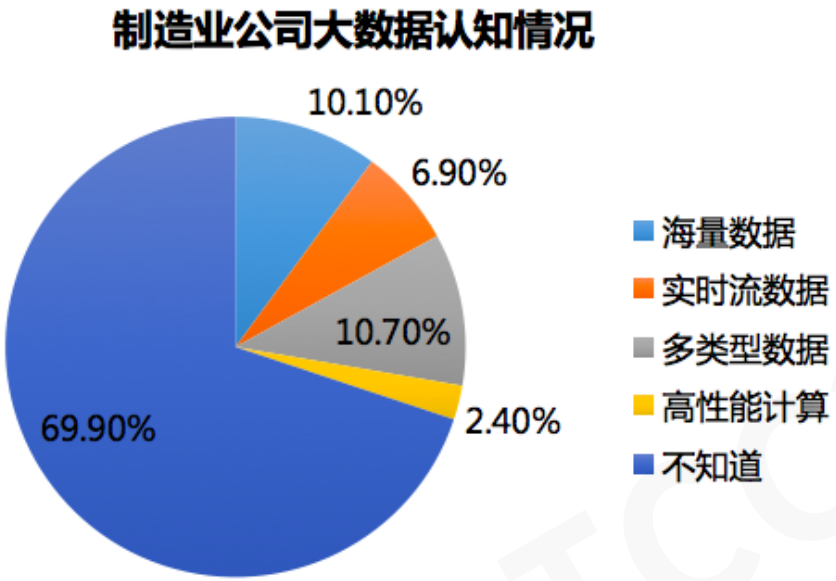
上汽集团数据湖

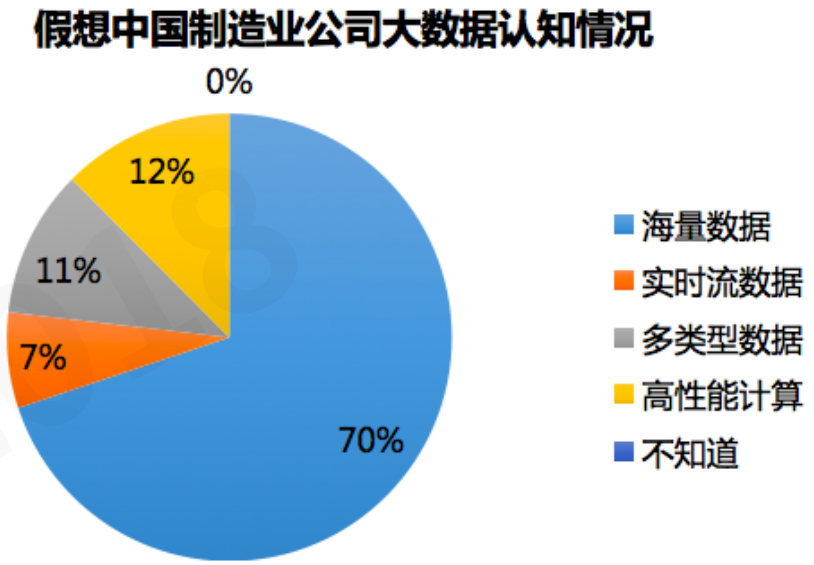
—— 实时工业大数据产品实践

侯松

- 微信公众号：**麻袋爸爸（madaiba）**
- 个人网站： http://www.housong.net
- 资深大数据架构师、Oracle ACE、PMP、北美寿险管理师， ACOUG社区、PG社区核心成员。
- 《高并发Oracle数据库系统的架构与设计》作者。
- 现就职于上海汽车集团股份有限公司，负责大数据中心数据服务、智能引擎、实时计算平台的研发，产品化及推广。

