开源大数据热力报告

2022

开放原子开源基金会
OPENATOM FOUNDATION

X-lab
开放实验室

阿里巴巴开源
Alibaba Open Source

报告目录

如何定量分析“后Hadoop时代”开源项目和技术趋势

研究目标

Hadoop 作为开源大数据技术的起源，兴起于2006年。我们收集从Hadoop 发展第10年，即2015年至今的相关公开数据进行关联分析，研究开源大数据进入新阶段后的技术趋势，以及开源社区的运作模式对技术走向的助推作用。

使用热力值进行定量分析

开源项目热力值，使用量化指标，刻画开源项目的开发迭代活跃度和受开发者欢迎程度。

具体来讲，主要来自于几个方面：
（1）开发者对开源项目的关注度，以及应用场景的广度和深度
（2）开发者参与项目开发的贡献活跃度
（3）开发者在开发过程中展现出的协作关联度
（4）项目和社区的可持续发展的健康度
这些和大数据技术发展趋势、开源项目的技术吸引力、开源社区治理水平以及项目传播力强相关。

本报告所呈现的开源大数据热力从全景、技术栈分类以及单项目角度对入围项目的热力表现进行可视化的多维度洞察，并将项目进程中的关键事件与热力表现进行关联分析，并引入开源基金会、知名开源项目等领域专家进行访谈，尝试找到项目健康发展一般规律，并对有效提升项目影响力的方法论进行了归纳总结。

数据来源

采集时间为2022年10月1日。
通过 GitHub log 获取2015年1月至2022年9月的公开数据（包括项目Id、Star、Issue、Open PR，Review Comment，Merge PR等）。
通过Jira api 获取2015年1月至2022年9月的公开数据（包括项目Id、Issue数量）。

热力值研究模型

热力值表征开源项目开发者参与热度

开发者参与开源项目一般遵从规律：关注项目（Star）->参与问题反馈（提 Issue）->参与开发协作（提交 PR 和 Review、活跃 Contributor 等）。因此，热力值由该规律中的3个关键指标加权而来。

- 项目关注：每年新增 Star 数量，来自于 GitHub 公开数据
- 社区反馈：每年新增 Issue 数量，来自于 GitHub 与 Jira 公开数据
- 开发协作：每年 OpenRank 值，OpenRank 由 GitHub 公开数据（Open PR，Review Comment，Merge PR）计算所得，算法来自X-Lab开放实验室

#说明：
因开源大数据项目中有超过40%的项目，使用Jira进行Issue提交和反馈，因此将 X-Lab 原有的 OpenRank 算法进行了修正，将Issue剔除出来，不参与 OpenRank 计算。而将 GitHub 与 Jira 公开数据中的 Issue 数单独列出作为社区反馈维度进行计算。

热力值计算公式

把2015年作为基期，2015年所有开源项目平均热力值作为基数，赋值为100。将三个关键指标做归一化处理，赋予对应权重比例，并由此确立了三个关键指标的归一化系数。详见下表：

2015年所有项目	原始值	归一化系数	归一化后的值	归一化后的权重
平均新增Star数	772.48	0.03	25	25%
平均新增Issue数	1137.80	0.03	35	35%
平均OpenRank值	14.37	2.78	40	40%
热力值	~	~	100	100%

某项目某年热力值=年度新增 Star 原始值 * 归一化系数 + 年度新增 Issue 原始值 * 归一化系数 + 年度 OpenRank 原始值 * 归一化系数

#说明：
所有大数据项目数据合计来看，新增 Star 数8年增长倍数为3，Issue 数8年增长倍数为1.8，OpenRank 值8年增长倍数为8。因此将基期中三个核心指标的权重比例设置为：25%：35%：40%。第8期的权重比例将变化为：15%：15%：70%。表征开发协作的比重大幅提升，这也与开源项目的生命源动力来自于更广泛的社区开发协作保持一致性。

热力值计算详见： *https://github.com/X-lab2017/open-digger/tree/master/cooperations*

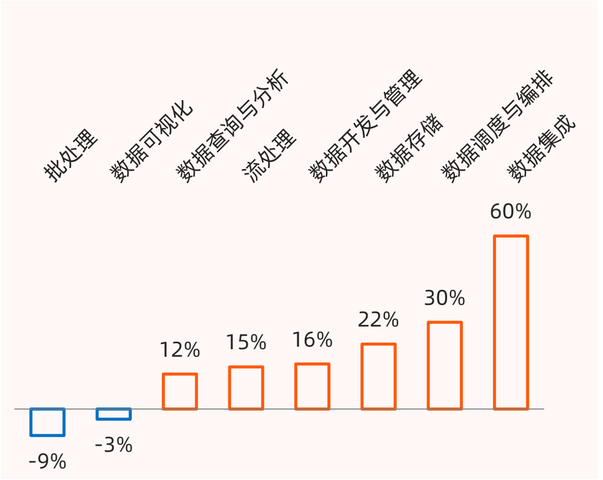
开源大数据热力的“摩尔定律”

