

OSC 原创会
年终盛典 2016

Apache Kylin ASF 顶级开源项目介绍

韩卿 | Luke Han

--Kyligence CEO

--Apache Kylin PMC Chair

- 关于Apache Kylin
- 技术原理及架构
- Kylin的开源之旅
- Q&A

Hadoop生产力未得到释放

OSC原创会
年终盛典 2016



Kylin: 用户和大数据间的桥梁

OSC 原创会
年终盛典 2016

应用层

统计报表

多维分析

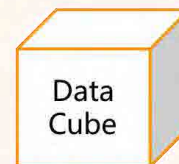
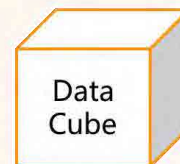
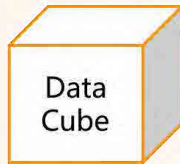
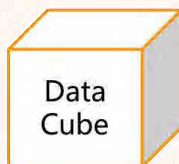
灵活查询

数据探索

数据探索

平台层

数据仓库
数据集市



数据湖



数据层



Apache Kylin的用户

OSC原创会
年终盛典 2016



全球的生产环境用户已经达到100+



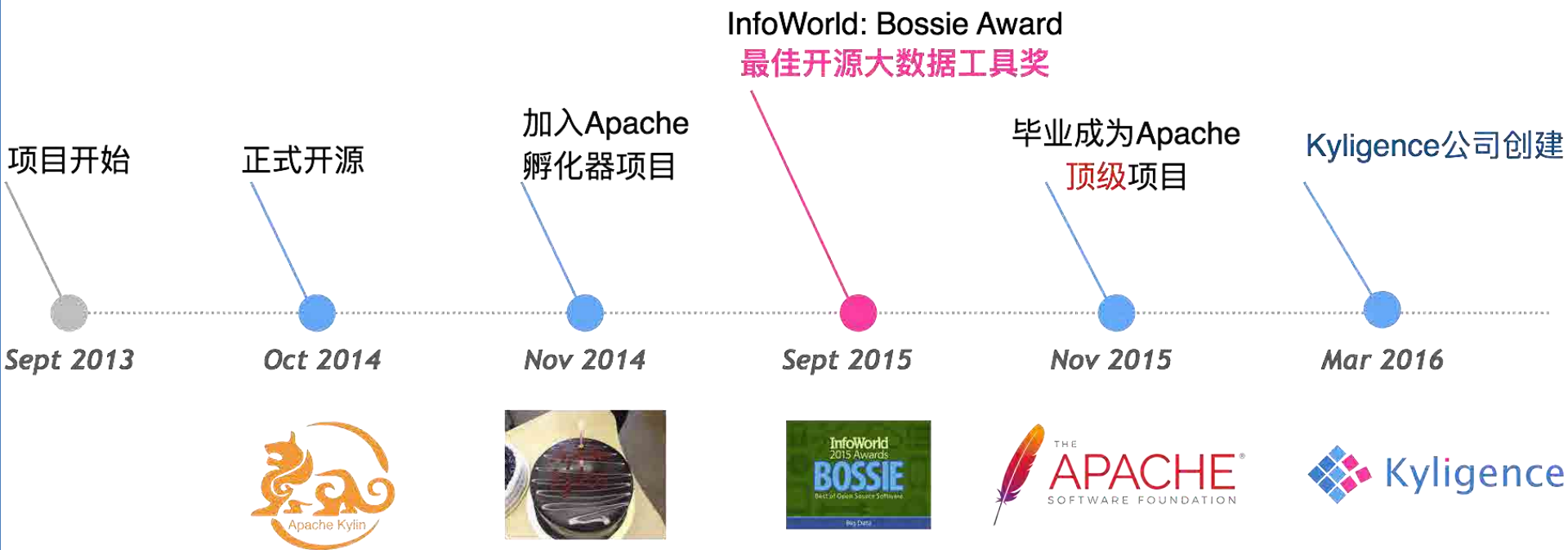
与Apache Kylin团队一起合作使Kylin通过孵化成为顶级项目对我而言非常激动人心，Kylin在技术方面当然是振奋人心的，但同样令人兴奋的是Kylin代表了亚洲国家，特别是中国，在开源社区中越来越高的参与度

