TRANSWARP
DATA HUB

江苏银行大数据技术平台选型分析

江苏银行大数据平台建设起步于2014年底，2015年年中初见成效。目前江苏银行利用大数据技术开发了一系列具有一定社会影响的大数据应用产品：如“e融”品牌下的“税e融”、“享e融”等线上贷款产品、基于内外部数据整合建模的对公资信服务报告、以实时风险预警为导向的在线交易反欺诈应用、基于柜员交易画面等半结构化数据的柜面交易行为检核系统等。

江苏银行股份有限公司信息科技部总经理 葛仁余

大数据应用的本质是对客户需求的认识和释放，应用效果取决于银行的综合运营服务意识，而选择一个合适的技术平台也是大数据成功应用的不可或缺的重要因素之一。江苏银行在大数据技术平台建设方面进行了大量探索和思考，本文重点介绍其大数据技术平台选型思路，以期与同业共同交流、分享、探讨大数据技术在银行业的应用实践。

一、为什么要建设大数据技术平台

截至2015年6月，江苏银行资产规模达到1.2万亿元，一方面，成立8年来，江苏银行积累了大量的内部数据，以往受限于高性能存储的成本和数据并行化处理能力，占总存储量80%以上的数据是“死”在系统里的。以对私客户的活期账户为例，一张拉链表的数据量就达数百GB，运行在IBMP6系列小型机上的Oracle数据库统计一下表的行数就要3个小时，若需要全量回算历史数据，为避免影响生产，需要将数据导出到另外的数据库上，花费几天时间。又如，诸如“柜员操作记录”这样的半结构化数据每天产生的数据量达几个GB，生产环境只能保留最近几天的数据，其他数据存储在磁带库上，使用时需花费大量的人力将数据从带库中导出。

另一方面，为减少贷前审查的录入成本，开发纯线上贷款产品等，江苏银行陆续引入税务、法院、工商、黑名单等外部数据。随着内外部数据量的快速增长，大规模数据处理和实时响应的需求使得传统的数据处理平台遭遇瓶颈，江苏银行急需探索新的数据架构，采用新的数据处理技术。

当前，银行业面临的挑战主要来自两个方面：利率市场化和互联网金融。利率市场化拉近了传统银行与实体经济的横向联系，要求银行快速提升数据洞察能力;互联网金融使得银行的数据应用不能局限于传统的查询统计分析应用，还需提供高效精准的营销，并具备实时风险防控能力。相较于大型商业银行，城商行的竞争更加激烈，传统的数据产品和应用服务已无法满足新形势下城商行应对市场竞争的需要。

二、大数据技术平台架构分析

经过对主要大数据处理平台的深入研究，江苏银行将关注点聚焦在两个方面：一是选择MPP还是Hadoop;二是选择开源版Hadoop还是发布版Hadoop。为此，江苏银行更近一步从数据容量和数据处理能力的线性关系分析传统数据平台、MPP和Hadoop的关系(如图1所示)。

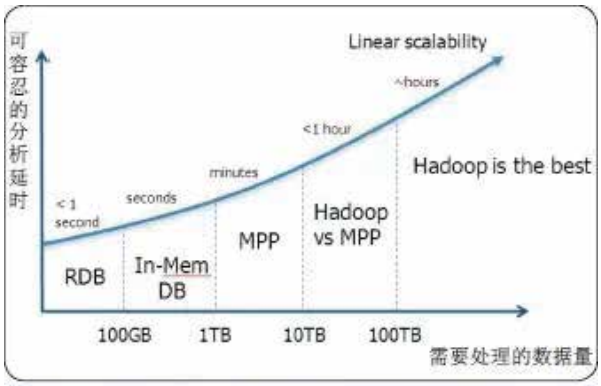

图 1　传统数据平台、MPP 和 Hadoop 在数据容量和数据处理能力的线性关系

传统观点认为，MPP的适用范围为1TB~100TB数据量，数据量超过100TB，Hadoop更具优势。当前，大中型城商行的数据量普遍在10TB级别，因此一些城商行选择MPP作为大数据处理平台。

然而，近年来随着Hadoop开源社区的不断发展，特别是Spark2.0的发布让Hadoop焕发了新的活力。Spark2.0具有RDD(ResilientDistributedDatasets)和DAG(有向无环图)两项核心技术，基于内存计算优化了任务流程，具有更低的框架开销，使得Hadoop在MPP擅长的100TB以下数据量的处理性能也大为改善。以目前的Hadoop技术，100GB以上的数据量处理性能不弱于传统关系型数据库和MPP，10TB以上性能优势更为明显。因此，图1所示混合架构的大数据处理平台模式逐渐淡出，形成如图2所示的新型应用模式。

图 2 大数据平台新型应用模式

江苏银行从经济成本和未来数据的非线性增长趋势的角度分析认为，传统的交易系统运用关系型数据库处理OLTP事务操作，产生的交易数据通过异构数据的批量复制方式或消息队列的准实时方式更新至Hadoop平台，Hadoop平台进行大体量数据的分析和挖掘，并提供基于大数据的应用系统实时检索的模式，与城市商业银行目前的数据架构相适应，决定选择Hadoop平台。