每隔40个月，热力值提升1倍

我们按照数据处理的生命周期，对开源大数据项目进行了技术分类，包括「数据集成」、「数据存储」、「批处理」、「流处理」、「数据查询与分析」、「数据可视化」、「数据调度与编排」、「数据开发与管理」8个类别。2022年开源大数据总热力值，增长到2015年的4倍。每隔40个月，热力值提升1倍。

热力变迁反映技术趋势

- 开发者对「数据查询与分析」保持了长期的开发热情，连续8年位于热力值榜首。
- 2017年「流处理」热力值超过「批处理」，大数据处理进入实时阶段。
- 数据规模越来越大，数据结构更多样化，「数据集成」从2020年开始爆发式增长。
- 近2年来，活跃的新兴项目为「数据调度与编排」、「数据开发与管理」注入新的活力。



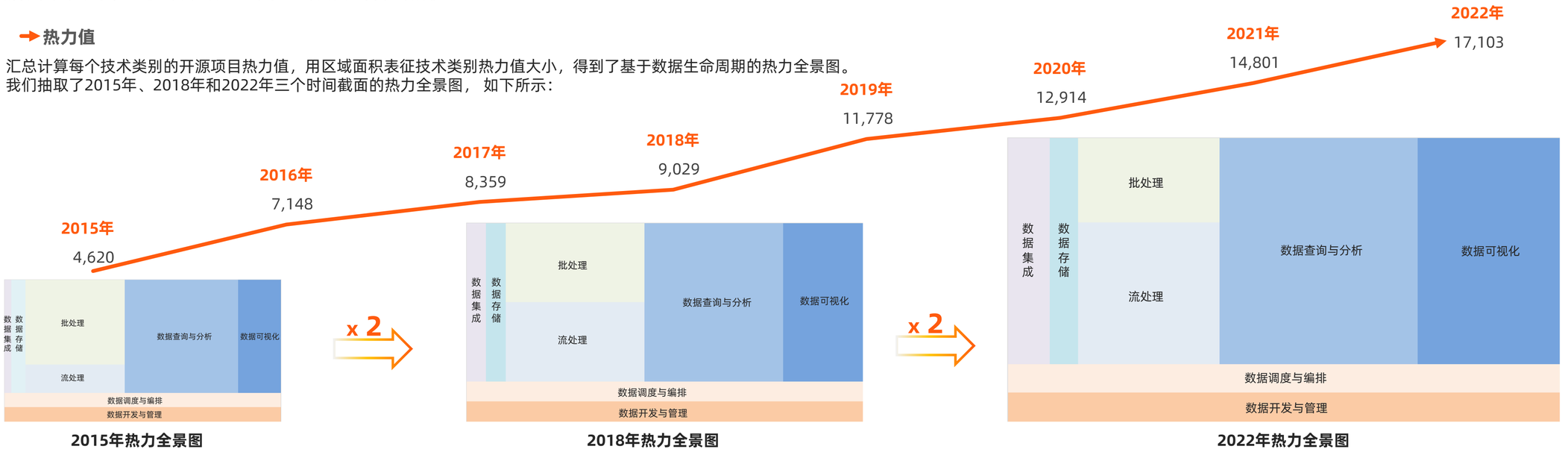
2022年热力值年同比增速

2025年总热力值将突破3万

按照目前热力增长趋势预测，到2025年，总热力值将突破3万，「数据集成」、「数据调度与编排」、「数据开发与管理」将成为强劲的热力增长点。

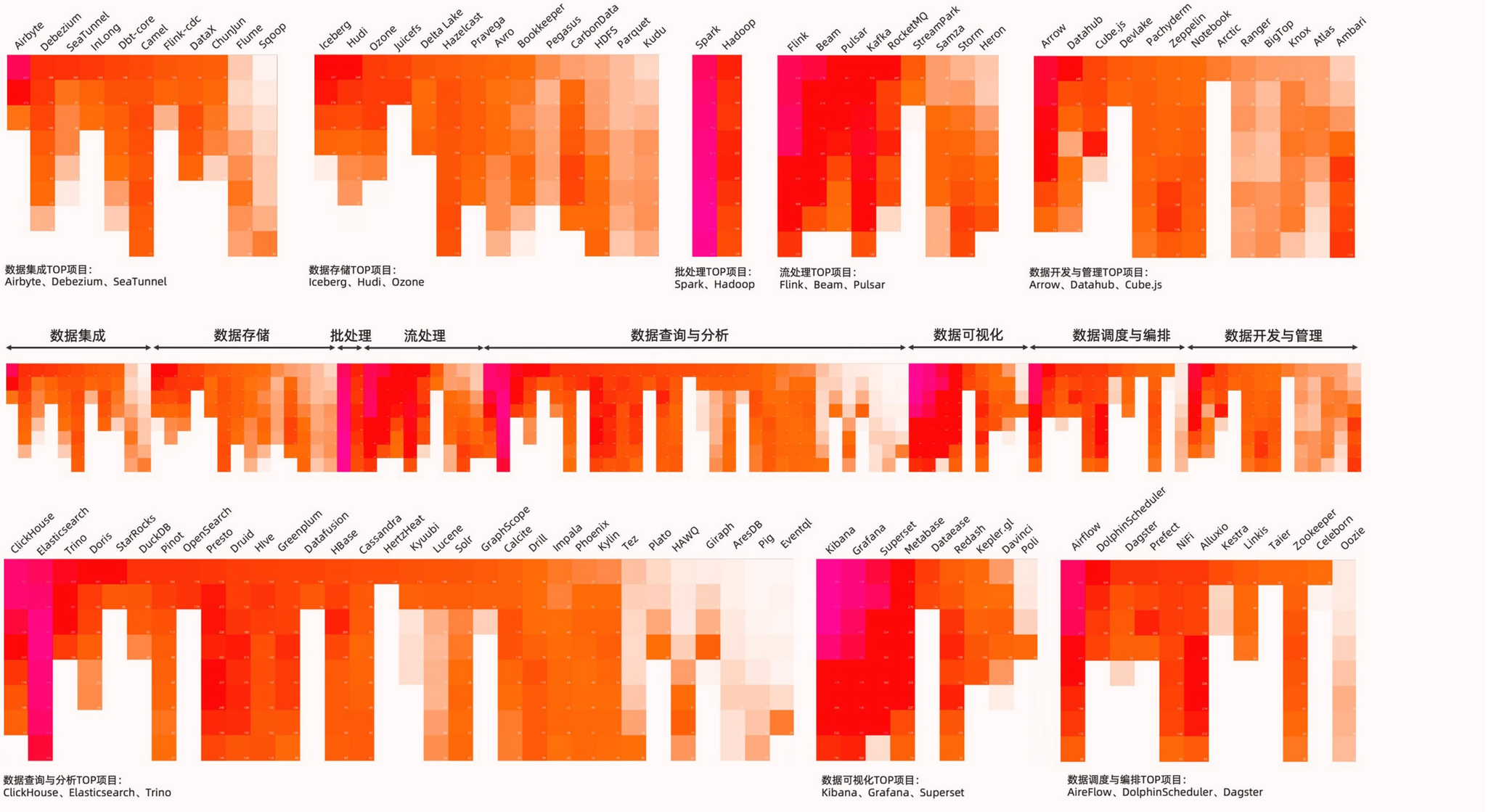
热力值

汇总计算每个技术类别的开源项目热力值，用区域面积表征技术类别热力值大小，得到了基于数据生命周期的热力全景图。
我们抽取了2015年、2018年和2022年三个时间截面的热力全景图， 如下所示：



2015年热力全景图　　2018年热力全景图　　2022年热力全景图

开源大数据热力图谱


Airbyte
Debezium
SeaTunnel
InLong
Dbt-core
Camel
Flink-cdc
DataX
Chunjun
Flume
Sqoop
数据集成TOP项目：
Airbyte、Debezium、SeaTunnel
Iceberg
Hudi
Ozone
Juicefs
Delta Lake
Hazelcast
Pravega
Avro
Bookkeeper
Pegasus
CarbonData
HDFS
Parquet
Kudu
数据存储TOP项目：
