HZ BOOKS
华章科技
大数据技术丛书

Kyligence

Apache Kylin 权威指南

Apache Kylin核心团队◎著

Apache Kylin是首个中国人贡献的
Apache顶级开源项目

机械工业出版社
China Machine Press

大数据技术丛书

Apache Kylin
权威指南

Apache Kylin核心团队◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Apache Kylin 权威指南 /Apache Kylin 核心团队著 . —北京：机械工业出版社，2017.1
（大数据技术丛书）

ISBN 978-7-111-55701-2

I. A… II. A… III. 互联网络 – 网络服务器 IV. TP368.5

中国版本图书馆 CIP 数据核字（2016）第 305395 号

Apache Kylin 权威指南

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：张梦玲 责任校对：董纪丽
印 刷：北京诚信伟业印刷有限公司 版 次：2017 年 1 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：12.75
书 号：ISBN 978-7-111-55701-2 定 价：49.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

华章科技
HZBOOKS | Science & Technology

Forword 推荐序一

2016 年早些时候，我曾经写过一篇有关联通 Hadoop 的文章，在其中的“展望篇”里谈到过 OLAP on Hadoop 的新技术 Apache Kylin。今天《 Apache Kylin 权威指南》一书即将出版，我也有幸受本书作者之一韩卿（Luke）的邀请来写推荐序。

联通集团的 BI 是 2010 年建设的，由于全国有 4 亿用户的明细数据需要集中处理，再加上对移动互联网用户流量日志的采集，使得数据量急增。截至 2013 年已达 PB 级规模，并仍以指数级速度增长，传统数据仓库不堪重负，数据的存储和批量处理成了瓶颈。另一方面 BI 上提供的面向用户的数据查询和多维分析服务，使得后台生产的 Cube 越来越多，几年下来已有七八千个。用户需求对某一维度的改变往往会造成一个新 Cube 的产生，耗费资源不说，也为管理带来了极大的不便。2013 年年底我们在传统数据仓库之外搭建了第一个 Hadoop 平台，节点数也从最初的几十个发展到了今天的 3500 个，大大提高了系统的存储及计算能力，为联通大数据对内对外的发展都起到了至关重要的作用。美中不足的是分布式存储和并行计算只解决了系统的性能问题，尽管我们也部署了像 Hive、Impala 这样的 SQL on Hadoop 技术，但在 Hadoop 体系上的多维联机分析（OLAP）却始终得不到满意的结果。Oracle + Hadoop 的混搭架构还因为有对 OLAP 的需求而继续维持着，零散的 Cube 数还在继续增长，架构师们还在继续寻找奇迹方案的出现。

Apache Kylin 就是在这种大背景下出现在我们的视野中的。一个好的产品首先要有一个清晰的定位，要有一套能够明确解决行业痛点的方案。Kylin 在这点上做得非常好，它把自己定义为 Hadoop 大数据平台上的一个开源 OLAP 引擎。三个关键词：Hadoop、开源、OLAP，使它的定位一目了然，不用过多地解释。同时，Kylin 也是透明的，不像许多产品把自己使用的技术搞得很神秘，Kylin 沿用了原来数据仓库技术中的 Cube 概念，把无限数据按有限的维度进行“预处理”，然后将结果（Cube）加载到 HBase 里，供用户查询使用，使得现有的分析师和业务人员能够快速理解和掌握。相比于 IOE 时代的 BI，它非常巧妙地使用

了 Hadoop 的分布式存储与并行计算能力，用横向可扩展的硬件资源来换取计算性能的极大提高。

为了能够将 Kylin 真正融入到联通的大数据架构中，我们正在紧锣密鼓地组织系统测试。比如对单用户级的数据查询、第三方可视化工具的集成、多维 Cube 建立的维度数极限等的测试。我们还计划用 Kafka 来导入数据，用 Spark 来加工 Cube，用其他产品来代替 HBase 进而提高数据读取性能，用 Kylin 的路由选择来桥接新老 Cube，等等。这时出版的《 Apache Kylin 权威指南》一书，对于我们来说无疑是雪中之炭，我们的许多疑惑都会在这本指南当中找到权威解答。

