DTCC

2016中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2016

数 据 定 义 未 来

SequeMedia 盛拓传媒　ITPUB.com　ChinaUnix　ITPUB

简介： 张粤磊 （Jackson)
邮箱：vzyuelei@126.com

@me：

微信：
vzyuelei

QQ:
416988515

- 飞谷云（www.feiguyun.com)创始人
- （2014-2016.3）平安付大数据平台架构师
- （2012-2014）外汇交易中心ETL项目开发经理
- （2010-2012）HP TRAM项目 ETL开发组长
- （2005-2010）DBA
- 10余年一线数据业务（制造，咨询服务，互联网金融）及数据处理技术实践经验

关注的技术产品及工具：

数据处理的哪些事

1. 传统数据仓库的数据处理技术及思考
2. 大数据环境下对于公共数据及行为数据的数据处理技术
3. 由传统数据仓库到大数据数据仓库的数据处理实践思考及建议

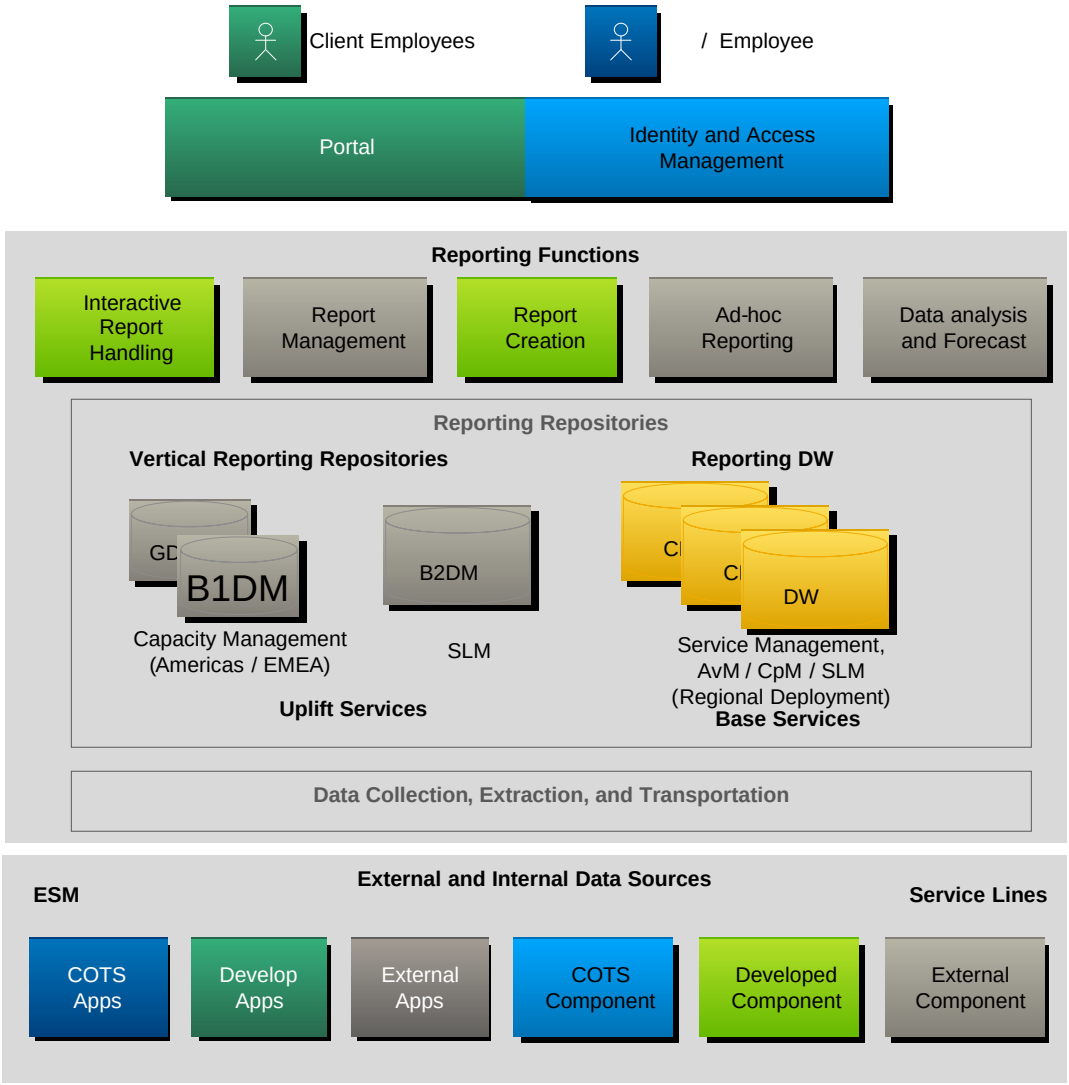
传统数据仓库的数据处理技术

传统数据仓库的数据处理技术是什么？

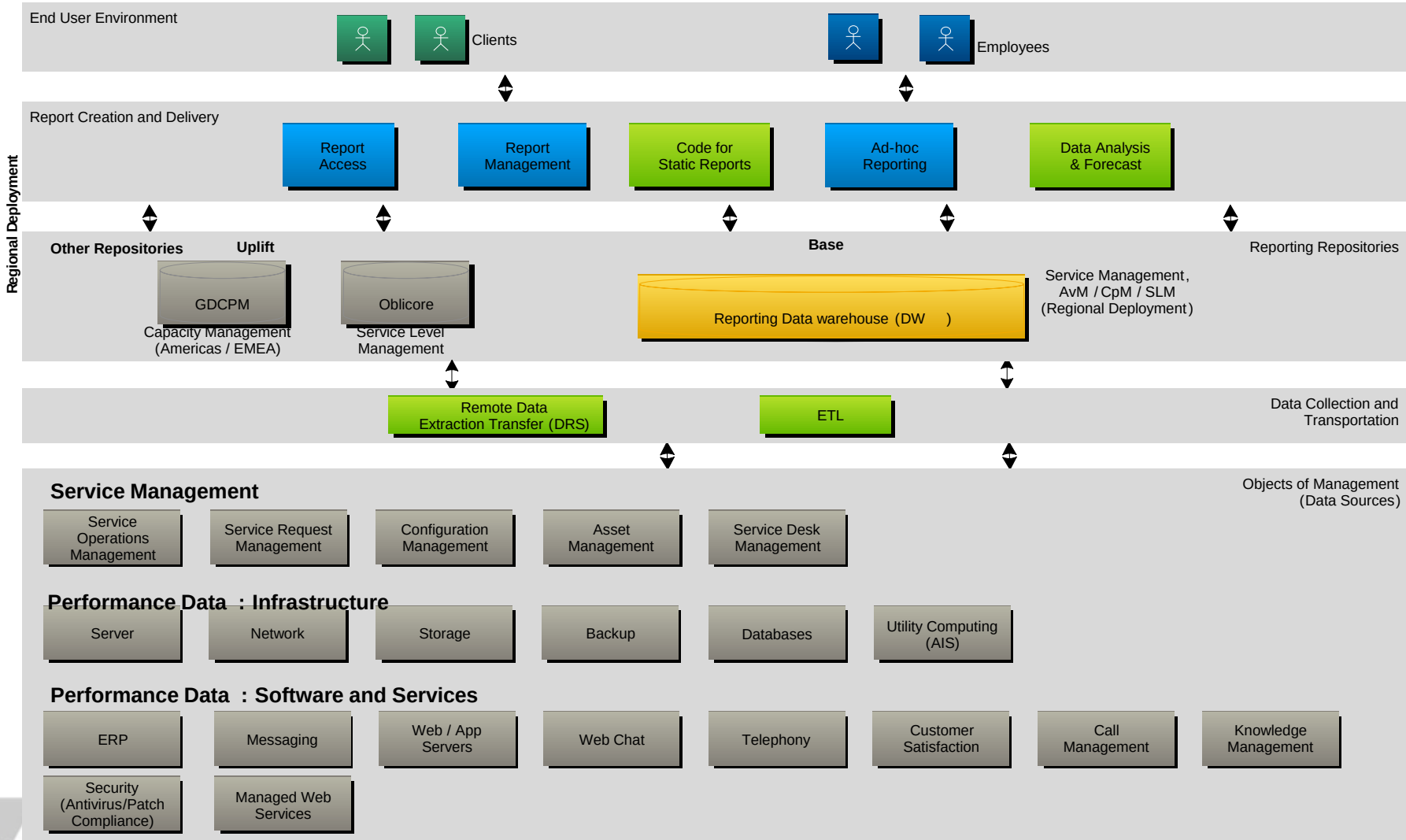