—Ted Dunning,
Apache 孵化项目副总裁

- 顶级项目
 - Apache Kylin, 中国唯一的Apache顶级开源项目，核心开发者及贡献者都在中国
- 行业认可
 - 连续两年荣获 InfoWorld “最佳开源大数据工具奖”，今年更是与 Google TensorFlow 一起获得该奖
- 用户认可
 - 国内外超过100多家大型公司正式使用Kylin作为大数据分析平台解决方案，分布各个行业

Apache Kylin的发展历程

OSC原创会
年终盛典 2016



并行计算

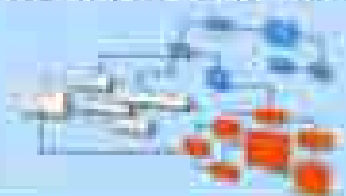
分布式/MPP



将计算过程化整为零
通过横向扩展增加计算能力

空间换时间

索引/物化视图/Cube



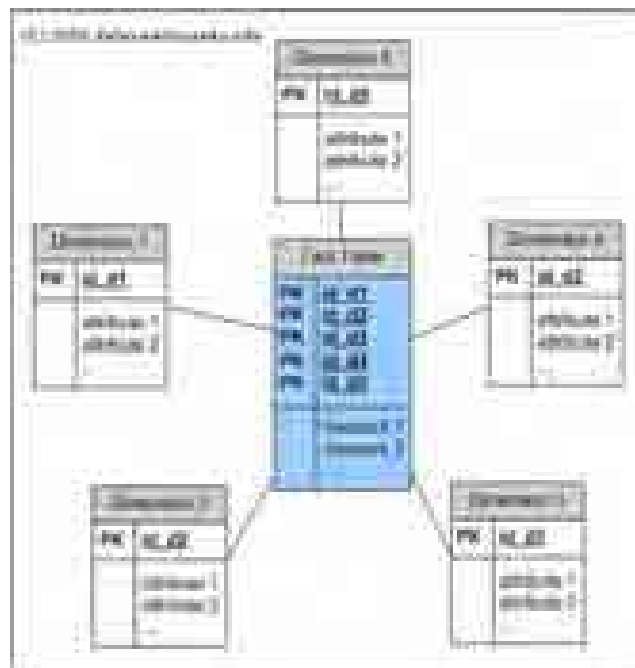
通过预计算生成结果
减少计算和I/O吞吐量

列式存储

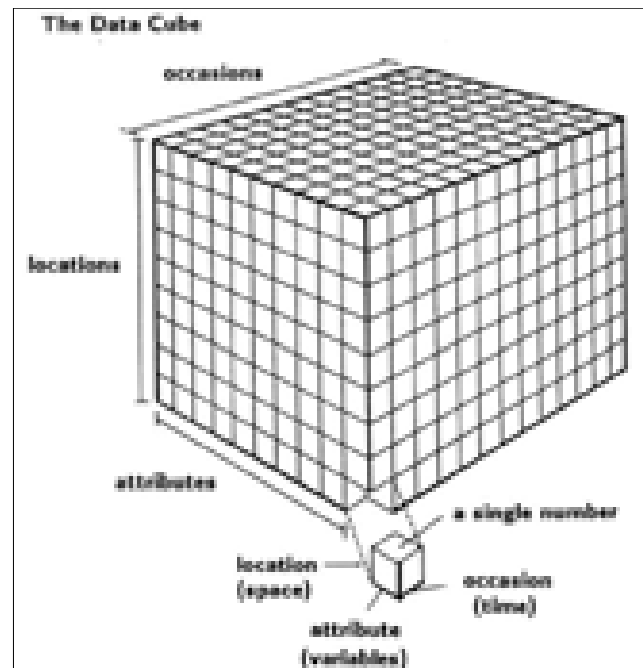
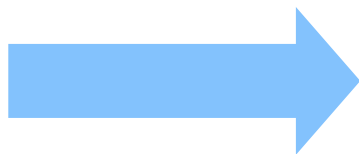


