iResearch
艾 瑞 咨 询

2024年开源大数据行业发展洞察报告

CONTENTS

目 录

01/ 大数据开源工具发展背景

大数据技术的行业应用

大数据技术应用广度与深度持续加大，成为决定企业竞争力的重要因素

十多年来，随着大数据技术的演进与成熟，其在经济领域中的应用也在拓展并持续深化。目前，在包括医疗保健、零售、金融服务、制造业、电信、能源与公共服务的各主要行业中，大数据技术在精细管理、趋势预测、风险识别、决策支持等场景中发挥着越来越重要的作用。数字时代背景下，数据已成为企业核心资产，而大数据技术则是对这项资产开发，利用，赋能企业的重要手段，越来越多的企业认识到用对、用好大数据技术将决定自身的行业竞争力。

大数据技术在各主要行业中的典型应用场景

医疗保健

预测分析用于病人护理：预测病人入院情况，优化资源分配
临床决策支持：通过数据聚合增强治疗建议
人群健康管理：分析数据以跟踪疾病爆发并针对性干预

零售

客户个性化检视：根据购买历史定制营销活动
库存管理：通过准确预测需求优化库存水平
价格优化：利用竞争者分析和市场分析动态定价产品

金融服务

欺诈检测：监控交易以识别和防止欺诈
风险管理：通过全面数据分析增强信用评分
客户细分：针对性分析客户，开发有针对性的产品

制造业

预测性维护：预测设备故障以减少停机时间
供应链优化：利用数据洞察改善物流和需求预测
质量控制：实时监控生产以确保产品质量

电信

客户流失预测：识别不满意的客户以降低流失率
网络优化：分析流量以更好地分配网络资源
欺诈预防：检测账单和使用数据中的异常情况

能源与公用服务

智能电网管理：通过需求预测改善负载平衡
预测性资产维护：安排维护以防止停电
可再生能源预测：优化可再生能源的接入电网

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

大数据工具的开源

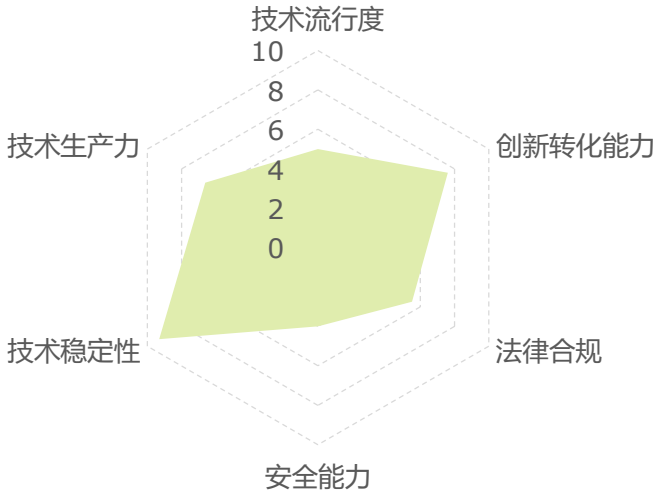
开源趋势下，大数据传统工具已经成熟，个性化新型工具不断加入

狭义上的开源大数据工具是指在开源大生态下，专注于解决海量、多类型数据的连接、存储、管理等功能的工具集合。但从搭建大数据平台角度出发，通常还需要加入AI类组件以帮助数据分析，云原生工具以实现容器编排，另外关系型及各类非关系型数据库被视为大数据的基础，由此得到广义上的大数据工具套件。本报告将以广义大数据工具为研究对象，对其进行分析。

开源生态下狭义与广义大数据工具

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

大数据技术领域开源生态成熟度雷达图



- 大数据领域具备较好的**技术稳健性**，以Hadoop、Spark、Flink等为代表的传统大数据产品已趋于成熟
- **新型开源大数据工具不断向个性化、定制化发展**，如大数据框架中加入AI类库，以及如Uber、Netflix、Spotify等企业根据自身特定业务贡献新的适用于具体应用场景的大数据工具

来源：中国信通院云计算开源产业联盟，中国通信标准化协会《全球开源生态洞察报告（2024年）》，艾瑞咨询研究院自主研究及绘制。

开源大数据工具的分类及功能

按功能类型分为5层11模块，合理的工具选型是搭建大数据平台的前提

大数据工具组件是大数据技术输出的载体，数字化与智能化时代下，一套完整的大数据工具可以分为基础层、数据连接层、编排与分析层、人工智能层、监控及可视化层共5层，包括储存格式、数据框架，数据库、数据管理、数据查询与连接、流处理与消息管理、数据编排、在线分析、机器学习运维、记录及监控、数据可视化11个模块。
大数据工具层级图是对大数据工具的总览，开源工具林林总总，企业应先解各个工具的定位与功能，根据自身需求牟定工具类型，再进行具体工具的选型。

开源大数据工具层级图

5 监控及可视化层

- 可视化展现系统健康情况，帮助理解数据洞察结论

记录及监控	数据可视化
监控数据基础设施健康情况，追踪指标表现	将数据洞察可视化展现

4 人工智能层

- 为以机器学习为代表的各类AI的运行提供基础服务

机器学习运维
自动化部署、监控及管理机器学习模型的运营平台

3 编排及分析层

- 实现大数据的自动化、实时处理及分析

数据编排	在线分析
数据流自动化编排、监控并处理，实现数据有效利用	大数据集的快速聚合、查询与实时分析

2 数据连接层

- 管理数据的连接、流动、查询等数据治理任务

数据管理	数据查询与连接	流处理与消息处理
数据组织，版本控制，数据治理	数据集间的查询、整合、控制	实时调控数据管道及事件流

1 基础层

- 数据存储、数据处理，确保基础设施的可扩展性

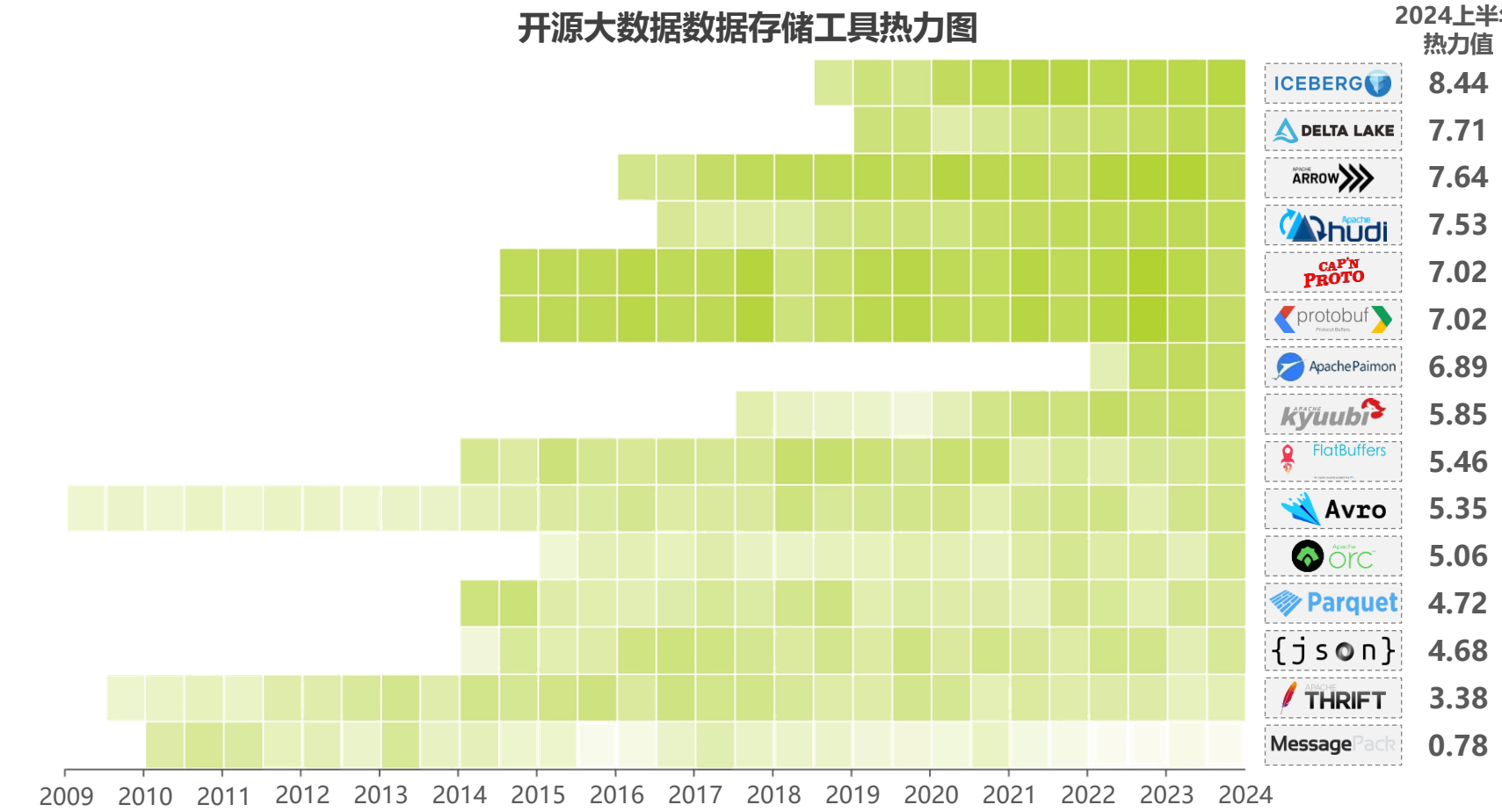
数据存储	数据框架	数据库
指导数据如何存储及序列化	为分布式数据处理提供核心能力	结构及非结构化数据主存储系统

