TERADATA UNIVERSE
BEIJING MAY 6 2016

数据仓库 开源融合 极致演绎

2016 | Teradata 大数据峰会

Hortonworks 与Apache开源社区

王健夫
Hortonworks 中国区 销售总监
13910191946

议程

- Apache开源社区
- Hortonworks的介绍
- Hortonworks的开源理念
- 案例分享
- 问答环节

Apache软件基金会

Apache软件基金会（也就是Apache Software Foundation，简称为ASF），是专门为支持开源软件项目而办的一个非盈利性组织。在它所支持的Apache项目与子项目中，所发行的软件产品都遵循Apache许可证（Apache License）

项目管理委员会
项目管理委员会（Project Management Committees，简称为PMC），主要负责保证一个或者多个开源社区的活动都能运转良好

Apache 图标

Apache社区的重要项目

TERADATA UNIVERSE BEIJING MAY 6 2016

- Apache Hadoop
  - 开源大数据技术的鼻祖，包括分布式文件系统，资源管理框架和计算框架。
- Apache NIFI
  - 数据流采集和传导数据, 可以将物联网各类型的数据安全传导各类数据库
- Apache Spark
  - UC Berkeley AMP lab所开源的类Hadoop MapReduce的通用并行框架
- Apache TEZ
  - Tez是Apache开源计算框架,它可以将多个有依赖的作业转换为一个作业从而大幅提升DAG作业的性能
- Apache Ambari
  - Apache Ambari是一种基于Web的工具,支持Apache Hadoop集群的供应、管理和监控

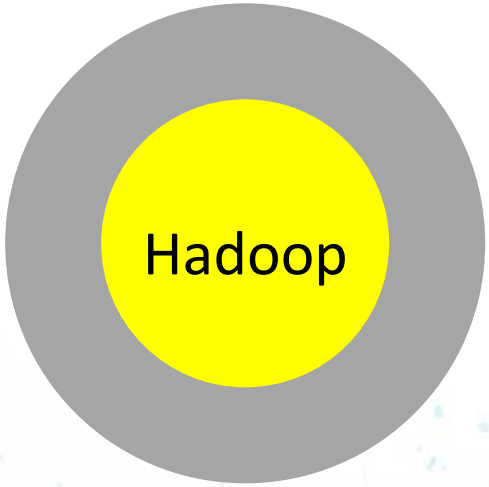
……

Apache社区对Hadoop发展的作用

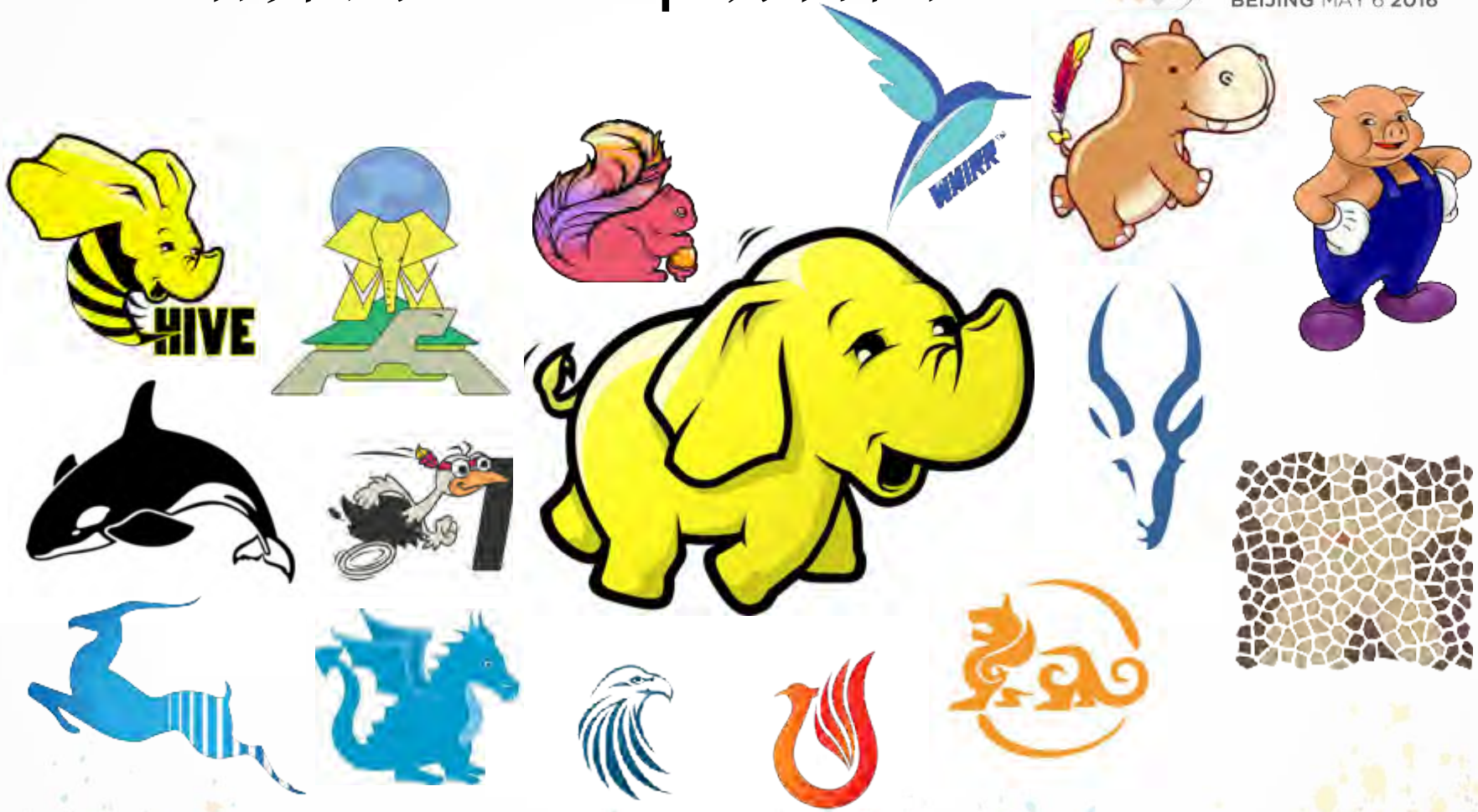
- 开源已经成为一种趋势
  - 随着计算机技术的发展，尤其是互联网技术和相关企业的兴起，开源软件在操作系统、编译工具链、数据库、WEB服务器、移动操作系统等各个方面已经成为企业的主流选择
- Apache开源社区在主导现代Hadoop数据架构
  - Hadoop的相关组件繁多,迭代迅速,任何单一的公司或者组织都无法全面的推动,需要大数据生态系统中的很多公司集体创新.
  - Apache社区中.汇聚了诸多优秀的公司与工程师,他们一起创新推动Apache社区的发展,时至今日, Hadoop技术的发展主要是由Apache社区推动。