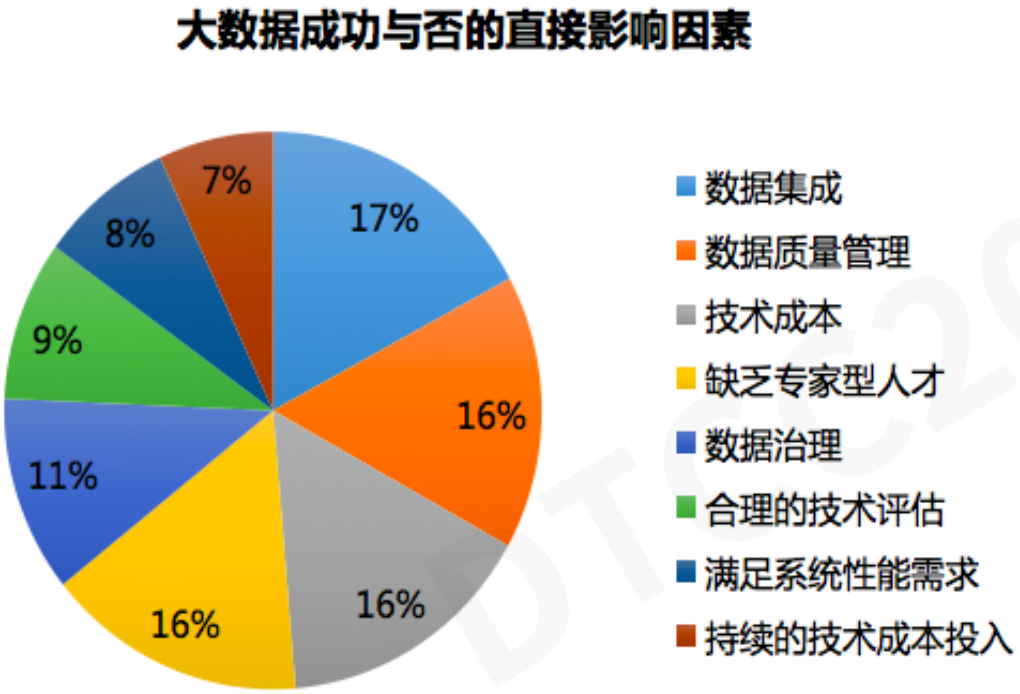
大数据的挑战……





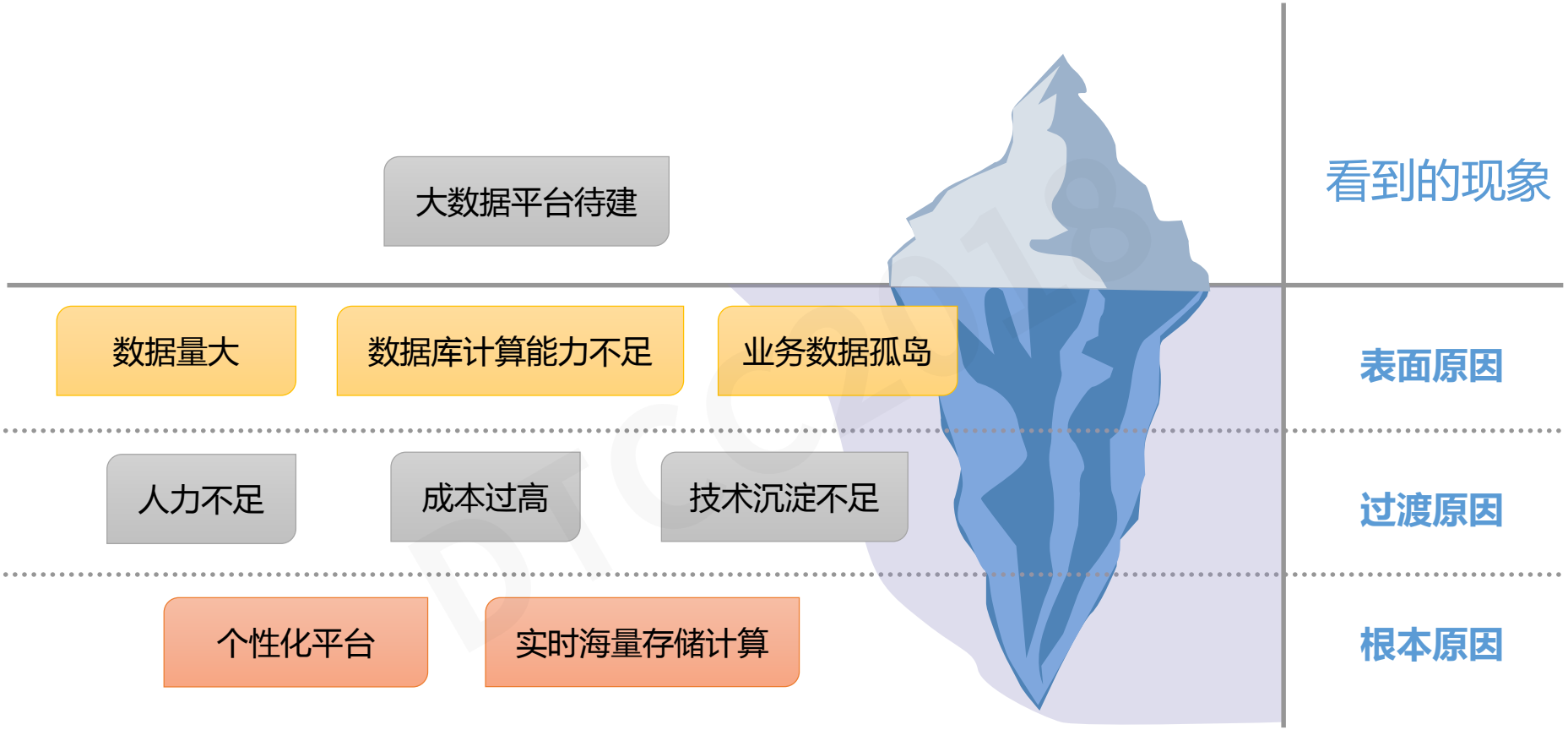
- I have a dream！有梦想固然是好的，但是任何不以落地为目的的梦想都是空想。
- 在明确实际的立足点之后，围绕着目标就要分清Want与Need了。Want是梦想，而Need是需求，是可以落地可以成为目标的实体。

大数据的挑战……



- 立足于制造行业来看，直接影响到大数据成功与否的因素大致如左图所示。
- 数据集成，也可以说是多方数据平台化的汇总吧。
- 对于制造行业的数据质量往往是不能得到足够保证的。

如何解决……


大数据平台待建
数据量大
数据库计算能力不足
业务数据孤岛
人力不足
成本过高
技术沉淀不足
个性化平台
实时海量存储计算
看到的现象
表面原因
过渡原因
根本原因

数据湖的概念

精细化规范体系建设

避免数据湖沦为数据沼泽

引流出多元化功能性支流

可便捷构建数据仓库

数据分析与科学计算

多元化数据源接入

多元化数据格式

低成本存储

简单模型弹性扩展

松耦合全量数据

更易于发掘数据本身的潜在价值

什么是数据湖（Data Lake）

- 数据湖并不是一个纯技术概念，而是数据管理的一种方法论。
- 数据湖实际上是一种利用低成本技术来**捕捉**，**提炼**，**储存**和**探索**大规模的长期的原始数据的方法与技术实现。

数据湖特征

- 数据存储：大容量低成本；
- 数据保真度：数据湖以原始的格式保存数据，具有高保真度；
- 数据使用：数据湖中的数据可以方便的被使用，进而引流到外围应用；
- 延迟绑定：不需要提前定义数据模型。