减少磁盘扫描范围
利用各种编码压缩技术

Kylin结合多种优化技术实现超高性能分析



预先进行
汇总、分类、排序

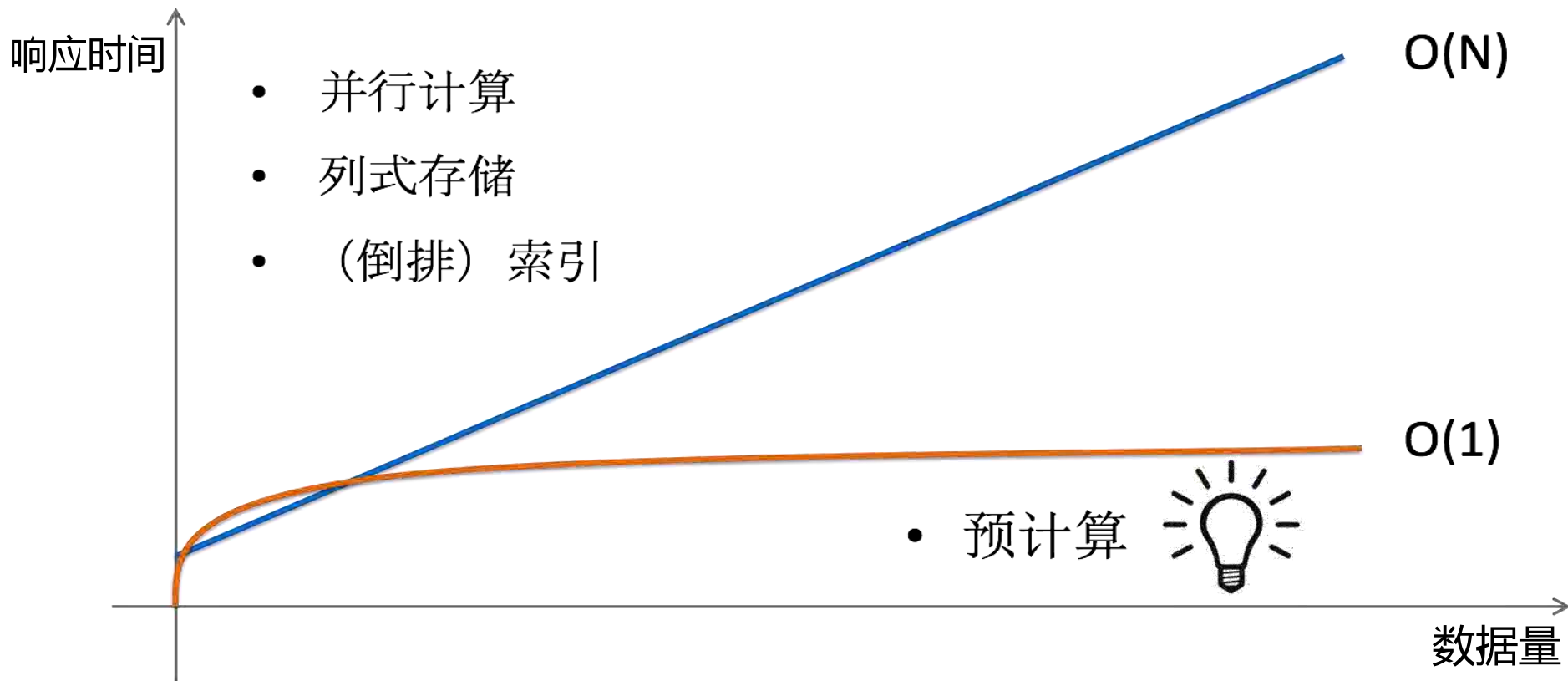


维度模型：数据仓库建设中的一种数据建模方式，按照事实表+维度表的方式来进行数据建模，星型模型应用最广泛的方法

数据立方体：一种多维分析的技术，通过预计算，将计算结果存储在某个多个维度值所映射的空间中，在查询时通过对Cube的再处理而快速获取结果。