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

02/ 大数据开源工具热力趋势

热力趋势（1/12）：数据存储

沿二进制存储、列存储、云上数据湖的路径演化，多样化容纳数据类型



① 二进制和结构化格式
- 针对数据序列化进行优化
- Avro, Thrift, Protocol Buffers

② 列存储格式
- 适应重任务下的数据分析查询
- Parquet, ORC

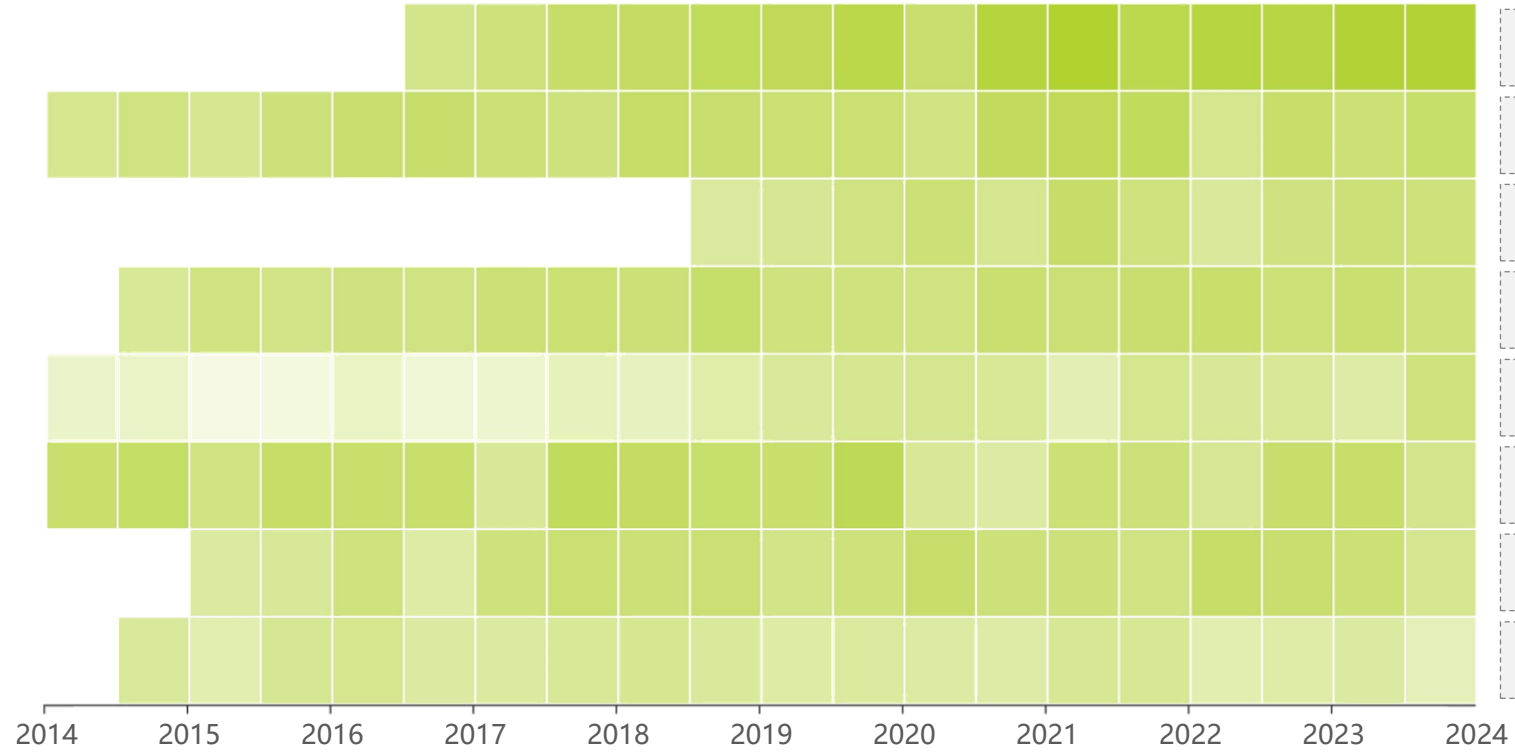
③ 云原生数据格式
- 云上数据湖
- Delta Lake, Iceberg, Hudi

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（2/12）：框架

大数据框架随数据量的扩大以及处理速度需求提升而迭代；进入大模型时代，大数据框架进而整合模型开发组件

开源大数据框架热力图



	2024上半年热力值
RAY	8.80
Apache Flink	6.50
yarn	5.78
APACHE hadoop	5.75
HELIX	5.63
APACHE Spark	4.97
dask	4.84
CDAP	3.05

① 分布式计算及存储
- 批处理
- Hadoop：HDFS+Mapreduce

② 实时计算、内存计算
- 流处理、批流一体
- Spark, Flink, Storm

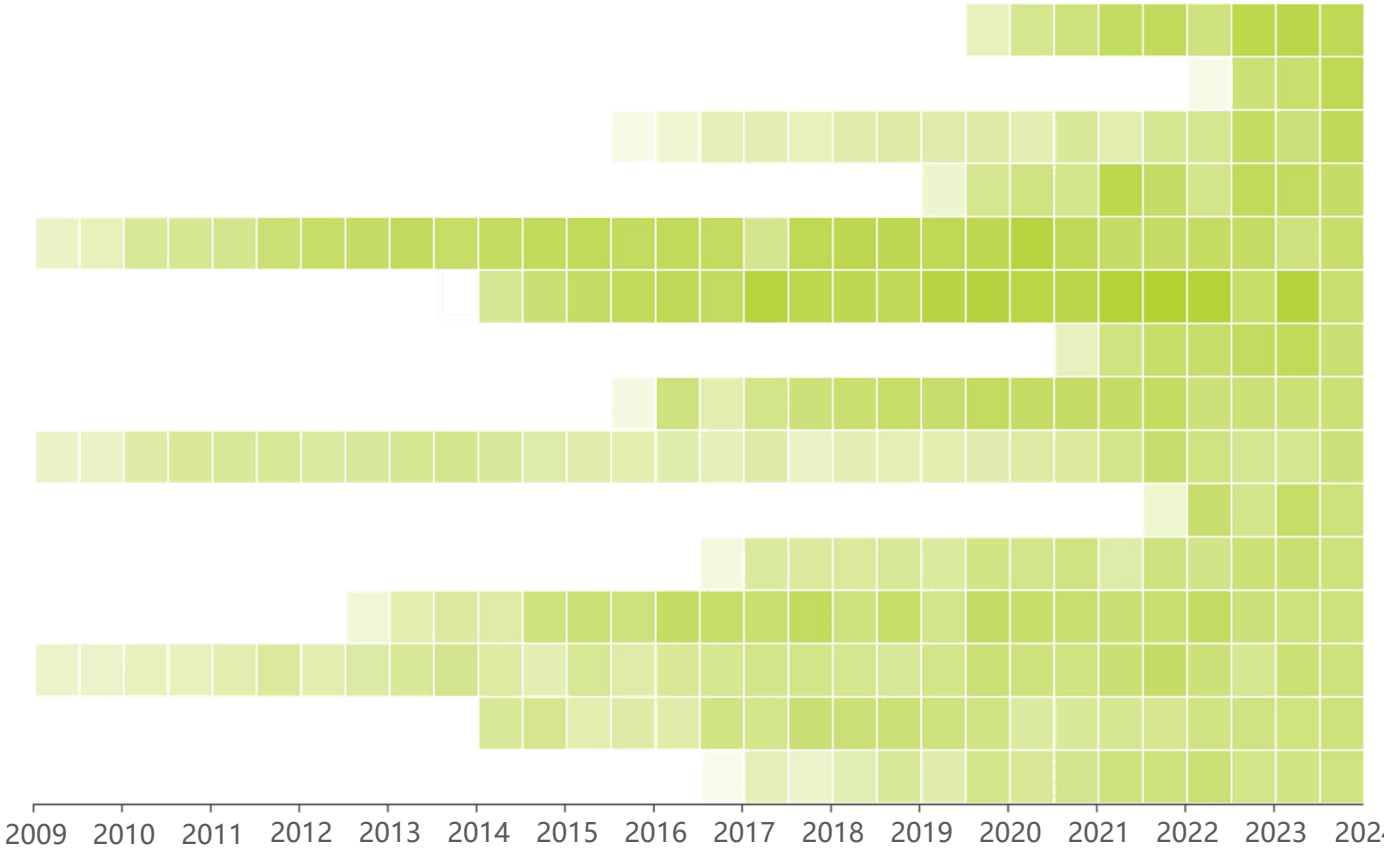
③ AI函数库
- 支持模型训练、微调
- Ray, MLlib(Spark)