数据湖的发展阶段

1 蛮荒期

- 企业各类数据分析通过传统数据仓库来实现

2 萌芽期

- 企业引入了大数据平台
- 企业的应用数据和大数据平台有交互

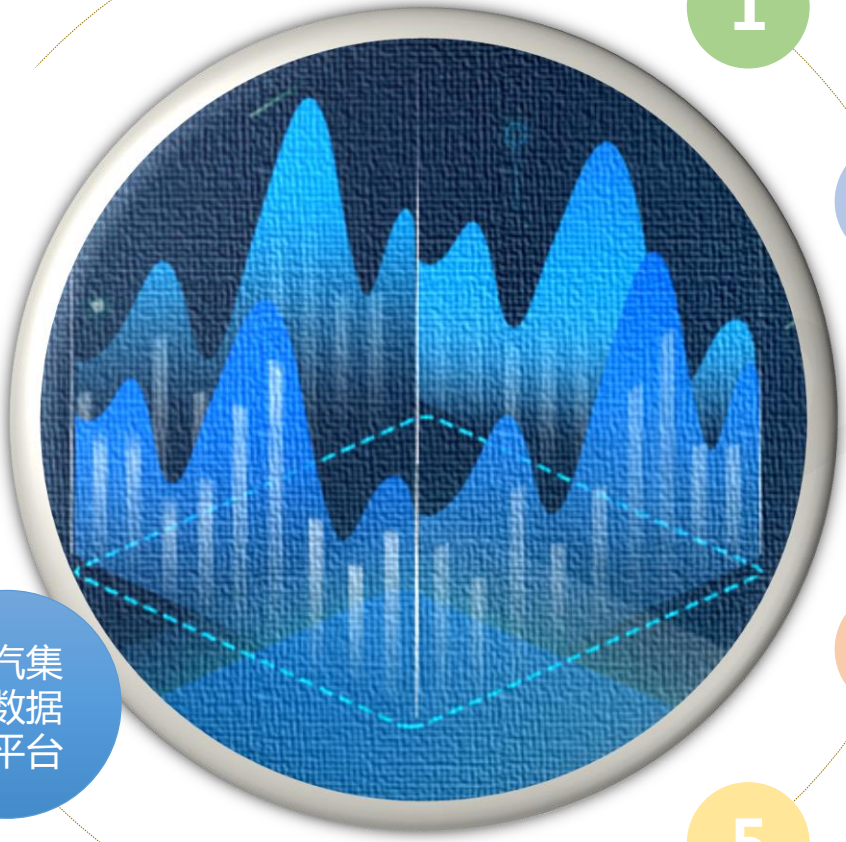
3 成长期

- 新的系统直接支持大数据平台
- 大数据平台成为缺省配置
- 数据仓库只在特定场景下使用
- 外部的数据也引入数据湖泊中

4 成熟期

- 数据湖和应用组件完善
- 大数据平台大量采用
- 加强其可靠性和安全性
- 对外提供丰富的应用接口
- 做到多租户的云服务

上汽集团数据湖产品

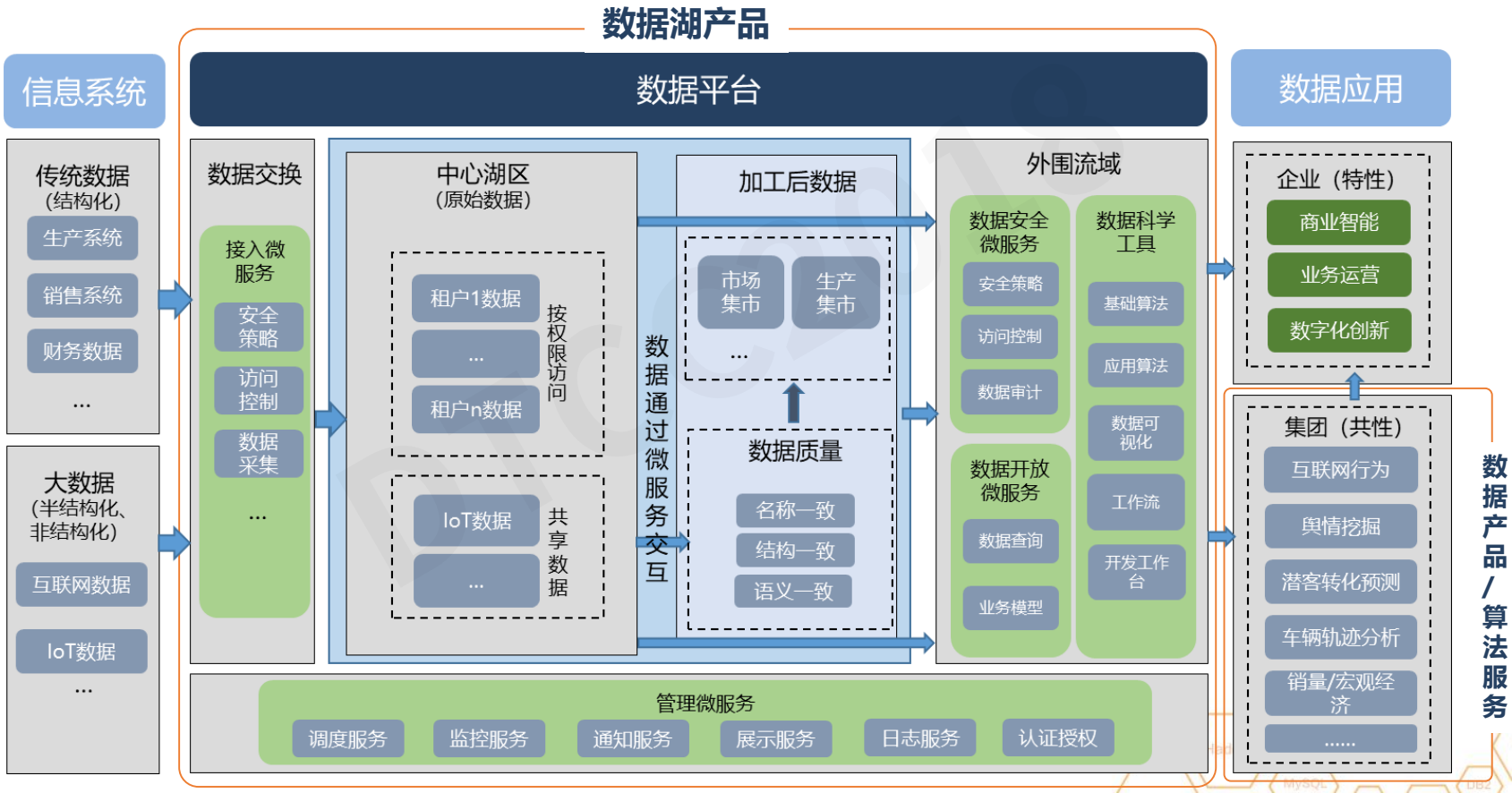


上汽集团数据湖平台

1. 数据库数据实时接入
   异构数据库数据融合
   每秒百万级数据接入
2. 数据备份及容灾功能
   数据快照及数据回溯
   百亿级数据亚秒级查询
3. 单位格级别统一权限管理
   金融级自动化数据加密
   敏感数据脱敏
4. 海量数据机器学习及数据挖掘系统
   海量小文件存储及检索
5. 无间断动态扩容
   高压缩比文件储存
   标准SQL接口，灵活扩展