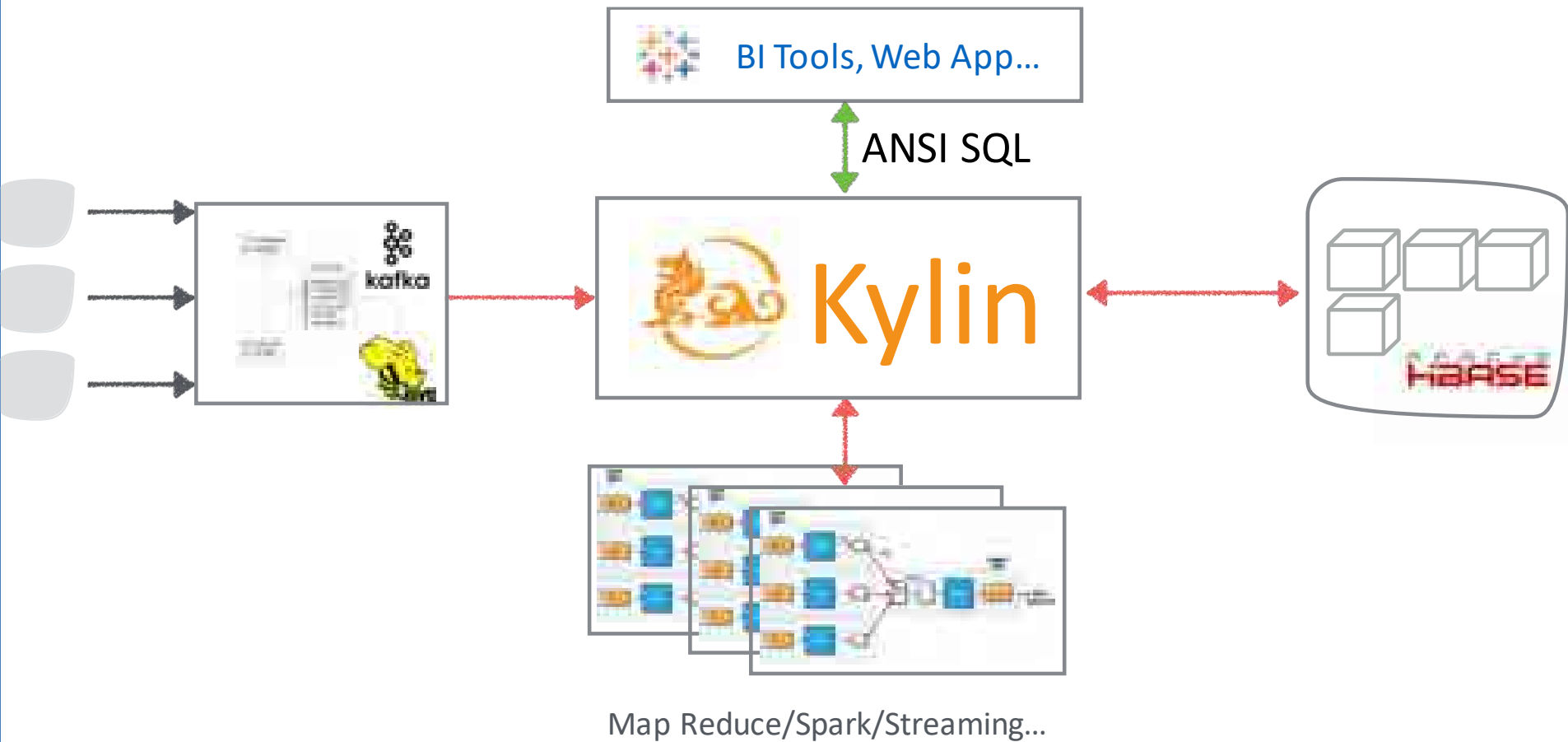
理论基础: $O(1)$ vs $O(N)$

OSC 原创会
年终盛典 2016



Apache Kylin 架构

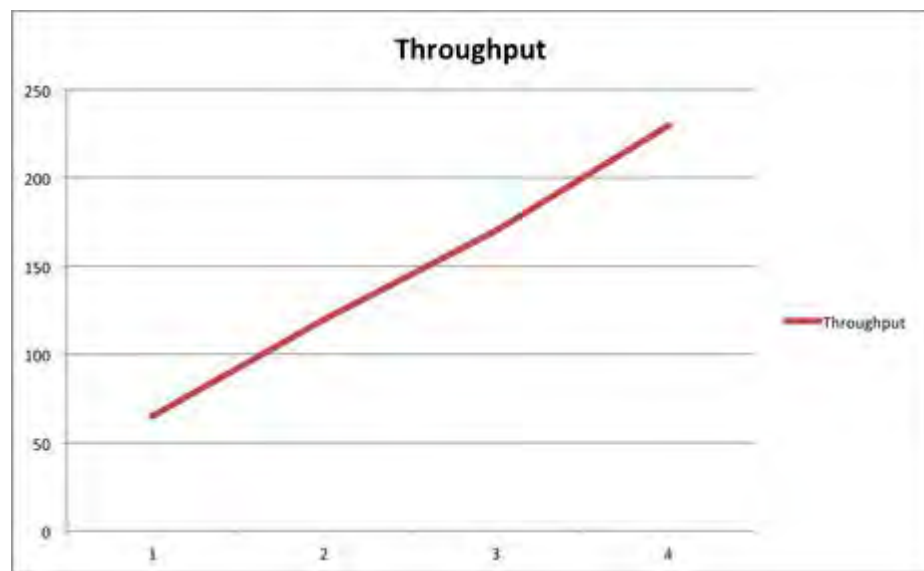
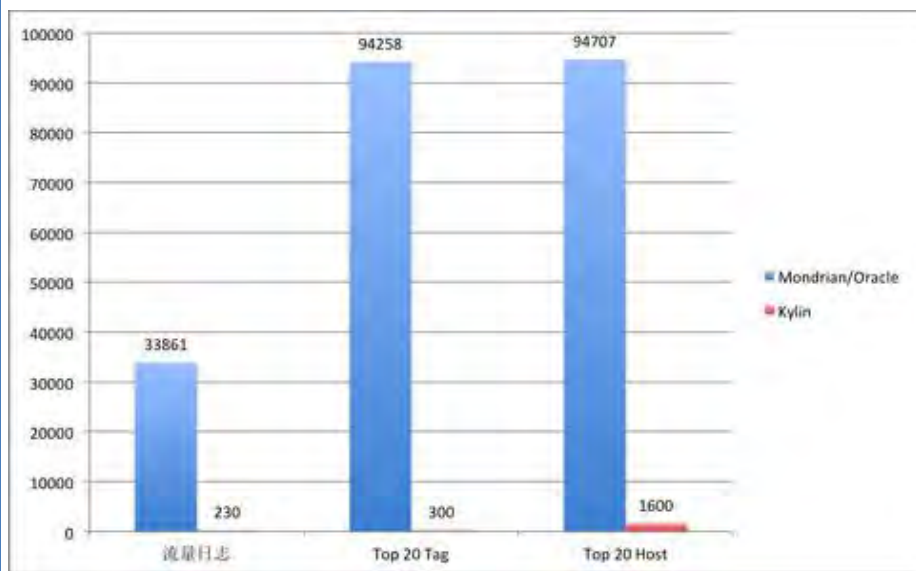
OSC 原创会
年终盛典 2016



超高性能的OLAP on Hadoop

OSC 原创会
年终盛典 2016

- 查询性能提升60~300倍，水平扩展支持高并发



By NetEase:

<http://www.bitstech.net/2016/01/04/kylin-olap/>

- ASF - Apache 软件基金会
 - 成立于1999年的非公益性组织
 - 近600位会员，350多个项目，160多个顶级项目
 - 所有的开发都志愿完成
 - 绝大部分大数据项目都在ASF
 - Community over Code