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（3/12）：数据库 - 之一

数据库种类逐渐丰富，支持云原生、大模型开发训练及实时分析

开源大数据数据库热力图（1-15）



- ■ 非关系型数据库
- ■ 管理处理半结构、非结构型数据
- ■ Cassandra, MongoDB, HBase

① ② ③

- ■ 云原生数据库
- ■ 为基于云的高性能数据分析优化
- ■ CockroachDB, TiDB

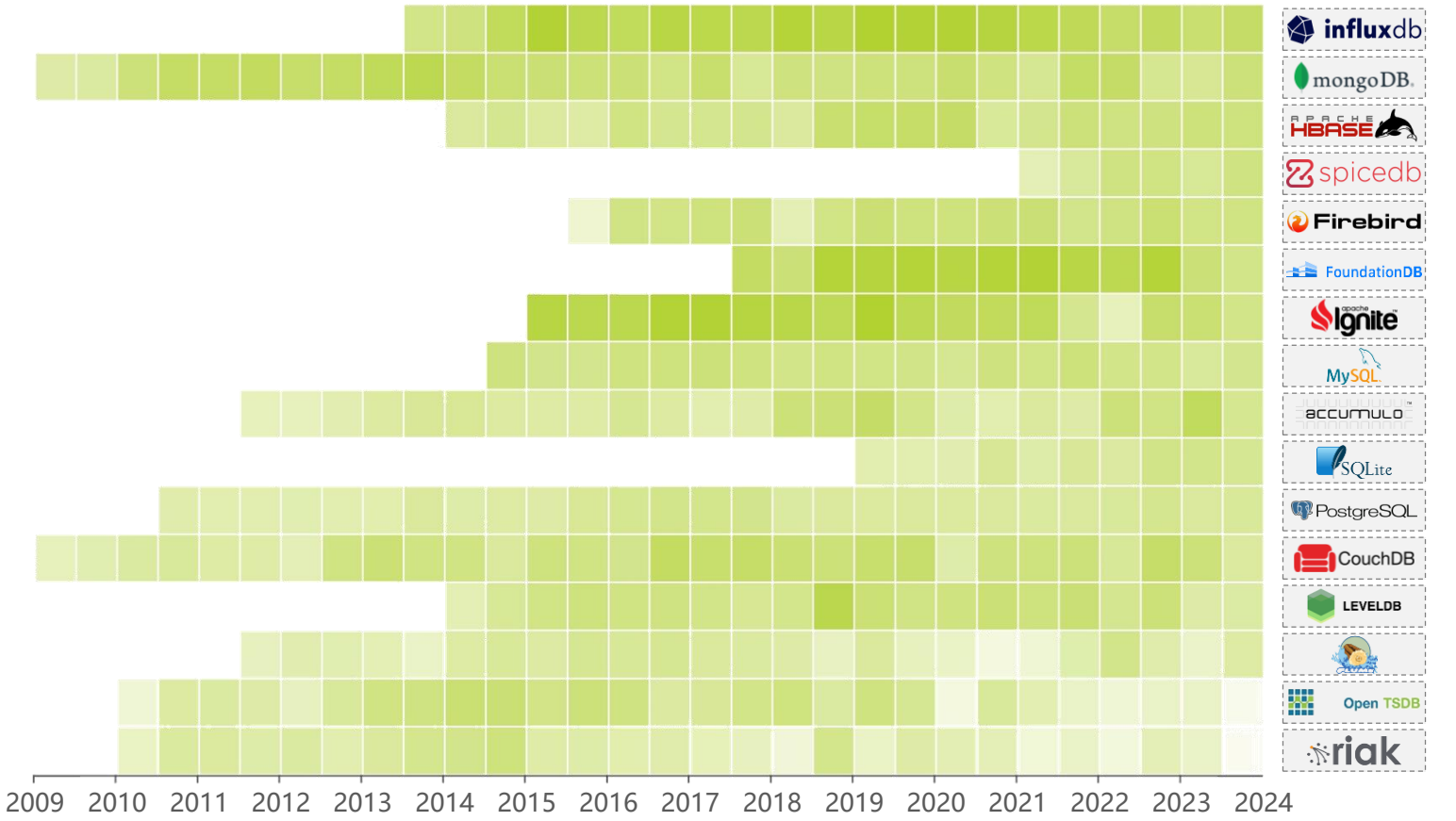
- ■ AI相关——向量数据库
- ■ 高效管理、查询嵌入向量
- ■ Milvus, Weaviate

