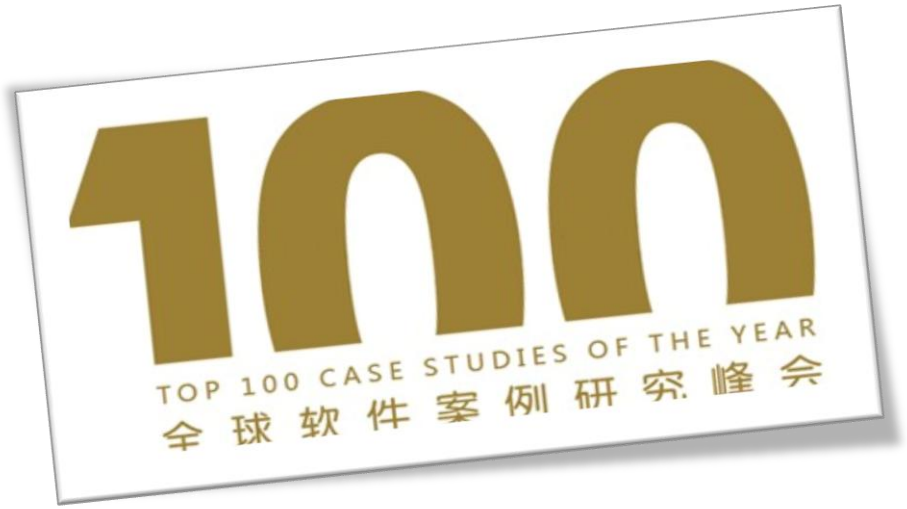
百度大数据质量保障方案探索

2014.11，钱承君



Baidu 百度

自我介绍

大型分布式系统

大数据

存储体系

分布式计算

机器学习

质量保障体系建设

架构师

管理者

Baidu 百度

百度的大数据在做什么

“BIG DATA”

is like teenage sex

everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Baidu 百度

当传统测试遭遇大数据

对系统或系统原件在特定条件下的运行结果进行观察或记录，并对系统和原件进行某些方面特性的评价

– IEEE 对“测试”的定义

构建输入与上下文

验证输出

传统测试

数据测

系统正确仅是第一步

还需诸多额外工作

百度地图的案例

Bai du 百度

百度搜索的案例

Bai du 百度

贪官落马

百度一下

百度首页 | 消息 | 设置

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约11,500,000个

贪官落马的最新相关信息

中央巡视高整改清单:看看哪些贪官落马了

新华网北京10月11日电 据新华社"新华视点"微信报道,中央纪委监察部网站目前正在陆续公布今年中央巡视组第一轮巡视整改情况,一大批已查处的案件及...

新华网上海频道 6天前

贪官落马"随带家属"将成反腐又一"风景"... 民主与法制网 1天前

落马贪官悔恨:多年没有听过批评的声音了 海外网 3天前

十八大后40名落马副厅以上贪官人均受贿14... 新浪四川 4天前

落马贪官家属怒斥煤老板:没良心只会送钱 扬子晚报网 5天前

贪官落马"随带家属"将成反腐又一"风景"?_博文推荐_民主与法制...

2014年10月17日 - 贪官落马时,贪官和他的家属们,岂能不成为"一根绳子上的蚂蚱"?要想让他不"随带家属"岂有可能? 中国有句话叫做"坏事变好事",虽然贪官落马"...

www.mzyfz.com/cms/boke... 2014-10-17 - 百度快照 - 评价 - 翻译此页

相关人物

卢嘉丽

史上最美的高官情妇

徐才厚

因收受贿赂被开除党籍

梁海玲

天上人间花魁之首

陈光明

巾帼英雄警界女杰

Bai du 百度

图片搜索的案例


javascript
百度一下
各种奇葩结果
JAVA
SCRIPT
Professional JavaScript for Web Developers
JavaScript
高级程序设计