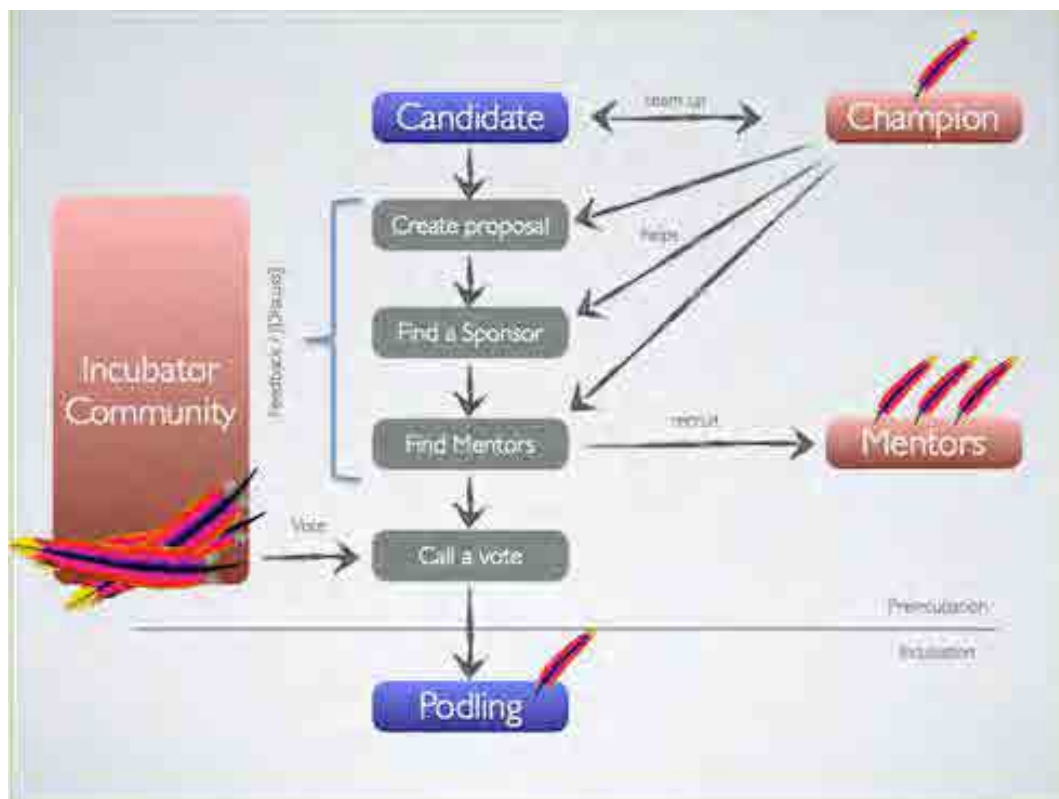


- ASF 会员 – ASF Member
- 董事会- Board
- 孵化器管理委员会 – IPMC
 - Podling
- 项目管理委员会 – PPMC
 - 项目管理委员会成员 – PMC Member
- 提交者 - Committer
- 贡献者 - Contributor
- 用户 - User



Apache 孵化项目流程

OSC 原创会
年终盛典 2016



- 投票机制

- +1 – 赞成
- +0 – 中立
- -1 – 反对

- 规则

- 最少净获得3票赞成
- 72小时 投票时间
- 欢迎每一个人投票
- Binding & no-binding

- Apache Kylin Mentors
 - Ted Dunning (IPMC Chair)
 - Henry Saputra
 - Julian Hydra
 - P. Taylor Goetz
 - Owen O'Malley



Apache Kylin

Abstract

Kylin is a distributed and scalable OLAP engine built on Hadoop to support extremely large datasets.

Proposal

Kylin is an open source Distributed Analytics Engine that provides multi-dimensional analysis (MOLAP) on Hadoop. Kylin is designed to accelerate analytics on Hadoop by allowing the use of SQL-compatible tools. Kylin provides a SQL interface and multi-dimensional analysis (MOLAP) on Hadoop to support extremely large datasets and tightly integrate with Hadoop ecosystem.

Overview of Kylin

Kylin platform has two parts of data processing and interactive. First, Kylin will read data from source, Hive, and run a set of tasks including Map Reduce job, shell script to pre-calculate results for a specified data model, then save the resulting OLAP cube into storage such as HBase. Once these OLAP cubes are ready, a user can submit a request from any SQL-based tool or third party applications to Kylin's REST server. The Server calls the Query Engine to determine if the target dataset already exists. If so, the engine directly accesses the target data in the form of a predefined cube, and returns the result with sub-second latency. Otherwise, the engine is designed to route non-matching queries to whichever SQL on Hadoop tool is already available on a Hadoop cluster, such as Hive.

Kylin实践：讨论及投票

OSC原创会
年终盛典 2016

[VOTE] Accept Kylin into the Apache Incubator

Luke Han
to: general
Following the discussion earlier in the thread
<http://mail-archives.apache.org/d2/mbx/incubator-comm/201411/mbx007330C46mOnOnO2+rrR+rdAa33CpCpL0h5L5w2MS3T-Ld.bv835Qp@mail.gmail.com/5.3e>

I would like to call a VOTE for accepting Kylin as a new incubator project.

The proposal is available at:

<https://wiki.apache.org/incubator/KylinProposal>

and posted the text of the proposal below also.