选择开源版本的Hadoop还是产品化的发布版Hadoop?众所周知，Hadoop的优势是没有额外的产品费用，技术更新快，开放程度高，应用服务集成商多。国内很多知名互联网企业在开源版本的Hadoop基础上优化形成了自己的大数据产品。为此，江苏银行考虑基于Hadoop开源框架自建大数据平台，但测试后发现此方法可行性不高，原因有三：

一是城商行科技力量有限，大部分力量投入在应用研发领域，在基础软件的研究和开发方面的专业能力远远比不上IT公司，即使只从事集成组件的工作也不一定能达到预估的效果;

二是深入研究平台技术需要一定的时间，城商行在起步阶段已经落后于互联网企业，来自互联网金融的激烈竞争留给城商行的时间远远不够;

三是行业监管机构对商业银行应用系统的安全性、稳定性和连续运营有着严格要求，开源产品一旦出现重大问题没有及时修复的保障。

经过慎重分析和实际测试，江苏银行将选择范围集中在符合银行应用需求的成熟的具有高效技术支持的Hadoop发布产品。

三、大数据平台选型要点

在对产品化的发布版Hadoop平台选型的过程中，江苏银行总结了以下需重点考量的内容。

1.性价比和扩展性

前期江苏银行在IOE传统架构上进行了大量投入，而城商行总体自主可控能力较弱、资产规模较小、盈利能力较低，因此，不论是从自主可控要求的目标出发，还是从降低软硬件成本投入的角度，都要求大数据产品须支持在x86虚拟化集群搭

建开放和高度并行化的处理平台，既要适应高并发低时延的移动互联网实时数据检索需求，又要满足大体量数据的统计分析与业务建模要求;要求总体技术方案具备高性价比，能够实现在同一服务器集群上针对不同应用动态灵活分配内存、CPU等硬件资源并支持动态扩展，在出现资源瓶颈时能够快速解决。Hadoop产品具有支持x86和可动态扩展的性能，但目前大多数Hadoop平台在不同应用间资源有效隔离方面存在一定缺陷。

2.对SQL的兼容性

开源Hadoop对标准SQL及PL/SQL支持程度不高，许多常用函数不支持，需要使用者编写程序实现。而银行以往数据集市、数据仓库等应用大都基于SQL开发，根据江苏银行的数据架构规划，数据集市、数据仓库将迁移至Hadoop平台，为避免少则几百行多则上万行的程序编写，SQL兼容性成为Hadoop平台选择不可或缺的考虑因素之一。

3.对于通用开发框架和工具的支持程度

江苏银行应用系统采用数据库+中间件+应用的三层模式，开发环境为JavaHibernate和Spring框架。为此要求Hadoop平台下的HDFS库、Hbase以及内存数据库等组件能够通过ODBC或JDBC连接，以实现数据库对应用开发人员透明，并支持诸如BI、ETL、数据挖掘等工具，数据源可以根据实际需要选择配置Oracle或Hadoop。

4.具备事务的基本特性

大数据平台不仅是关系型数据库数据转存储和统计分析工具，更是一些新型应用，如客户线上行为等的原始数据库。为确保数据准确性，数据操作必须具备事务的基本特性：原子性、一致性、隔离性和持久性。Hadoop分布式计算的特点，决定其本身不具备事务的基本特性，必须借助插件实现。

5.图分析与流处理能力

银行的实时营销和实时风险预警场景需要大数据平台具有历史数据快速统计、窗口时间内的信息流和触发事件及模型匹配、百毫秒级事件响应等性能，流处理技术是关键。目前Hadoop平台通用的流处理引擎主要为SparkStreaming和Storm，两者各有千秋，SparkStreaming由时间窗口内批量事件流触发，Storm由单个事件触发，单笔交易延迟方面SparkStreaming高于Storm，但在整体吞吐量方面SparkStreaming略有提升。在进行Hadoop产品选型时江苏银行主要考量了经过优化的流处理引擎是否能够在流上实现统计类挖掘算法。

6.数据存储形式的多样性

要求Hadoop产品至少支持3种数据存储形式：一是行式存储，用于数据由传统数据库向Hadoop数据库过渡;二是基于键值对的存储，用于大体量、高并发数据的实时查询;三是内存式数据库，用于交互式数据分析和挖掘，可通过构建分布式cube加速性能，也可部分使用SSD替代，程序自动选择存储层。

7.多用户多数据库的隔离

商业银行对数据安全非常重视，要求不同来源的数据在Hadoop平台上分库存放，并且为不同用户针对库、表、行访问分配不同的权限。开源Hadoop平台不具有用户权限概念，许多使用者在Hadoop平台只建一个库，所有应用使用同一个用户名访问资源，数据资源完全开放。这种方式存在严重的安全隐患，预计随着平台重要性的提升，拆分数据库细分用户权限的需求也将越来越迫切，为避免因前期规划不合理导致的后期巨大的拆分工作量，江苏银行在大数据平台选型之初就将多用户多数据库的隔离作为重点考量的因素。