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（3/12）：数据库 - 之二

数据库种类逐渐丰富，支持云原生、大模型开发训练及实时分析

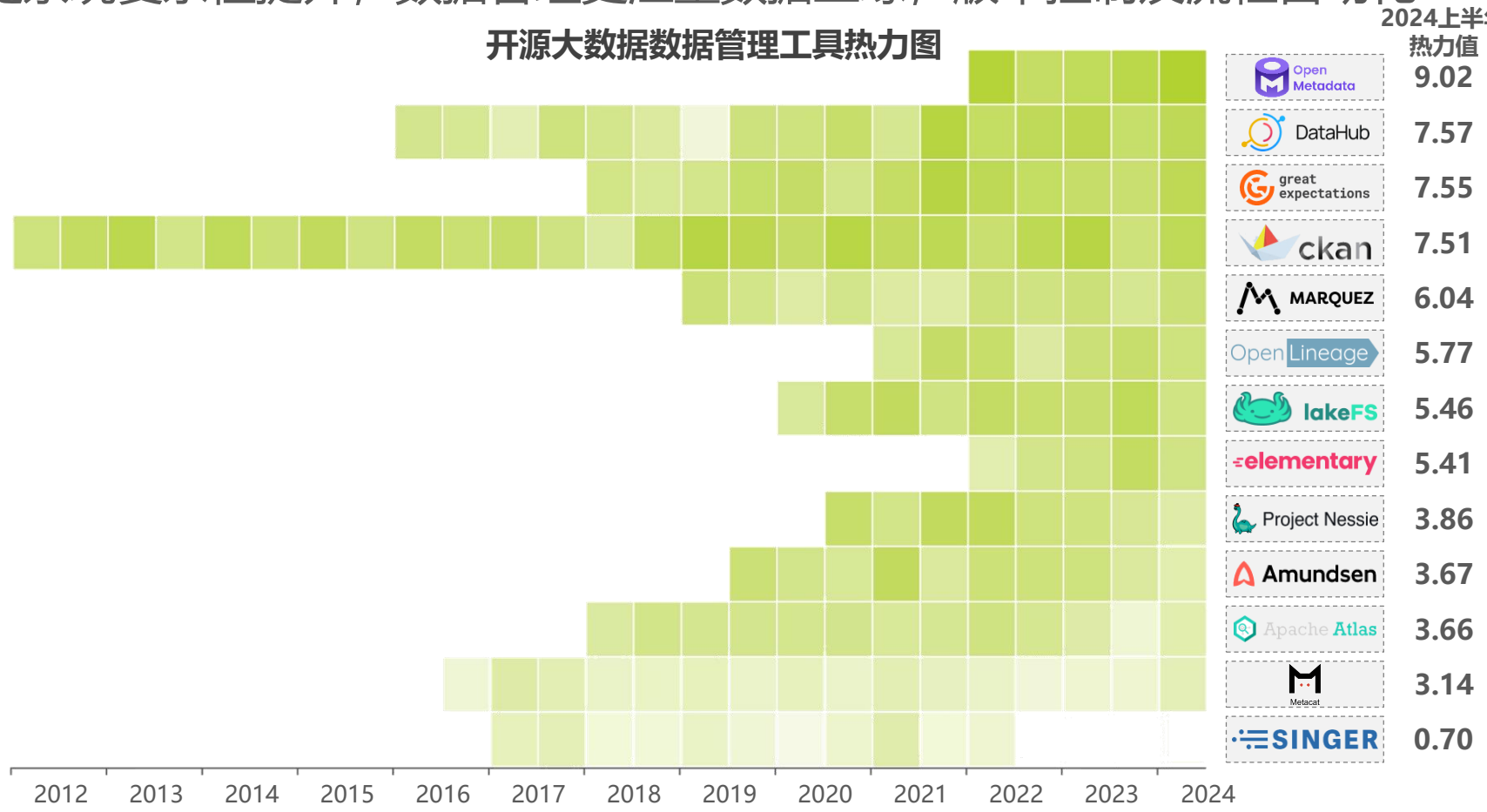
开源大数据数据库热力图（16-31）



来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（4/12）：数据管理

随系统复杂性提升，数据管理更注重数据血缘，版本控制及流程自动化



- ■ 数据目录及数据治理
- ■ 快速精准查找、正确使用数据资产
- ■ CKAN, Metacat

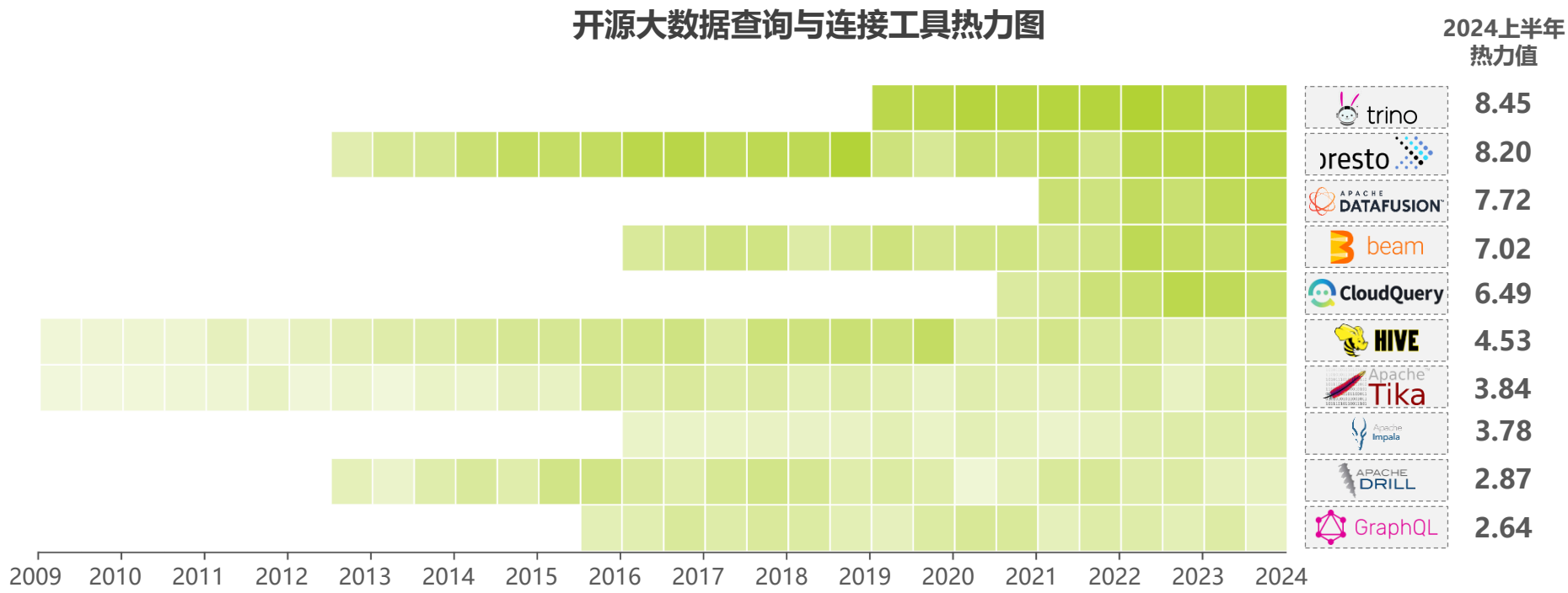
① → ■ 元数据治理，数据血缘 ■ 洞察数据关系，数据价值挖掘 ■ Amundsen, DataHub, Atlas

② → ■ 数据质量/一致性保证、版本控制 ■ 自动化验证，可回溯 ■ Great_Expectations, LakeFS ③

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（5/12）：查询与连接

从批量到实时，从单一数据源到跨系统多元数据，从关系型数据到非关系型数据，工具的进化让数据查询更迅速、更灵活、更丝滑



① 基于Hadoop框架的大数据查询
使用SQL语句进行低延时批量查询
Hive, Pig, Presto

② 对于分布式数据的快速查询做优化
实时查询，实时分析
Druid, Impala

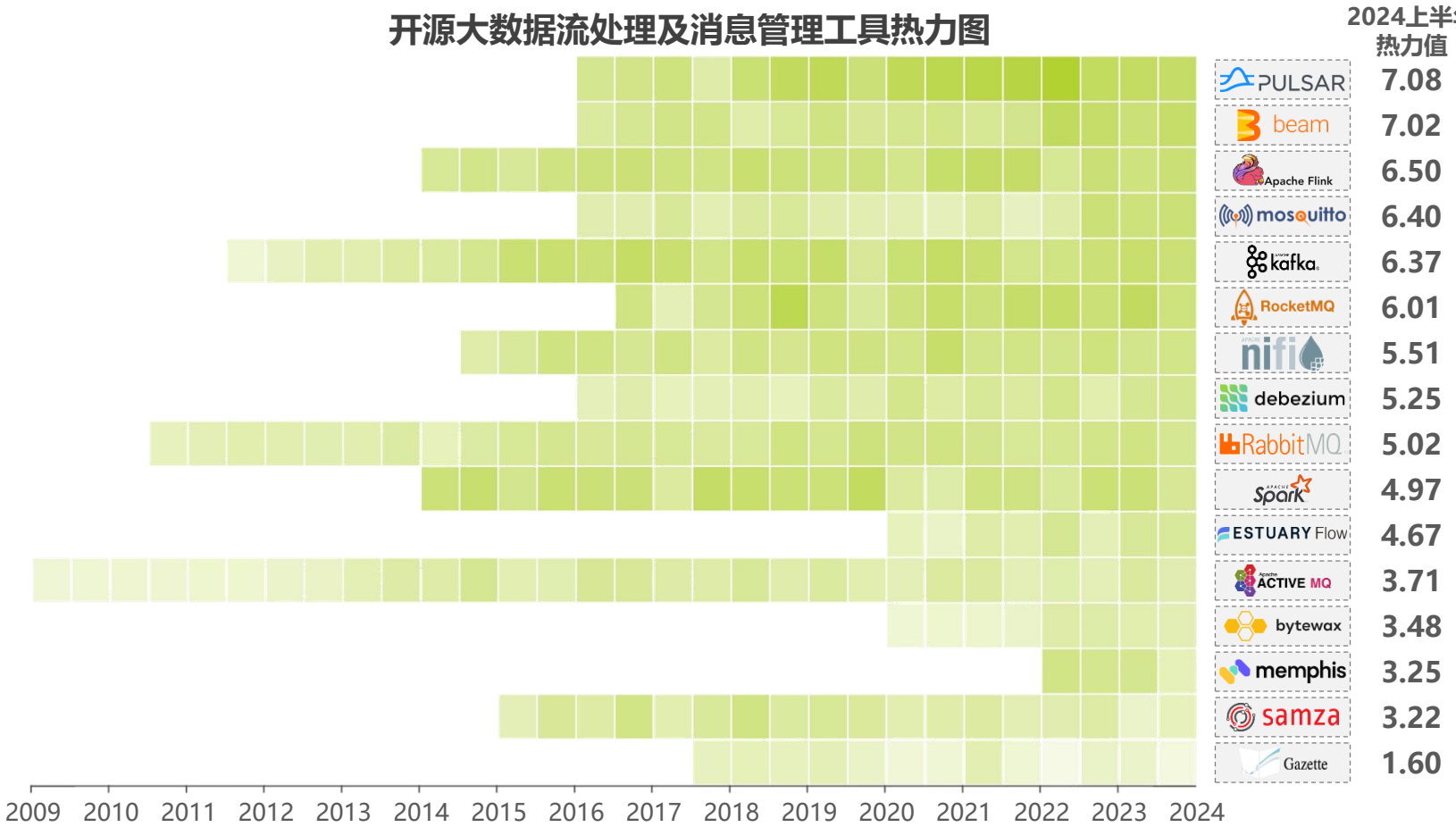
③ 多数据源多数据类型统一联合查询
使用一套查询语句及统一界面
Beam, Trino, Drill