从我参与某大型数据仓库项目经历为例来分享：

- 该过程涵盖传统数据仓库的标准流程和数据处理规范
- 数据处理方法和实践同样适用于数据平台数据处理

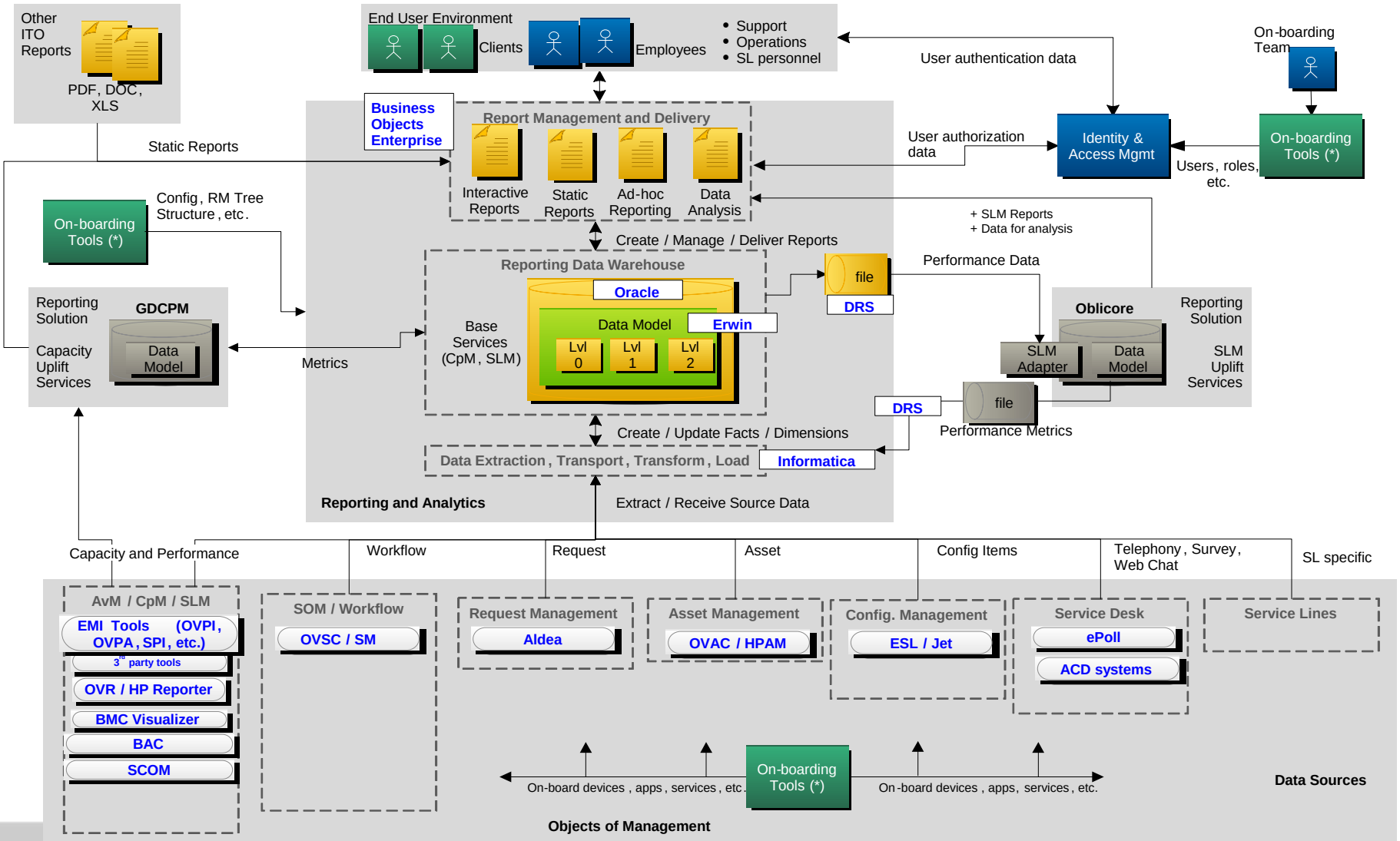
传统数据仓库的数据处理技术-概念定义



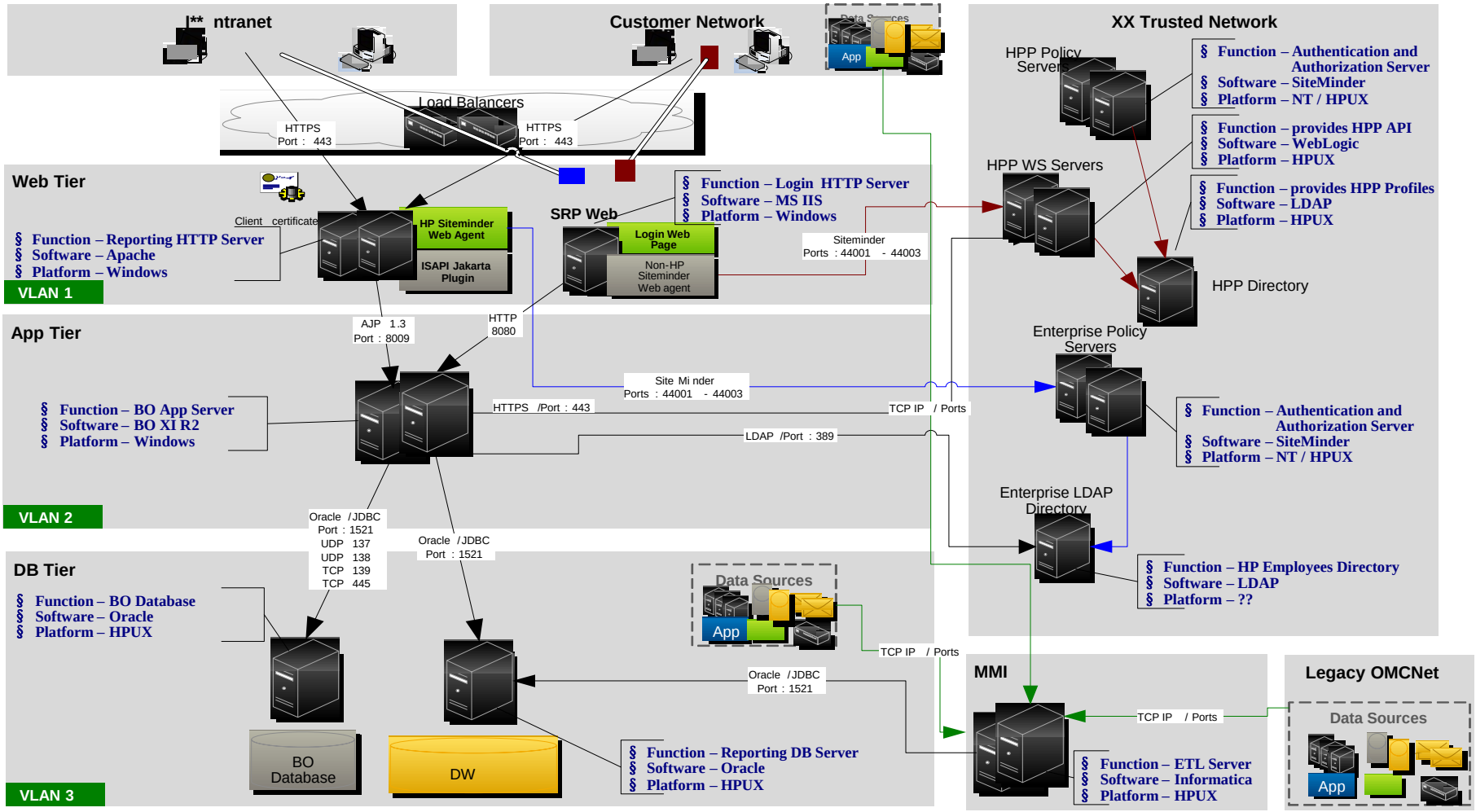
传统数据仓库的数据处理技术-业务定义



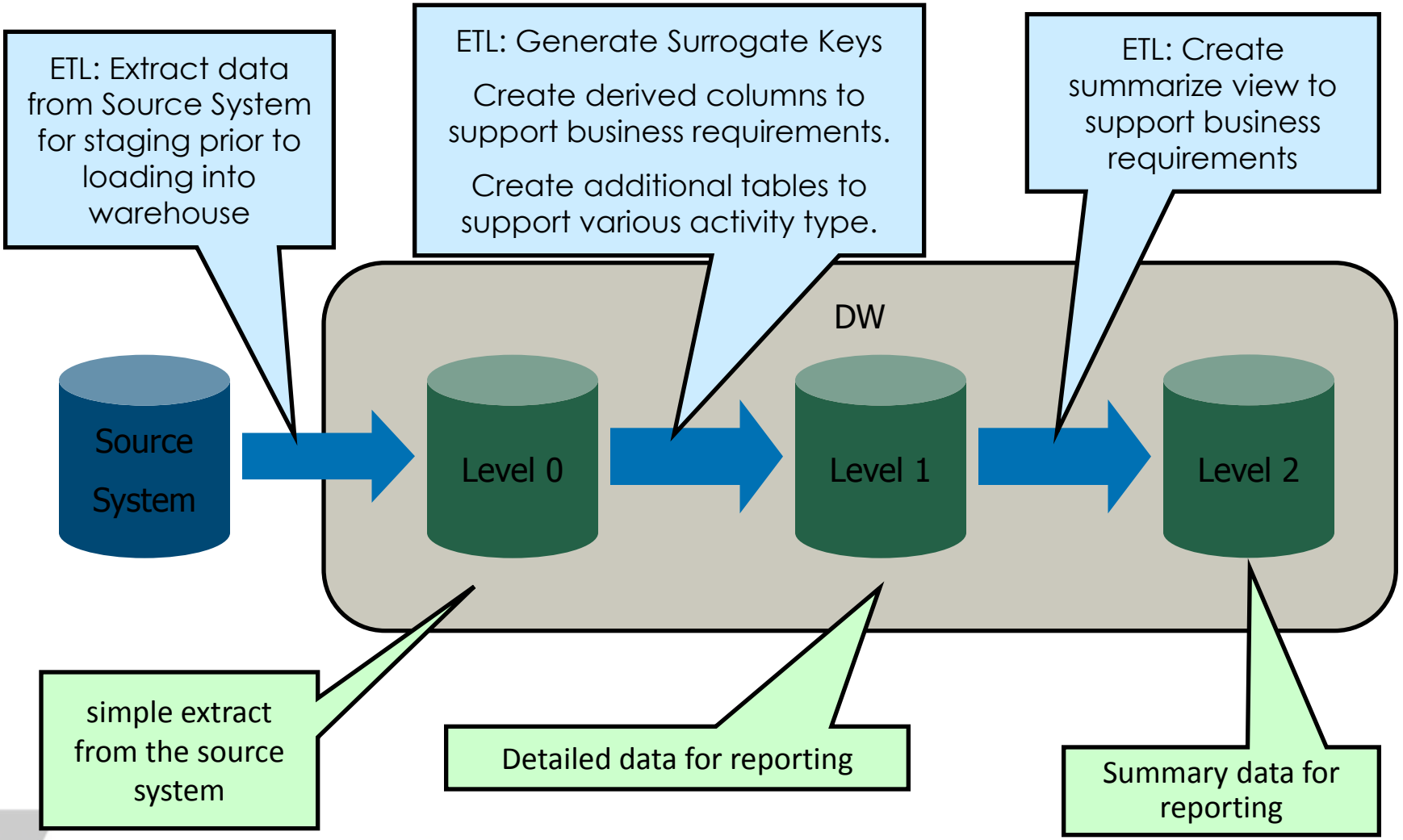
传统数据仓库的数据处理技术-逻辑定义



传统数据仓库的数据处理技术-物理定义



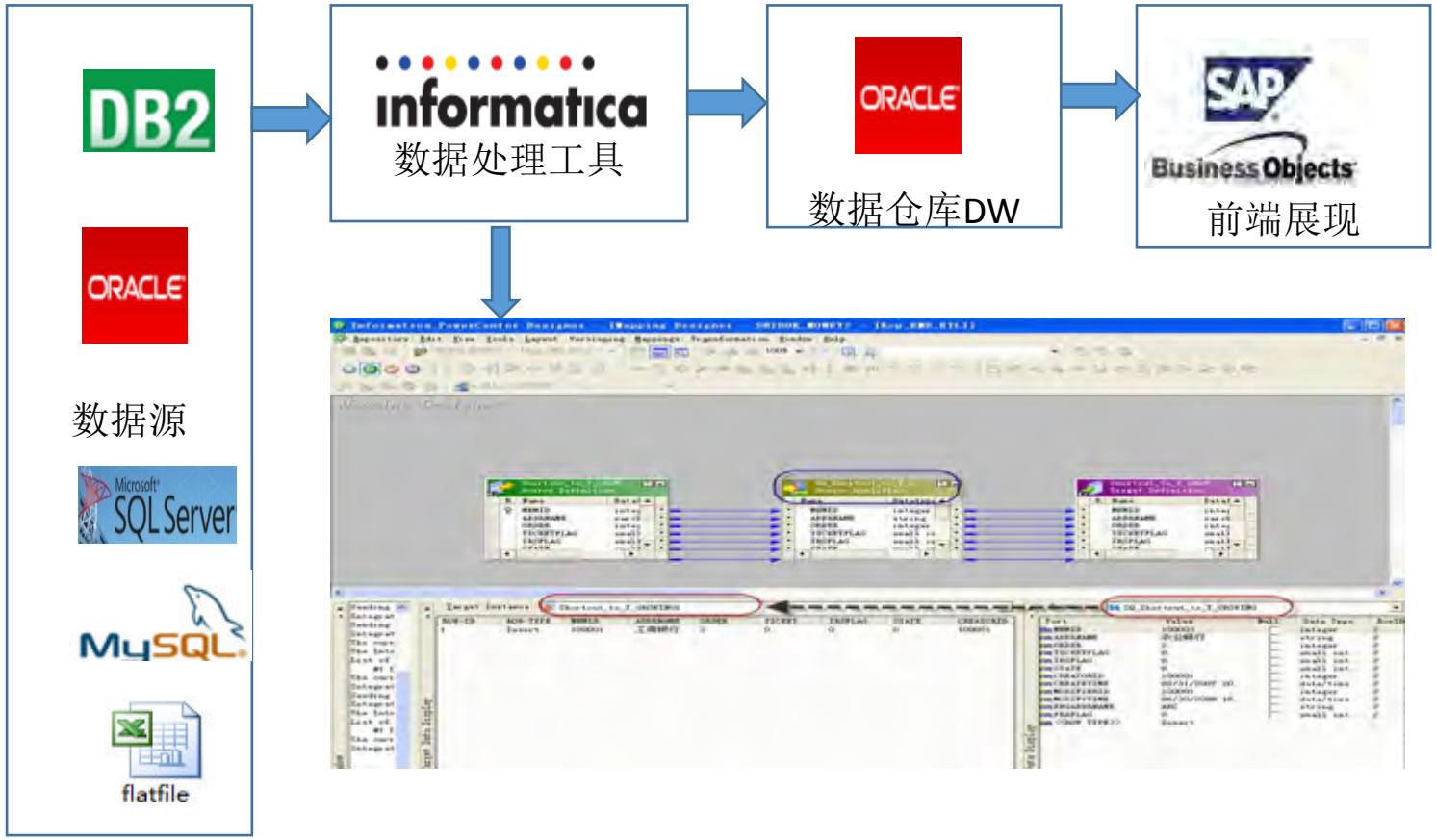
传统数据仓库的数据处理技术-规则定义



传统数据仓库的数据处理技术-设计定义

A	B	C	D	E	F	G	H	I	J	K	L	M	O	P	Q	R	S
	源系统信息(Source System Information)								目标系统信息(Target System Information)				ETL抽取 (PowerCe				
序号(No)	源Schema (Source Schema)	表名 (Table Name)	标志位读取方式（表表直接生成序有标志位时，需要填写：标志为涤，标志位名称，标志位数值	