Hadoop(概念拆分)

- 一般人们会对Hadoop有两种不同的解释
- Apache Hadoop项目
- 其他Apache Hadoop相关的项目集合



ASF培养的Hadoop动物园



一头神兽？

Hadoop 周期表

TERADATA UNIVERSE BEIJING MAY 6 2016

核心
处理
分析
管理
数据管理
安全
其他

1 M MAPREDUCE	2 H HDFS	26 Y YARN									
0 So SOLR	28 Te TEZ	33 St STORM	6 Ma MAHOUT	39 Da DATAFU	5 Z ZOOKEEPER	22 Am AMBARI	17 Sq SQOOP	40 Pa PARQUET	32 Se SENTRY	20 O OOZIE	
3 Hb HBASE	30 Sp SPARK	4 P PIG	8 Hi HIVE	59 Im IMPALA	11 W WHIRR	16 F FLUME	19 K KAFKA	27 Kn KNOX	13 Hu HUE	38 Sl SLIDER	

Apache Hadoop有多个分支版本

Hadoop及其相关元素表

其他Hadoop相关的ASF项目以及非ASF项目

7 Ha HAMA	31 Sa SAMZA									61 Ku KUDU	15 Mu MRUNIT	EMC ISILON	
9 C CASSANDRA	37 Fl FLINK	48 As ASTERIXDB	53 Ap APEX	14 T TAJO	36 Mq MRQL	54 Hq HAWQ	12 Ch CHUKWA	10 A AVRO	45 N NIFI	64 Ar ARROW	18 B BIGTOP	CLOUDERA MANAGER	
21 G GIRAPH	43 I IGNITE	50 Ge GEODE	58 Sg S2GRAPH	25 D DRILL	44 K KYLIN	55 Md MADLIB	29 Fa FALCON	24 Cr CRUNCH	51 At ATLAS	42 R RANGER	34 Tw TWILL	MAPR-FS	
23 Ac ACCUMULO	47 Ti TINKERPOP	53 Tr TRAFODION	63 Be BEAM	35 Ph PHOENIX	46 Z ZEPPELIN	59 Sm SYSTEMML	49 My MYRIAD	41 Ca CALCITE	56 Ry RYA	57 Ea EAGLE	62 Me METRON	IBM BIG SQL	
												AMAZON S3	

核心	分析	数据处理	其他
处理	管理	安全	非ASF

议程

- Apache开源社区
- Hortonworks的介绍
- Hortonworks的开源理念
- 案例分享
- 问答环节

关于Hortonworks

成立于 2011 年

由Yahoo最初的24名

原创Hadoop的架构师及软件工程师创立

800+
雇员

1500+
合作伙伴

客户概要

- 800+客户（截至2015年底）
- 2015年Q3一个季度增加152位客户
- 纳斯达克上市企业, 代码：HDP

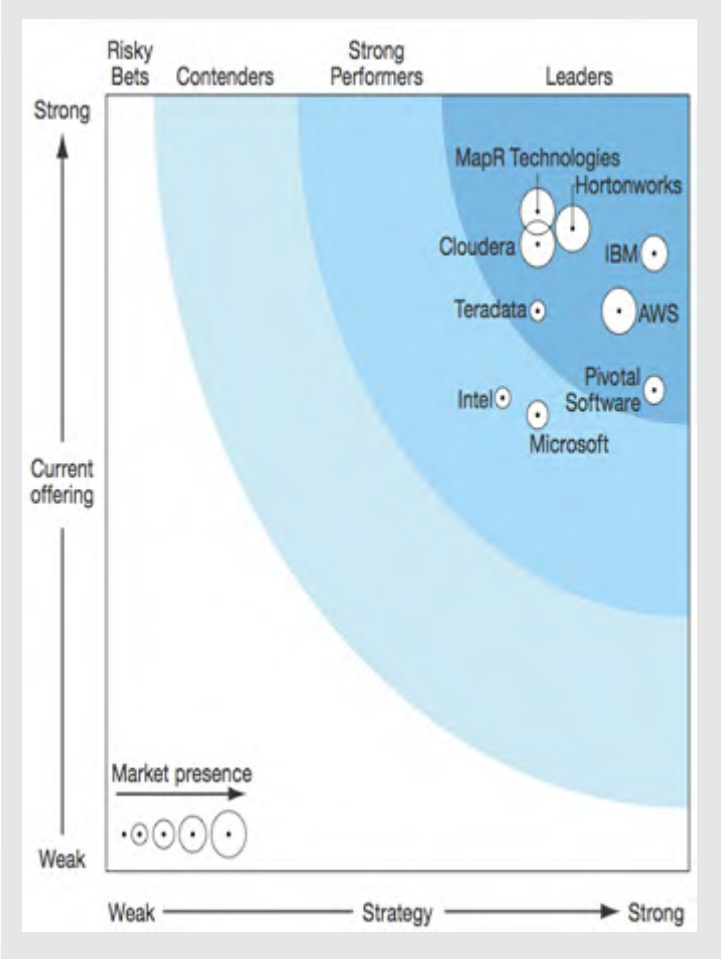
Hortonworks 数据平台

- 一个适用于任意程序和数据且完全开放的多租户数据服务平台
- 具备了一个稳定的企业级平台服务所应具有的安全性，可操作性和可管理性

协助客户获得成功的伙伴

- 一个致力于满足企业级需求的开源社区的领导者
- 灵活的订阅模式为企业提供Hadoop技术支持服务

Hortonworks 引领市场



"The Forrester Wave™: Big Data Hadoop Solutions"

Hortonworks领导者

在Hadoop市场，Hortonworks为40%的全球100强提供支持，包括:

- 75% 的全球电信运营100强
- 65% 的保险公司
- 55% 的全球制造100强
- 46% 的全球零售100强
- 40% 的全球健康保险100强

"Hortonworks热爱开源创新，并以此为生。"

Hortonworks的营收

TERADATA UNIVERSE BEIJING MAY 6 2016

Hortonworks: $121.9 million in 2015 revenues

BARCLAYS

Hortonworks is the fastest software company ever to reach $100 million in revenues.