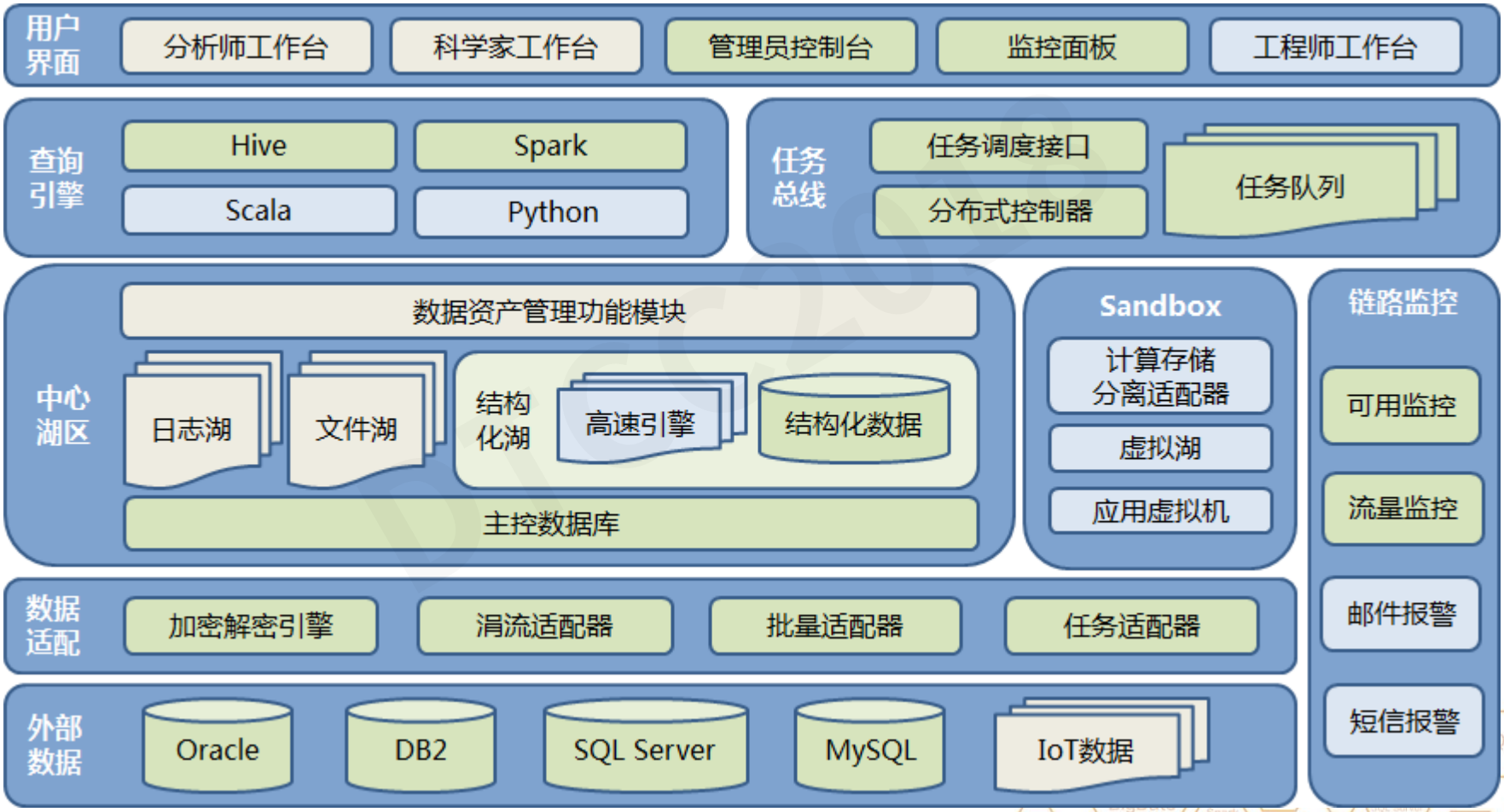
数据湖建设目的

集团数据平台在**统一规划和运营的基础上**，可根据用户的能力和需求，提供**灵活、多样、敏捷的服务**，协助企业建立自身大数据应用能力。目前，集团数据平台技术已经逐步产品化，并计划向合资企业和外部企业输出。同时，在数据应用项目的过程中平台将积累共性需求，形成数据产品、算法服务。