启动这一项目的背景

年度平衡记分卡（BSC）关键行动项

行业中大数据倍受关注，希望从质量保障的角度，抓住机会，更多介入这一领域

百度关键领域的工程架构日趋成熟，业务衍进越来越多地依赖数据、算法、策略

百度测试团队完成基础技术积累（测试设计、领域积累、工具、自动化、持续集成），需要探寻新的突破

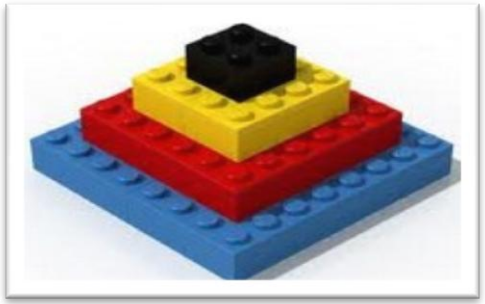
项目的总体目标

当传统测试团队遇到数据项目，怎么办？

我们希望提供“数据测试体系建设”的解决方案

定义数据质量标准

提供实施案例参照



形成体系快速复用

Baidu 百度

大数据带来的测试挑战

复杂算法

无验收标准

复杂数据流

超大数据量

平台与应用

基础架构

Baidu 百度

算法测试的常用手段

常规功能测试

- 功能测试，数据驱动、蜕变
- 异常测试，容错、抗压、死锁、健壮性
- 算法特性，例如线性递增性

非功能性测试

- 基本指标，例如吞吐、并发、时延
- 伸缩性，例如算法复杂度、性能拐点
- 资源损耗，计算密集型还是存储密集型

其他常用方法

- 同类算法的交错验证
- 引入类似真实场景，对算法系统端对端测试
- 建设获取大数据样本的能力

无验收标准的大数据应用

JD.京东 多·快·好·省

大数据时代 搜索

热门搜索：手机节 办公打印 美菱冰箱 无肉不欢 万件好货 iPhone6 哈利波特 爱心东东

热门推荐

大数据时代：推开财政数
25条 (100%好评)
¥31.60 [7.9折]

赤裸裸的未来·大数据时代
0条 (100%好评)
¥33.10 [6.9折]

大数据时代必读套装：大
20257条 (96%好评)
¥78.80 [7.9折]

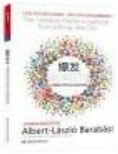
数据化决策：大数据时代
323条 (94%好评)
¥44.70 [7.8折]

数据化决策：大数据时代
0条 (100%好评)
¥44.80 [8折]

大数据时代
19438条 (96%好评)
¥39.40 [7.9折]

热门关注

大数据：正在到来的数据
1849条 (96%好评)
¥37.50 [7.6折]



爆发：大数据时代预见未
1789条 (95%好评)
¥39.90 [6.7折]

互联网金融：框架与实践
782条 (95%好评)
¥47.60 [7.4折]

3D打印：从想象到现实
1673条 (97%好评)
¥38.70 [7.9折]

给你一部手机，你能怎么
24条 (83%好评)
¥30.20 [6.8折]

浅薄-互联网如何毒化了我
103条 (95%好评)
¥28.00 [6.7折]

Bai du 百度

大数据应用的质量保障

推荐、预测、数据挖掘、机器学习等

质量标准：相关性、重复度、品类覆盖、排序

持续评估：低成本例行评估，采样、众包

小流量实验支持

研发过程支持，全流程工具链建设

运营支持，数据分析、竞品分析

基础数据的质量保障

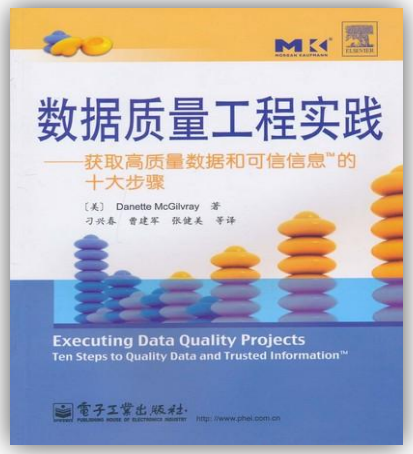
WHY

- 最终产品结果的正确性
- 大数据应用可更好逼近理想值上限

WHAT

- 上游变更，例如重启重传、扩容、数据升级
- 数据碎片化，例如非归一化、时钟边缘切割
- 不满足场景，例如画像与数据分析的混用

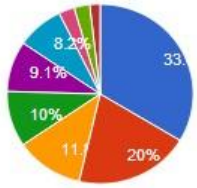
数据质量是一个独立的细分行业



Baidu 百度

数据质量利器：数据剖析（Data Profiling）

- 数据理解与规则挖掘
- 数据异常诊断
- 数据问题排查
- 数据后置校验
- 数据监控迁移
- 数据归一化梳理

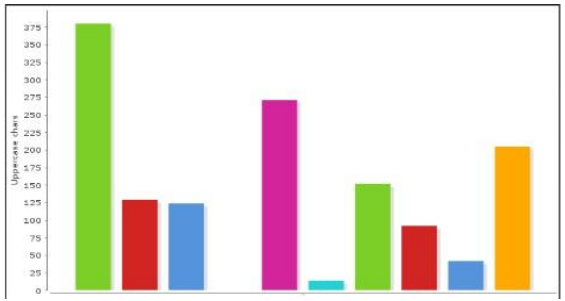


Classic Cars
Vintage Cars
Motorcycles
Planes
Trucks and Buses
Ships
1/2

Classic Cars	37
Vintage Cars	22
Motorcycles	13
Planes	11
Trucks and Buses	10
Ships	9
Trains	3
Vintage cars	2
Classiccars	1
Plane	1
Trux and Buses	1