FIGURE 1

Fastest To Reach $100mn in Revenue

Company	$100mn Year	Founded In	Years to $100mn	Company	$100mn Year	Founded In	Years to $100mn
Hortonworks*	2015	2011	4	FireEye	2013	2004	9
Salesforce	2004	1999	5	Intuit	1992	1983	9
Palo Alto	2011	2005	6	Microsoft	1984	1975	9
Workday	2011	2005	6	NetSuite	2007	1998	9
Informatica	2000	1993	7	Tableau	2012	2003	9
LogMeIn	2010	2003	7	Varonis	2014	2005	9
MobileIron	2014	2007	7	Red Hat	2003	1993	10
Citrix	1997	1989	8	Cornerstone	2012	1999	13
ServiceNow	2011	2003	8	Five9	2014	2001	13
Splunk	2011	2003	8	Paycom	2013	1998	15
Demandware	2013	2004	9	QlikTech	2008	1993	15

Source: Company Data, Barclays Research. Note: Hortonworks represents a potential future event

Hortonworks 产品

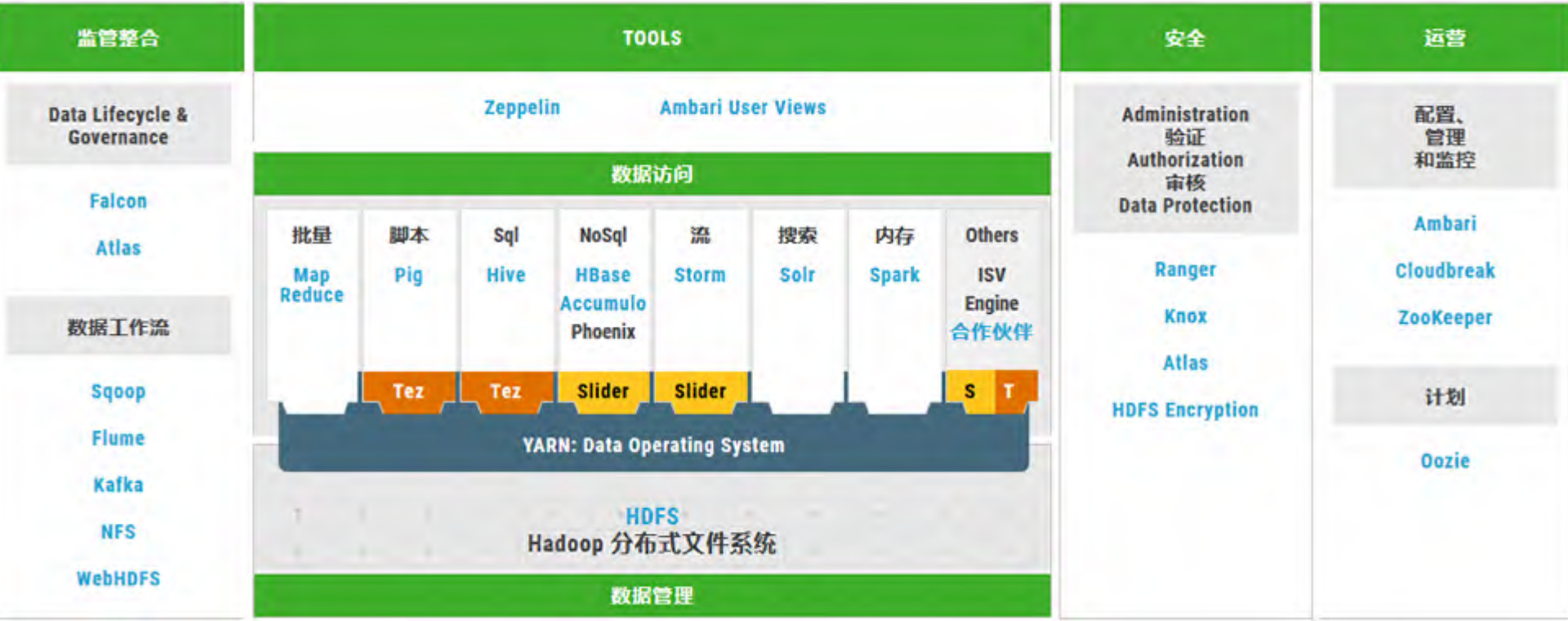
TERADATA UNIVERSE
BEIJING MAY 6 2016

Hortonworks数据平台

- 数据管理
- 数据访问
- 数据治理
- 数据安全
- 平台运维
- 云端部署

由Apache Hadoop驱动

Hortonworks数据平台



HDP 是业内唯一一款基于集中化架构 (YARN)、真正安全、可用于企业的开源 Apache™ Hadoop® 分布式系统。HDP 可满足静态数据的全部需求，助力实时客户应用程序，并提供可加速决策和创新进程的可靠分析