Iceberg、Hudi、Ozone
Spark
Hadoop
批处理TOP项目：
Spark、Hadoop
Flink
Beam
Pulsar
Kafka
RocketMQ
StreamPark
Samza
Storm
Heron
流处理TOP项目：
Flink、Beam、Pulsar
Arrow
Datahub
Cube.js
Devlake
Pachyderm
Zeppelin
Notebook
Arctic
Ranger
BigTop
Knox
Atlas
Ambari
数据开发与管理TOP项目：
Arrow、Datahub、Cube.js
数据集成
数据存储
批处理
流处理
数据查询与分析
数据可视化
数据调度与编排
数据开发与管理
2022
2021
2020
2019
2018
2017
2016
2015
ClickHouse
Elasticsearch
Trino
Doris
StarRocks
DuckDB
Pinot
OpenSearch
Presto
Druid
Hive
Greenplum
Datafusion
HBase
Cassandra
HertzHeat
Kyuubi
Lucene
Solr
GraphScope
Calcite
Drill
Impala
Phoenix
Kylin
Tez
Plato
HAWQ
Giraph
AresDB
Pig
Eventql
数据查询与分析TOP项目：
ClickHouse、Elasticsearch、Trino
Kibana
Grafana
Superset
Metabase
Dataease
Redash
Kepler.gl
Davinci
Poli
数据可视化TOP项目：
Kibana、Grafana、Superset
Airflow
DolphinScheduler
Dagster
Prefect
NiFi
Alluxio
Kestra
Linkis
Taier
Zookeeper
Celeborn
Oozie
数据调度与编排TOP项目：
AireFlow、DolphinScheduler、Dagster

