

Apache Kylin项目管理委员会主席 韩卿
大数据技术专家、智慧城市建设架构师 杨正洪
万达网络科技集团大数据中心副总经理 龚少成
大数据架构师、大数据技术实践者 贾传青



—— Extreme OLAP Engine for Big Data ——

基于 Apache Kylin 构建大数据分析平台

蒋守壮 著



清华大学出版社



基于 Apache Kylin 构建大数据分析平台

蒋守壮 著

清华大学出版社
北京

内 容 简 介

Apache Kylin 是一个开源的分布式分析引擎, 提供 Hadoop 之上的 SQL 查询接口及多维分析 (OLAP) 能力以支持超大规模数据, 最初由 eBay 公司开发并贡献至开源社区。它能在亚秒内查询巨大的 Hive 表。

本书分为 21 章, 详细讲解 Apache Kylin 概念、安装、配置、部署, 让读者对 Apache Kylin 构建大数据分析平台有一个感性认识。同时, 本书从应用角度, 结合 Dome 和实例介绍了用于多维分析的 Cube 算法的创建、配置与优化。最后还介绍了 Kyligence 公司发布 KAP 大数据分析平台, 对读者有极大的参考价值。

本书适合大数据技术初学者、大数据分析人员、大数据架构师等, 也适用于高等院校和培训学校相关专业师生教学参考。

本书封面贴有清华大学出版社防伪标签, 无标签者不得销售

版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目 (CIP) 数据

基于 Apache Kylin 构建大数据分析平台 / 蒋守壮著. — 北京: 清华大学出版社, 2017
ISBN 978-7-302-45452-6

I. ①基… II. ①蒋… III. ①互联网络—网络服务器②数据处理 IV. ①TP368.5②TP274

中国版本图书馆 CIP 数据核字 (2016) 第 274718 号

责任编辑: 夏毓彦

封面设计: 王 翔

责任校对: 闫秀华

责任印制: 刘海龙

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-62770175

邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 刷 者: 清华大学印刷厂

装 订 者: 三河市漂源装订厂

经 销: 全国新华书店

开 本: 190mm×260mm

印 张: 17

字 数: 435 千字

版 次: 2017 年 1 月第 1 版

印 次: 2017 年 1 月第 1 次印刷

印 数: 1~3500

定 价: 69.00 元

产品编号: 072388-01

推荐序

Apache Kylin 将传统的数据仓库及商务智能分析能力带入到了大数据时代，作为新兴的技术已经被广大用户所使用。作为创始作者，我非常欣喜能看到关于 Apache Kylin 相关书籍的出版，这无疑对中国用户更好地使用 Kylin，解决实际的大数据分析架构及业务问题有很大帮助。

韩卿

Kyligence 联合创始人兼 CEO，Apache Kylin 项目管理委员会主席（PMC Chair）

伴随着大数据发展的三条主线是大数据技术、大数据思维和大数据实践。

因为 RDBMS 很难处理单表 10 亿行数据，所以大数据技术应运而生。大数据技术从最初的解决海量数据的快速存储和读取，到今天的海量数据的 OLAP，当中衍生出众多的技术产品，Apache Kylin 就是其中的一个优秀产品，目标是解决大数据范畴中的 OLAP。

第二条主线是大数据思维。数据处理的最近几十年都被 RDBMS 的思想所束缚，小表、多表、表的连接、过分注重冗余性的坏处，等等，这些都限制了海量数据上的处理与分析。大数据技术出来之后，随之而来的大数据思路，给我们带来了海量数据处理的新思维。这个新思维的核心就是突破表的概念，而采用面向对象的数据模型在数据层上实现。Apache Kylin 的 Cube 模型就是在逐步体现大数据的思维。

最后一条主线是大数据实践。大数据实践分为数据梳理、数据建模、数据采集、数据管控、数据服务、数据可视化和数据分析。这是一环套一环的步骤，不能跳过。Apache Kylin 作为数据分析环节的技术产品，一定要同数据管理的优秀产品相结合，才能充分发挥出分析的功效。

蒋守壮是业界知名的 Apache Kylin 专家。《基于 Apache Kylin 构建大数据分析平台》一书浅显易懂、实操性强，是目前 Apache Kylin 界不可多得的技术资料，值得细读和研究。

