**八大行业Hadoop大数据应用回顾和展望**

任何新技术的发展都会经历一个从被公众了解到最终普遍应用的过程。大数据技术作为一个新兴的数据处理技术，经过了近十年的发展，刚刚开始在各个行业得到应用。但从媒体和公众视野中，大数据技术总是带有神秘的色彩，似乎有着挖掘财富和预测未来的神奇力量。广泛流传的大数据应用案例包括Target超市根据女孩的购物历史判断是否怀孕，信用卡公司根据用户在不同时空的购物行为预测客户的下一个购买行为，等等。大数据技术也为我们描绘了一个个如“智慧城市”，“智慧交通”和“智慧医疗”等等的美好梦想。这些描述让大家对大数据技术充满了憧憬和美好预期。

我从2014年的大数据应用中总结了两个重要的现象或者说应用趋势。第一个现象是大数据技术优先应用在采用SQL进行结构化数据处理上，来解决数据量增大带来的处理能力的挑战;这与很多人宣传的大数据技术最适合处理非结构化数据(而不适合结构化数据处理)相反。我们发现企业面临两方面的挑战，一方面是累积的数据量越来越大，从GB增长到了TB(拥有PB级的企业客户也有，但是少数)，另一方面，随着应用的增多和复杂化，计算能力越来越不能满足要求。大多数企业多年来根据业务需求在传统的关系数据库如DB2或者Oracle上开发自己的应用，数据量和应用的数量都在快速增加，传统数据库运行这些应用花的时间越来越长，即使只有1TB的数据，由于业务逻辑的复杂性，在传统关系数据库上运行统计业务，也从以前的日报(每日统计)降低到只能做现在的周报了。这样的时效性已经大大限制了企业的生产力。在IT系统日益成为企业业务本身的大趋势下，IT系统效率的低下严重影响了企业的竞争力。这些待处理的数据都是企业的结构化业务数据，现有的应用也都是基于SQL的。这是分布式的SQL on Hadoop技术发展的客观原因，也是星环科技致力于提升SQL运行性能和SQL支持完整程度的一个现实需求。

第二个现象或者需求是对实时时序数据的处理需求日益强烈，特别是随着传感器和监控设备等电子仪器的普及，企业有越来越多的实时数据。传统处理方法是将电子仪器产生的数据存入数据库后再统一分析。随着设备的增多和数据的增长，传统方案的延时越来越高。利用流处理技术在数据产生的时候就进行实时处理可以极大地提高企业的反应速度和工作效率。2014年星环科技部署了较多的流处理集群，来处理从用户产生的实时数据到传感器产生的数据。

这两个应用趋势我认为在2015年会变得更加强烈。下文简单总结了过去一年中在运营商、金融、物流、工商、交通、能源、广电和电商等多个领域中的大数据应用。

**电信运营商**

移动互联网时代的运营商面临着许多新挑战。微信等手机通讯APP的出现侵蚀了运营商的语音和短信收入，流量业务显得更加重要。另一方面，无线网络服务是运营商的核心竞争力。近年来，运营商正在投入大量资金建设网络来大力发展4G。4G网络的覆盖率不高或者质量不高导致的4G回落到3G或者2G会大大降低客户满意度。

经过最近一两年的探索，运营商在大数据平台建设方面总结了两个方向，一是利用大数据技术来提升运营效率，同时探索新的商业模式和数据运营方式。在过去的一年中，大数据在运营效率提升方面得到验证，而新的商业模式仍然在探索中。我们在广东移动的经营数据分析中应用星环的内存计算技术成功地将800多个指标的计算从原先Oracle的30小时减少到了4小时，在上海移动成功地将流量经营系统从DB2完整地迁移到了星环的TDH上，运行效率比原先的集群有5倍左右的提升。我们对SQL的完整支持使得应用系统迁移成为可能，而之前合作伙伴曾经尝试过将应用系统迁移到某知名Hadoop发行版上但没有成功。我们正在参与某省电信和某直辖市移动的4G网络优化项目，在这些项目中，我们的合作伙伴在用更高性能的星环TDH代替传统MPP数据库进行网络优化模型的建立和高速的模型运算，一方面发现网络中存在的问题，例如信号回落的问题，帮助运营商快速找出有问题的区域。另一方面通过TDH提供的完整SQL结合统计和机器学习算法，找到最佳的优化模型和参数，对网络进行细粒度的精确调整，以提高网络的覆盖度以及信号的质量。