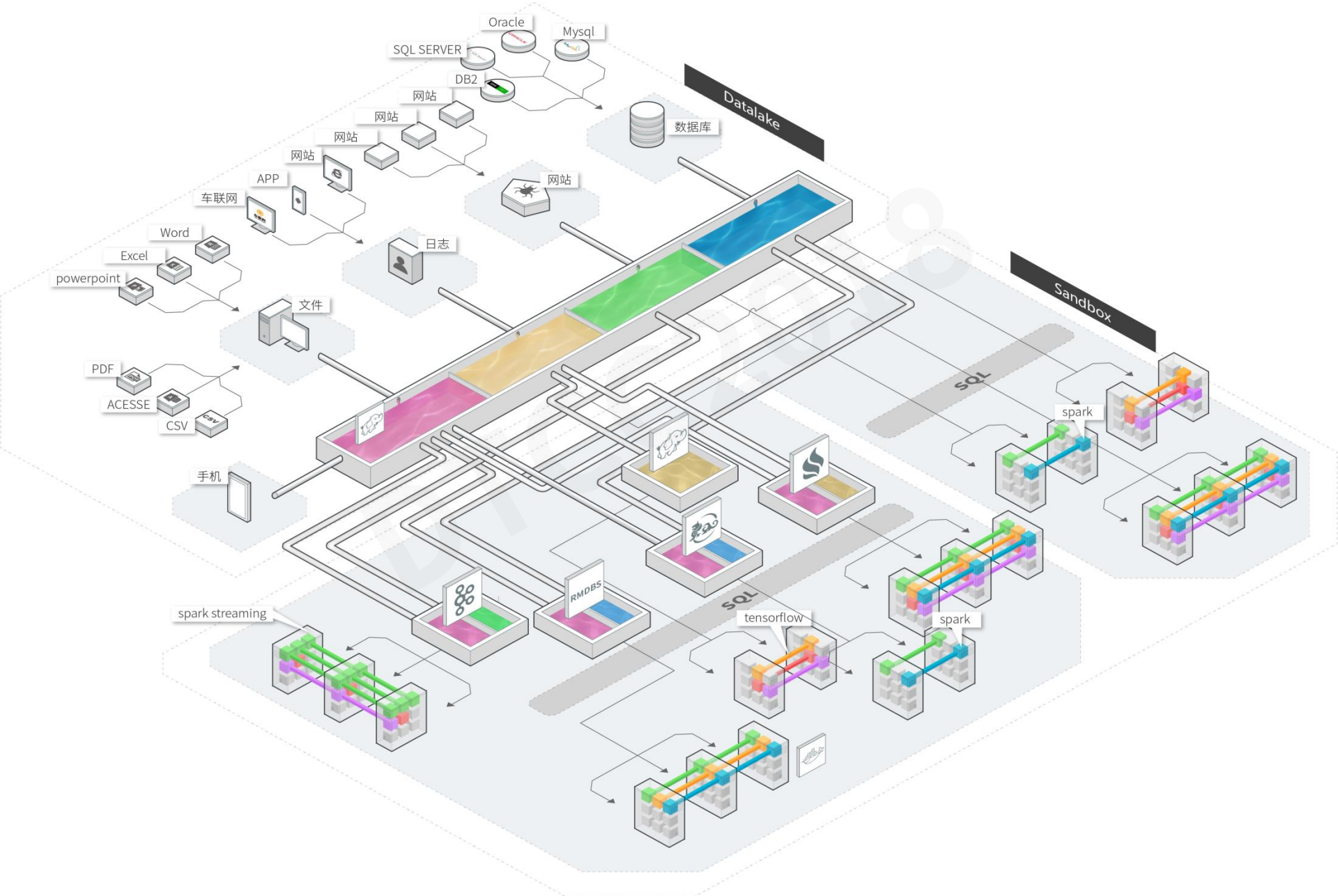
数据湖产品框架

采用**开源软件架构**，构建的**实时大数据集成平台**。降低企业使用大数据技术的成本，为数据分析师、业务分析师们提供更高效易用的工具，加速数据应用的建设和推广，并提供全字段金融等级3DES加密，自动无感知的密钥更新，防止密钥泄露。单元格级别权限控制和数据脱敏访问。**为集团大数据平台一体化打下基础。**