杨正洪

武汉市云升科技发展有限公司董事长

Apache Kylin 是一个由国人主导开发并在大数据领域真正进入全球主流应用的开源项目。作为国人软件开发的骄傲之作，市面上却缺少一本系统性介绍该项目的书籍。

万达科技集团大数据中心蒋守壮同学在项目诞生之初就一直跟踪 Kylin 的进展，深入研究项目的技术原理，并将其运用在许多实际项目中。无论您是大数据技术爱好者，抑或您正在考虑引入 Kylin 这样杰出的大数据处理工具，该书都将是您很好的参考指南！

龚少成

万达网络科技集团大数据中心副总经理，《Spark 高级数据分析》中文版译者

Apache Kylin 是基于 MOLAP 的实时大数据引擎，与 Hadoop 生态系统结合更加紧密，先天的优势注定了其支持更大的数据规模、更好的扩展性，独有的中国血统较其他开源软件更具本地化优势，更符合中国国情。本书包含了守壮多年的实践经验，系统全面地介绍了 Apache Kylin 技术，值得推荐。

贾传青

数据架构师，IT 脱口秀（清风那个吹）创始人

Apache Kylin 是基于大数据技术的一类 OLAP 实现，其根据 OLAP 原理、利用 MapReduce 框架构建 CUBE，并将预计算结果存储在 HBase 中，实现多维分析和查询的秒级响应。Apache Kylin 虽属于 MOLAP 范畴，但还是有别于传统的 MOLAP，它充分利用了 Hadoop 分布式计算的精髓，是分布式 OLAP（DOLAP: Distributed OLAP）的一个具体实现，在 TB、PB 级数据集上体现出卓越的性能表现，自开源以来就备受各界关注。

作为一位技术达人，蒋守壮依托自身深厚的技术功底，结合实际工作对 Kylin 做了许多研究工作。从各种部署环境的搭建、实际工作案例开发测试到各类问题的分析及解决，作者深入分析了 Kylin 的源代码，也给 Kylin 社区反馈了很多缺陷，被 Kylin 社区确认并在新的版本中加以完善。《基于 Apache Kylin 构建大数据分析平台》这本书即是蒋守壮对自己研究工作的总结和升华，是当前第一本系统介绍 Kylin 的实用书籍。

项同德

平安科技（深圳）有限公司高级经理

目前在企业级市场上主流的 BI 产品有 Oracle 的 BIEE、IBM 的 Cognos、SAP 的 BO 等，这些产品主要是基于传统的关系型数据进行报表开发和数据分析，虽然可以通过提高服务器性能来提升数据处理的能力，但受限于其自身的架构，在处理大数据（TB 级及以上）上就显得缓慢，而 Kylin 是一款专为大数据而生的开源产品。相对于传统大厂商主导的 BI 产品，Kylin 是一个开源的分布式分析引擎，提供 Hadoop 之上的 SQL 查询接口及多维分析（OLAP）能力以支持超大规模数据，其最初由 eBay 公司开发并贡献至开源社区，它能在亚秒内查询巨大的 Hive 表，并且在不断地完善和进化。

Kylin 作为 Apache 顶级项目，在社区备受推崇，但一直缺少一本实用、可操作的技术书籍让普通的开发人员将其部署实施，应用于企业的发展，产生经济价值。蒋守壮的这本书系统而全面地介绍了 Kylin 的架构、搭建及应用，能让有一定技术功底的人员，快速实施部署，对于目前苦于大数据处理的人员来讲，无疑是久旱逢甘霖。

蒋守壮一直专注于大数据的研究和应用，技术出众，尤其是擅长解决各类疑难问题。这本书融合了作者多年的技术积累和实战经验，相信对您，无论是学习还是实战都是大有益处。

万文兵

万达网络科技集团有限公司大数据资深项目经理

前言

自 2011 年下半年开始，我就一直关注 Apache 开源社区，侧重点放在大数据方面的成熟框架和产品。在这期间，陆续研究过 Hadoop、Hive、HBase、Mahout、Kafka、Flume、Storm，以及近两年很火的 Spark 和 Flink 等，和很多从事大数据的朋友一样，经历过无数的夜晚，对着电脑屏幕逐行研究这些源代码，同时也看到无数的开源爱好者和技术专家加入 Hadoop 开源社区，贡献自己的力量，日复一日，乐此不疲。