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（6/12）：流处理及消息管理

由简单的消息处理功能发展为功能复杂适应混合场景的数据管理工具

开源大数据流处理及消息管理工具热力图



- ■ 消息队列、订阅/发布、日志聚合
- ■ 简单消息系统中处理少量实时数据
- ■ RabbitMQ, ActiveMQ

① →

- ■ 分布式架构
- ■ 实时数据+高吞吐量+低容错率
- ■ Kafka, NiFi, Debezium

② →

- ■ 云原生、事件驱动架构
- ■ 混合负载+多租户+地域复制
- ■ Pulsar, Memphis

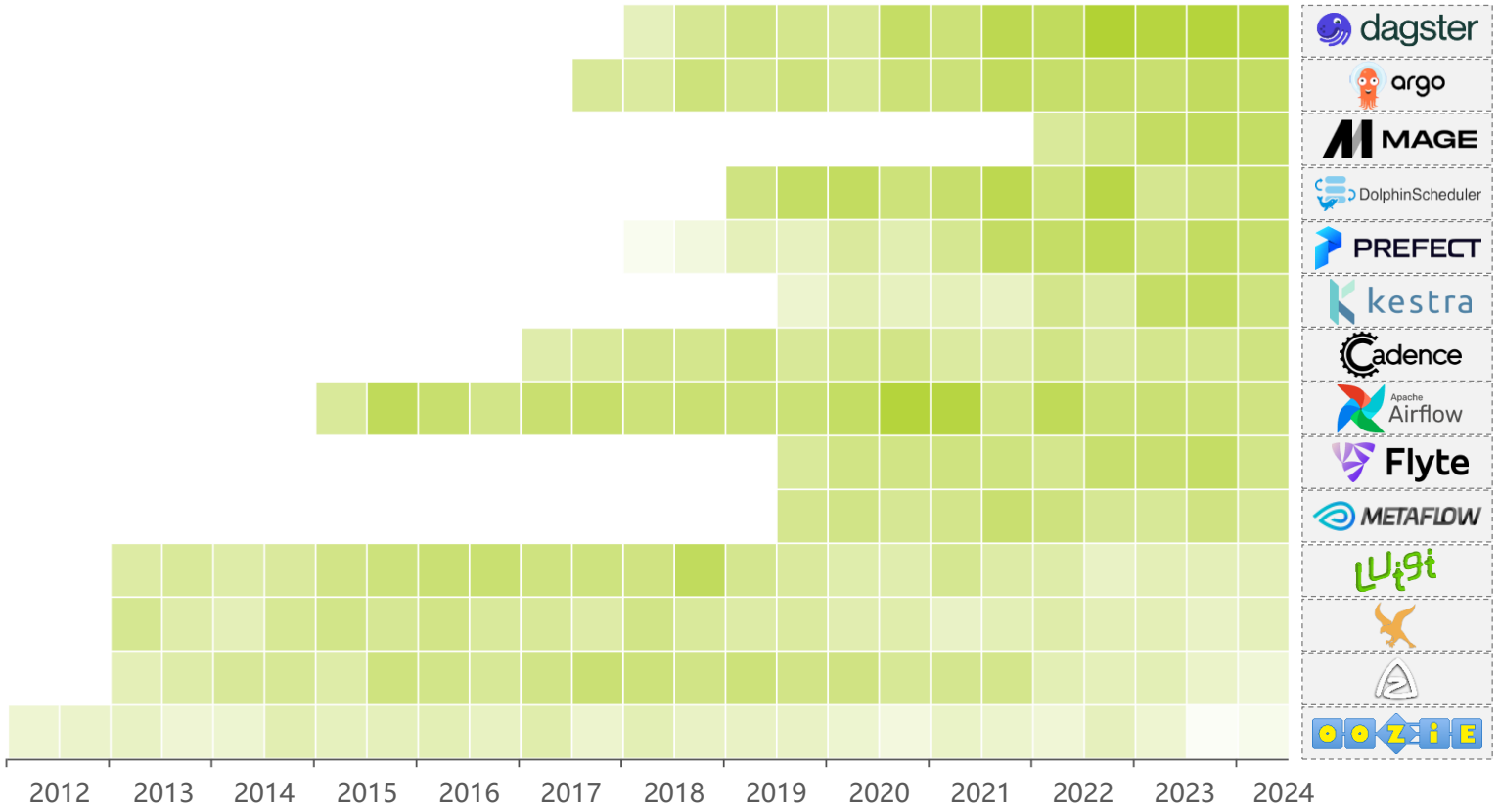
③

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（7/12）：编排

大数据编排工具的演变反映了数据工作流不断变化的需求和复杂性

开源大数据编排工具热力图



- ■ 批处理过程、简单任务依赖
- ■ 集中式调度器管理任务的执行
- ■ Luigi

① → ■ 基于有向无环图构建任务关系
- ■ 模块化架构并与云服务集成
- ■ Airflow, argo

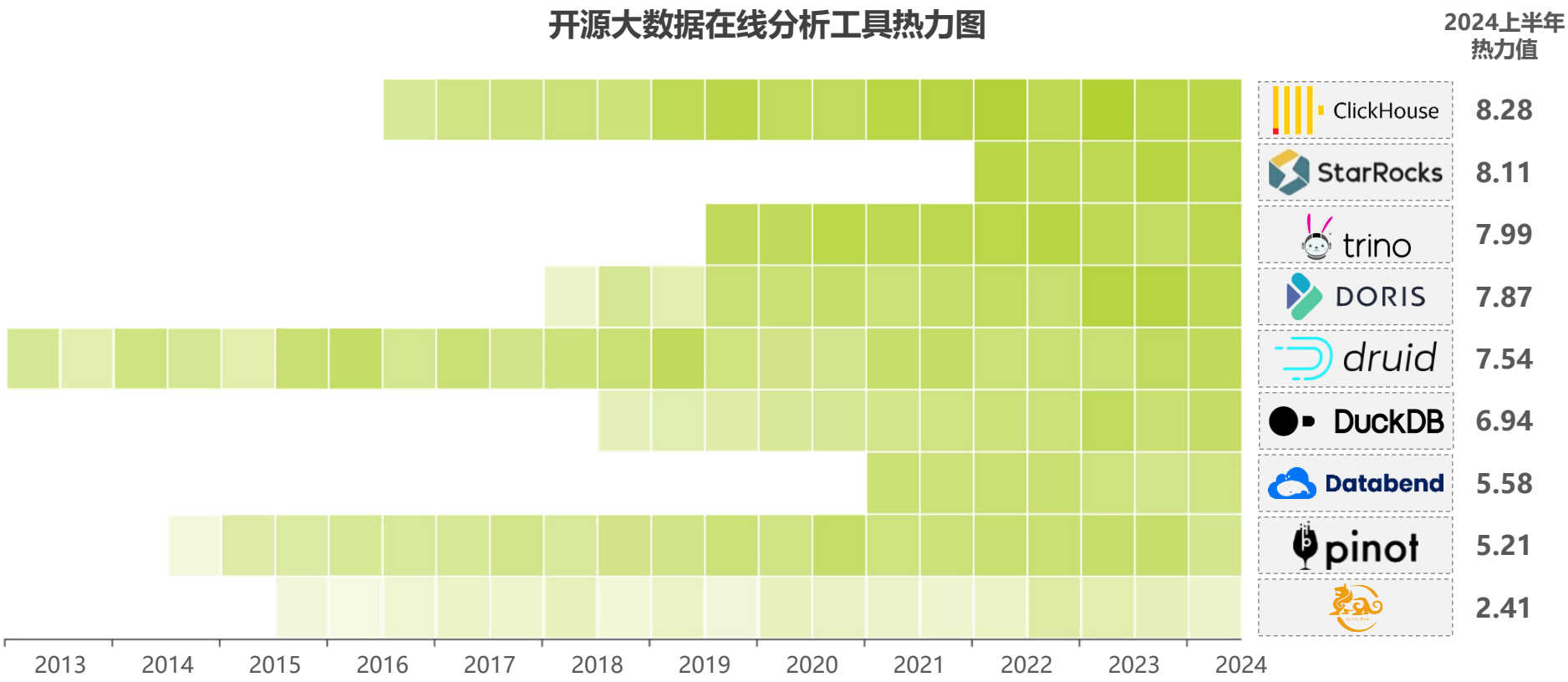
② → ■ 将数据管道视为软件资产
- ■ 数据血缘追踪，推动团队协作
- ■ Dagster, DolphinScheduler

③

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（8/12）：在线分析

由对数据的批量抓取分析发展为云原生可处理高并发的实时数据分析



① 查询处理结构化、预聚合数据；准实时抓取查询数据，分布式结构；Druid, Pinot, Kylin

② 简化查询处理过程，实时动态分析；列存储，矢量化执行；ClickHouse, Trino, Doris

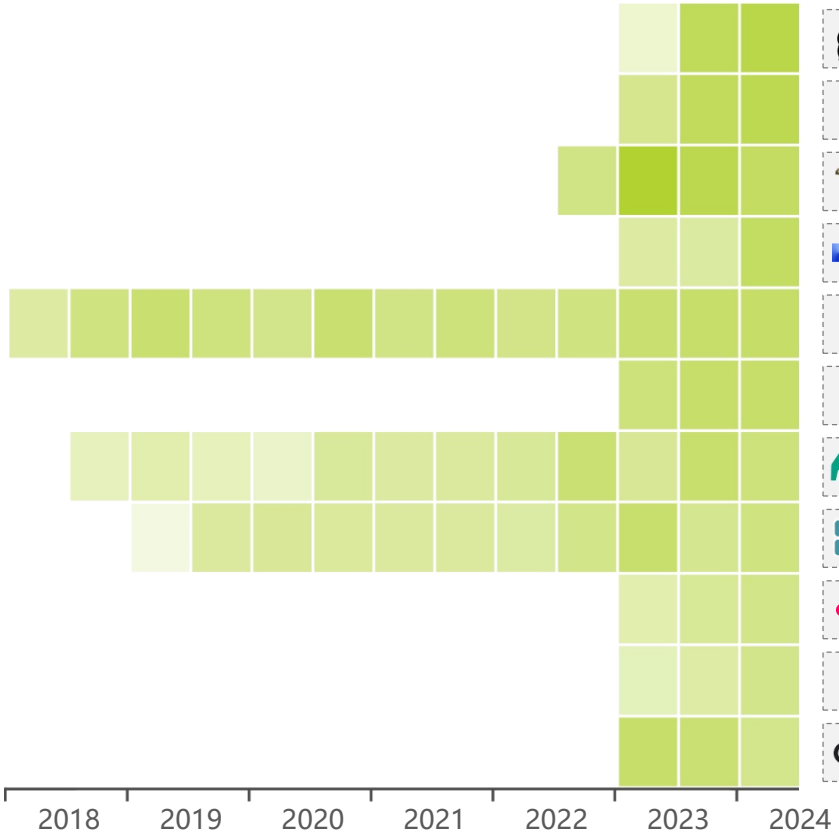
③ 云原生架构，内存计算；实时高并发数据分析；StarRocks, Databend, DuckDB

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（9/12）：机器学习运维 - 之一

由基础开发生命管理发展为以AI专有性能指标为核心设置的工具生态体系

开源大数据机器学习运维工具热力图（1-11）



	2024上半年热力值
ollama	8.81
vLLM	8.44
LangChain	7.65
Dify. Define & Modify Do It For You	7.59
mlflow	7.29
FlowiseAI	7.17
mindsdb	6.41
MediaPipe	6.13
Chainlit	5.77
Gorilla	5.62
PRIVATE GPT	5.54