表当前的数据量 (Row Count)	表每月增量 (Month Inc)	标识增量的时间戳字段 (Record Increment Timestamp Field	该表是否有物理删除操作 (Is delete Action)	数据生成时间窗 (Data Creation	目标Schema (Target Schema)	目标表名 (Target Table Name)	加载类型 (Load Type)	数据到达时间需求 (Required Data Arrival Time)	ETL抽取时间	ETL抽取频度	标志位等待时间窗口（等	标志位等待循环时间间隔	删除方式（对目标表的操作）
1	CRM	V_CNY_ENTY_MASTER_LOG							etluser	TRDX_ENTY_MASTER_LOG		每5分钟	每5分钟	每5分钟			任何时候不删
2	CRM	V_CNY_ENTY_TYPE_MASTER_LOG							etluser	TRDX_ENTY_TYPE_MASTER_L		每5分钟	每5分钟	每5分钟			任何时候不删
3	CRM	V_CNY_ENTY_TYPE_DTLS_LOG							etluser	TRDX_ENTY_TYPE_DTLS_LOG		每5分钟	每5分钟	每5分钟			任何时候不删
4	CRM	V_CNY_ENTY_ALT_CODE_MASTE_LOG							etluser	TRDX_ENTY_ALT_CODE_MAS		每5分钟	每5分钟	每5分钟			任何时候不删
5	CRM	V_CNY_RGN_CNFG_MASTER_LOG							etluser	TRDX_RGN_CNFG_MASTER_L		每5分钟	每5分钟	每5分钟			任何时候不删
6	CRM	V_CNY_MKT_PRMTR_DTLS_LOG							etluser	TRDX_MKT_PRMTR_DTLS_LO		每5分钟	每5分钟	每5分钟			任何时候不删
7	CRM	V_CNY_ENTY_TRMNL_CNT_HSTR_LOG							etluser	TRDX_ENTY_TRMNL_CNT_HS	UPD	每5分钟	每5分钟	每5分钟			任何时候不删
8	CRM	V_CNY_ENTY_ELGBLT_DTLS_LOG							etluser	TRDX_ENTY_ELGBLT_DTLS_L		每5分钟	每5分钟	每5分钟			任何时候不删
9	CRM	V_CNY_ENTY_MKT_MAKING_DTLS_LOG							etluser	TRDX_ENTY_MKT_MAKING_D		每5分钟	每5分钟	每5分钟			任何时候不删
