基于 OLAP 技术的物联网客户数据平台构建

郭悦，龙岳，张勋，屈阳
（中国联合网络通信有限公司研究院，北京 100176）

摘　要：现阶段物联网客户业务飞速发展，行业中逐渐暴露出客户运营能力较弱的管理问题，这导致客户分析处理（AP, analytical processing）值普遍较低、盈利低、运营成本高，因此，急需增强物联网海量数据分析能力，为业务发展提供数据指引。以提升客户价值为目标，利用大数据技术、多维分析技术、海量查询技术，对物联网企业客户进行画像，实现运营服务精细化、高效化的客户运营平台构建，从技术实现方面探讨了如何进行数据运营平台的构建。

关键词：大数据；物联网客户运营；联机分析处理
中图分类号：TP39
文献标识码：A
doi: 10.11959/j.issn.2096−3750.2020.00141

Construction of Internet of things customer data platform based on OLAP technology

GUO Yue, LONG Yue, ZHANG Xun, QU Yang
China Unicom Research Institute, Beijing 100176, China

Abstract: At this stage, the Internet of things (IoT) customer business is developing rapidly. The management problems of weak customer operation ability have gradually exposed in the IoT industry, resulting in the low AP (analytical processing) value of customers, the low profitability and high operating cost. Therefore, it is urgent to enhance the massive data analysis capabilities of the IoT to provide data guidance for business development. To improve the customer value, big data technology, multi-dimensional analysis technology and massive query technology were used to make portraits of IoT enterprise customers, and realize the construction of the customer data platform with refined and efficient operation services. It was introduced how to construct the data operation platform from the aspect of technology implementation.

Key words: big data, IoT customer operation, on line analytical processing (OLAP)

1　引言

随着物联网产业技术日趋成熟，商业模式逐步形成，市场需求快速增长，低效的企业客户运营管理逐渐成为阻碍物联网市场发展的一道屏障。本文以物联网客户运营为场景，介绍如何将大数据分析技术、高性能海量数据查询技术、联机分析处理（OLAP, on line analytical processing）技术应用到物联网客户运营管理中。

在现阶段，物联网业务仍然以做大连接服务规模、推动应用和部件服务突破为手段，来实现做大做强物联网业务的目的，这将使物联网客户运营面临巨大的技术挑战[1]，主要体现在以下 3 点。

1) 现有的运营指标加工后无法直观地产生市场价值；运营系统需要对指标进行二次个性化加工或缺少二次加工；运营系统缺少可视化数据展示、用户画像及数据建模功能。

2) 现有的物联网大数据分析能力无法满足物联网业务快速发展的应用需求。

3) 相比传统移动设备，物联网设备具有覆盖密度大、数据实效性高、穿透性强等特征，传统经营分析系统不具备承载海量物联网数据的能力。

收稿日期：2019−06−25；修回日期：2020−02−03

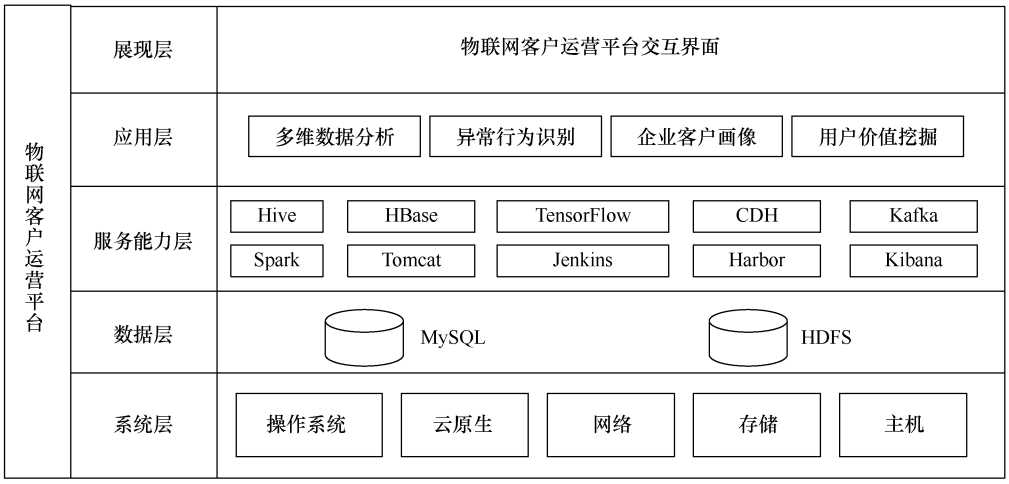


图 1　客户运营平台技术架构

2　物联网客户运营平台方案

物联网客户运营平台使用大数据技术代替业务部门的传统经营分析手段，建设高效、可靠的物联网客户数据运营支撑体系。平台采用 5 层架构，分别为展现层、应用层、服务能力层、数据层和系统层，客户运营平台技术架构如图 1 所示。