① 基础模型开发跟踪、可视化及部署
- 未与云融合，编排与自动化能力有限
- Mlaflow, DVC, Pachyderm

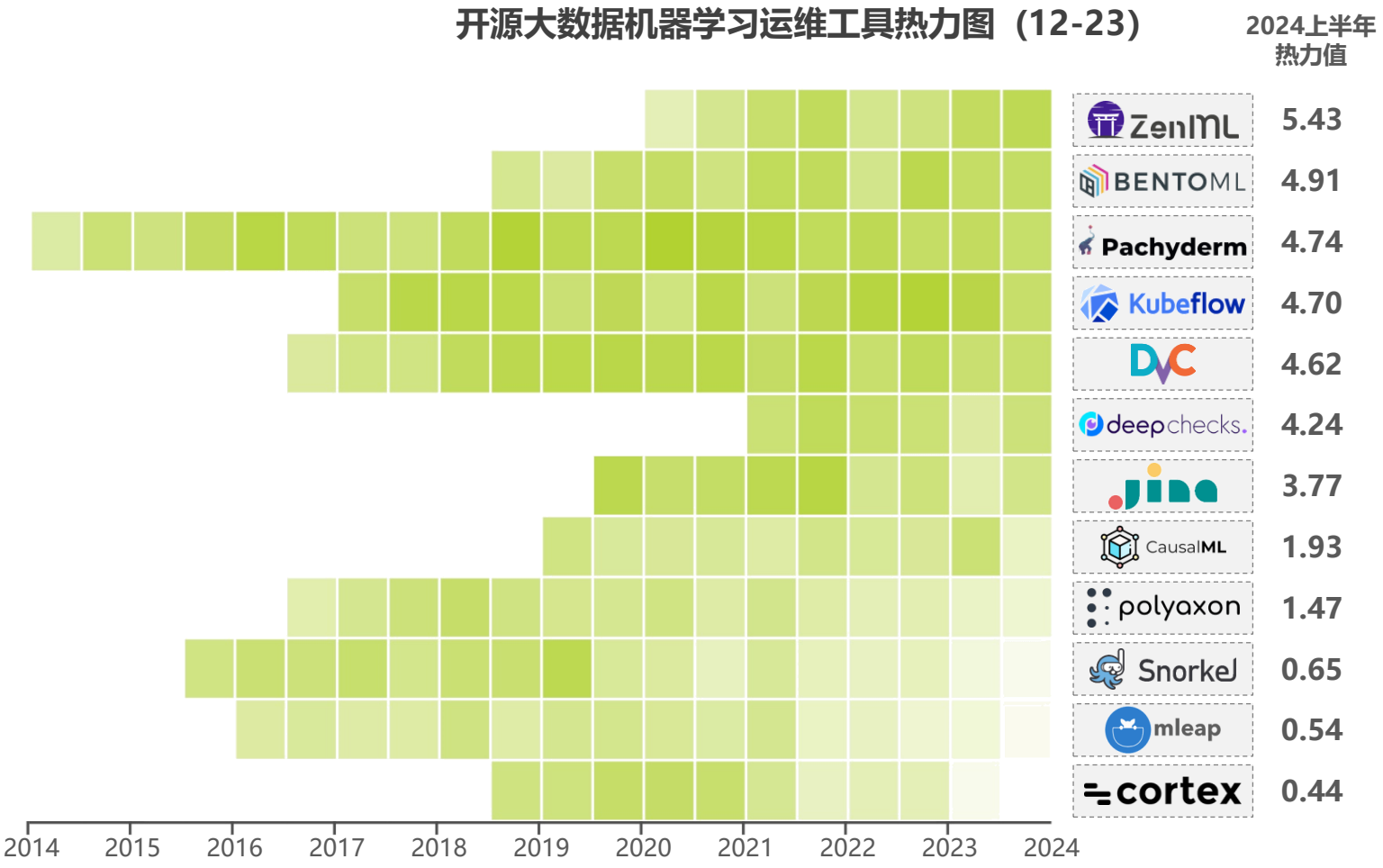
② 端到端的ML流程编排与自动化
- 支持本地与云环境
- Kubeflow, Polyaxon