Hortonworks数据流
采集, 传导 & 转换

采集所有类型的数据

通过高度安全的轻量型代理进行数据采集。

传导数据可信度高

传导点对点，双向信息流

转换数据

在保留其来源及变化过程的元数据的同时转换数据。

过程完全透明可控

追踪数据的流，图形到动态的调整。

数据安全并加密

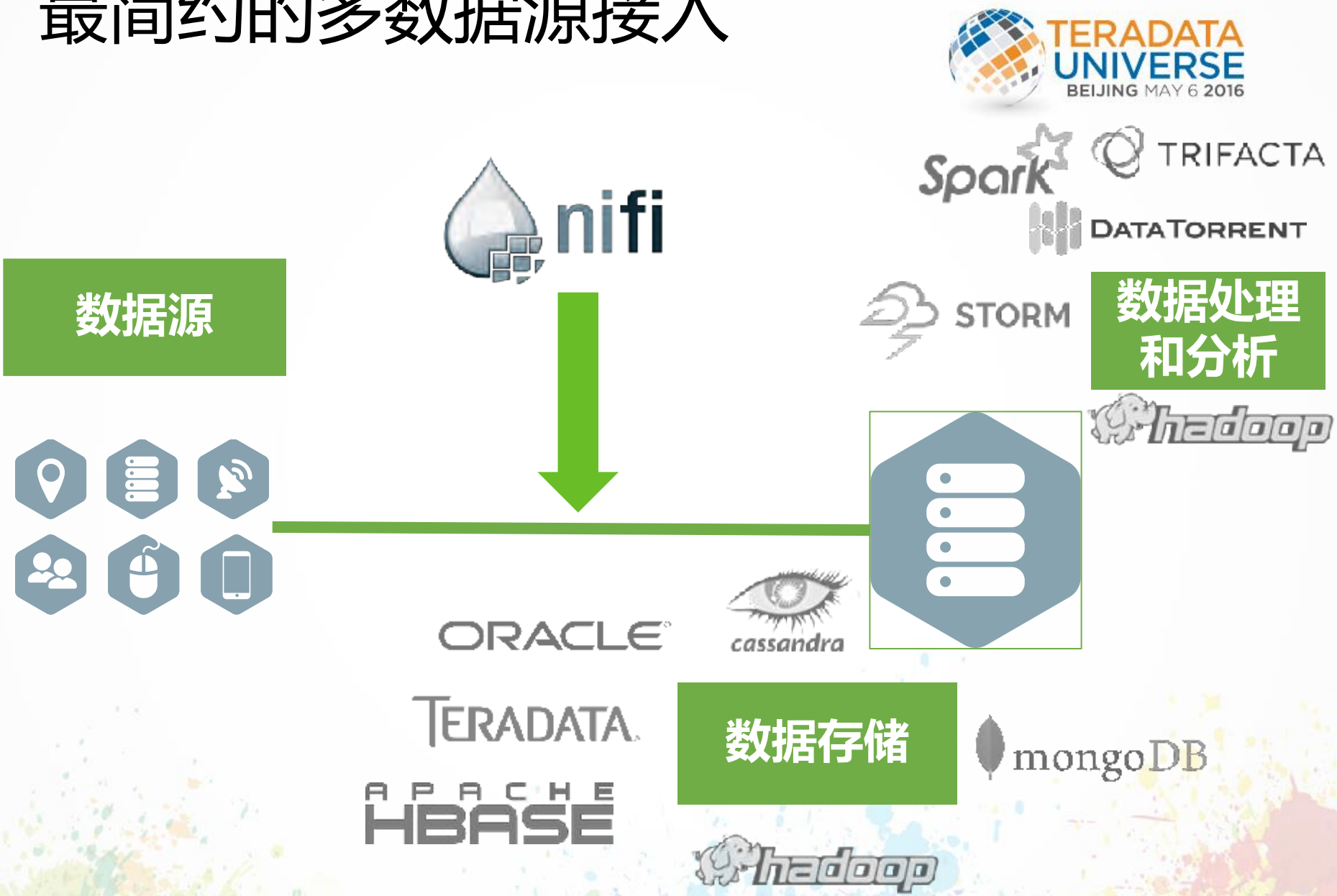
企业级的授权服务，能够频繁改变授权

灵活安全

可以将IOT的数据 安全可视化传到到任何数据源中

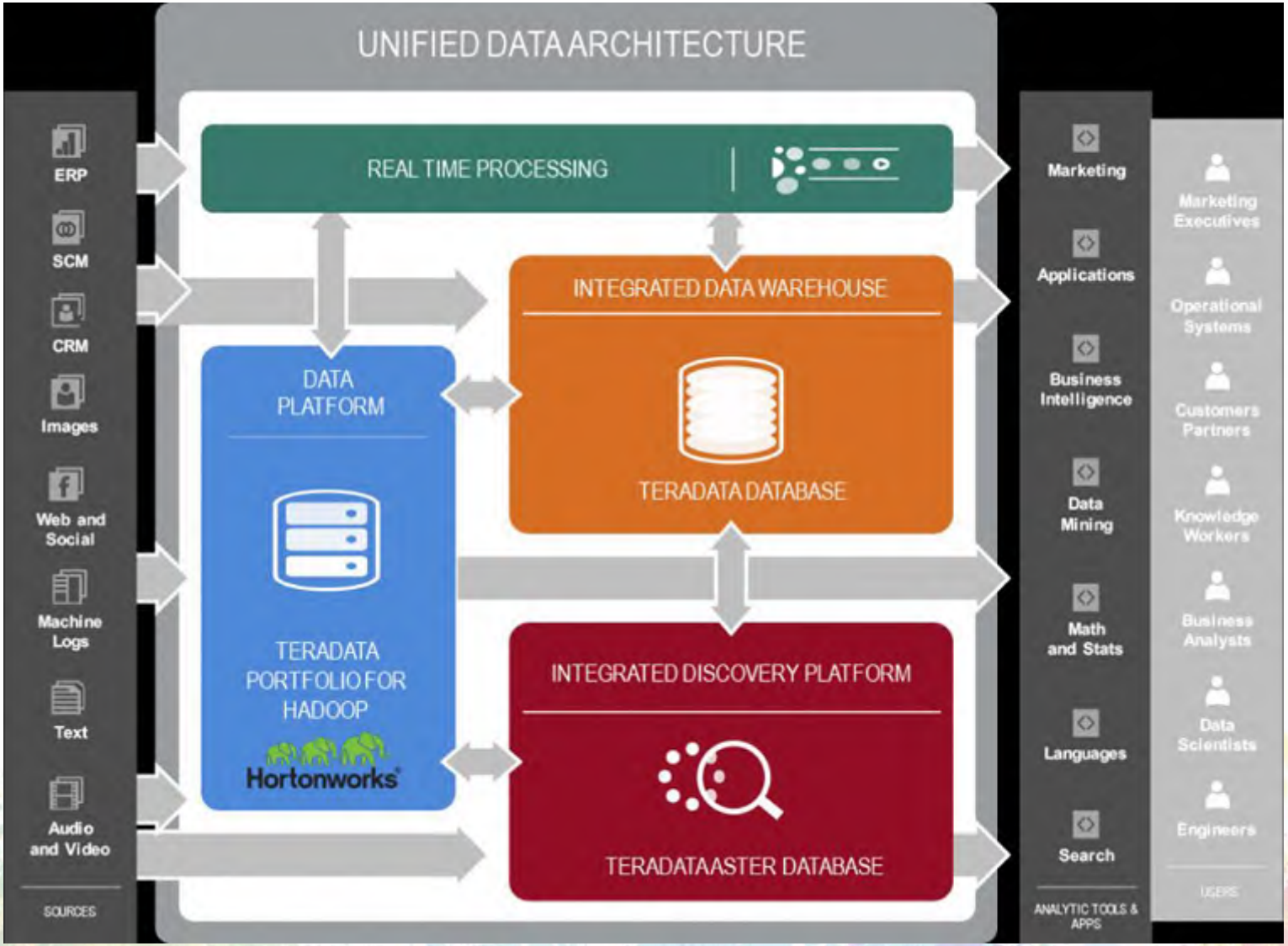
由Apache NiFi驱动

最简约的多数据源接入


TERADATA UNIVERSE
BEIJING MAY 6 2016
nifi
Spark
TRIFACTA
DATATORRENT
STORM
数据处理和分析
hadoop
数据源
ORACLE
cassandra
TERADATA
数据存储
mongoDB
APACHE HBASE
hadoop