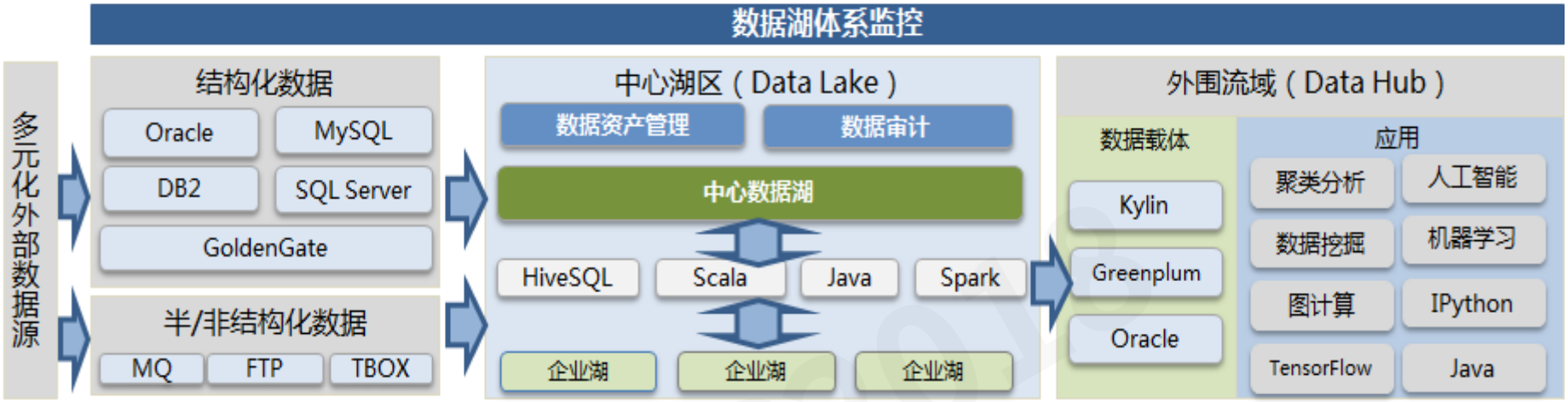


数据湖产品框架

数据业务部
Data service department



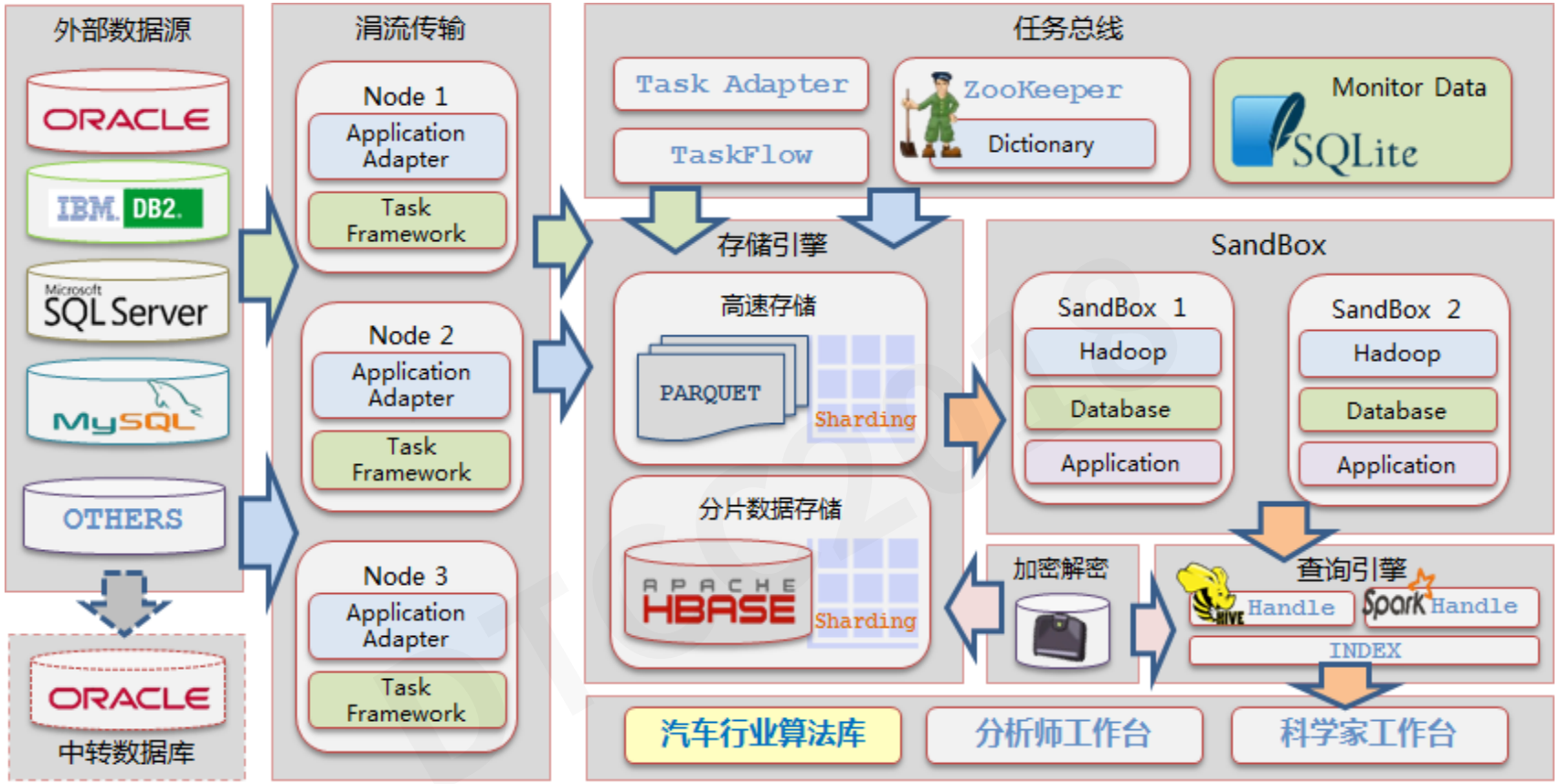
数据湖物理架构



- 整个数据湖体系分为三个部分：多源数据接入、中心湖群、外围流域。
- 多源数据接入：可分为结构化数据（需保证强一致性的数据库数据）、半/非结构化数据（不需要保证一致性的日志、音频数据）。
- 中心湖区：由核心业务对应的中心湖区和其他功能湖组成。集团湖与企业湖之间通过统一的数据交换层实现数据交换。中心湖的数据受到严格监管，包括：数据资产管理、数据审计等。
- 外围流域：从中心湖区通过统一的数据交换层，将数据引流到多元化的数据载体中，提供各类型的数据分析与科学计算应用服务。

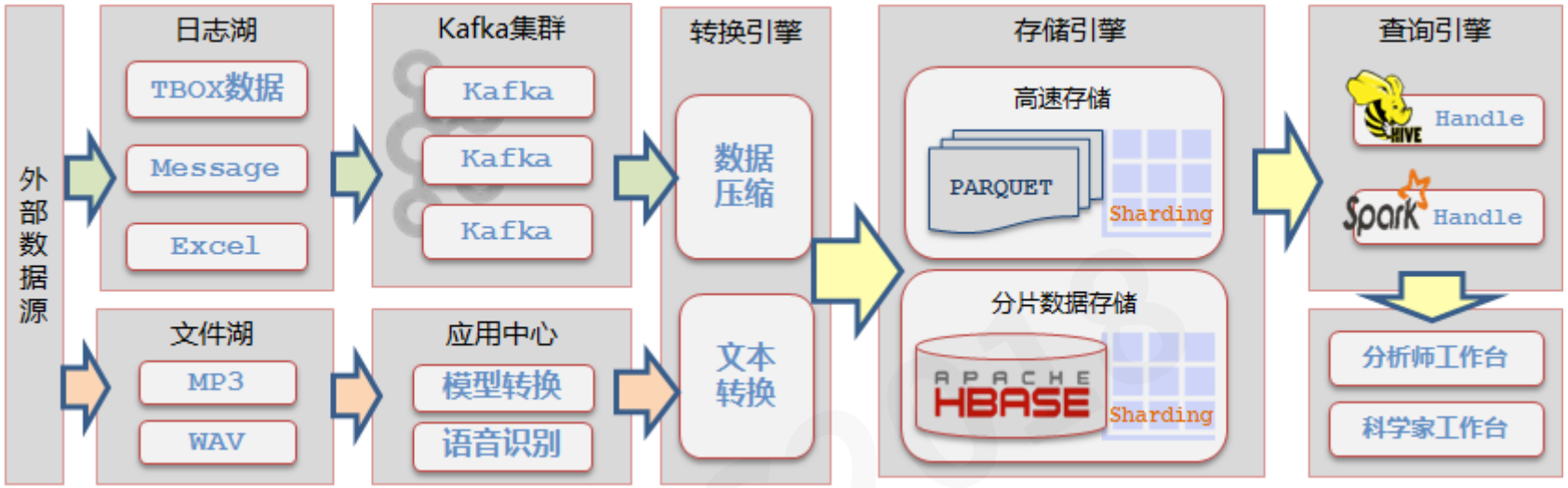
结构化数据湖概览图

数据业务部
Data service department



- 分布式涓流传输集群，完美融合存量数据高速并发导入与增量数据导入。
- 任务总线控制涓流数据加密后入库到HBase数据库分片数据存储，同时记录metastore。基于Hive和Spark的定制版Handle提供HiveSQL和SparkSQL接口，同时完成数据出库的解密。
- 在定制化工作台内，植入汽车行业相关业务的智能算法库，实现拖曳式智能算法应用。新增文件湖和日志湖的架构，以支持车联网数据的承接与应用。