1) 展现层：面向物联网业务工作者的交互页面，通过多维度聚合可视化指标，分层级、分价值地展现企业客户。

2) 应用层：包括数据挖掘、数据查询、数据转码、数据清洗、数据归一化融合等物联网大数据分析能力。

3) 服务能力层：包括大数据基础能力、数据订阅能力、镜像管理能力、消息转发能力、负载均衡管理能力和日志检索能力。

4) 数据层：包括关系型数据库及分布式文件系统。

5) 系统层：包括硬件系统和软件系统。

2.1　通过多维数据分析引擎实现模型加速

物联网客户运营平台为适应物联网分析的实效性高、数据回传量级大、穿透性强等特征，结合主流 Hadoop 集群形成较成熟的技术架构，采用麒麟（Kylin）为多维数据分析引擎，采用 Presto 为高性能查询引擎，完成了从数据源的数据挖掘、数据查询、数据转码、数据清洗、数据归一化融合到自身的存储引擎等一系列工作，提供了 REST 数据接口服务，OLAP 技术架构如图 2 所示。

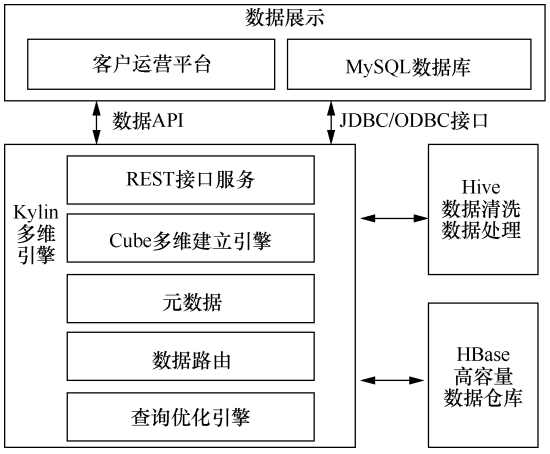


图 2　OLAP 技术架构

Presto 和 Kylin 两种 OLAP 引擎在互联网企业中都有广泛的应用，如在 Facebook、Comcast、京东、美团等公司中都有比较成熟的应用方案。现在将 Presto 与 Kylin 进行比较，分析这两种引擎在实际应用中的异同。在数据分析中，常用的操作是计数、求和与维度筛选，因此，挑选 3 个典型查询场景测试 Presto、Kylin1.5 和 Kylin2.0 的查询性能。性能测试用例如表 1 所示。

千万级测试结果如图 3 所示，可以看到在计数、求和两项查询中，3 种方法的响应时间差距较小。但在增加维度后，Kylin 预计算的优势得到充分体现，Kylin 在维度较多的情况下查询性能非常好。

亿级测试结果如图 4 所示，可以看到 Kylin 的表现更稳定。经过测试，对各引擎进行了全方位的对比。各引擎性能全方位对比如图 5 所示。

表 1　　**性能测试用例**

测试用例编号	GN-01		
测试用例名称	各个 OLAP 引擎性能对比		
测试目的	测试 Presto、Kylin1.5、Kylin2.0 在数据量分别为千万规模和亿规模下的查询性能		
场景描述	计数、求和与 10 个维度的过滤		
预设条件	Kylin 已经进行了预计算		
测试过程	序号	测试步骤	检查点
	1	在 Presto、Kylin1.5、Kylin2.0 上执行计数运算	记录各个引擎的运算时间
	2	在 Presto、Kylin1.5、Kylin2.0 上执行求和运算	记录各个引擎的运算时间
	3	在 where 后面增加 10 个维度条件	记录各个引擎的运算时间