**金融**

在2013年到2014年间，国有银行以及部分股份制银行或多或少地进行了大数据技术应用方面的探索，但是早期的应用局限在简单的历史交易查询以及非结构化数据的存储和检索上，并没有对银行的关键业务产生影响。而大数据技术在银行的应用前景被广泛传播，通过综合处理银行自有结构化交易数据以及外部互联网/政府数据，可以提升精细化客户管理水平以及进行大数据征信降低风险等。这些远景在2014年没有成为现实，2015年预计将是应用探索的一年。但我们在2014年在银行中实践了一些务实的应用。在这些应用中，TDH作为数据仓库的补充，用于提升数据分析的效率。同样受益于我们对SQL的完整支持，某股份制银行开始把一些复杂的贷款风险控制逻辑迁移到TDH Hadoop平台上进行运算。这些风控模型客户此前在多个MPP数据库和Hadoop发行版上进行过尝试，性能或者功能都没能满足他们的要求。从技术角度来看，这些分析涉及的数据量只有几个TB，但是分析业务极其复杂，涉及到近百张事实表和维度表，有些表的宽度甚至超过了几万字节。这个案例说明传统关系型数据库或者MPP数据库对于大数据场景下的复杂计算变得越来越捉襟见肘，银行需要一个更高效的数据处理工具。

**快递**

快递行业IT系统产生的数据量和承载的压力过去一直没有得到大家的关注。近年来，快递行业的规模随着电子商务的高速发展出现了快速的扩张。巨大的市场需求给快递公司带来了前所未有的挑战，每年的“双十一”会给快递公司的处理能力施加远高于平时的压力。因此，怎样缓解“双十一”的爆仓、避免快件变“慢件”是每个快递公司的难题。

如何通过大数据的分析对快递流程进行改善和优化成为一个值得研究的问题，也是快递业提高竞争力的一个重要手段。快递的每一个生产环节都会产生的大量数据，监控这些数据进而对全国各处理中心的收寄和运载能力、出班投递计划做实时优化调整，公司就能降低成本。分析这些数据来对业务发展的趋势做出预测，公司就能做好准备应对暴涨的需求。然而，快递生产环节中的数据具有数据量大、并发性高、类型复杂的特点，上层应用对实时性要求很高，传统数据库在这样的情况下捉襟见肘。

我们和华胜天成合作一起为中国邮政EMS速递部门部署了大数据平台，对它在全国的揽投部、处理中心和集散中心的数据(包括已接收、留存件、已下段、未下段、已投递、未投递、揽收员、地址、已封发、已发运、未发运等等)进行处理。大数据平台将ESB(企业生产总线)流来的数据实时动态加载进流处理集群以及实时数据库，进行实时统计和指标监测，并且实现实时数据查询。这次部署给了客户简单易用的工具来对业务的每个环节实时监控，使得他们在海量的快递业务中都可以快速精准地发现问题，如快件的积压、遗失、破损等，从而提高服务质量。这个大数据平台平稳支撑了2014年“双十一”的数据处理压力。未来该平台也可根据最新的生产数据帮助快递公司调整和优化投递计划，为公司降低成本。

**工商**

工商部门在建设国家的“经济户籍库”积累了大量的市场主体信息、年检情况、执法数据和12315投诉等数据。对这些数据的统计分析可以帮助工商部门理解市场与经济形势。

大数据技术的其中一个简单应用是用在数据质量管理和统计分析上。由于是人工录入数据，不可避免存在一定出错的概率，虽然概率不大。同时企业和个人的基本信息被分散在几十张关系表中，信息存在一定程度的交叉关联。通过对数据进行大规模交叉比对和统计，可以发现数据中隐藏的错误并及时得到更正。这个应用使用了星环的内存计算技术，全量数据的校验和统计可在十分钟内完成，极大提高了工作效率。