Hortonworks支持UDA

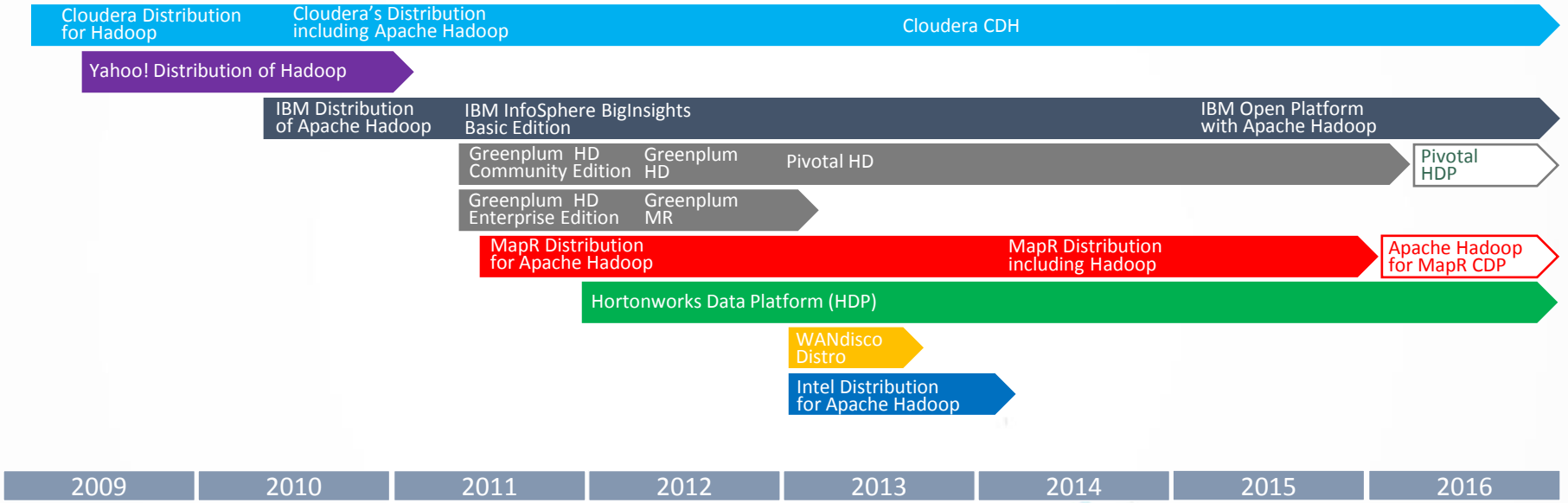
TERADATA UNIVERSE BEIJING MAY 6 2016



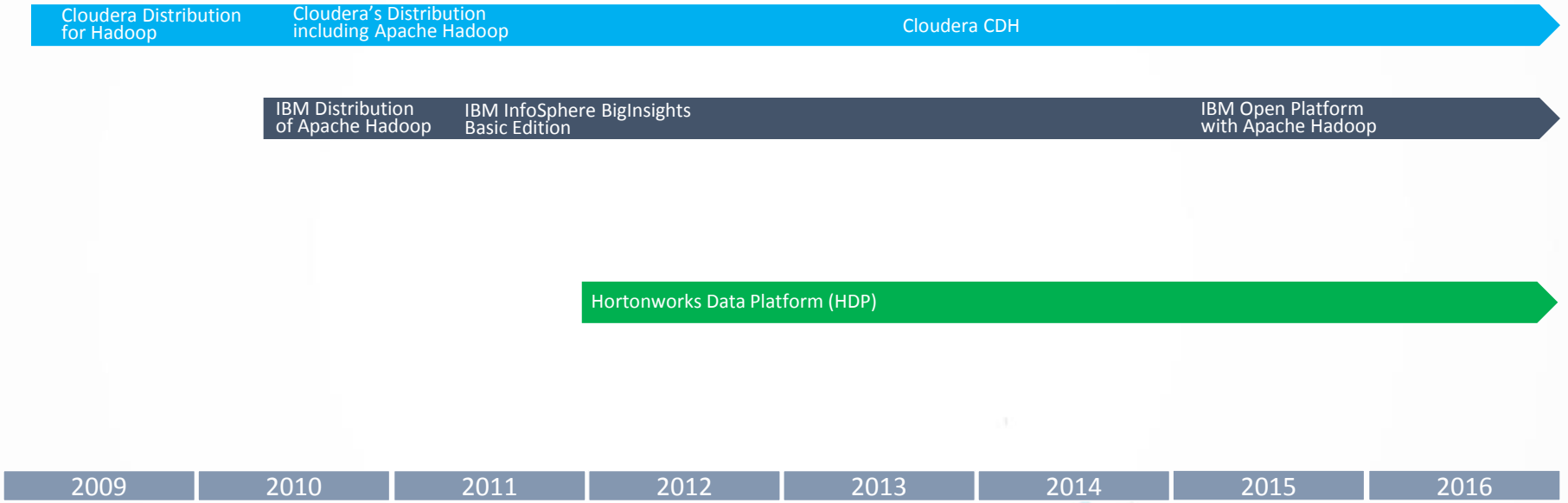
议程

- Apache开源社区
- Hortonworks的介绍
- Hortonworks的开源理念
- 案例分享
- 问答环节

Hadoop 发行版时间轴



最终的Hadoop 发行版厂商



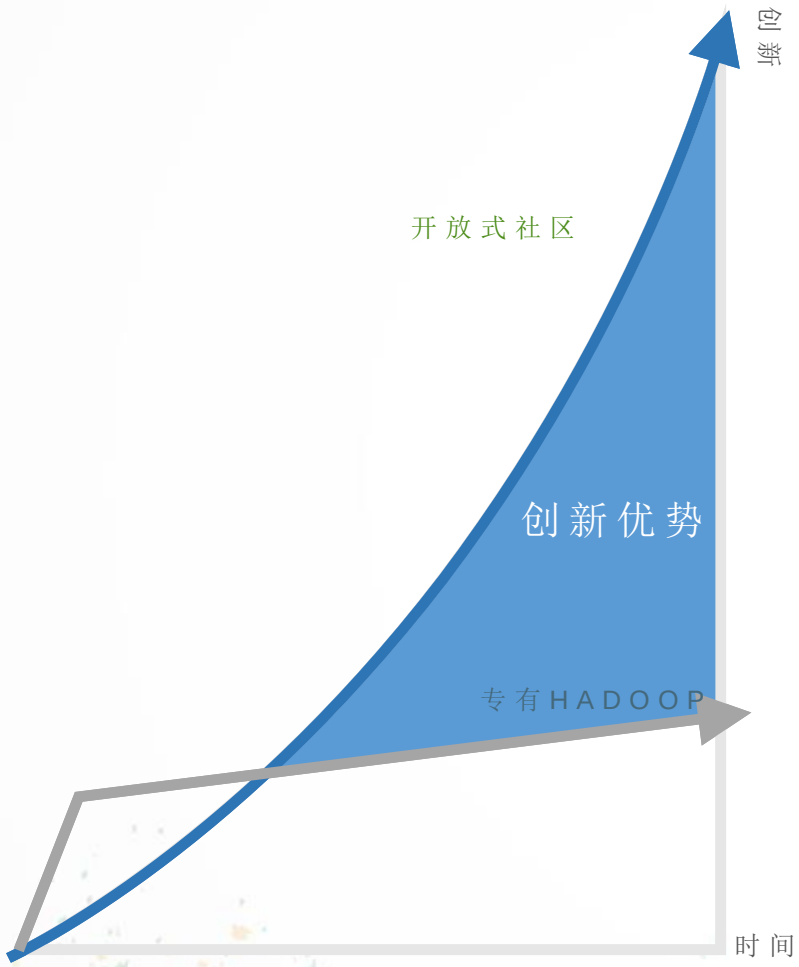
HDP的Apache项目

HDP 2.4 包含且仅包含了27个ASF项目

核心 分析 数据处理 其他

处理 管理 安全

Hortonworks100%的开放



社区创新最大化