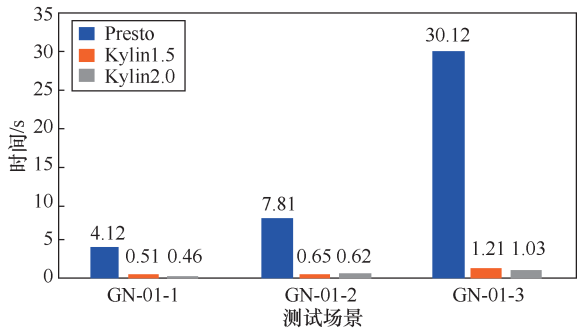


图 3　千万级测试结果

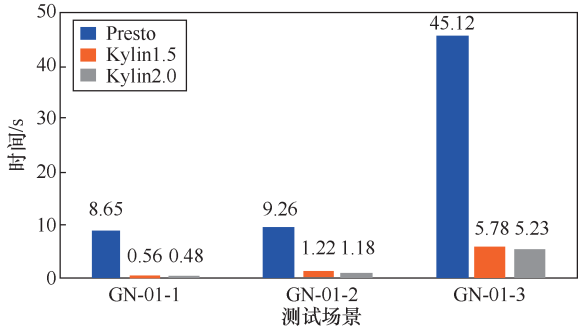


图 4　亿级测试结果

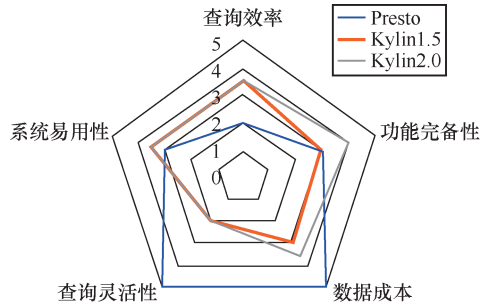


图 5　各引擎性能全方位对比

1) 查询效率：Kylin2.0 和 Kylin1.5 的查询效率差不多，Kyline2.0 稍好一些；Presto 的查询效率最差。

2) 系统易用性：各引擎区别不大，都支撑标准的结构化查询语言（SQL, structured query language），学习成本都不是很高。

3) 数据成本：因为 Presto 几乎不需要在业务系统或自身的管理系统中提前做特殊处理，Hive 能读的数据 Presto 都可以读，所以 Presto 的查询成本最有优势；而 Kylin 需要对数据进行预计算，当维度数量特别多的时候，处理更复杂。

4) 查询的灵活性：Presto 对 SQL 全面支持；Kylin 必须对数据进行预计算，如果 SQL 里包括没有预计算的维度，就会报错。

5) 功能完备性：Kylin2.0 版本增加了 Spark Cubing 功能，大幅度减少了预计算的时间，功能得到进一步加强。

Presto 与 Kylin 的优缺点对比非常明显，Presto 的优点包括：

1) 纯内存计算；

2) 支持高并发查询；

3) 数据存储在 HDFS（Hadoop distributed file system）上；

4) 支持 Hive（即基于 Hadoop 的一种数据仓库工具）的 Metadata（元数据）；

5) 计算节点之间互相无依赖。

Presto 的缺点包括：

1) 进行大数据量 join 计算时，会出现内存溢出的问题；

2) 由于每次查询都需要拉取数据查询，当用户反复发起相同条件的查询时，存在资源浪费的情况。

Kylin 的优点包括：

1) 通过预建 Cube（指数据立方体），提高常用数据查询的响应速度；

2) 支持高并发查询；

3) 支持数据存储在 HDFS 上；

4) 支持 Hive 的 Metadata。

Kylin 的缺点包括：

1) 设计和构建 Cube 过程比较复杂；

2) 维度数量不能过多。

多维数据分析引擎的多维计算能力主要体现在 OLAP Cube 的计算上。单个 Cube 可包含多个 Cuboid（指原始数据聚合的数据集），所以创建 Cube 等价于在导入原始数据时进行了预计算的过程。Kylin 最大的优势在于充分利用了 Hadoop 的 MapReduce 并行处理的能力，以此来实现高效处理导入的数据，具体处理流程如下。

首先根据 Cube 定义的事实表和维度表，利用 Hive 创建一张宽表，抽取事实表上维度的数量，将事实表上的维度以字典树方式压缩编码转化成目录，将维度表以字典树的方式编码。然后通过 Hadoop 的 MapReduce 把上一步中的宽表文件作为输入，创建 N-Dimension Cuboid，持续迭代，根据前一次的结果串行生成下一次的输入文件。最终形成 *n* 维结果集，为后续用户画像提供数据基础。同时，根据生成的 Cuboid 数据量计算 HTable（指面向 HBase 查询的视图）的 Region 分割策略，创建 HTable，将 HFile 导入多维引擎。