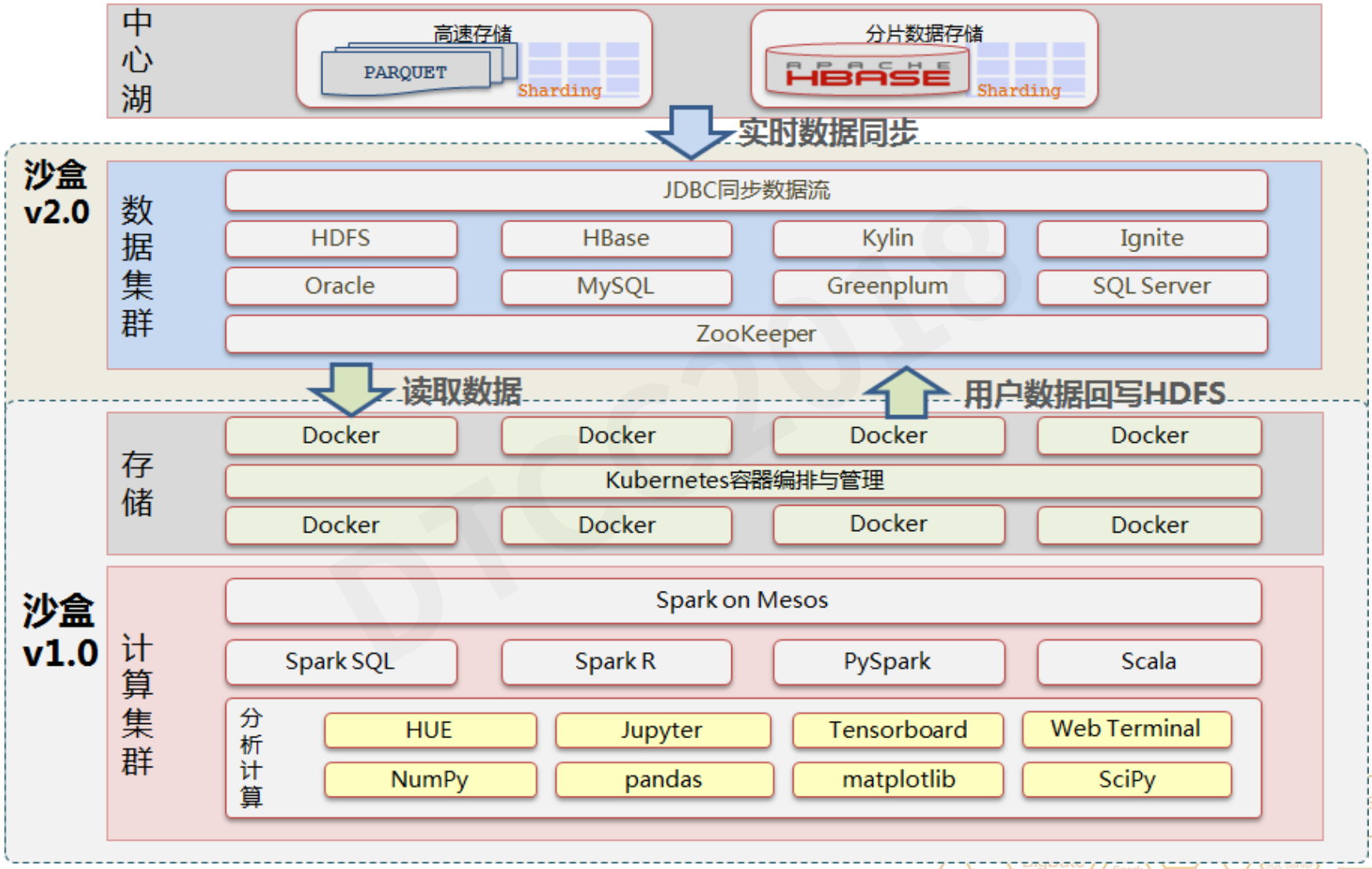
日志湖与文件湖概览图

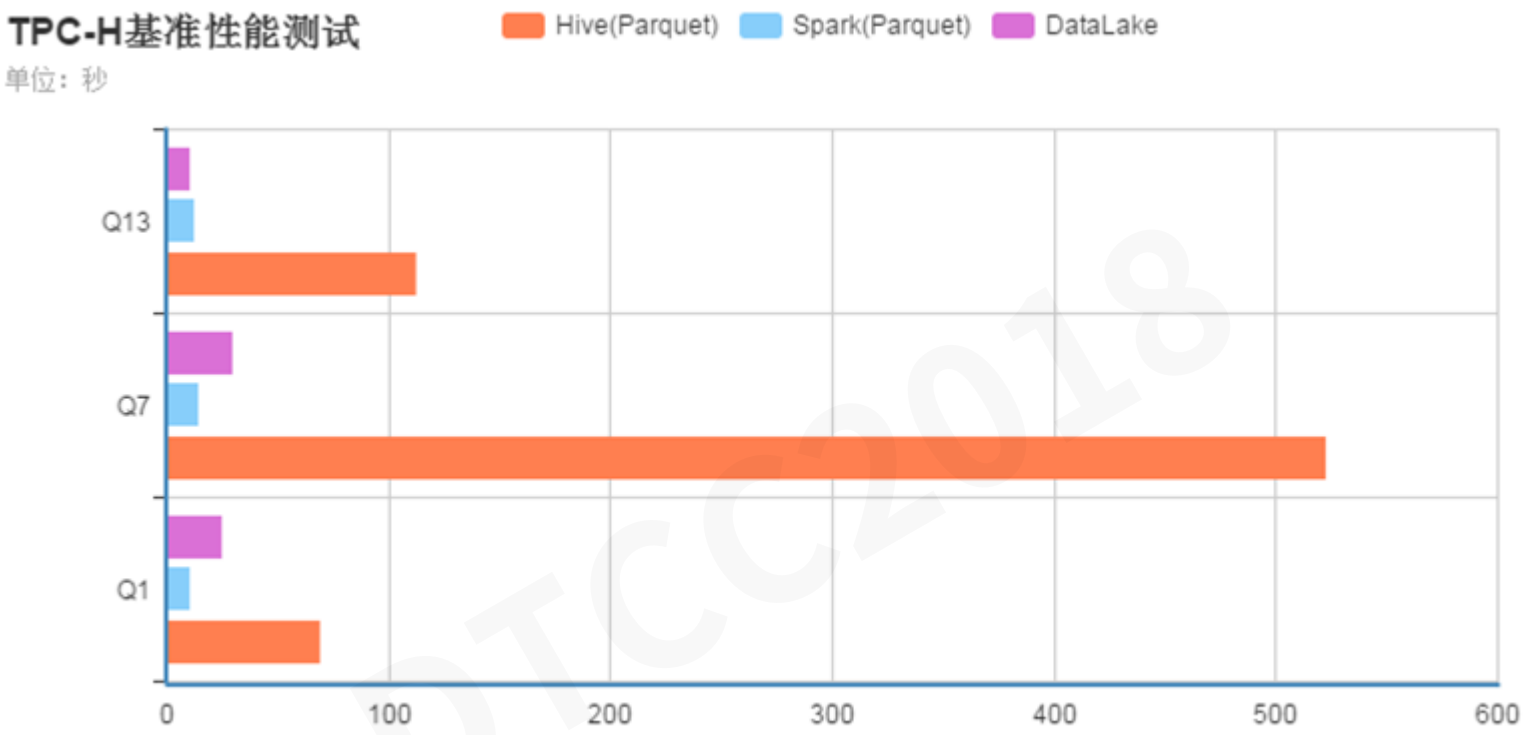


- 日志湖区和文件湖区往往数据量非常大，且价值密度较低。对于这类数据不要求强一致性，故而可不进行数据审计和定期数据一致性校验。
- 日志湖和文件湖多以半/非结构化数据为主，需要进行关联分析的进行模型转换，并将其导入到集团湖的HDFS或HBASE中。
- TBOX数据和用户网页行为分析的数据，数据产生并发度高，数据流量大，需要用Kafka集群进行数据承接，承接过程中需要进行一定比例的数据压缩，之后直接存储到HDFS中，通过HIVE外部表的形式进行访问，以降低集群负载。
- 对于文件中心的音频文件，推荐进行语音识别，将其转换为文本之后，再行入库。

BigData on Docker

数据业务部
Data service department

