TRANSWARP
DATA HUB

大数据技术在江苏邮储银行的创新应用

邮储银行江苏省分行现有数据下载平台系统共包含超过1200张数据表，内容涵盖储蓄、汇兑、理财、个人信贷及对公业务等邮储银行各项主营业务。近五年来共支持完成十多项主题案例分析以及大量日常（临时与例行）数据提取，为经营管理、业务营销及风险防控等工作提供了强有力的数据信息支持。

然而近年来随着省内数据下载平台数据范围的不断扩展以及日常加载数据量的不断积累，一些问题逐渐积累并显现出来，其中以下面几个问题尤为突出：

首先是存储空间紧张，目前下载平台共约1200张数据表，占据约12T存储空间，并且以每日增加约10G的速度快速增长。经过多次清理，磁盘使用率仍然接近警戒线，经常需要不定期突击清理，以避免影响源业务数据的正常加载。

其次是数据质量参差不齐，在数据信息服务过程中常常出现表间字段异常匹配、数据字典缺乏等问题。

最后是由于生产系统数据表结构限制，对于大部分数据表无法保留并追溯其历史变动，无法获取特定时点的状态，导致需要经常向总行申请数据或者在月初时点手动备份数据，严重影响数据信息服务效率。

前述三项问题目前较为严重，已经影响到数据下载平台的日常运行维护，并给数据分析团队的日常工作带来了较大压力，是目前亟待解决的主要问题。

同时，当前省级机构建立数据中心的必要性也是一直都被考虑的问题。以往的模式不便于数据的加工处理，数据统一集中管理与自主掌握数据能使机构具有更高的主动性。而在时代和数据不断变化的前提下，自主掌握数据仓库也能极大提高平台与机构对数据的适应力。

自建数据中心，可以做到更加快速的响应，更好的解决地市的数据需求。地市及以下机构的科技力量薄弱，无法完全依赖总行，需要省行据有一定的“开发能力”以支撑地市分行、支行的业务发展。

由此背景，希望建成以省分行数据集市为核心，具备数据存储、数据处理、信息加工、信息发布和数据安全管理等功能的企业级数据分析平台。并且以历史数据集中管理平台项目（数据下载系统部分）项目建设为切入点，缓解目前下载平台数据存储压力，建立数据分层管理，强化数据质量管理，着重解决存储空间紧张、数据质量参差不齐、数据表无法追溯历史变动、提高数据安全等近期亟待解决的几项主要问题，同时也解决档案管理系统、云盘系统的数据存储问题。

在平台建设上，历史数据集中管理平台，通过统一的数据控制和管理平面，面向上层应用提供数据存储和查询接口，提供标准的SQL接口并保证分布式事务处理一致性，业务操作人员也可以通过RStudio工具直接连接到历史数据集中管理平台进行分析挖掘，同时通过分布式ETL工具从下载平台和其他系统完成数据采集，并根据数据存储和分析要求，选择HDFS、TDH Hyperbase进行数据存储。

与传统数据仓库相比，一般Hadoop大数据平台更适合从价值密度低的数据中挖掘金子，更适合作为数据仓库和OLAP分析体系最基础的平台构建。但是，TDH Inceptor基于Hadoop平台通过对于SQL 2003以及PL/SQL的高度支持以及内置高性能的内存计算引擎，并能够支持分布式事务处理保证CRUD操作的ACID特性，是新一代数据仓库的代表产品。

尤其是这个应用场景中，增量的从总行同步数据，要求大数据系统能够支持分布式事务处理保证CRUD操作的ACID特性，来做到增量同步，保证数据一致性，这个技术能力至关重要。

而大数据技术平台的应用优势具体可体现在四大方面。

首先是扩展性上，平台可无缝扩展，支持不停机扩容，满足企业不同时期数据增长对数据平台的应用需求，这种扩展对上层应用完全透明。

其次是多样性，除了结构化数据外，Hadoop还能够对非结构化数据进行处理和分析，例如weblog、syslog、音视频等，Hadoop对数据类型不敏感，为了海量数据的分析应用所专门设计。

完整性方面，Hadoop可以存储完整的原始详单，提供高并发低延时检索查询，同时可以在TDH Discover中进行分析挖掘以及在Inceptor中进行数仓类应用,并能结合分布式内存列式存储进行交互式分析，能够挖掘更多有价值的信息，回溯分析、趋势预测。此外,Hadoop提供了完整的数据库导入导出和各类ETL工具。

最后则是高性能，Hadoop被百度、Google、阿里等互联网公司广泛应用，主要在于基于Hadoop提供了一个整合的数据平台，使得计算更靠近存储，同时所有的任务都可以并行执行，并结合Inceptor分布式内存计算引擎，大大提升数据分析挖掘的性能。

平台分层的架构设计如下图所示。

相比以往的数据平台，有三方面的改变。首先是各类数据库导入，包括通过ETL、数据采集工具等进行批量导入，导入过程中结构化数据、非结构化数据、流水数据直接存储到HDFS中，在Hadoop平台中利用其高性能计算，进行数据清洗转换整合，变传统的ETL为ELT。