10	CRM	V_CNY_ENTY_MKT_STATUS_DTLS_LOG							etluser	TRDX_ENTY_MKT_STATUS_D		每5分钟	每5分钟	每5分钟			任何时候不删
11	CRM	V_CNY_ENTY_TYPE_RLTN_MAST_LOG							etluser	TRDX_ENTY_TYPE_RLTN_MA		每5分钟	每5分钟	每5分钟			任何时候不删
12	CRM	V_CNY_ENTY_RLTN_DTLS_LOG							etluser	TRDX_ENTY_RLTN_DTLS_LO		每5分钟	每5分钟	每5分钟			任何时候不删

传统数据仓库的数据处理技术-开发实现



传统数据仓库的数据处理技术-开发实现

传统数据仓库的数据处理技术-数据治理思考

完整性

哪些数据丢失了或者哪些数据不可用?

无论选择任何一种RDBMS，都无法涵盖大量的非结构化业务数据

准确性

哪些数据和信息是不正确的，或者数据是超期的?

不同RDBMS对数据类型的定义精度各有区别

规范性

哪些数据未按统一格式存储?

基于RDBMS的数据存储并不能真实反映业务数据本源格式，文本视频，邮件在DB中的存储

唯一性

哪些数据是重复数据或者数据的哪些属性是重复的

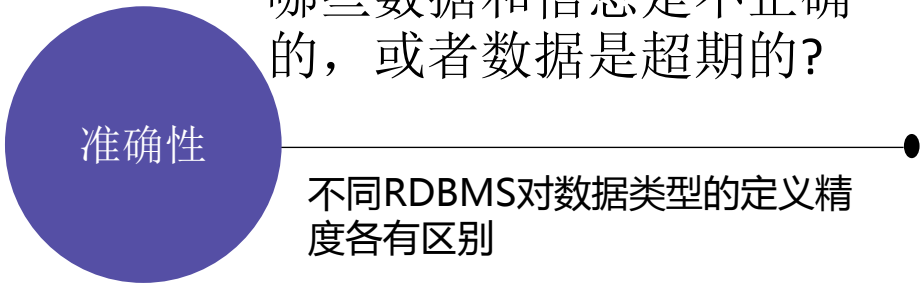
一致性

哪些数据的值在信息含义上是冲突的?