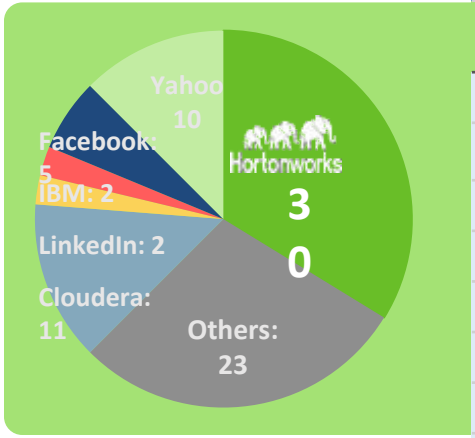
- 消除风险
  - 通过提供100% Apache开源技术避免锁定供应商风险。
- 社区创新最大化
  - 逾千家公司的逾千名开发者
- **无缝集成**
  - 与其他业内领先的技术合作，共同进行工程开发

Hortonworks的开源力量

TERADATA UNIVERSE BEIJING MAY 6 2016

Hortonworks的支持团队工作在 开源社区上，与Hadoop架构师、建设者和操作员直接互动。

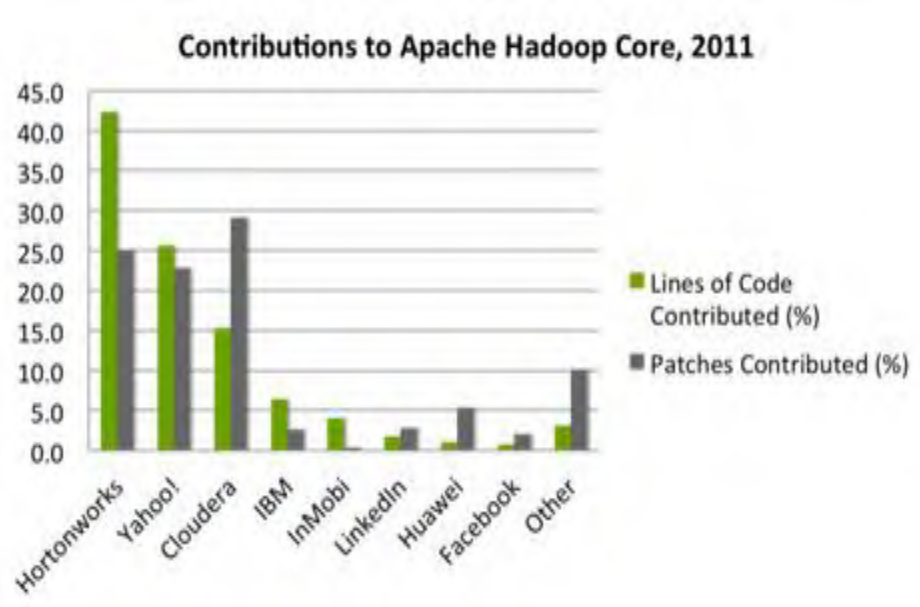
- 最优秀的团队解决最复杂的Hadoop问题
- 全球丰富的Hadoop经验
- 作为领军者和创新者，只有Hortonworks能确保最新开放源代码产品的成功
- 与全球社区共同产生Hadoop路线图