Total count	110
Distinct count	11

占比分析



离群分析

Baidu 百度

我们很快发布了数据质量平台

产品线	产出完成时间	数据质量
iknow	2013-12-10 07:34:09	7 BAD,18 GOOD
appsearch	2013-12-10 07:06:10	6 BAD,8 GOOD
map	2013-12-10 05:46:37	4 BAD,18 GOOD
jiasule	2013-12-10 07:32:30	2 BAD,0 GOOD
CLOUD	2013-12-10 01:34:42	2 BAD,0 GOOD
image	2013-12-10 04:11:33	2 BAD,7 GOOD
ps	2013-12-10 02:26:11	2 BAD,6 GOOD

表名	产出完成时间	BAD数据分区	操作
iknow_mobile_app_submit	2013-12-10 01:46:41	2013-12-09 00:00:00 2013-12-09 01:00:00 2013-12-09 02:00:00 2013-12-09 03:00:00 2013-12-09 04:00:00 2013-12-09 05:00:00 2013-12-09 06:00:00 2013-12-09 07:00:00 2013-12-09 08:00:00	查看详情
iknow_pc_web_other	2013-12-10 03:24:06	2013-12-09 18:00:00 2013-12-09 19:00:00 2013-12-09 20:00:00	查看详情
iknow_pc_web_view	2013-12-10 03:24:06	2013-12-09 00:00:00	查看详情
iknow_wap_common_other	2013-12-10 01:58:10	2013-12-09 23:00:00	查看详情
iknow_wap_common_submit	2013-12-10 01:58:10	2013-12-09 04:00:00 2013-12-09 22:00:00 2013-12-09 23:00:00 2013-12-10 00:00:00	查看详情

表名	分区	产出时间	字段	规则类型	占比	波动质量
iknow_mobile_app_submit	2013-12-09 00:00:00	2013-12-09 01:46:32	event_cuid	incomplete	79.3096%	BAD
iknow_mobile_app_submit	2013-12-09 00:00:00	2013-12-09 01:46:32	event_urlpath	wrongformat	0%	GOOD
iknow_mobile_app_submit	2013-12-09 00:00:00	2013-12-09 01:46:32	event_urlparams	incomplete	0%	GOOD

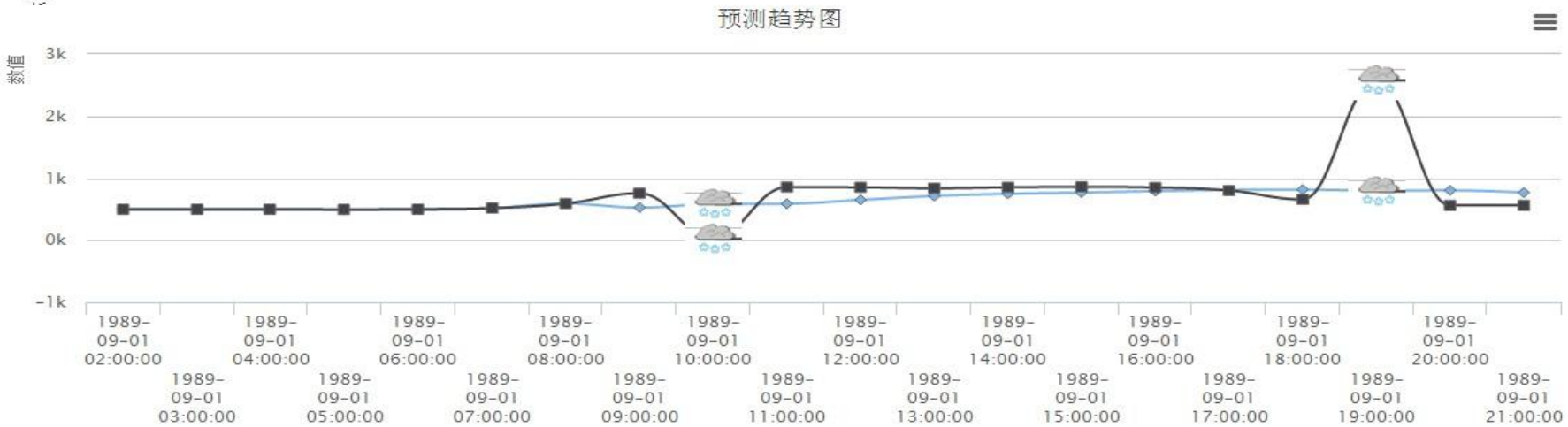
Bai du 百度

利用算法作一致性拟合消除过多报警

数据样例

数据来源 - "The Great Energy Predictor Shootout" - The First Building Data Analysis And Prediction Competition; ASHRAE Meeting; Denver, Colorado; June, 1993