③ 实时模型服务，AI优先功能：可解释性、公平性、漂移检测
- BentoML, ZenML, Ollama

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（9/12）：机器学习运维 - 之一

由基础开发生命管理发展为以AI专有性能指标为核心设置的工具生态体系

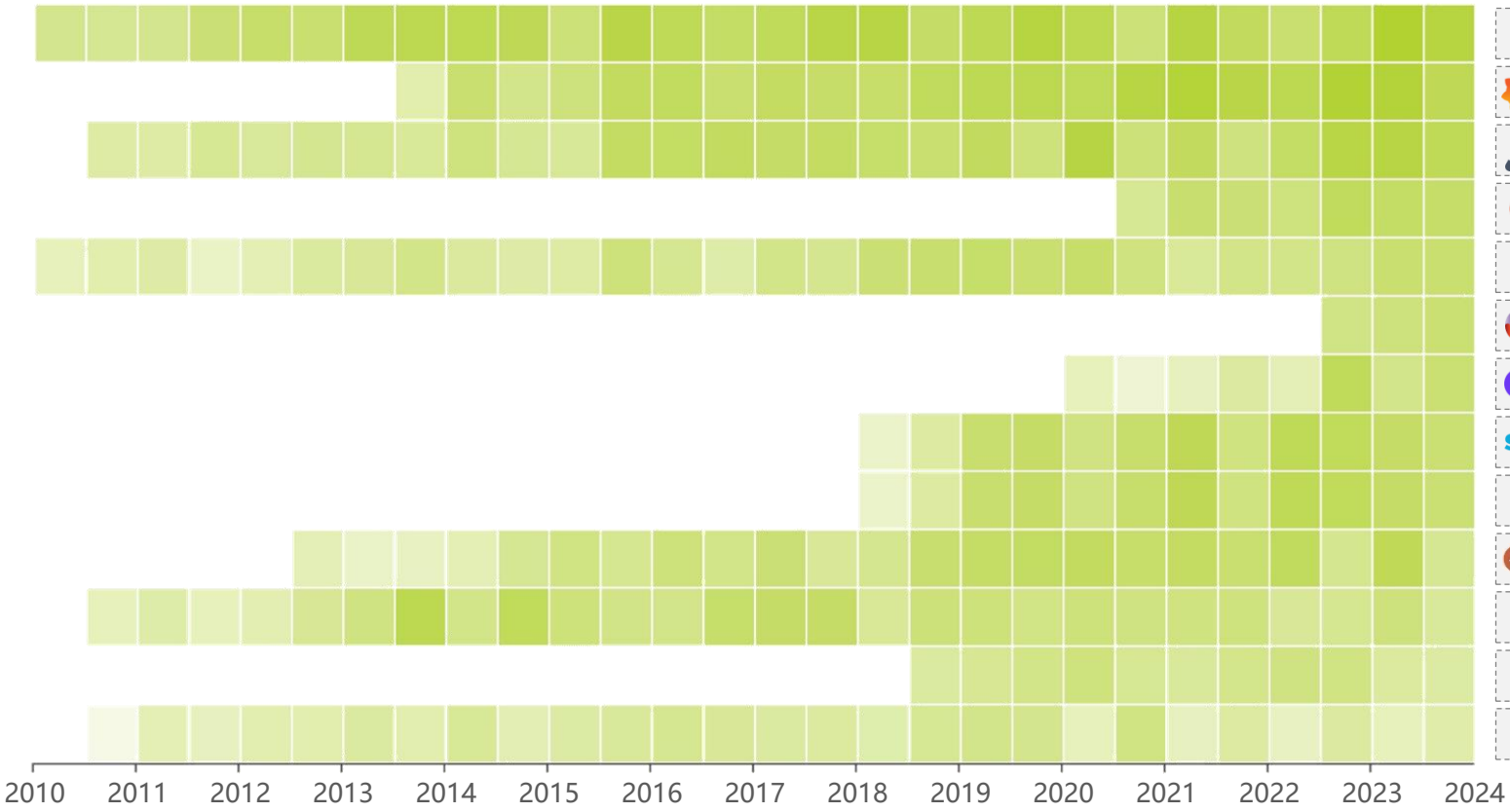


来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（10/12）：记录与监测

由简单的日志管理及可视化发展为集日志、指标、追踪为一体数据观测栈

开源大数据记录与监测工具热力图



① 集中式日志管理与分析
- 提供日志搜索能力及可视化界面
- Elasticsearch, Logstash, Graylog

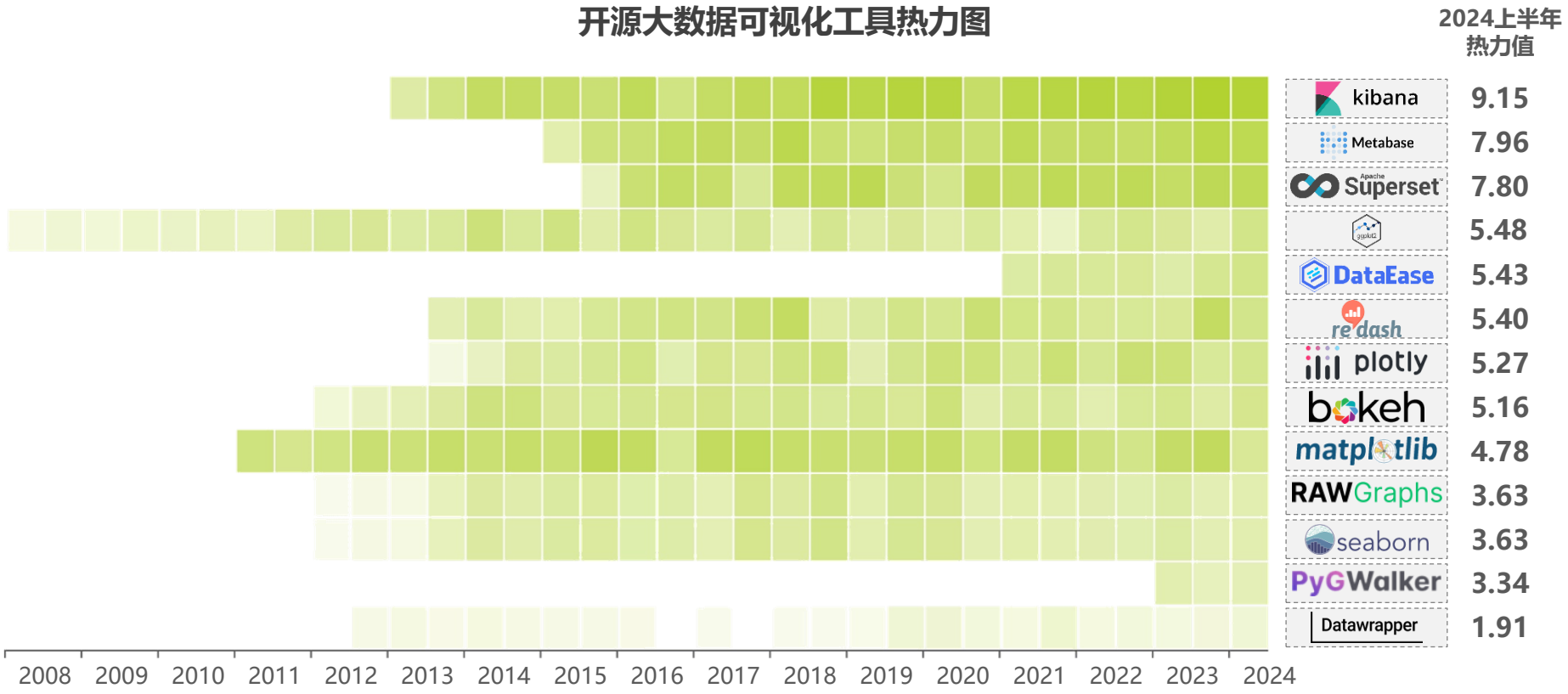
② 构建更强大的指标评估系统
- 实时、主动监测与预警
- Prometheus, Grafana

③ 扩展性更强，效率更优
- 与其他大数据处理组件无缝结合
- SigNoz, OpenTelemetry

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（11/12）：可视化

由静态、本地化解决方案向高互动性、云化、融合AI能力的方向演进



① ■ 静态可视化，基础绘图 ■ 与桌面环境或某些编程语言整合 ■ ggplot2, Matplotlib, Seaborn

② ■ 互动性可视化、仪表盘 ■ 与数据实时互动，基于网络部署 ■ Metabase, Bokeh, Plotly

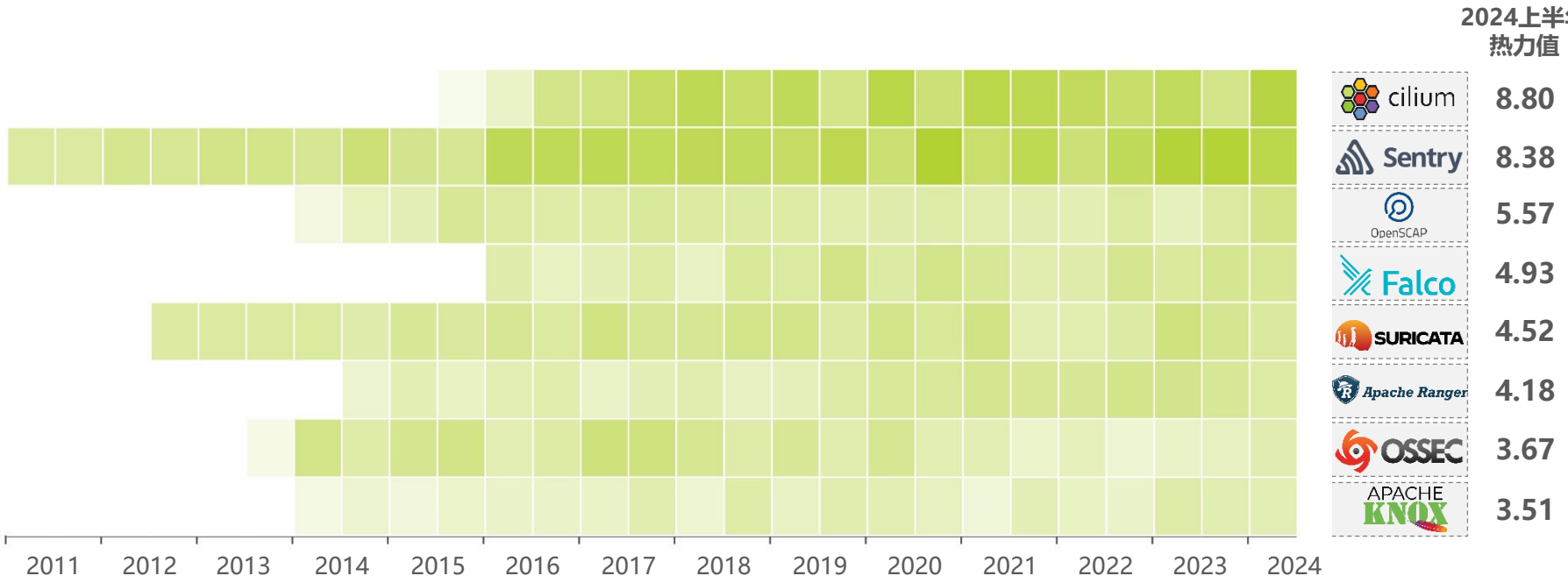
③ ■ 支持多用户协作，加入AI能力 ■ 与大数据架构、数仓深度融合 ■ Superset, Kibana, Redash

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

热力趋势（12/12）：数据安全

从基础安全和监控能力发展到高级威胁检测，最终实现全面的访问管理和数据治理

开源大数据安全组件工具热力图



① 日志分析与事件关联
■ 基本的入侵检测能力（日志监控）
■ OSSEC, Sentry

② 实时威胁检测和响应
■ 网络流量的深度包检测
■ Falco, Suricata

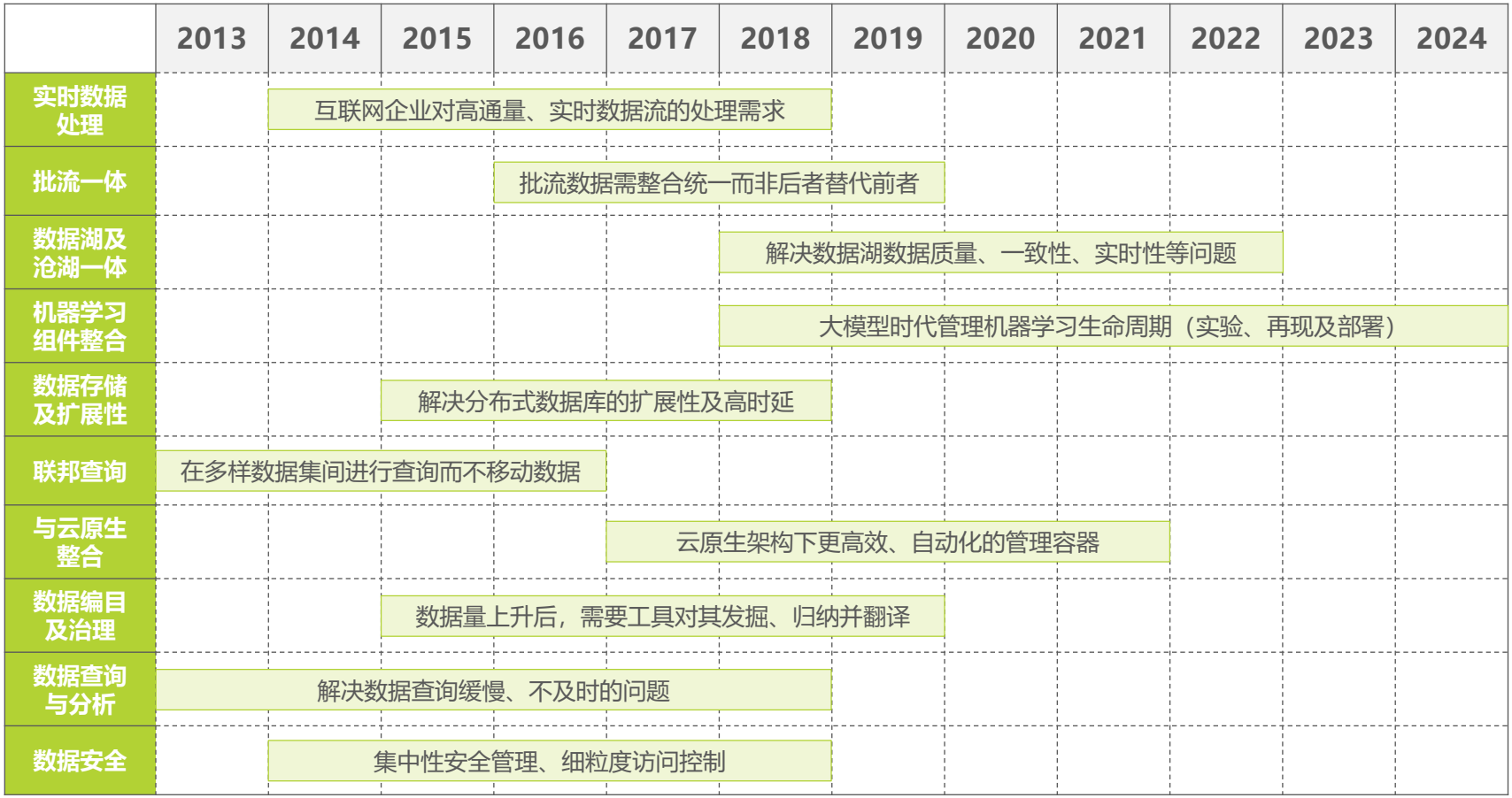
③ 细粒度的访问控制策略
■ 集中的安全策略管理
■ Cilium, Ranger, Knox