联通公司现在经历的这些过程很多企业都会遇到，“坑”我们愿意去填，路希望大家来走。在向读者推荐《 Apache Kylin 权威指南》一书的同时，我们真诚期望 Kylin（作为 Apache 开源社区第一个由中国人开发并主导的产品）能够成功，能够在不断的实践中提高自己，能够充分利用中国这个占世界数据量 20% 的大市场，把自己打造成大数据领域的一只独角兽。

范济安
国家千人计划专家
中国联通集团信息化部 CTO

Forword 推荐序二

我是一个开源软件的爱好者，算是开源届的一名老兵。从 1995 年到美国留学起，就开始接触开源软件，当时的 GNU、Linux、FreeBSD 和 Emacs 等自由软件让刚出国门的我感到惊艳万分。从那时开始，我就再没有和自由软件、开源软件分开过：从读博士期间一直参与研发自由软件 XSB、因个人爱好参与贡献 GNU Emacs、在 IBM 工作期间基于一系列开源软件为团队开发 DocBook 文档写作工具链，到后来在 LinkedIn 工作期间研究作为 5 个核心成员开源的分布式实时搜索系统 SenseiDB，再到近几年在小米大力推动开源战略，打造基于开源软件的小米云计算、大数据和机器学习技术及团队。20 多年来，对开源软件的热爱，让我逐渐从一名早期的自由软件爱好者、信仰者、贡献者和管理者，变成了一名坚定的开源软件倡导者。在这期间，我见证了开源技术的萌芽、兴起和今天的繁荣，也经历了国内外不同文化下的开源发展历程。

作为一名参与开源软件较早的中国人，我也深深地感受到了最初西方世界对中国人使用开源技术、参与开源软件开发的质疑和冷落。因为互联网和自由软件进入我国较晚，也因为中国人在英语上的不足和东西方文化的差异，还因为早期国内的一些开源爱好者对开源软件的理解不足，使得在开源方面较为领先的西方开源人士对国人在开源上的使用和贡献存在极大偏见。中国开源力量融入国际开源社区的过程是缓慢和艰苦的，幸运的是，近四五年来，随着 GitHub 的兴起和多个开源社区的迅猛发展，中国每年产生的计算机人才也多了起来，中国越来越多的互联网公司开始正确地拥抱开源，中国工程师在国际开源社区的贡献和影响力也越来越大（比如，作为一个很年轻的创业公司，小米就在不到一年半的时间里推出了 3 个 HBase committer），这确实不是一件容易的事。但是，今天不管是在云计算、大数据，还是容器等诸多开源技术领域，真正由中国人自己主导、从零开始、自主研发、最后贡献到国际开源社区并成为顶级开源项目的，应该就只有 Apache Kylin 一个。Apache Kylin 是 2013 年由 eBay 在上海的一个中国工程师团队发起的、基于 Hadoop 大数据平台的开源 OLAP 引

擎，它利用空间换时间的方法，把很多分钟级别乃至小时级别的大数据查询速度一下子提升到了亚秒级别，极大地提高了数据分析的效率，填补了业界在这方面的空白。

我非常高兴能够看到一个来自国内的团队开源一个项目，并在短短不到一年的时间里顺利使其毕业，也使其成为 Apache 软件基金会的顶级项目，取得了可以和 Hadoop、Spark 等重大开源软件相提并论的成就。一支来自国内的工程师团队能够快速融入国际开源社区，被全球最大的开源软件基金会接纳并成功占领一席之地，这是一件非常不容易的事情，足以让国人欣慰和骄傲。这一切都和 Apache Kylin 项目背后的负责人韩卿（Luke）密不可分。我是在 QCon 北京 2014 全球软件开发大会上认识韩卿的，并由此第一次知道了 Kylin 这个项目，和韩卿开始交谈不久，我就觉得他是当时国内为数不多的、真正懂得开源软件打法的一个人。那次的交谈非常愉快，从此我也开始关注这个项目并极度看好它。