Vote is open until 24th November 2014, 20:58:03 UTC.

+1 accept Kylin in the incubator
as
-1 because...

Thanks

Luke

Kylin Proposal

incubator.apache.org

Accepted

Kylin is a distributed and scalable OLAP engine built on Hadoop to support extremely large datasets.



Luke Han
to: general

11/20/14

The vote has passed: 4/0 For/0 No/0 Abstain/0.

For/0 No/0
Abstain/0
For/0 No/0
Abstain/0
For/0 No/0
Abstain/0
For/0 No/0
Abstain/0
For/0 No/0
Abstain/0

For/0 No/0
Abstain/0

Thanks everyone for voting. We will proceed with the next steps as per the RFC guidelines.

Thanks
Luke

- 社区的核心是参与的人
- 始终在邮件列表里沟通
- 使用英语
- 积极解答
- 线下活动
- 全球技术大会
- 培养用户，挖掘苗子，不断吸引贡献者
- 与其他项目形成良好互动

Kylin实践：毕业成为顶级项目

OSC原创会
年终盛典 2016

[DISCUSS] Graduate Kylin from the Apache Incubator

Luke Han 10/17/15

The Kylin community and project made significant advances during the incubating (from Nov 2014) and believes it is ready to graduate as a top-level project.

The Apache Kylin is very active. The IPMC doubted in size (added 6 contributors and 2 reporters) and increased diversity in the past year. Released 3 versions in the past 8 months. There were presentations about Kylin at most of the big conferences of the world (including Strata-Hadoop World London, Hadoop Summit San Jose, ApacheCon EU, Big Data Technology China, Database Technology Conference China) and some meetups (Bay Area, Beijing and one is coming in this weekend in Shanghai), and many talks around the world. The user mailing list is growing very much, about 500+ topics per month now. The community created 1000+ JIRA issues, many patches from contributors/committees have been merged into code base.

A vote passed unanimously on the dev list (27 +1 votes). Please find below references to the graduation preparation articles:

- * discussion on dev list [1]
- * vote thread [2]
- * polling name search call in progress [3]
- * incubator status [4]
- * proposed resolution below

We believe Apache Kylin is ready to become a top-level project and if the IPMC

[Result][VOTE] Graduate Apache Kylin from the Apache Incubator

Luke Han 10/22/15

The vote for Apache Kylin to become a top-level project has passed with 27 +1 votes and no 0 or -1 votes.

Thank you everyone for taking that time to review and cast your vote.

We will now prepare a resolution for the next Board meeting.

10 Voting

- * Henry Sapkota
- * Andrew Poyat
- * Bertrand Delacour
- * Julian Hyde
- * P. Taylor Goss
- * Ted Dunning
- * Steven J. Yoon
- * Alexander Beiderman
- * John D. Aylward
- * Owen O'Malley

17 non-voting

- * Luke Han
- * Shaozhong Gu
- * Jason Zhong
- * Guanhao Zhou
- * Dong Li
- * Delany Fu
- * Xueping Bai (vacate)

Kylin实践：毕业成为顶级项目

OSC 原创会
年终盛典 2016



The Apache Software Foundation Blog

December 8, 2016

The Apache Software Foundation Announces Apache™ Kylin™ as a Top-Level Project

Open Source, completely scalable Big Data Distributed Analytics Engine to use as plug, Connector, ADQuery, Metrics, Mining(MML) and NoSQL, among others.

Forrest HUI, MD - 8 December 2016- The Apache Software Foundation (ASF), the international foundation, develops and incubates of more than 200 Open Source projects and initiatives, announced today that Apache™ Kylin™ has graduated from the Apache Incubator to become a Top-Level Project (TLP), signifying that the project's community and technical base have been well-governed under the ASF's membership process and principles.

Apache Kylin is an Open Source Distributed Analytics Engine designed to provide a SQL interface and multi-dimensional queries (OLAP) on Apache Hadoop™, supporting extremely large datasets.

"Apache Kylin is a crucial journey has demonstrated the value of Open Source governance at ASF and the power of building an open source community and ecosystem around the project," said Lee Hong, Vice President of Apache Kylin. "Our community is engaging the whole Hadoop ecosystem community in alignment with the Apache Way."