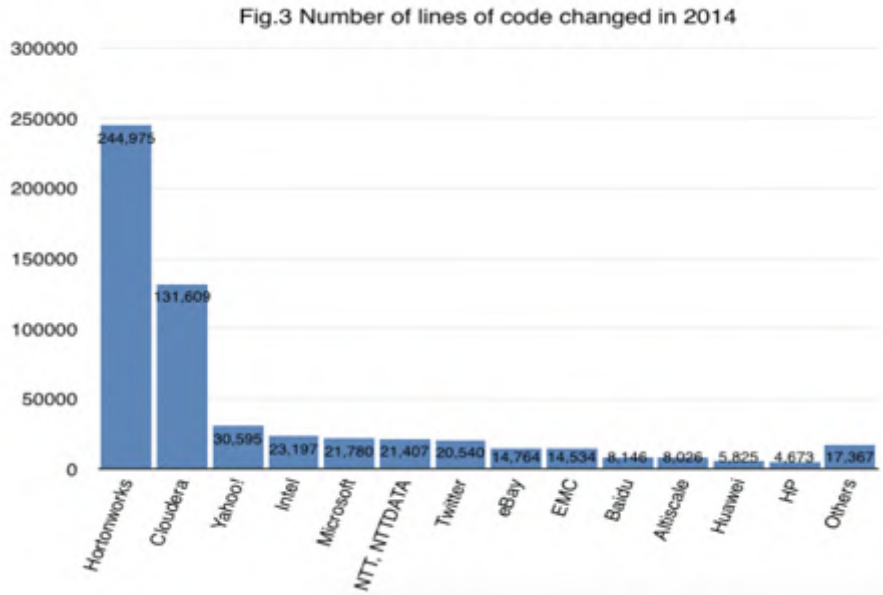


Apache Project	Committers	PMC Members
Hadoop	30	25
Pig	5	5
Hive	18	6
Tez	16	15
HBase	6	4
Phoenix	4	4
Accumulo	2	2
Storm	3	2
Slider	11	11
Falcon	5	3
Flume	1	1
Sqoop	1	1
Ambari	36	28
Oozie	3	2
Zookeeper	2	1
Knox	13	3
Ranger	11	n/a
Spark	2	n/a
TOTAL	169	113

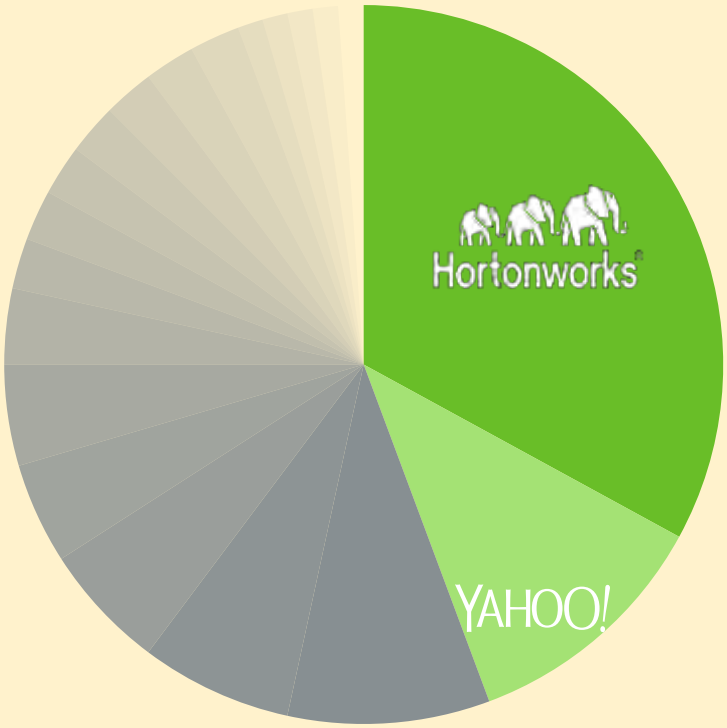
THE APACHE SOFTWARE FOUNDATION

Hortonworks社区的代码贡献率





Hortonworks引领着Apache社区



APACHE HADOOP 提交者们

我们雇佣内容提交者。

-- 三分之一的Apache® Hadoop™ 项目内容提交者, 以及其他重要项目中的大部分。

我们的提交者富于创新

同时创新并扩展开放式企业Hadoop和 Apache NiFi

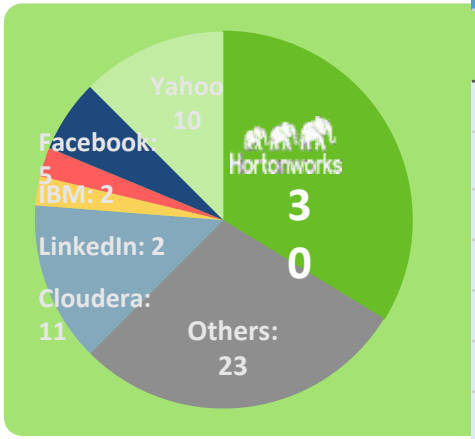
我们影响着Hadoop的产品演变

通过我们在行业内的领导背景，我们跟社区就关键要求保持沟通。

Hortonworks的开源理念

Hortonworks的支持团队工作在 开源社区上，与Hadoop架构师、建设者和操作员直接互动。

- 与社区共同成长
- 注重分享意识，强调共赢
- 最优秀的团队解决最复杂的Hadoop问题
- 重点创新,广泛整合
- 与全球社区共同产生Hadoop路线图。



Apache Project	Committers	PMC Members
Hadoop	30	25
Pig	5	5
Hive	18	6
Tez	16	15
HBase	6	4
Phoenix	4	4
Accumulo	2	2
Storm	3	2
Slider	11	11
Falcon	5	3
Flume	1	1
Sqoop	1	1
Ambari	36	28
Oozie	3	2
Zookeeper	2	1
Knox	13	3
Ranger	11	n/a
Spark	2	n/a
TOTAL	169	113