开源项目，并不是将代码公开就完事了，团队需要做更多艰苦的工作来不断推广技术、经营社区和营销品牌，使得项目能够被广泛接纳和使用。韩卿及 Kylin 团队在这方面做得非常出色，在各种国内外的技术大会上、很多开源社区里都可以看到他们忙碌的身影。在短短的两年时间里，我就看到 Kylin 项目从 Apache 孵化器项目毕业成为顶级项目，也看到这个团队离开 eBay 并创立了 Kyligence 这家创业公司。今天，很多成功的重大开源项目背后都有一两个伟大的创业公司：Hadoop 背后是 Cloudera 和 Hortonworks、Spark 后面是 Databricks，等等。我也看好 Apache Kylin 后面的 Kyligence！

小米不仅仅是一家手机公司，更是一个大数据公司，公司内部的很多产品和业务都深度依赖大数据分析，我们所面对的数据量、挑战和困难都是空前的。Apache Kylin 独特的数据查询性能优势在小米中有很多应用场景，我希望将来我们能够更多地用到 Apache Kylin 技术，也希望和 Kyligence 能有深度的技术合作。

今年，深度学习和大数据引发了人工智能的热潮，人工智能的热潮反过来也会推动大数据领域相关技术的发展和演进，大数据领域必将诞生更多的新技术和新产品。相信在不久的未来，会有更多的、类似于 Apache Kylin 的、由中国人主导的项目从实际需求中产生、开源并被贡献到国际开源社区，向世界输出我们的技术实力。在将本书推荐给读者的同时，我也希望更多的读者、团队和公司能一起参与、贡献和拥抱开源，努力提高我国技术人员在国际开源社区的影响力。Apache Kylin 项目相关的经验也非常值得其他技术人员学习和借鉴！

崔宝秋

小米首席架构师

小米云平台负责人

Forword 推荐序三

在大数据处理技术领域，用户最普遍的诉求就是希望以很简易的方式从大数据平台上快速获取查询结果，同时也希望传统的商务智能工具能够直接和大数据平台连接起来，以便使用这些工具做数据分析。目前已经出现了很多优秀的 SQL on Hadoop 引擎，包括 Hive、Impala 及 SparkSQL 等，这些技术的出现和应用极大地降低了用户使用 Hadoop 平台的难度。为了进一步满足“在高并发、大数据量的情况下，使用标准 SQL 查询聚合结果集能够达到毫秒级”这一应用场景，Apache Kylin 应运而生，在 eBay 孵化并最终贡献给开源社区。Apache Kylin 是一种分布式分析引擎，提供 Hadoop 之上的标准 SQL 查询接口及多维分析（OLAP）功能。

Apache Kylin 通过空间换时间的方式，实现在亚秒级别延迟的情况下，对 Hadoop 上的大规模数据集进行交互式查询；Kylin 通过预计算，把计算结果集保存在 HBase 中，原有的基于行的关系模型被转换成基于键值对的列式存储；通过维度组合作为 HBase 的 Rowkey，在查询访问时不再需要昂贵的表扫描，这为高速高并发分析带来了可能；Kylin 提供了标准 SQL 查询接口，支持大多数的 SQL 函数，同时也支持 ODBC/JDBC 的方式和主流的 BI 产品无缝集成。

同时，Apache Kylin 是目前国内少有的几个通过了 Cloudera 公司产品工程认证的大数据分析和查询引擎。Cloudera 公司相信，作为唯一一个来自中国的 Apache 顶级开源项目，Apache Kylin 不仅仅代表了中国对国际开源社区的参与，同时也将为我国及全球企业用户探索大数据的价值的进程做出卓越的贡献。

在过去的一年中，我们有机会与 Kyligence 公司合作，共同为国内的企业客户提供基于 Cloudera Hadoop 平台上的大数据应用。本书的出版为开发人员和数据分析人员利用这一技术提供了极大的便利。更重要的是，这本书不仅能够指导开发人员安装和使用 Apache Kylin，而且还深入探讨了 Apache Kylin 的核心技术架构，并且通过丰富的案例展示了如何通过优化

来提升大数据的应用性能。本书的作者之一韩卿先生是 Apache Kylin 的主要创建者和项目委员会主席（PMC chair），对于 Kylin 的技术架构、应用及未来发展都有深刻的理解。我相信本书对于 Kylin 使用者和开发者来说，是及时的且不可或缺的。