关联性

哪些关联的数据缺失或者未建立索引

传统数据仓库的数据处理技术-数据治理思考



当源系统与目标系统属于不同RDBMS或字符集等情况，可能存在字符类型不兼容问题，如：Oracle 的date数据类型有时分秒而db2 的date数据类型不含时分秒；oracle的Integer数据类型是8字节38位精度，db2 的Integer数据类型是4字节10位精度等等。

数据处理的哪些事

1. 传统数据仓库的数据处理技术及思考
2. 大数据环境下对于公共数据及行为数据的数据处理技术
3. 由传统数据仓库到大数据数据仓库的数据处理实践思考及建议

公共数据及行为数据的数据处理技术

公共数据
python
Scrapy link

Scrapy page

load

埋点数据 Flume/kafka/storm

用户交易数据 sqoop

LEVE0:HIVE 落地表
STG表

应用层调用

HBASE API

LEVE2:分析报表层
RPT,DM表

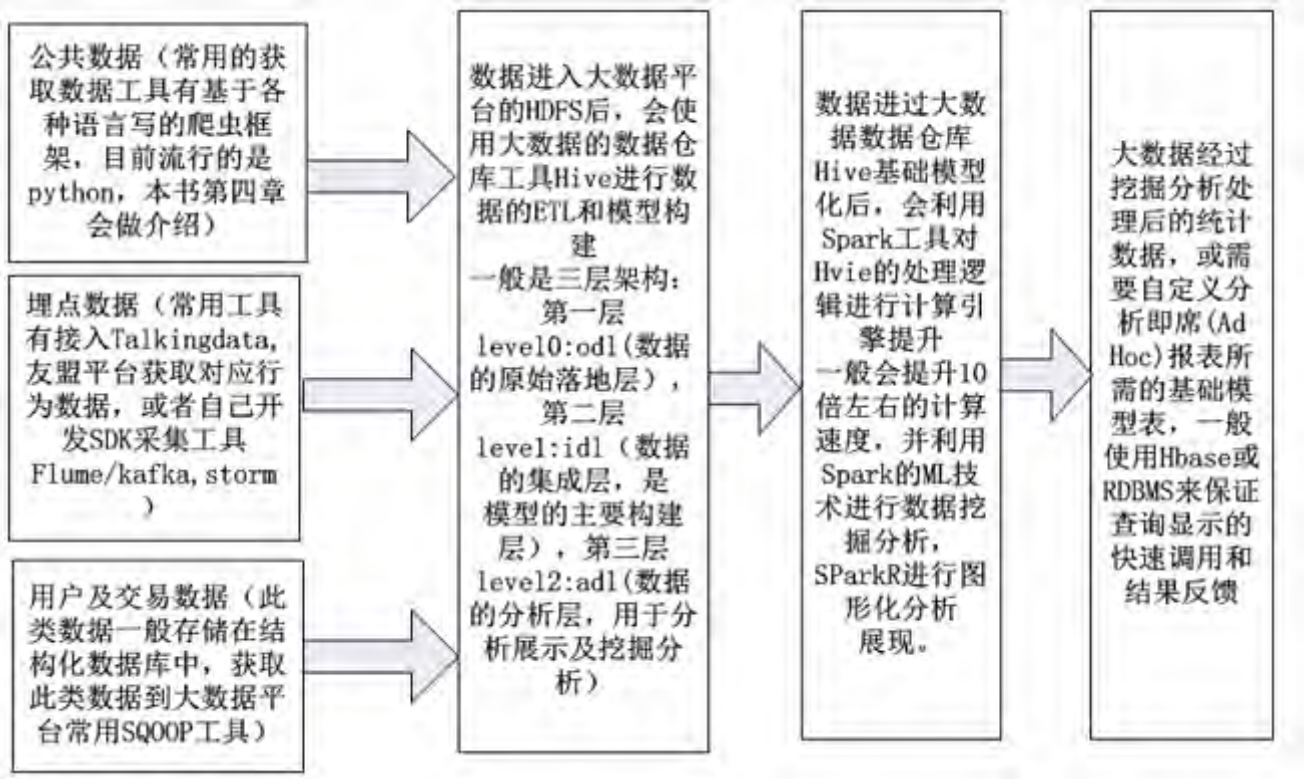
Hql:聚合指标
标签生成

LEVEL1:HIVE模型
表DW,DIM表

SPARK SQL

Hql:清洗标
准化

公共数据及行为数据的数据处理技术



公共数据及行为数据的数据处理技术

按数据特征分类

■结构化数据

定义：目前其实专指的是关系模型数据，即以关系型数据库表形式管理的数据。绝大多数的企业业务数据都以此格式进行存放。

简析：虽然从专业角度讲，结构化就是关系模型的说法并不准确。但针对目前业内现状，还是将其定义为关系模型数据为最为妥当，因为它清晰而准确地代表了我们传统上最熟悉的企业业务数据，基本没有歧义。

■半结构化数据

定义：半结构化与非结构化常常一同被提及，两者其实专指所有其他“非”结构化数据。但如果想更加清晰地描述，可以将“半结构化数据”定义为：那些非关系模型的、有基本固定结构模式的数据，例如应用日志文件、XML文档、JSON文档和电子邮件等。