谈起大数据，不得不提 Hadoop，如今其早已发展成为了大数据处理的事实标准。Hadoop 诞生于 2005 年，其受到 Google 的两篇论文（GFS 和 MapReduce）的启发。起初，Hadoop 只是用来支撑 Nutch 搜索引擎的项目，从 2006 年开始，Hadoop 脱离了 Nutch，成为了 Apache 的顶级项目，无论是在学术界还是工业界都得到了迅猛的发展。

如今已是 2016 年了，Hadoop 十周岁了，这十年期间围绕其核心组件（HDFS、MapReduce、Yarn）陆续出现了一批工具，用来丰富 Hadoop 生态圈，解决大数据各方面的问题，这其中就包括 Apache Kylin。

Apache Kylin（麒麟）是由 eBay 研发并贡献给开源社区的 Hadoop 上的分布式大规模联机分析（OLAP）平台，它提供 Hadoop 之上的 SQL 查询接口及多维分析能力以支持大规模数据，能够处理 TB 乃至 PB 级别的分析任务，能够在亚秒级查询巨大的 Hive 表，并支持高并发。Apache Kylin 于 2014 年 10 月开源，并于当年 11 月成为 Apache 孵化器项目，是 eBay 第一个贡献给 Apache 软件基金会的项目，也是第一个由中国团队完整贡献到 Apache 的项目，在这里对 Apache Kylin 的中国团队表示感谢，感谢贡献如此出色的大数据分析平台。

从去年开始接触 Apache Kylin，我感觉很亲切，也很惊喜。当前研究的版本为 0.7.1，也就是 Kylin 加入 Apache 孵化器项目后的第一个 Apache 发行版本，虽然当时的 Kylin 存在一些问题，但是其基于 Hadoop 设计的框架还是很有创意和特色的。经过一年多的发展，截至目前，Apache Kylin 的版本已经发展到 1.5.3，并且从 1.5 版本开始，Apache Kylin 进行了重构，支持可扩展架构，支持更多的数据源、构建引擎和存储引擎，构建算法不断优化，支持与更多的可视化工具集成等。

如今，Apache Kylin 已被应用在 eBay、Exponential、京东、美团、明略数据、网易及其他公司。越来越多的大数据团队开始选择 Apache Kylin 作为公司大数据分析平台的组成部分，满足其海量数据的多维指标实时查询分析。通过很多社区的交流分享，我发现不少朋友对

Apache Kylin 没有一个整体的认识，在使用过程中出现各种各样的问题，打击自信心，他们急切希望能有一本全面介绍 Apache Kylin 的书籍。因为我经常在博客和社区分享 Apache Kylin 实战方面的一些经验，所以很多朋友鼓励我能够写一本比较全面介绍 Apache Kylin 的书籍，帮助更多的爱好者更好地加入 Apache Kylin 的社区，并在生产环境中进行实践。刚开始比较犹豫，毕竟写书需要花费大量的时间和精力，而且要对读者负责，容不得半点马虎。后来有社区的几个朋友给我打电话劝说，以及清华大学出版社的夏毓彦编辑一再鼓励，还有家人的支持，我就下定决心写这本书，目的只有一个，就是希望读者能够通过这本书，对 Apache Kylin 有一个完整的认识，掌握各方面的技能，并最终应用在自己公司的生产环境中。

本书内容

这是一本全面介绍 Apache Kylin 的书籍，包括环境搭建、案例实战演示、源码分析、Cube 优化等，此外还会涉及数据仓库、数据模型、OLAP、数据立方体等方面的知识。通过本书系统性学习和实战操作，朋友们将能够达到基于 Apache Kylin 搭建企业级大数据分析平台，并熟练掌握使用 Apache Kylin 多维度地分析海量数据，最终通过可视化工具展示结果。

受众人群

本书适合从事 Hadoop、HBase、Hive 和 Kylin 等方面工作的人员参考阅读，最好能掌握一点 OLAP、数据立方体等数据仓库方面的知识。但是我相信这本书也适合任何想从事大数据方面工作的程序员和架构师。

代码规范和下载

本书中会涉及大量的 Linux Shell 命令，这些命令都是在 CentOS 操作系统上执行成功的，对于其他的一些 Linux 系统也同样适用，如有不适用的，可以查阅资料，修改命令以符合对应的操作系统。