凌琦

Cloudera 全球副总裁兼大中华区总经理

Foreword 推荐序四

大数据在近几年已经成为一个火爆的名词，而企业针对数据的分析也从未停止过。从早些年传统企业的数据仓库、BI，到近些年互联网公司的广告推荐、产品分析，再到现在基于IoT硬件的线下用户行为画像，无论是互联网企业还是传统企业，一直都在尝试通过数据帮助企业或企业的用户提升工作效率和体验。从过去的决策支持，到现在普及的精准推荐，乃至未来的基于实时分析的AI交互，大数据及相关技术将一直是这些业务发展的基石，因而在最近的10年，大数据技术有了日新月异的发展。

从海量数据的批量计算到实时分析，从精准推荐到OLAP查询，业界涌现了大量优秀的开源项目。Apache Kylin就是其中一颗由国人研发的璀璨的明星，是国内第一个Apache顶级开源项目（与Kafka、Spark齐名），它解决了海量数据下OLAP查询的关键技术。大数据本身并不能产生价值，针对数据的分析和运用才可以产生价值，而OLAP是企业对数据做深度分析必用的组件。在过去，它能帮助企业从不同维度汇总、下钻看到企业不同部门、地区的差异及发展趋势；现在，它能帮助企业针对不同用户画像的人群做相关行为分析、排行，也可以针对不同的点击事件深入分析不同渠道的转化率、客单价。OLAP技术曾经在百亿数据集、PB级别规模的时候，遇到了很大的瓶颈，无论是并行计算还是近似计算，都对I/O、CPU和查询时长带来了挑战。Kylin运用它独有的技术，在数据存储不产生指数级增长的情况下，采用预计算技术以空间换回了时间，在百亿甚至万亿级别数据集上实现了毫秒级的查询响应速度。同时也利用了模糊计算等技术在允许一定误差的情况下，对10亿级别用户、几千种用户行为标签的数据实现了用户行为的即时查询，帮助企业极大地降低了大数据OLAP实施的门槛，也降低了大数据平台实施的TCO，是企业建设大数据平台的优质OLAP引擎。本书可以帮助企业的技术管理者、开发者详细了解Kylin并将应用部署到自己的企业当中，规避其中的实施风险、提高部署与处理效率。

数据是一种新的能源，它与石油、电力不同，产生于企业和用户的行为，能通过不断地

深入使用和反复分析利用来帮助企业增收、节支、提效、避险，其中各个环节都要有适用的工具，Apache Kylin 就是其中之一。大数据从过去的批量数据处理发展到现在的实时数据分析，我非常高兴地看到 Kylin 也支持了相关特性，让数据不止是用于实时计算，还可以帮助管理者看到实时的联机分析处理结果。当然，数据的实时 OLAP 只是实时分析中的一种，要结合数据实时采集、数据实时计算、数据流挖掘、实时场景引擎等技术，才可以让企业从 T+1 的分析发展到实时数据分析，进而实现实时决策与反馈，最终实现企业自身的智能分析与交互。数据的实时分析是每个企业实现 AI 的必经之路，而数据实时分析的应用又离不开 Kylin 这样的 OLAP 引擎。

最后，很荣幸可以为本书写推荐序，本书作者之一韩卿（Luke）也是我多年的好友，从他在 eBay 之时我们就有很多交流，我也有幸看着 Apache Kylin 项目逐步成为国际著名的开源项目。大数据的发展不是一个项目或一个企业就可以独立推动的，也希望更多的人才和企业加入大数据分析的行业中来，使得我国能够涌现出更多像 Apache Kylin 一样的优秀项目，让数据成为每一个企业的再生能源！

郭炜

易观 CTO

Preface 前　言

“麒麟出没，必有祥瑞。”

——中国古谚语

“于我而言，与 Apache Kylin 团队一起合作使 Kylin 通过孵化成为顶级项目是非常激动人心的，诚然，Kylin 在技术方面非常振奋人心，但同样令人兴奋的是 Kylin 代表了亚洲国家，特别是中国，在开源社区中越来越高的参与度。”