2.2　大数据企业客户画像分析

物联网客户运营平台对数据仓库中的数据进行多维度分析，将其拥有的大量客户数据和可视化数据工具连接起来，根据不同的客户业务场景，挖掘客户数据的深层价值，保证用户能够根据需要随时自主地分析不同客户特征，快速洞察客户需求[2-3]。

从数据源到企业客户画像展现分成如下 4 层[4]。

1) 数据源：包括来自集客系统、订单系统以及 Jasper 平台的数据源，数据内容包括用户使用行为数据、购买行为数据和客户基本信息数据 3 个方面。

2) 数据建模：构建用户画像评价体系，输出数据模型。

3) 数据处理：过滤无效数据，根据业务需求把数据从源端抽取、交叉分析、转码转存至数据仓库，按照数据集的方式将数据进行高效压缩，并根据评价体系聚类拆分客户类群标签。分析计算时，采用分布式计算完成高容量数据处理，最终将结果送往用户画像可视化分析层做展现。

4) 用户画像可视化分析：采用价值象限图、多维雷达图、用户标签体系及建议系统，根据特定用户的不同纬度数据指标的具体数值，自动对应用户标签，并给定特定的维系建议，用户画像特征如表 2 所示。

表 2　　用户画像特征

展示内容	数据指标	评价分数
运营建议评价体系	高活跃、高价值客户	9～10
	高价值客户	7～8
	普通客户	5～6
	待维系客户	3～4
	低价值客户	1～2
	僵尸客户	0
客户经营特征	连接数	—
	连接数上月增长	—
	活跃率上月增长	—
	上月出账金额	—
客户行为特征	激活率	—
	订购情况	—
	累计离网率	—
	API 使用次数	—
	活跃率	—
客户基本属性	客户名称	—
	主要行业	—
	重点产品	—
	细分产品	—
	客户编号	—
	系统内标识	—

2.3　高性能查询引擎实现海量行为数据融合

Kylin 作为多维数据分析引擎，在系统中承担物联网海量客户数据识别分析的工作，但物联网业务具有跨行业、跨平台的业务特性，客户行为数据具有分析维度不确定的问题。同时，数据引擎无法实时、高效地完成极端场景下超过 200 个数据维度

的复杂客户行为数据分析。为了分析极端场景下客户复杂行为数据，系统中应引入高性能查询引擎（Presto）来实现海量行为数据的维度融合，以提高多维引擎的分析效率。Presto 的任务需要放在每个工作节点上执行，每个任务执行完成后，数据会存放在内存中，而 MapReduce 则需要写入磁盘。当多个任务之间需要进行数据交换（如 shuffle 流程）时，直接在内存中处理数据。

高性能查询引擎流程如图 6 所示。首先，在原有的框架中增加了 SQL 分析记录模块，该模块负责记录客户运营分析的 SQL 和客户数据融合的 SQL，在 Presto 或 Kylin 里执行运算。通过 SQL 分析记录模块将 SQL 的统计结果记录在远程字典服务（Redis, remote dictionary server）中，Redis 负责管理并记录分析步骤。通过预计算，定期将融合成功的热数据快速响应反馈回 Kylin，有效降低了 Presto 反复拉取数据产生的资源消耗。

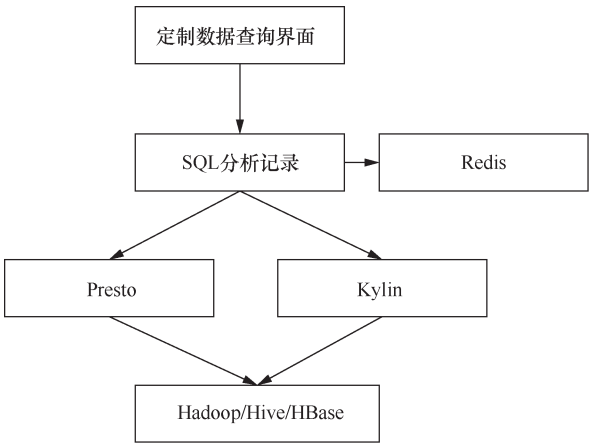


图 6　高性能查询引擎流程

Presto 将需要融合的客户行为按照类别切分数据维度集，并通过工作节点进行快速运算，融合结果以数据集的形式堆放在系统内存中，通过使用高性能查询引擎，将原有的从属行业、主要订购产品、短信播发数、使用数据量、日离网数和连接行为明细等 247 种客户行为维度，降低为 87 种有效行为维度，使得多维数据引擎的处理速度提升了 25%[5-6]。