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

开源大数据工具热力趋势总结

由于不同时期的技术挑战与应用需求促使大数据工具的迭代与丰富

开源大数据工具发展时间图



来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

云厂商开源大数据工具支持度比较

基础设施覆盖度、云计算成本及效用以及开源配套服务是影响客户在利用开源工具自建大数据平台时选型底层云平台的主要因素

基础设施覆盖度：云厂商更广阔的基础设施覆盖度意味着客户在进行大数据处理时的延迟时间更少，并可以选择本地化的部署方式，这对于需要低延时以及数据驻留合规性要求更为严格的国际化用户尤为重要。

云计算成本与效用：大数据的处理需要耗费海量计算资源，因此计算效率与成本效益是客户的重要考量因素。定制化核心基础硬件能够从底层增强云计算效率，从成本及能耗角度看也会带来显著提升。

开源配套服务：云平台对于开源大数据工具更广泛的配套服务以及更深度的融合决定了客户利用开源工具构建大数据平台的难易度与开发成本，客户更倾向于使用开源友好度高的云平台服务。

综合比较亚马逊云科技，Azure与GCP三大全球性云厂商，亚马逊云科技在基础设施覆盖的**广度**、云计算优化的**深度**、以及生态中开源配套服务的**丰富度**上均有一定优势，与当下处理复杂数据类型、重分析呈现的大数据热点开发组件契合度较高，是大数据云基础平台的优质选择。

	基础设施覆盖度	云计算成本与效用	开源配套服务
亚马逊云科技	➢ 在34个地理区域内运营108个可用区 ➢ 计划在墨西哥、新西兰、沙特阿拉伯王国、泰国、中国台湾和亚马逊云科技欧盟主权云增加18个可用区和6个亚马逊云科技区域 ➢ 拥有超过410个边缘站点与本地区域	➢ 自研ARM架构Graviton处理器 ➢ 为云原生工作任务高度定制，使亚马逊云科技更具成本效益、更节能、更高效 ➢ 相较于x86芯片，Graviton3可达到60%的能耗提升， Graviton2可达到最高30%的性能提升	➢ 对开源大数据工具提供广泛支持，为主流大数据框架提供托管服务 ➢ 亚马逊云科技生态中的如EMR，MSK等服务与大数据开源工具无缝结合 ➢ 亚马逊云科技兼容各类开源数据仓库与数据湖，通过Glue与Athena可以轻松查询或转换各类开源格式的数据
Azure	➢ 服务范围涵盖包括64个区域在内的共140个国家 ➢ 共有175个边缘节点	➢ 基于Intel及AMD芯片搭建云服务	➢ 通过HDInsight与Databricks将开源大数据工具与Azure整合，为用户处理结构及非结构化数据提供统一平台 ➢ Azure Synapse也将基于Spark的分析原生地整合进来
Google Cloud Platform	➢ 在38个区域中的115个节点运营（包括本地节点与边缘节点） ➢ 正在另外13个区域加紧布局，但总体上在欧美之外的区域布局较少	➢ 主要基于Intel及AMD芯片搭建云服务 ➢ 提供第三方厂商Ampere设计的Altra ARM架构芯片 ➢ 第三方芯片尚未能与自身云平台进行深度整合	➢ BigQuery是完全托管的数据仓库，支持多种开源数据格式，同时支持与开源框架整合进行高阶数据分析 ➢ 通过GKE支持容器化工作负载，在Kubernetes集群上调度开源大数据应用，管理复杂的大数据管道

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

03/ 大数据工具热力值说明

说明（1/2）：热力值意义及数据采集

热力值意义

本报告中所指热力趋势是从开发者视角所做的研究判断，通过对开发者围绕开源社区相关行为的定量分析，综合得到热力值，是开发者对该开源大数据工具的关注、参与、讨论、贡献的综合体现。

因此开源大数据工具的热力值越高，代表该工具能够更快速的迭代，受到更精细的优化打磨。从应用视角看，该开源工具更易被使用，并在应用场景中被广泛推开，即热力值由开发者端传导至应用端。事实上，许多开源大数据工具的应用者同时也是开发者，他们针对实践中的问题持续优化大数据工具，将解决方案回馈至开发社区。

基础数据

【数据来源】GH Archive: https://www.gharchive.org/； Github Stars Explorer: https://emanuelef.github.io/daily-stars-explorer

【数据采集时间】起始时间为最早有记录时间，终止时间为2024年6月30日

【数据采集对象】开源大数据工具所对应的Github代码仓（Repository），而非对应的Github项目（Project）

核心指标

【选取范围及指标意义】指标选取范围为GH Archive可提供的17类Github事件，事件定义遵循GH Archive中对应的属性说明。

【指标选定逻辑】基于开发者在开源社区（Github）中的基础行为，选取**Star、Fork、Issue、Commit、Pull Request**五项核心指标，其他Github事件或为此五类事件的从属事件，或其本身一般性属性较低。

以下表格为GH Archive中所列举的17类事件，标色事件为本报告选取的五项基础指标。事件具体定义请参考Github文档：https://docs.github.com/zh/rest/using-the-rest-api/github-event-types。

CommitComment Event	CreateEvent	DeleteEvent	**ForkEvent**	GollumEvent	IssueComment Event
IssuesEvent	MemberEvent	PublicEvent	**PullRequestEvent**	PullRequestReview Event	PullRequestReview CommentEvent
PullRequestReview ThreadEvent	**PushEvent**	ReleaseEvent	SponsorshipEvent	**WatchEvent**	

注：GH Archive中的事件命名可能与一般认知不同。如，WatchEvent对应Star，PushEvent对应Commit。

来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

说明（2/2）：热力值计算方式

计算方式

【观察值提取】以半年为计算的标准时段，根据获取的时点基础数据，计算每半年指标变动值。即当年6月30日相对于上一年12月31日的变动值，以及当年12月31相对于当年6月30日的变动值。

【核心指标标准化处理】采用对数函数非线性标准化方式，通过指标极值确定阈值，对指标的观察值做进行无量纲化处理，便于不同数量级指标间进行综合分析和比较。

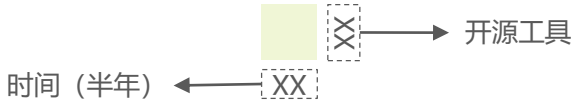
$$标准化值=Log_{10}（1+观察值）/Log_{10}（1+阈值）$$

【AHP层次分析法加权】结合定量与定性分析，通过多位专家判断五项核心指标的相互重要程度，取几何平均后，确定偏好矩阵，再经过一致性检验后确定指标对热力值影响，即指标在计算热力值中所占权重。

	Star	Fork	Issue	Commit	PR
Star					
Fork					
Issue					
Commit					
PR					

- 左侧为专家根据行业经验填写的偏好矩阵，采用10分制，绿色部分为打分区域；
- 从指标意义来看，Star、Fork、Issue、Commit、PR是渐进发展的，代表开发者参与开源社区由浅入深的过程。因此，尽管专家的矩阵打分各有不同，但总体上遵循由Star至PR逐渐升高这一规律；
- Star数量长期以来存在着“刷星”等数据虚假问题，因此其在热力值中所占的权重最小。

【热力值计算及展现】根据各指标权重及该指标中开源大数据工具的标准化值，加权计算该开源大数据工具热力值。以半年为基础热力区间，展示热力图。热力图中每一格对应时间（横坐标）与开源工具（纵坐标），颜色深浅代表热力值大小。



来源：公开资料整理，艾瑞咨询研究院自主研究及绘制。

BUSINESS
COOPERATION

业务合作

联系我们

400 - 026 - 2099

ask@iresearch.com.cn

www.idigital.com.cn　　www.iresearch.com.cn

官　网

微信公众号

新浪微博

企业微信

LEGAL STATEMENT

法律声明

版权声明

免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，部分文字和数据采集于公开信息，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，艾瑞咨询对该等信息的准确性、完整性或可靠性作尽最大努力的追求，但不作任何保证。在任何情况下，本报告中的信息或所表述的观点均不构成任何建议。

本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。

iResearch
艾瑞咨询

THANKS

艾瑞咨询为商业决策赋能