简析：此部分数据可以用程序化格式解析处理，公共数据，行为数据多以此种格式

■非结构化数据

定义：除去结构化与半结构化的所有数据，即没有固定结构模式的数据，例如WORD、PDF、PPT、EXL文档，以及各种格式的图片和视频等。

简析：区分半结构化与非结构化数据的意义在于，目前在企业内对两者的处理方法（包括存储、访问与分析）是不同的。非结构化数据大多采用内容管理的方法，展示上需要采用对应的组件工具。

公共数据处理的注意点

接口定义加入接口规范变更版本及内容到数据字段中

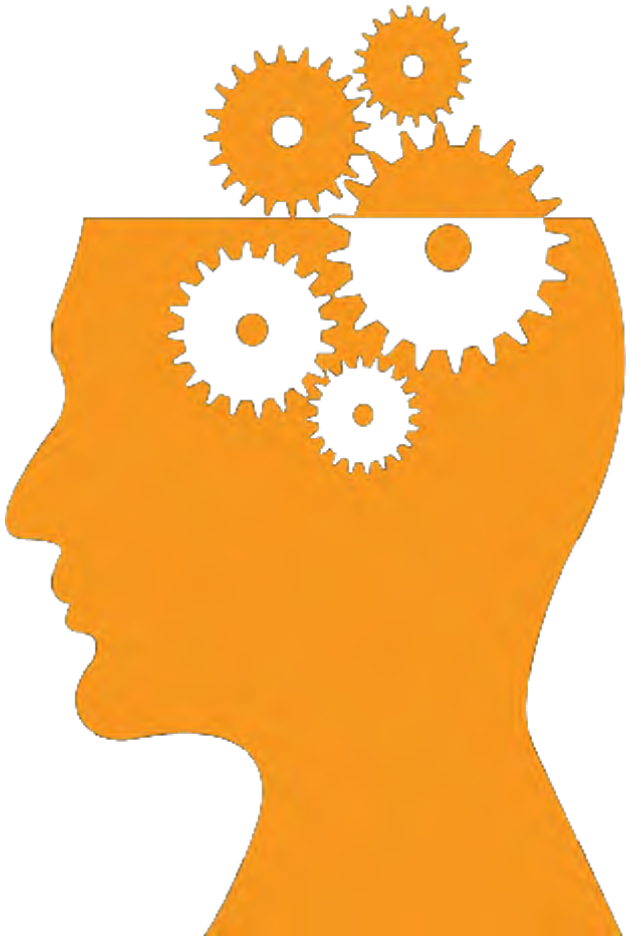
对于网站抓取或接口调用的变化版本记录有利于对数据准确和完整性的可追溯

落地后的文件时间和成功标志信息同样参与数据处理

合并到数据落地层（LEVEL0）后数据的落地时间和数据大小行数记录到数据监控表中

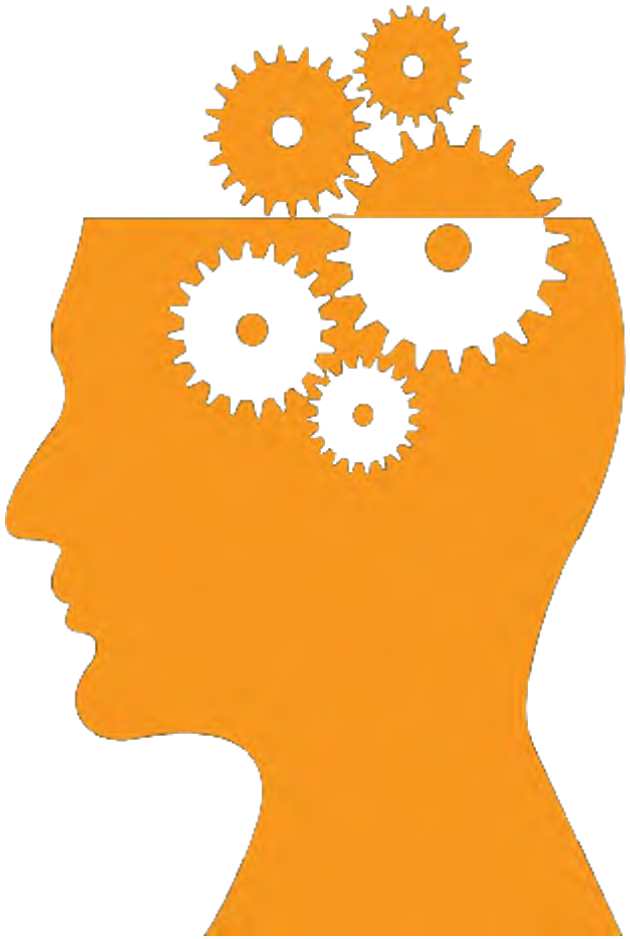
在数据仓库处理和分析展示中添加数据处理的可追溯信息

对于核心指标及对应元数据显示和监控，确保对于数据的理解和定义全局一致

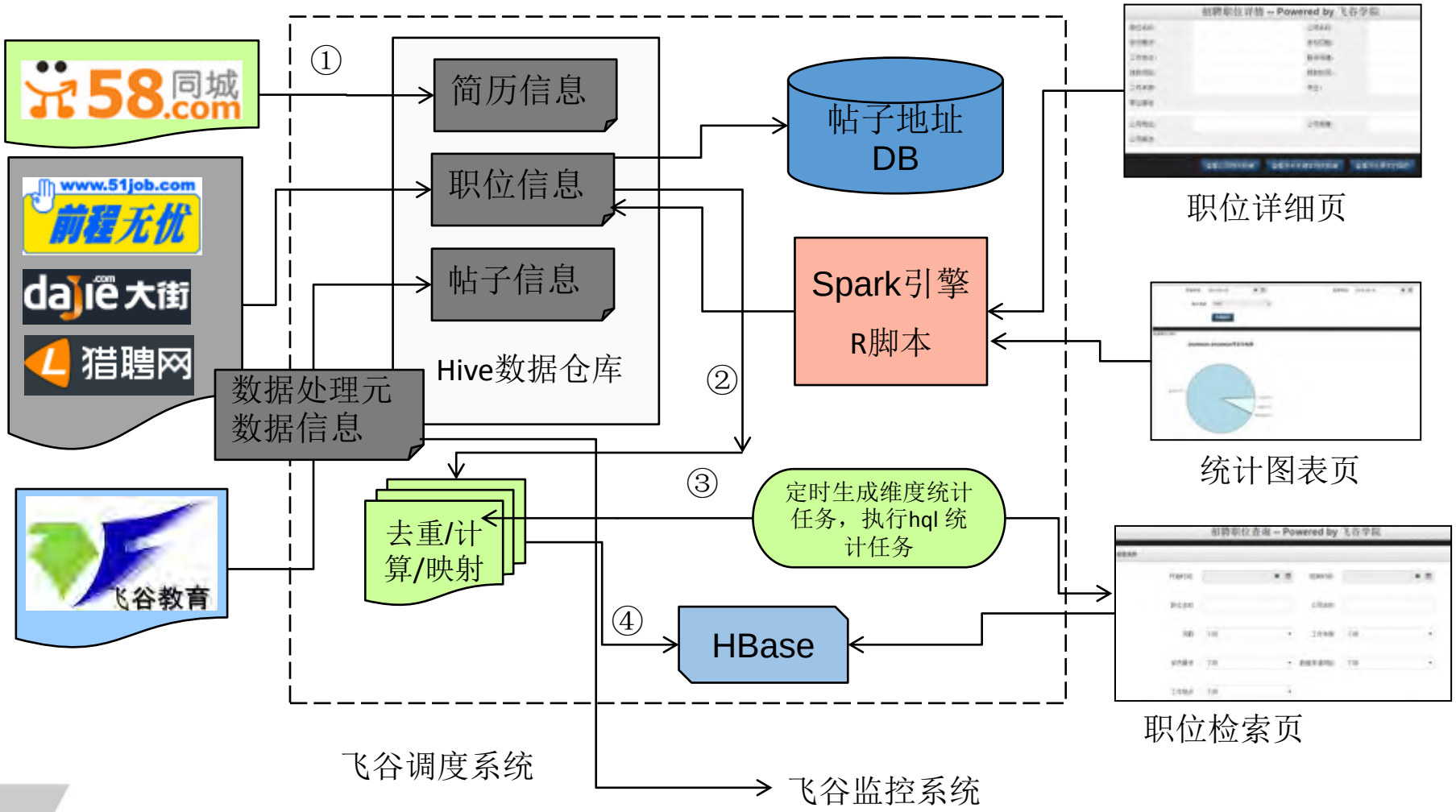


行为数据处理的注意点

- 埋点数据一定要符合业务数据信息流才能保证数据处理的完整性和确保数据的业务可用性
- 行为数据的标识健（UID，DID）要与其它数据源统一关联健和对应时间周期，确保数据的一致性和关联性。
- 行为数据的元数据信息尽可能从源头以字段化方式植入数据处理的数据文件中



公共数据及行为数据的数据处理技术案例图



数据处理的哪些事

1. 传统数据仓库的数据处理技术及思考
2. 大数据环境下对于公共数据及行为数据的数据处理技术
3. 由传统数据仓库到大数据数据仓库的数据处理实践思考及建议

传统数仓到大数据数仓的数据处理

大数据平台的迁移与构建

传统数据仓库

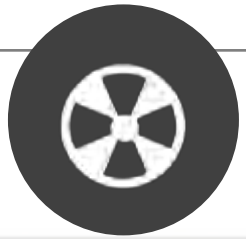
1 以RDBMS为主要的数据处理存储层。
2 数据处理采用通用的ETL产品工具