另外，大数据技术也用于市场主体信息的查询系统中，可以应对上亿用户并发查询并在几百毫秒内就返回查询或搜索结果。对企业历史快照的查询可以让用户跟踪企业变更信息，掌握企业生命周期的变化规律。在解决了存储和查询问题的基础上，我们还帮助客户利用图计算引擎快速发现企业之间和企业相关人员之间的关联。通过对全库数据进行扫描，确认这些企业基于股权、任职等方面的关联关系，建立企业关联关系信息库。

**电力**

随着电力企业信息化快速建设和智能电力系统的全面建成，电力数据的增长速度将远远超出电力企业的预期。从发电侧为例，电力生产自动化控制程度的提高，对诸如压力、流量和温度等指标的监测精度，频度和准确度更高，对海量数据采集处理提出了更高的要求。就用电侧而言，一次采集频度的提升就会带来数据体量的“指数级”变化。电力数据量的增长已经远远超过某电力部门原先使用的关系数据库的处理能力。

我们在2014年主要帮助电力部门处理用电侧的数据。我们意外地发现电力数据的统计分析涉及非常复杂的SQL运算，从技术角度来看，大量使用了Oracle的PL/SQL扩展语法，包括存储过程/控制流/异常处理/增删查改/事务处理等。从应用角度来看，这些SQL逻辑主要用于用电量的历史统计和用电趋势的分析，以及对线路损耗的计算。我们协助客户通过机器学习的方法进行分析，发现用电量跟宏观经济走势以及气候有一定的相关性，同时也跟每个行业以及每个企业的经营状况密切相关。通过对企业用电量的统计以及它所处行业的用电水平的对比，可以发现企业的节能情况，通过对用电历史数据的分析，可以发现企业生产活动的变化或者节能措施的效果。某南方供电局采用TDH的平台统计找出节能环保的企业和用电大户，并对节能环保的企业给予补贴，目的是对全社会节能减排观念进行引导，推动工业由高耗能的粗放发展方式向低耗能、高效率的绿色和谐发展方式转变。

我们还为某电力部门部署了一个试验性的故障处理系统，我们和合作伙伴建立了统一的配电网供电拓扑模型，利用图数据库存储从用户到变电站的整个供电拓扑网络数据，利用流处理系统进行实时告警，并实时查询电网拓扑图，快速研判停电事故发生的地点以及影响的范围。在此基础上，可以将停电事件通知抢修班组，及时恢复供电。同时可以主动告知用户，加强与用户互动，全面且直观的掌握全网的停电分布情况。

**交通**

随着经济迅猛发展，机动车辆不断增加，全国性的交通拥堵现象也越来越严重，如何通过信息化手段提高交通管理水平和保障道路安全已经成为一个重要的课题。

目前常用的方式是在道路卡口部署数字监控设备，这些设备7×24小时不间断捕获图像和视频数据，并进行识别，一个省或直辖市每日产生的过车数据有几千万条记录。这些数据主要用来为交通管理部门提供实时的路况信息，这些信息未来可以发布给公众作为出行的参考信息。同时协助管理部门进行交通管理，包括对重点营运车辆的监控，违法车辆的识别和布控，区间测速、套牌分析等实时性的分析应用。 我们和合作伙伴为某省公安厅交通管理部门部署了全省范围的交通监控系统，采用分布式队列实时采集全省各个交通卡口的车辆信息，使用流式计算集群对过车记录进行实时统计和监测，并实现上述多种实时分析应用，系统处理信息的端到端延时在2秒以内，较好地提高了交通管理的效率。

当然，交通行业的大数据应用还处于起步阶段，刚刚开始或者即将完成大数据的集中收集。利用大数据技术的强大分析和挖掘能力，未来可以显著提高交通信息的实时透明度，提升交通和拥堵管理的水平，降低事故的发生率，并为城市规划提供参考。