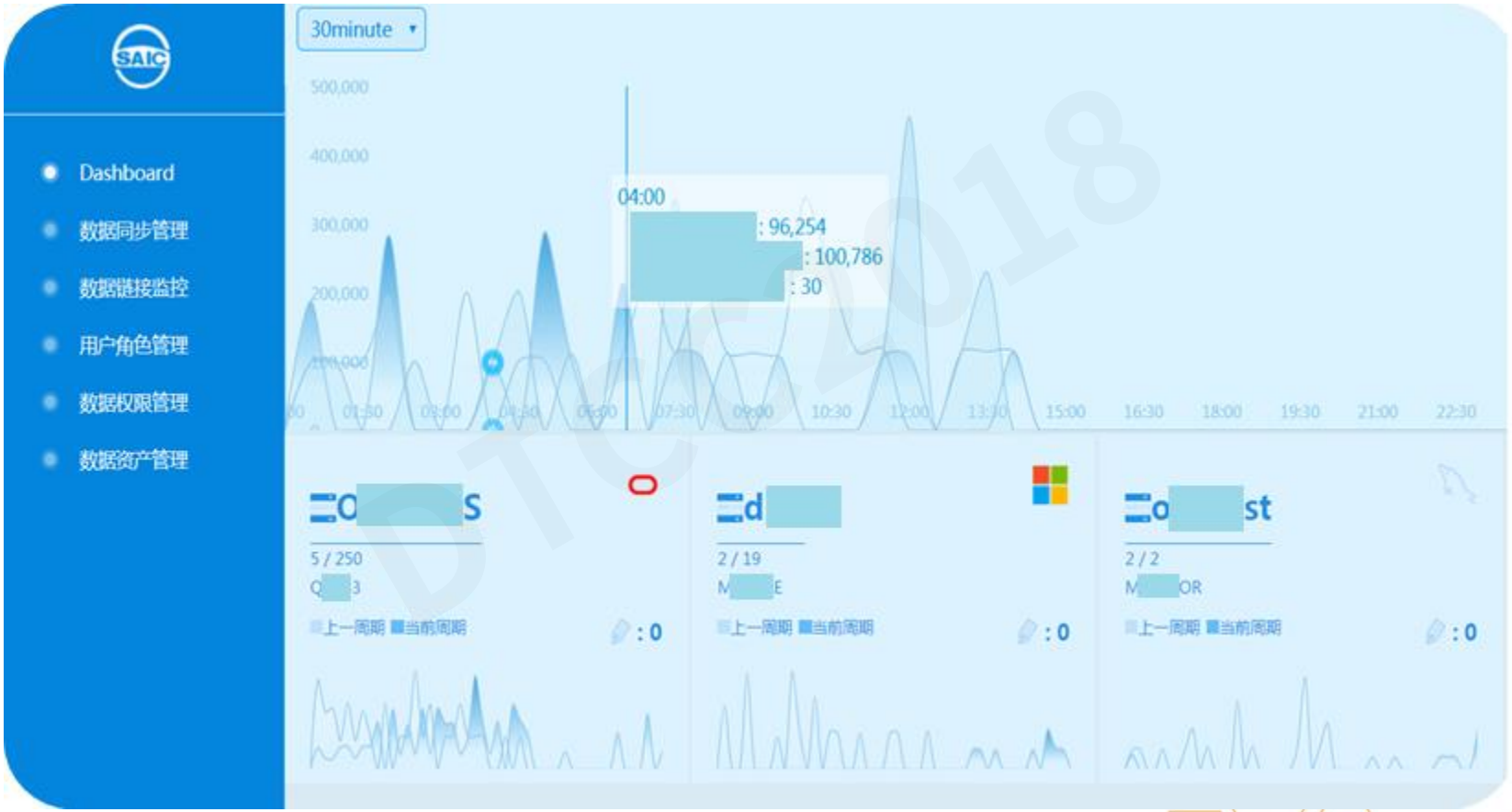




- 涓流复制传输平均速度：3万行/min。
- 数据湖在查询性能上，约为Hive（Parquet）的10~20倍，且与Spark（Parquet）相差无几。

数据湖平台UI —— 看板

- 如下图所示，完成数据湖主页面装载；
- 成功接入Oracle、MySQL、SQL Server三个数据库的实时。



数据湖平台UI —— 安全管理

- 数据安全管理页面，可以完成加密方式、脱敏控制、列访问权限、行查询权限的设置。
- 以表INVOICE_DOC为例，针对INV_TYPE列，组合四种安全选项的设置，达到单元格级别的加密和权限控制。

生产环境用户遍布上汽集团





数据业务部
Data service department

爱上汽车 畅行天下 欢迎您加入我们





THANKS

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168