As leading OLAP on Hadoop solution, Apache Kylin fills the gap between Big Data ecosystem and human use, enabling traditional analysts to interact datasets with sub-second latency for analysis, and users, dashboards, and data structures. With these capabilities, Apache Kylin brings from business intelligence (BI) to Apache Hadoop to further the value of Big Data. Kylin is signed as plug, and was released by the Apache Incubator in December 2014.

Apache 基金会宣布 Apache Kylin 成为顶级项目

2016年12月8日 星期五 14:00

Apache Kylin 是 Hadoop 生态系统中一个重要的组件，它提供了在 Hadoop 上运行 OLAP 查询的能力。

Apache Kylin 是一个开源的分布式分析引擎，它提供了在 Hadoop 上运行 OLAP 查询的能力。它支持多种数据源，包括 Hadoop、Hive、Presto、Impala 等。它支持多种查询语言，包括 SQL、Pig、MapReduce 等。它支持多种部署方式，包括单机、集群、云等。

Apache Kylin 是一个开源的分布式分析引擎，它提供了在 Hadoop 上运行 OLAP 查询的能力。它支持多种数据源，包括 Hadoop、Hive、Presto、Impala 等。它支持多种查询语言，包括 SQL、Pig、MapReduce 等。它支持多种部署方式，包括单机、集群、云等。

Apache Kylin 是一个开源的分布式分析引擎，它提供了在 Hadoop 上运行 OLAP 查询的能力。它支持多种数据源，包括 Hadoop、Hive、Presto、Impala 等。它支持多种查询语言，包括 SQL、Pig、MapReduce 等。它支持多种部署方式，包括单机、集群、云等。

"Apache Kylin 是一个开源的分布式分析引擎，它提供了在 Hadoop 上运行 OLAP 查询的能力。它支持多种数据源，包括 Hadoop、Hive、Presto、Impala 等。它支持多种查询语言，包括 SQL、Pig、MapReduce 等。它支持多种部署方式，包括单机、集群、云等。"

作为一个重要的组件，Apache Kylin 在 Hadoop 生态系统中扮演着重要的角色。它提供了在 Hadoop 上运行 OLAP 查询的能力，支持多种数据源、查询语言和部署方式。Apache Kylin 的社区正在积极扩展，与整个 Hadoop 生态系统社区保持一致。

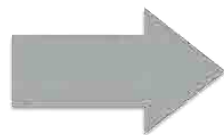
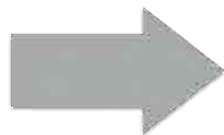
"Apache Kylin 是一个开源的分布式分析引擎，它提供了在 Hadoop 上运行 OLAP 查询的能力。它支持多种数据源，包括 Hadoop、Hive、Presto、Impala 等。它支持多种查询语言，包括 SQL、Pig、MapReduce 等。它支持多种部署方式，包括单机、集群、云等。"

"Apache Kylin 是一个开源的分布式分析引擎，它提供了在 Hadoop 上运行 OLAP 查询的能力。它支持多种数据源，包括 Hadoop、Hive、Presto、Impala 等。它支持多种查询语言，包括 SQL、Pig、MapReduce 等。它支持多种部署方式，包括单机、集群、云等。"

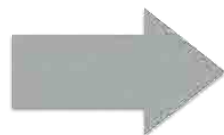
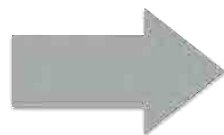
Community Over Code

成功开源项目背后的创业公司

OSC 原创会
年终盛典 2016



Apache Kafka





- ✓ 公司由Apache Kylin的**核心研发团队**组建
- ✓ 公司拥有**6名**Kylin PMC
- ✓ 贡献了**90%+**的Kylin源代码
- ✓ **KAP** : 企业级大数据分析产品
- ✓ **KyBot** : 企业级在线支持云服务
- ✓ 提供官方的培训及认证
- ✓ 提供5x8或者7x24的专家支持
- ✓ 提供专业的企业服务能力

谢谢

OSC 原创会
年终盛典 2016

- 关于Apache Kylin
 - <http://kylin.apache.org>
 - 微信公众号: @ApacheKylin
- 关于Kyligence
 - <http://kyligence.io>
 - info@kyligence.io