要下载本书章节中的样例代码，请到 <http://github.com/jiangshouzhuang> 下载。

读者服务

由于本人的写作能力有限，可能有些章节内容考虑并不全面，或者版本升级导致某些章节部分内容不是最新的。为了更好地为读者服务，我特意建立了一个 QQ 群：118152802，读者有关本书的任何问题，我都会及时给朋友们答复，谢谢支持。

致谢

这本书的面世，得到了很多朋友的鼎力相助，在这里感谢所有帮助我完成这本书的人。

感谢公司的同事们，特别感谢项同德和万文兵两位项目经理给予的支持和鼓励，感谢施健给予的技术支持和帮助。

感谢 CSDN 和 cnblogs 博客中优秀的文章给予的技术支持。

感谢清华大学出版社所有为本书的出版和发行付出了辛勤劳动的人们。

最后，我要感谢我的家人，给予我的不懈支持。感谢父母帮我们照顾调皮捣蛋的宝宝；感谢妻子一如既往地照顾我的生活，给予我充足的时间用来写作。没有家人的支持和照顾，是不可能完成这本书。

作者

2016 年 10 月

目 录

第一部分 Apache Kylin 基础部分

第 1 章	Apache Kylin 前世今生	3
1.1	Apache Kylin 的背景	3
1.2	Apache Kylin 的应用场景	3
1.3	Apache Kylin 的发展历程	4
第 2 章	Apache Kylin 前奏	7
2.1	事实表和维表	7
2.2	星型模型和雪花型模型	7
2.2.1	星型模型	7
2.2.2	雪花型模型	8
2.2.3	星型模型示例	8
2.3	OLAP	9
2.3.1	OLAP 分类	9
2.3.2	OLAP 的基本操作	10
2.4	数据立方体 (Data Cube)	11
第 3 章	Apache Kylin 工作原理和体系架构	12
3.1	Kylin 工作原理	12
3.2	Kylin 体系架构	13
3.3	Kylin 中的核心部分: Cube 构建	15
3.4	Kylin 的 SQL 查询	16
3.5	Kylin 的特性和生态圈	16

第 4 章	搭建 CDH 大数据平台	18
4.1	系统环境和安装包	19
4.1.1	系统环境	19
4.1.2	安装包的下载	20
4.2	准备工作：系统环境搭建	21
4.2.1	网络配置(CDH 集群所有节点)	21
4.2.2	打通 SSH，设置 ssh 无密码登录（所有节点）	21
4.3	正式安装 CDH：准备工作	29
4.4	正式安装 CDH5：安装配置	30
4.4.1	CDH5 的安装配置	30
4.4.2	对 Hive、HBase 执行简单操作	39
第 5 章	使用 Kylin 构建企业大数据分析平台的 4 种部署方式	41
5.1	Kylin 部署的架构	41
5.2	Kylin 的四种典型部署方式	42
第 6 章	单独为 Kylin 部署 HBase 集群	44
第 7 章	部署 Kylin 集群环境	58
7.1	部署 Kylin 的先决条件	58
7.2	部署 Kylin 集群环境	61
7.3	为 Kylin 集群搭建负载均衡器	70
7.3.1	搭建 Nginx 环境	70
7.3.2	配置 Nginx 实现 Kylin 的负载均衡	73
第二部分 Apache Kylin 进阶部分		
第 8 章	Demo 案例实战	77
8.1	Sample Cube 案例描述	77
8.2	Sample Cube 案例实战	78
8.2.1	准备数据	78
8.2.2	构建 Cube	81
第 9 章	多维分析的 Cube 创建实战	89
9.1	Cube 模型	89

9.2 创建 Cube 的流程.....	90
9.2.1 步骤一: Hive 中事实表, 以及多张维表的处理	90
9.2.2 步骤二: Kylin 中建立项目 (Project)	95
9.2.3 步骤三: Kylin 中建立数据源 (Data Source)	95
9.2.4 步骤四: Kylin 中建立数据模型 (Model)	98
9.2.5 步骤五: Kylin 中建立 Cube.....	104
9.2.6 步骤六: Build Cube.....	114
9.2.7 步骤七: 查询 Cube.....	118
第 10 章 Build Cube 的来龙去脉.....	120
10.1 流程分析	120
10.2 小结	134
第三部分 Apache Kylin 高级部分	
第 11 章 Cube 优化	137
第 12 章 备份 Kylin 的 Metadata	142
12.1 Kylin 的元数据	142
12.2 备份元数据	143
12.3 恢复元数据	146
第 13 章 使用 Hive 视图	147
13.1 使用 Hive 视图	147
13.2 使用视图实战	149
第 14 章 Kylin 的垃圾清理	153
14.1 清理元数据	153
14.2 清理存储器数据	154
第 15 章 JDBC 访问方式.....	157
第 16 章 通过 RESTful 访问 Kylin	161
第 17 章 Kylin 版本之间升级	179
17.1 从 1.5.2 升级到最新版本 1.5.3	179