3　结果验证

验证并比较物联网客户运营平台与传统经营分析工具的性能，验证初始时，客户接入量较小，经营分析工具和平台处理速度的差距不大，经营分析工具的处理速度优于平台，多维引擎未体现加载预先计算的优势。当同时接入的企业客户数量超过 30 000 个、物联网设备超过 1.2 亿部时，执行客户价值统计查询，平台依靠多维引擎和海量查询引擎的优异查询性能，处理性能提高了 87%，大幅度超过了经营分析工具的处理性能[7-8]。

硬件环境主要是指功能测试的服务器硬件环境，服务器硬件环境如表 3 所示。软件环境包含服务器软件环境和测试工具，服务器软件环境如表 4 所示[9]。测试项及测试结果如表 5 所示。

表 4　　服务器软件环境

软件名称	版本	与实际差异
操作系统	CentOS 7	一致
数据库系统	MySQL 5.7	一致
查询工具	Hive 3.1.0	一致
Hadoop 集群	Cloudera CDH V5.7	一致
数据仓库	HBase V1.4.10	一致
Kylin	V2.6	一致
Presto	V0.225	一致

表 3　　服务器硬件环境

名称	配置	与实际差异	备注
机架式服务器	E5-2620V4 8 Core×2 128 G(8×16 G、4×32 G) 系统盘：240 G s3510 以上 SSD 硬盘：4 T(6×12) 网卡：10 GE×4 GE×2 GE×1(IPMI) RAID 卡	模拟	实验室环境性能弱于实际环境
负载均衡器	Nginx	模拟	
光纤交换机	—	无法模拟	

表 5　　**测试项及测试结果**

测试项	测试语句	测试结果	耗时
数据量全分区查询	select count(*) from dm_custom_2018	总记录条数：16 718 581 条	1 min 7 s
全量数据多维度查询	SELECT BUS_ACC_YEAR,AREA,PROVINCE FROM dm_custom_2018 a WHERE STR_TO_DATE(BUS_ACC_YEAR, '%Y-%m-%d')<=STR_TO_DATE('20180801', '%Y%m%d') AND STR_TO_DATE(BUS_ACC_YEAR , '%Y-%m-%d')>= (SELECT DATE_ SUB (STR_TO_DATE ('20180831', '%Y%m%d') GROUP BY a.BUS_ACC_YEAR,AREA,PROVINCE	总记录条数：122 536 条	测试 3 次：1.845 s 1.731 s 1.703 s
全量数据多维度聚合查询	SELECT BUS_ACC_YEAR, AREA, PROVINCE, SUM (a.CONNECTIONS) AS VALUE FROM dm_custom_2018 a WHERE STR_TO_DATE(BUS_ACC_YEAR, '%Y-%m-%d')<= STR_TO_DATE ('20180801', '%Y%m%d') AND STR_TO_DATE(BUS_ACC_YEAR , '%Y-%m-%d')>= (SELECT DATE_SUB (STR_TO_ DATE('20180831', '%Y%m%d') GROUP BY a.BUS_ACC_YEAR, AREA,PROVINCE	总记录条数：122 536 条	测试 3 次：2.166 s 1.932 s 2.021 s
分区数据多维度聚合查询	SELECT a.'拍照月融合类型', a.sumUserG, a.sumUserP,(a.sumUserG/ a.sumUserP)as rateY from (SELECT ls.'拍照月融合类型', sum(ls.`拍照月用户数 1`) as sumUserG, sum (ls.'拍照月用户数') as sumUserP from 'dm_201801` ls where ls.'省份'='北京' and ls. '拍照月融合类型'!='2I'GROUP BY ls. '拍照月融合类型') as a SELECT a. '拍照月融合类型', a.sumUserG, a.sumUserP,(a.sumUserG/ a.sumUserP)as rateY from (SELECT ls.'拍照月融合类型',sum(ls.'拍照月用户数 1') as sumUserG ,sum(ls.'拍照月用户数') as sumUserP from 'dm_201801' ls where ls.'省份'='北京' and ls.'入网渠道' ='集团渠道' GROUP BY ls.'拍照月融合类型') as a;-	总记录条数：1 167 533 条	测试 3 次：3.776 s 2.932 s 3.221 s