3报表层是报表产品通过标准的数据库连接驱动连接到数据仓库DB中。
4 数据库安全级别可以通过RDBMS安全管理

数据同步与脱敏

1数据处理的重点是全样本数据的基于业务时间周期同步落地
2基于RDBMS的敏感数据在进入大数据平台后进行脱敏处理，确保数据安全

大数据平台

1 以HDFS为主要的数据处理存储层
2 数据处理根据数据来源采用不同工具，其中同步RDBMS数据的SQOOP，实时数据采集的Flume，kafka,storm，及公共数据的接口API
3 报表层产品采用自主开发或支持大数据平台的工具tablelu等
4 数据安全管理难度大，需要全局设计控制



由传统数据仓库到大数据数据仓库的数据处理思考及建议

01 **数据基因定义完整准确**

构建数据平台时的数据基因一定要准确完整，这是整个数据平台的根基

02 **数据血缘设计清晰可溯**

数据平台的数据处理开始就需要同业务数据流一同设计数据的元数据血缘流，确保业务数据断点可查可控

03 **数据安全机制原子化**

对数据平台的分层数据做到基于存储机制的原子化安全控制，确保从底层实现数据的安全分层控制。主数据及业务权限数据等

04 **核心指标及元数据做到可视化和监控自动化**

可视化设计时除了正常的业务数据报表外，对于主线重要的业务元数据及技术元数据的信息同样要做可视化设计，并加入自动化监控内容中。

关于飞谷云

飞谷云是大数据爱好者的家园，是共同有着‘诚信进取协同分享’文化的码农们聚在一起共同打造的大数据学习实践云平台,旨在帮助大学生或需要职业技能提升的码农们通过飞谷云平台（老师，实战环境，大数据生产项目）达到企业大数据相关岗位的技能要求。我们正在进行各大学和中小企业的免费公益交流活动，欢迎各大学社团或院系组织及企业联系我们，一起交流合作，一起让大数据落地！

大数据公益班

大数据生产项目

大数据人才服务

@飞谷云邮箱：feigu@vip.163.com

飞谷教育

THANKS

SequeMedia 盛拓传媒 IT168.com ChinaUnix ITPUB