时间标识	1989-09-01 02:00:00	1989-09-01 03:00:00	1989-09-01 04:00:00	1989-09-01 05:00:00	1989-09-01 06:00:00	1989-09-01 07:00:00	1989-09-01 08:00:00	1989-09-01 09:00:00	1989-09-01 10:00:00	1989-09-01 11:00:00	1989-09-01 12:00:00	1989-09-01 13:00:00	1989-09-01 14:00:00	1989-09-01 15:00:00	1989-09-01 16:00:00
能	496.07	497.06	496.67	494.54	498.09	516.96	589.44	754.65	31.1	853.42	850.58	834.82	851.32	860.25	847.71



构建闭环反馈机制

考察指标：误报率、召回率、应答率、应答时延

激励，对靠谱值班人进行物质奖励

负向激励，引入考评、引入问责

超时自动填充，加强问责

补充策略与产品机制，降低成本

ID	产品线名	异常类型	报告时间	反馈情况	反馈时间	反馈rd	操作
2609	map	字段					
2814		数据					
2607	jiasule	字段					
2606	news	字段异常					
2608	tieba	字段异常					
2604	iknow	字段异常					
2605	iknow	字段异常					
2602	wenku	字段异常	2014-03-26 08:00:00	已反馈[忽略]	2014-03-27 00:00:00	zhenghuajiong	查看反馈
2603	mtj	字段异常	2014-03-26 00:00:00	已反馈[忽略]	2014-03-27 00:00:00	zhenghuajiong	查看反馈
2809		数据量异常	2014-03-25 22:00:00	已反馈[忽略]	2014-03-27 10:21:33	zhenghuajiong	查看反馈