ODP开放标准数据平台

TERADATA UNIVERSE BEIJING MAY 6 2016

议程

- Apache开源社区
- Hortonworks的介绍
- Hortonworks的开源理念
- 案例分享
- 问答环节

Hortonworks:多维度欺诈检测

项目：美国跨国银行与金融服务公司

客户：服务超过5000万客户和小企业

200K +员工

问题：

✓传统的欺诈检测模型只考虑金融事件

✓评分模型大部分是基于规则的和已发生的事件信息

✓正确的对客户的综合的角度评价需要很多非金融事件信息

✓传统的存储和平台架的限制了基于大规模历史数据的规则模拟

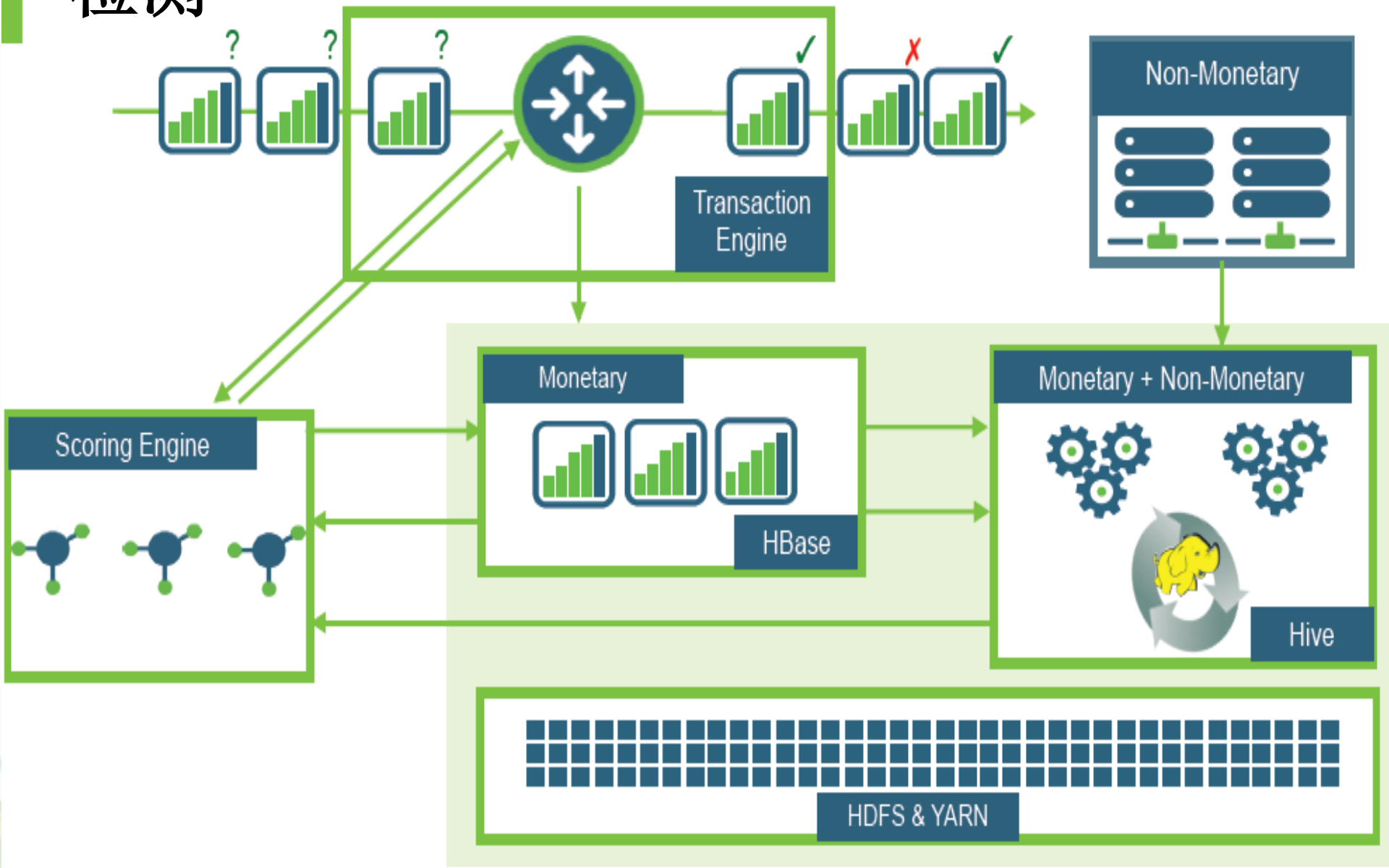
解决方案：

✓金融和非金融性事件对客户进行多角度综合历史视图，进行闭环分析，不断提高检测精度

✓基于历史活动数据规则模拟测试

✓机器学习可以基于历史活动数据对欺诈活动做进一步分析

Hortonworks: 多维度欺诈检测



议程

- Apache开源社区
- Hortonworks的介绍
- Hortonworks的开源理念
- 案例分享
- 问答环节

Apache社区投入是否值得

- 开源社区带来的收益
- 快速创新，快速推出产品
- 共享技术，回报社区，提升能力
- 吸引人才
- 吸收整个社区的贡献
- 提高软件质量
- 提升社区的影响力，引导开源技术标准

私有化Hadoop的误区

- 私有化Hadoop技术的误区
- 很多Hadoop发行商借助Apache社区的技术,发展某些自己私有化的技术,在某些应用下确实能够解决一部分问题,但是这种私有化的技术由于只有自己去发展,技术会越来越重,而且越来越少人支持,所以离Hadoop的主要发展会越来越远

Hortonworks

谢谢！

2016/5/6