17.2	从 1.5.1 升级到 1.5.2 版本	180
17.3	从 Kylin 1.5.2.1 升级到 Kylin 1.5.3 实战	181
17.4	补充内容	187
第 18 章	大数据可视化实践	189
18.1	可视化工具简述	189
18.2	安装 Kylin ODBC 驱动	190
18.3	通过 Excel 访问 Kylin	192
18.4	通过 Power BI 访问 Kylin	194
18.4.1	安装配置 Power BI	194
18.4.2	实战操作	198
18.5	通过 Tableau 访问 Kylin	199
18.6	Kylin + Mondrian + Saiku	205
18.7	实战演练：通过 Saiku 访问 Kylin	211
18.7.1	第一个 Schema 例子：myproject_pvuv_cube 的演示	211
18.7.2	第二个 Schema 例子：kylin_sales_cube 的演示	219
18.7.3	Saiku 使用的一些问题	223
18.8	通过 Apache Zeppelin 访问 Kylin	229
18.9	通过 Kylin 的 “Insight” 查询	232
第 19 章	使用 Streaming Table 构建准实时 Cube	236
第 20 章	快速数据立方算法	251
20.1	快速数据立方算法概述	251
20.2	快速数据立方算法优点和缺点	253
20.3	获取 Fast Cubing 算法的优势	254
第四部分 Apache Kylin 的扩展部分		
第 21 章	大数据智能分析平台 KAP	257
21.1	大数据智能分析平台 KAP 概述	257
21.2	KAP 的安装部署	259

第一部分

Apache Kylin 基础部分



第 1 章

◀ Apache Kylin前世今生 ▶

1.1 Apache Kylin 的背景

在现在的大数据时代，Hadoop 已经成为大数据事实上的标准规范，一大批工具陆陆续续围绕 Hadoop 平台来构建，用来解决不同场景下的需求。

比如 Hive 是基于 Hadoop 的一个用来做企业数据仓库的工具，可以将存储在 HDFS 分布式文件系统上的数据文件映射为一张数据库表，并提供 SQL 查询功能，Hive 执行引擎可以将 SQL 转换为 MapReduce 任务来进行运行，非常适合数据仓库的数据分析。

再比如 HBase 是基于 Hadoop，实现高可用性、高性能、面向列、可伸缩的分布式存储系统，Hadoop 架构中的 HDFS 为 HBase 提供了高可靠性的底层存储支持。

但是缺少一个基于 Hadoop 的分布式分析引擎，虽然目前存在业务分析工具，如 Tableau 等，但是它们往往存在很大的局限，比如难以水平扩展、无法处理超大规模数据，同时也缺少 Hadoop 的支持。此外，Hadoop 以及相关大数据技术的出现提供了一个几近无限扩展的数据平台，在相关技术的支持下，各个应用的数据已突破了传统 OLAP 所能支持的容量上界。每天千万、数亿条的数据，提供若干维度的分析模型，大数据 OLAP 最迫切所要解决的问题就是大量实时运算导致的响应时间迟滞。

Apache Kylin（中文：麒麟）的出现，能够基于 Hadoop 很好地解决上面的问题。Apache Kylin 是一个开源的分布式存储引擎，最初由 eBay 开发贡献至开源社区。它提供 Hadoop 之上的 SQL 查询接口及多维分析（OLAP）能力以支持大规模数据，能够处理 TB 乃至 PB 级别的分析任务，能够在亚秒级查询巨大的 Hive 表，并支持高并发。

1.2 Apache Kylin 的应用场景

（1）假如你的数据存在于 Hadoop 的 HDFS 分布式文件系统中，并且你使用 Hive 来基于 HDFS 构建数据仓库系统，并进行数据分析，但是数据量巨大，比如 PB 级别。