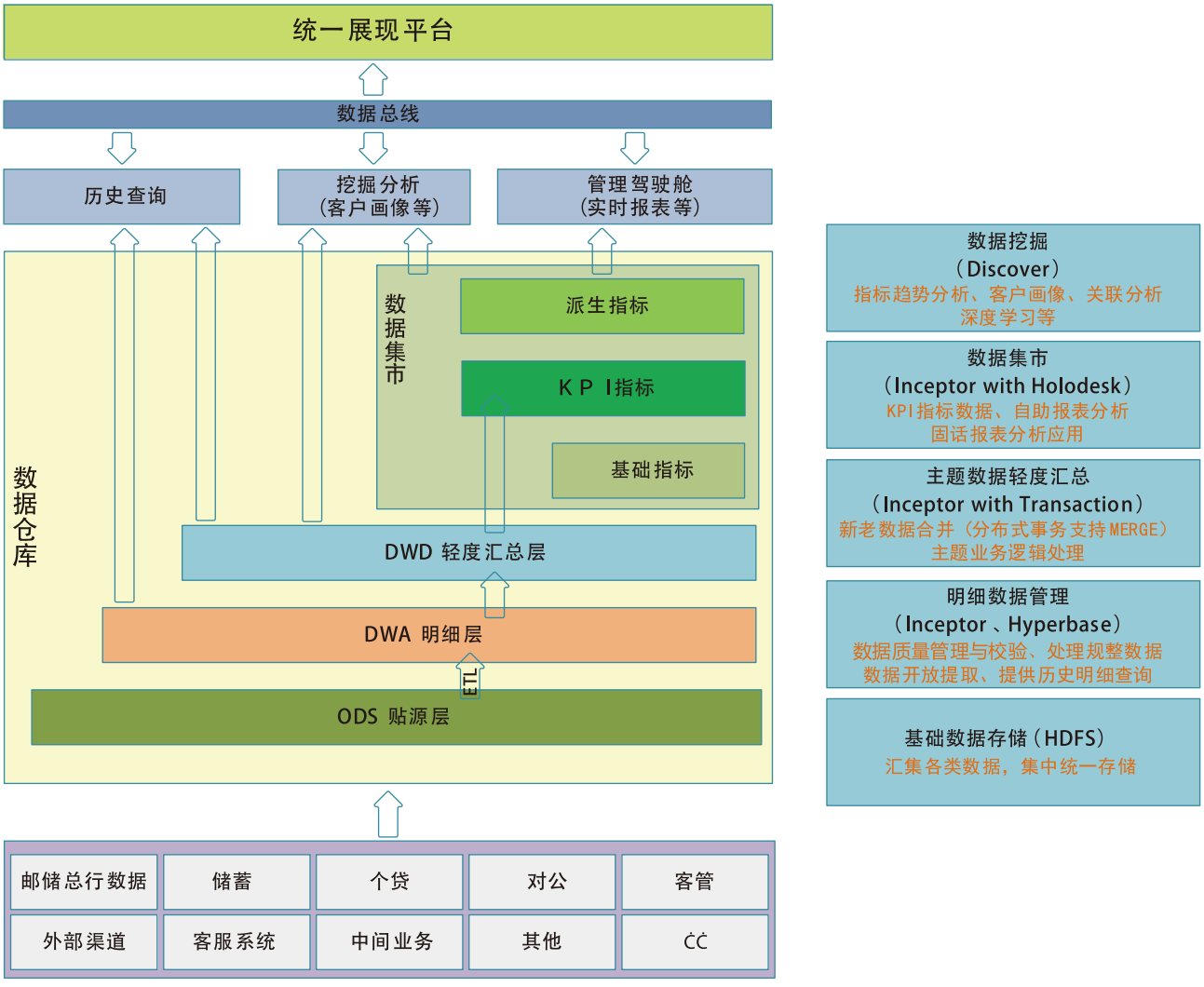
主要的数据来源为从邮储总行增量的同步数据，在数据校验后，将总行同步数据，通过Inceptor中对于分布式事务处理的支持保证新老数据批量合并过程中的数据一致性。在MERGE INTO、INSERT、UPDATE等CRUD语法对于数据可能同时多样的操作中，必须保证整个事务操作的ACID特性（原子性、一致性、隔离性以及持久性）来保证整个数据仓库中

的数据最终一致性。如果不具备分布式事务处理特性，就无法上线数据仓库业务，所以目前开源Hadoop产品无法有效应用于真正数据仓库领域。在处理规整后的数据，可方便的通过SQL Bulkload批量加载到Hyperbase中，同时建立索引，提供检索查询；也可以通过各类业务逻辑进行进一步数据处理与汇总。

其次，基于Inceptor计算框架，对于Hyperbase的中数据，支持建立二级索引，通过SQL提供高并发低延时的检索查询；对于Inceptor数据仓库中事务表进行数据整合汇总，同时可以将汇总数据供数给数据集市；通过简单的SQL语句将数据仓库中处理数据抽取到Holodesk分布式内存列式存储中，提供秒级上亿数据的交互式探索。相比较于，传统的数据集市，可以提供不确定模型的即席秒级分析，业务人员通过报表工具随意拖拽业务维度，后台秒级完成计算，交互式进行数据探索。

同时，Discover中提供统计类和机器学习类函数和算法，并与R语言良好结合，提供各类算法的R语言接口，完成各类数据的挖掘探索。对于各类全量数据的挖掘分析计算，当通过R提交统计分析算法时，系统自动转换成分布式任务并执行。

这三方面的改变使得平台对操作员来说更容易掌握与运维。



本期平台按照200TB的整体容量进行规划，考虑数据存储三个副本，在线数据采用Snappy压缩，近线数据采用Erasure code进行压缩，数仓从ODS到DWA的收敛比为20%，DWA到DWD的收敛比为10%，这里考虑原始数据在ODS层尽可能存储较长的时间，并且预留30%的数据膨胀空间 。预期在未来五年达到54T的增量。

平台的建成将消除各源业务系统的信息孤岛，有效减少数据冗余，保证数据的准确性与唯一性，并能够实现安全的数据资源共享，为各业务部门和各地市分行提供高效、立体、及时的数据信息支持。。

星环信息科技（上海）有限公司

地址：上海市徐汇区桂平路481号18幢3层301室（漕河泾新兴技术开发区）

邮编：200233　　电话：4008 079 976

网址：www.transwarp.io