**广电**

在中国，广电系统正经历着数字化浪潮的冲击，基于网络化的影视播放给传统广电运营商很大挑战。在此背景下，华数传媒敏锐意识到，要想获得未来网络化传媒的生存与竞争优势，现在就必需向用户倾斜，打造“精准型”广电内容及传播运营商。华数传媒需要的数据基础架构需要能够满足海量、多来源、多样性数据的存储、管理要求，支持平台硬件的线性扩展，并提供快速实时的数据分析结果，迅速作用于业务。华数传媒选择了我们为其部署了大数据平台，在其之上开发了数字电视分析系统。该系统可以提供基于全量数据的实时榜单。以时间(小时/天/周)、用户等维度，对点播节目、直播节目、节目类别、搜索关键词等进行排名分析、同比环比分析、趋势分析等。系统还可以从时间、频道、影片类型、剧集等维度，根据在看数量、新增数量、结束观看数量、完整看完等分析用户走向。另外，通过对用户行为数据的采集分析，华数传媒可以对客户进行精准画像，使用智能推荐引擎，系统可以先于观众知道他们需求，预知将受到追捧的电视，为每一个用户量身定做推荐节目，以提高了产品的到达率，增强用户忠诚度。另外，系统还可通过观众对演员、情节、基调、类型等元数据的标签化，来了解受众偏好，从而进行分析观测，为后续的影视制作等内容开发做好准备。得益于基于大数据平台的数字电视分析系统，华数传媒正在进行从内容传输到内容制造的“华丽转身”。

**电子商务**

在电子商务领域，大数据可以说已经成为业务支撑的关键技术，在营销推广、客户关怀等众多环节发挥重要作用。我们和锦江电商合作，利用大数据平台为该电商打造了产品推荐系统。我们基于大数据平台建设了客户标签体系。依托该电商大量的会员和访客，深度学习和挖掘客户的行为数据，依据RFM模型和客户信息，形成客户消费喜好、客户年龄、家庭状况、甚至星座、属相、消费频次、金额、出行方式等等信息计入客户标签。再将客户标签聚类分析，形成客户分群。如此，便能精准获取客户群体，实施精准营销。同时，我们还协助客户建设了产品标签体系。依据酒店与旅游等各类型产品特征，建设和挖掘产品标签，并经过一定的机器学习挖掘过程，将客户标签和产品标签对接，根据各类标签分析权重，建设智能化推荐系统。

该推荐系统可以智能化推荐产品，正逐步成为针对电商的会员关怀体系和精准服务体系中重要的基础环节。

**总结和展望**

总结了一下2014年的Hadoop大数据行业应用，有些应用可能是大家之前没有预想到的简单应用，有些则是复杂的数据分析和挖掘类应用。大数据技术本身是一个全新的数据处理和分析技术，拥有超过现有技术的强大处理能力和深度挖掘数据的能力，然而技术本身带来的价值需要通过上层应用来展现，因此如何应用这些能力来解决现实的问题是各个行业都在探索的课题。在2015年预计会有大量的基于大数据技术的创新应用涌现出来。

同时在过去的一年中，大数据技术已经证明可以大幅提升运营效率，我们预计将来的一年内，利用SQL on Hadoop技术解决企业面临的数据量大难算的难题会成为一个普遍的应用趋势，随着SQL支持程度的不断完善和性能的不断提升，企业应用大数据技术进行结构化数据处理，以提升运营效率和解放生产力，将得到立竿见影的效果。

2014年是大数据技术开始落地的一年，我们看到了市场对大数据技术和产品的巨大需求。我们对2015年和以后的大数据发展都十分看好。大数据快速发展的趋势会持续很长一段时间，数据中还有太多价值没有被挖掘出来，会有越来越多的企业、政府机构和公益组织需要大数据解决方案。普及优秀的大数据产品帮助大众解决数据处理的难题，让我们共同努力!