数据类项目研发流程的考虑

平台与应用共存

渐进放大数据量

基于模型生成与模糊数据

上线后的持续校验

Tuning Box

Fuzz Tool
Data Generation

渐进式验证，关注流程的衔接、问题定位与回退

复杂系统的特殊考虑

不稳定场景

- 多线程并发、竞争冒险
- 异步乱序

异常场景

- 硬件故障，文件破损、磁头老化、磁盘坏道
- 网络故障，延迟阻塞、丢包、重包、分割
- 分布式异常，节点增删、状态不一致

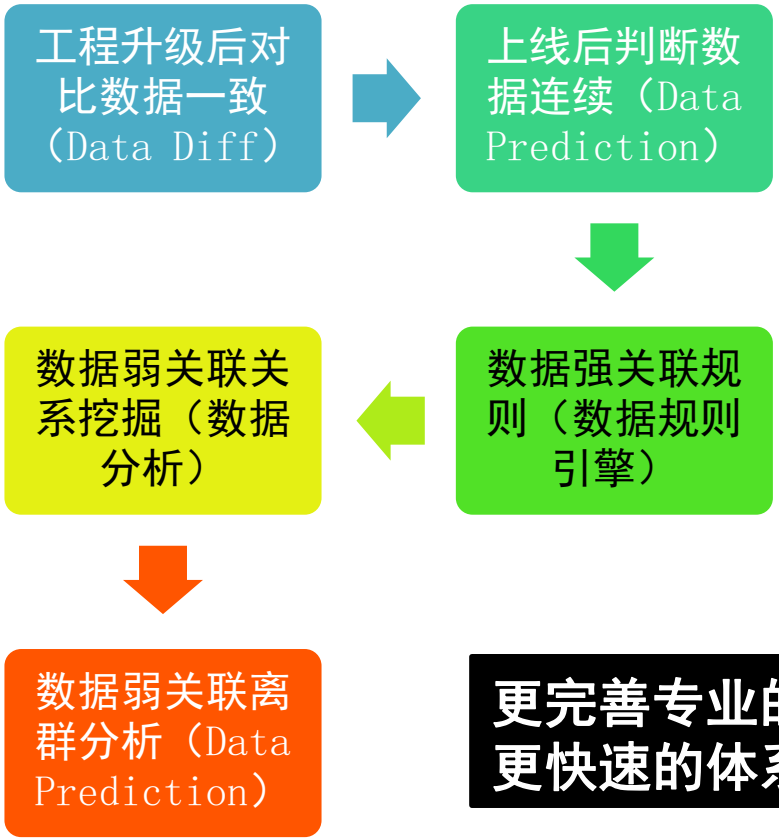
系统环境

- 注意测试环境与真实场景的差异
- 注意系统的极限与拐点，负载均衡、雪崩
- 特殊情况，例如核心交换机压力过载

总结：技术与工具

Input Generation	• Fuzz tool • DGL、Model-based data generation • Metamorphic testing
Output Verification	• Consistency check • Data profiling、data quality、data clearance • Prediction、alert center
System	• Mock for racing condition • Collision test for large distributed environment • Robustness、fuzz injection
Environment and Process	• Tuning box for every single developer • Full cycle automation • Tracing and quick debugging

例子：快速拼装测试体系



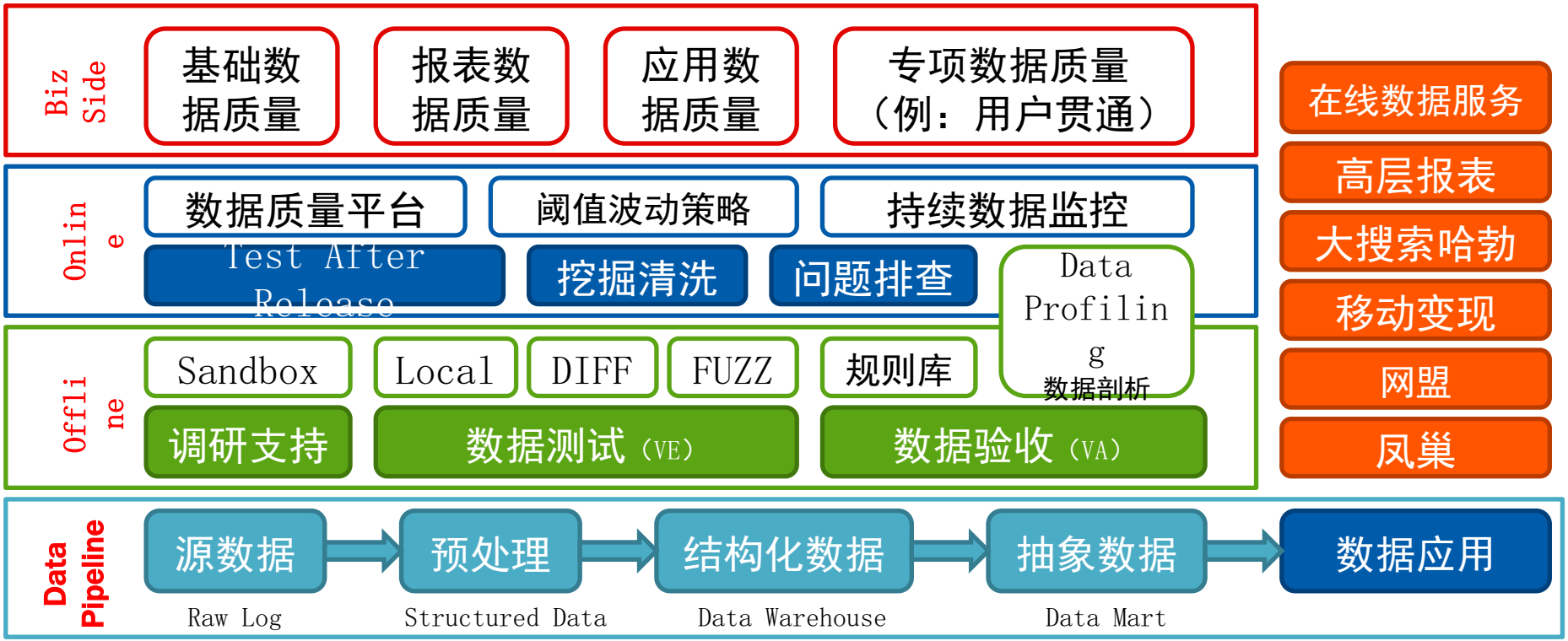
用户为何质疑报表正确性？

1. 信息不连续
2. 信息与其它渠道冲突
3. 信息与领域认知违背

更完善专业的解决方案
更快速的体系建设

Baidu 百度

数据质量体系建设的一个实施案例



大数据技术在质量领域的应用刚刚起步

分类器　关联分析　异点分析

预测技术　标注获取

Bai du 百度

Thank You

Baidu 百度