——Ted Dunning Apache 孵化项目副总裁，MapR 首席应用架构师

今天，随着移动互联网、物联网、AI 等技术的快速兴起，数据成为了所有这些技术背后最重要，也是最有价值的“资产”。如何从数据中获得有价值的信息？这个问题驱动了相关技术的发展，从最初的基于文件的检索、分析程序，到数据仓库理念的诞生，再到基于数据库的商业智能分析。而现在，这一问题已经变成了如何从海量的超大规模数据中快速获取有价值的信息，新的时代、新的挑战、新的技术必然应运而生。

在数据分析领域，大部分的技术都诞生在国外，特别是美国，从最初的数据库，到以 Hadoop 为首的大数据技术，再到今天各种 DL（Deep Learning）、AI，等等。但我国拥有着世界上独一无二的“大”数据，最多的人口、最多的移动设备、最活跃的应用市场、最复杂的网络环境等，应对这些挑战，我们需要有自己的核心技术，特别是在基础领域的突破和研发方面。今天，以 Apache Kylin 为首的各种来自中国的先进技术不断涌现，甚至在很多方面都大大超越了国外的其他技术，这一点也彰显了中国的技术实力。

自 Hadoop 选取大象伊始，上百个项目，以动物居之者为多，而其中唯有 Apache Kylin（麒麟）来自中国，在众多项目中分外突出。在全球最大的开源基金会——Apache 软件基金会（Apache Software Foundation，ASF）的 160 多个顶级项目中，Apache Kylin 是唯一一个来自中国的顶级开源项目，与 Apache Hadoop、Apache Spark、Apache Kafka、Apache Tomcat、

Apache Struts、Apache Maven 等顶级项目一起以 The Apache Way 构建了开源大数据领域的国际社区，并拓展了生态系统。

大数据与传统技术最大的区别就在于数据的体量对查询带来的巨大挑战。从最早使用大数据技术来做批量处理，到现在越来越多地需要大数据平台也能够如传统数据仓库技术一样支持交互式分析。随着数据量的不断膨胀，数据平民化的不断推进，低延迟、高并发地在 Hadoop 之上提供标准 SQL 查询的能力成为必须要攻破的技术难题。而 Apache Kylin 的诞生正是基于这个背景，并成功地完成了很多人认为不可能实现的突破。Apache Kylin 最初诞生于 eBay 中国研发中心（坐落于上海浦东新区的德国中心），在 2013 年 9 月底，eBay 中国研发中心的技术人员开始对此进行 POC 并组建团队，经过一年的艰苦开发和测试，于 2014 年 9 月 30 日使其正式上线，并在第二天（2014 年 10 月 1 日）正式开源。

在这个过程中，使用何种技术，如何进行架构，如何突破那些看似无法完成的挑战，整个开发团队和用户一起经历了一段艰难的历程。今天呈现出的 Apache Kylin 已经经历了上千亿乃至上万亿规模数据量的分析请求，以及上百家公司的实际生产环境的检验，成为各个公司大数据分析平台不可替代的重要部分。本书将从 Apache Kylin 的架构和设计、各个模块的使用、与第三方的整合、二次开发及开源实践等方面进行讲解，为各位读者呈现最核心的设计理念和哲学、算法和技术等。

Apache Kylin 社区的发展不易，自 2014 年 10 月开源到今天已有两年，从最初的几个人发展到今天的几十个贡献者，国内外上百家公司在正式使用，连续两年获得 InfoWorld Bossie Awards 最佳开源大数据工具奖。来自核心团队、贡献者、用户、导师、基金会等的帮助和无私的奉献铸就了这个活跃的社区，也使得 Apache Kylin 得以在越来越多的场景下发挥作用。现在，由 Apache Kylin 核心团队撰写了本书，相信能更好地将相关的理论、设计、技术、架构等展现给各位朋友，希望能够让更多的朋友更加充分地理解 Kylin 的优点和使用的场景，更多地挖掘出 Kylin 的潜力。同时也希望本书能够鼓励并吸引更多的人参与 Kylin 项目和开源项目，影响更多人贡献更多的项目和技术到开源世界来。

韩卿

Apache Kylin 联合创建者及项目委员会主席

2016 年 10 月

Contents 目　录