8.平台的研发能力和开放性

Hadoop作为创新型技术，与传统数据库相比，技术成熟度不够。江苏银行选择使用产品化的Hadoop，目的在于借助专业技术厂商的强大的自主研发和服务支持能力，快速修复技术缺陷，在充分理解银行数据应用复杂需求的基础上，充分

发挥产品特性，支持银行业务创新。

9.不同数据规模和应用场景下的性能表现

银行业的应用场景及需求较其他行业更为复杂，一些典型的应用场景和主要技术包括以下几个。

①用户行为采集分析：数据探头(JS、SDK，Nginx、ICE)、数据分发(Kafka)、离线数据存储及处理(HBase)、运营分析结果展现(MySQL)。

②跨部门数据整合：数据桥接(Sqoop)、日志接入(Flume)、数据分发(FTP)、离线数据存储及处理(HBase、ES)。

③离线用户画像和用户洞察(支持营销)：离线数据存储及处理(HBase、ES)。

④实时用户画像及推荐：实时数据处理(Storm、Spark)、数据存储(Redis、MongoDB)。

⑤实时反欺诈：数据接口(API)、数据分发(MQ)、实时数据处理(Storm)。

此外，风险管理领域的应用场景包括实时反欺诈、反洗钱，实时风险识别、在线授信等;渠道领域的场景包括全渠道实时监测、资源动态优化配置等;用户管理和服务领域的场景包括在线和柜面服务优化、客户流失预警及挽留、个性化推荐、个性化定价等;营销领域的场景包括(基于互联网用户行为的)事件式营销、差异化广告投放与推广等。

10.并行数据挖掘能力与R语言支持

目前江苏银行已经采购SAS数据挖掘工具，在风险管控、市场营销、产品定价等领域开展了一系列的模型开发和策略设计等业务应用，随着Hadoop大数据平台的引入，江苏银行开始积极探索基于并行数据处理技术下R语言运用，R语言可以直接访问Hadoop数据，为全表、全字段立体式的数据挖掘提供了坚实的技术保障。利用R语言的机器学习算法，如深度学习算法可以快速从风险、市场营销、差别化服务等角度对客户进行细分。Hadoop平台通常只支持单机版R，在选型时，江苏银行重点考虑了R算法的支持度问题，要求所选Hadoop平台对R算法支持超过70种以上。

11.非结构化数据处理能力

当前国内各银行已建有数据仓库或数据集市平台，大数据平台的引入往往独立于数据仓库，对于某些场景，将结构化数据与非结构化数据整体应用具有更好的分析效果。大数据平台和传统数据仓库应如何有效整合?

首先需明确“结构化”和“非结构化”数据概念。狭义的理解，结构化数据指关系型数据，其余都是非结构化数据。广义的理解，结构化数据是相对于某一个程序来讲的，如视频对于播放器来说显然是结构化的，但是对于文本编辑器来说就是非结构化的。

基于上述理解，江苏银行认为，无论是语音、影像还是其他“狭义”的非结构化数据，只要和银行的经营管理、业务发展有关，就可以作为大数据应用的一个数据源，技术上借助特定工具对其进行处理即可使用，如通常HTML网页被认为是非结构化数据，因为难以从中提取结构化字段，如电商网页上的商品名称、产品价格等，但借助网页抓取工具，可将上述页面信息转化为结构化字段，那么后续按照结构化数据处理即可。语音、影像也是一样，关键是我们期望从中提取什么信息，用什么工具提取，一旦提取成功，即可整合到大数据应用中。

在实践中，江苏银行大数据平台已实现网页、文本、JSON、XML等非结构化数据整合以及部分图像和语音数据的整合，并应用到了业务分析中。

产品化Hadoop独立于开源框架，却不能完全脱离开源框架，对开源框架的兼容和支持，有助于提升平台的开放性，过于独立的产品不利于在市场上寻找更多的合作伙伴。

江苏银行大数据应用从起步到取得多项成果效，经历了9个多月的时间，其中平台选型和技术调研花费了近半年时间。然而磨刀不误砍柴工，找对技术方向，后续的整合数据、建立模型、应用开发就成了水到渠成的事情。

星环科技是目前国内极少数掌握大数据核心技术的高科技公司，专注于企业级大数据核心平台数据库软件的研发与服务。公司产品Transwarp Data Hub (TDH)以其业界最完整的SQL on Hadoop支持; 独特的对分布式ACID数据一致性支持；以及对SSD优化提高集群性价比等特点，比肩硅谷同行。产品的功能和性能在业界处于领先水平。在全球去IOE的大背景下，TDH已成为在数据仓库，数据集市等领域替代传统数据库公认的大数据产品。

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