（2）同时你的 Hadoop 平台也使用 HBase 来进行数据存储和利用 HBase 的行键实现数据的

- 快速查询等应用。
- (3) 你的 Hadoop 平台的数据量逐日累增。
 - (4) 对于数据分析的维度大概 10 个左右。

如果你的应用类似上面，那么非常适合采用 Apache Kylin 来做大数据量的多维数据分析。

Apache Kylin 的核心思想是利用空间换时间，将计算好的多维数据结果存入 HBase，实现数据的快速查询。同时，由于 Apache Kylin 在查询方面制定了多种灵活的策略，进一步提高空间的利用率，使得这样的平衡策略在应用中值得采用。

1.3 Apache Kylin 的发展历程

Apache Kylin 于 2014 年 10 月在 github 开源，并很快在 2014 年 11 月加入 Apache 孵化器，2015 年 9 月，Apache Kylin 与 Spark、HBase、Kafka 等并列荣膺 InfoWorld 2015 年 Bossie 最佳开源大数据工具奖。这也是国人项目第一次获得该国际大奖，于 2015 年 11 月正式毕业，成为 Apache 顶级项目，也成为首个完全由中国团队设计开发的 Apache 顶级项目，如图 1-1 所示。

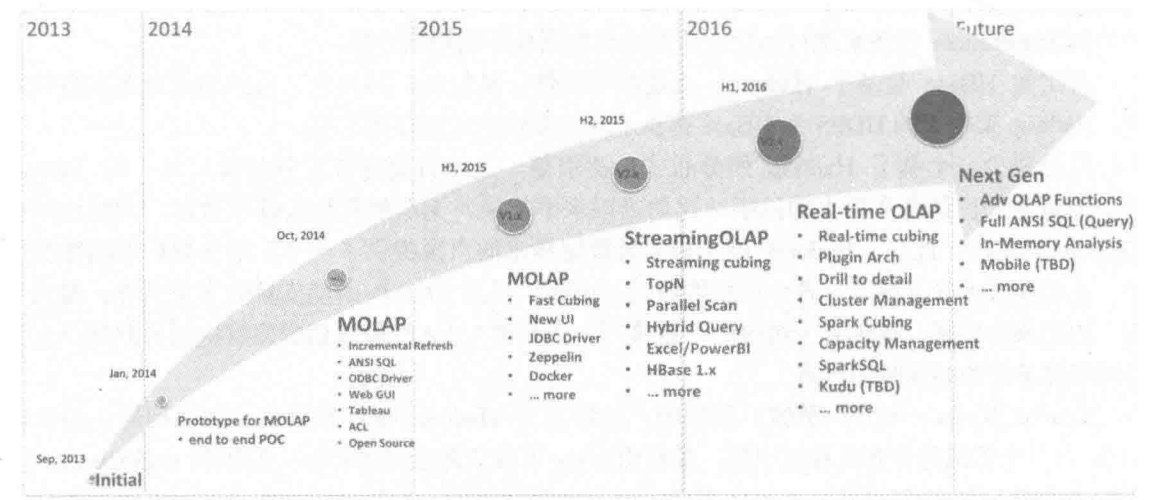


图 1-1

Apache Kylin 在大数据分析领域应用广泛，获得了快速的推广。国内外一线的互联网、金融、电信等公司越来越多地采用 Apache Kylin 作为其大数据分析平台。

Apache Kylin 的官网为 <http://kylin.apache.org>，如图 1-2 所示。



图 1-2

于 2016 年 3 月, Apache Kylin 核心开发成员在上海创建 Kyligence 公司, Kyligence 是一家专注于大数据分析领域创新的数据科技公司, 致力于为用户提供基于 Apache Kylin 的智能分析平台及产品, 提供领先的企业级商业分析解决方案, 以使用户在超大规模数据集上获取极速的洞见能力, 驱动业务增长。Kyligence 也是首家在国内由 Apache 顶级项目核心贡献者团队组建的创业公司, 公司将致力于进一步推动 Apache Kylin 开源项目的发展和演进, 提供基于 Apache Kylin 的大数据分析产品和服务, 拓展全球用户社区, 构建更为丰富的生态系统。

公司官网为 <http://kyligence.io>, 如图 1-3 所示。



图 1-3

为了获取更好的发展, 2016 年 4 月, 大数据公司 Kyligence 跬智科技宣布获得了数百万美元的天使轮投资。

Kyligence 于 2016 年 8 月 3 日, 在北京宣布正式发布其企业级大数据智能分析平台 KAP