4　结束语

物联网客户运营分析系统旨在通过人工智能、多维数据引擎、大数据挖掘等创新技术的应用，解决实际应用中物联网企业客户运营面临的难题，并实现以下两点创新。

1) 应用高速数据处理能力，形成客户画像运营体系。针对物联网客户运营，可利用目前业界流行的开源大数据和高性能查询引擎技术进行研发，使用多维数据分析引擎实现数据的收集、存储、加工、融合的全流程管理，大幅度提高了数据开发效率，降低了研发人工成本。同时，随时自主地分析不同用户特征形成画像，快速洞察用户需求，提出了运营建议。

2) 利用高性能查询引擎实现了海量行为数据的有效性融合，解决了多维度引擎应用到物联网客户行为数据分析时，因维度过多导致膨胀率过大的效率问题，进一步提高了多维分析的快速处理能力[10]。

参考文献：

[1] 丁飞. 物联网开放平台[M]. 北京: 电子工业出版社, 2018.
DING F. Internet of things open platform[M]. Beijing: Publishing House of Electronics Industry, 2018.

[2] SANDY R, URI L, SEAN O, 等. Spark 高级数据分析[M]. 北京: 人民邮电出版社, 2017.
SANDY R, URI L, SEAN O, et al. Advanced analysis with Spark[M]. Beijing: Posts and Telecom Press, 2017.

[3] ISMAIL B I, GOORTANI E M, KARIM M B A, et al. Evaluation of docker as edge computing platform[C]//IEEE Conference on Open Systems. IEEE, 2015: 130-135.

[4] TAKASE W, NAKAMURA T, WATASE Y, et al. A solution for secure use of Kibana and Elasticsearch in multi-user environment[J]. arXiv: 1706.10040, 2017.

[5] 苏敏坚. 大数据是运营商复兴的“起搏器”[J]. 中国电信业, 2017(4): 40-41.
SU M J. Big data is the pacemaker of operators' renaissance[J]. China Telecommunication Trade, 2017(4): 40-41.

[6] 范铭. 运营商进入调速换挡关键时期大数据创造新机遇[J]. 通信世界, 2017(16): 54.
FAN M. Operators enter the key period of speed shift and big data create new opportunities[J]. Communications World, 2017(16): 54.

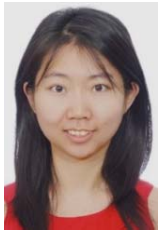
[7] JD-Presto 研发团队. Presto 技术内幕[M]. 北京: 电子工业出版社, 2016.
JD-Presto Core Team. Presto technology insider[M]. Beijing: Publishing House of Electronics Industry, 2016.

[8] 蒋守壮. 基于 Apache Kylin 构建大数据分析平台[M]. 北京: 清华大学出版社, 2017.
JIANG S Z. Building a big data analytics platform based on Apache Kylin[M]. Beijing: Tsinghua University Press, 2017.

[9] Apache Kylin 核心团队. Apache Kylin 权威指南[M]. 北京: 机械工业出版社, 2007.
Apache Kylin Core Team. Apache Kylin authoritative guide[M]. Beijing: China Machine Press, 2007.

[10] 王建民. 工业大数据技术综述[J]. 大数据, 2017, 3(6): 3-14.
WANG J M. Survey on industrial big data[J]. Big Data Research, 2017, 3(6): 3-14.

[作者简介]



郭悦（1986− ），女，辽宁沈阳人，中国联合网络通信有限公司研究院大数据研究中心软件开发工程师，主要研究方向为开源技术和前端开发等。

龙岳（1987− ），女，湖南株洲人，中国联合网络通信有限公司研究院大数据研究中心软件开发工程师，主要研究方向为大数据框架新技术和开发等。

张勋（1990− ），男，北京人，中国联合网络通信有限公司研究院大数据研究中心软件开发工程师，主要研究方向为开源技术和容器技术等。

屈阳（1985− ），女，河北石家庄人，中国联合网络通信有限公司研究院工程师，主要研究方向为大数据相关技术和开发等。