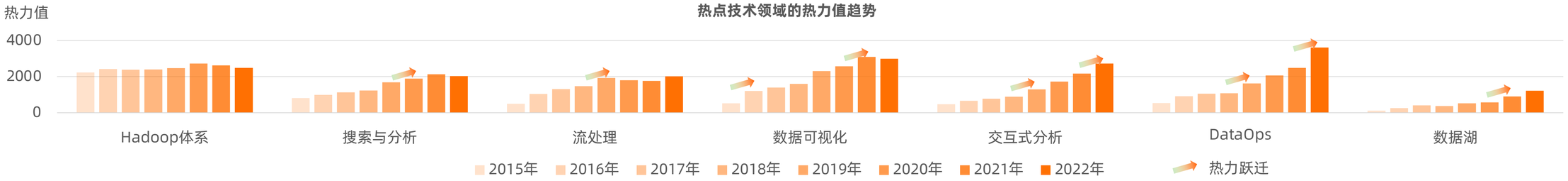
热力趋势一：用户需求多样化推动技术多元化

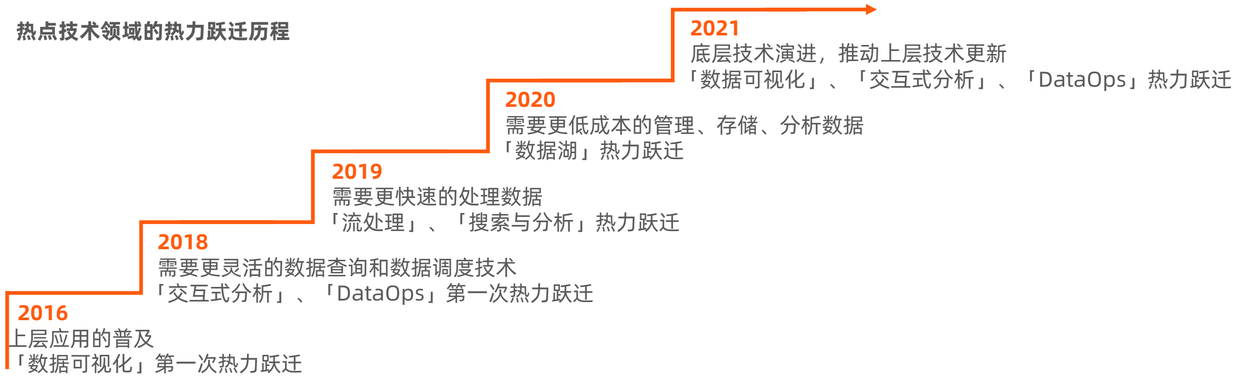
一套复杂体系分化为六大热点技术

经过10年发展，以Hadoop为核心的开源大数据体系，从2015年开始，转变为多元化技术并行发展，开发者的热情分别涌向「搜索与分析」、「流处理」、「数据可视化」、「交互式分析」、「DataOps」、「数据湖」六大技术热点领域，每个热点领域集中解决某个特定场景问题。其中，「数据湖」以34%的热力值年均复合增长率高居第一位，「交互式分析」、「DataOps」紧随其后，分列第二、三位 。而原有Hadoop体系的产品迭代则趋于稳定，热力值年均复合增长率为1%。部分Hadoop生态项目（如HDFS）成为其他新兴技术的基础依赖，另一部分项目（如Sqoop ）则逐渐退出舞台。

热力跃迁更加频繁，彼此交替推动

与大数据应用场景和规模变化趋势相呼应，热点领域的热力跃迁（热力值大幅度跳变）遵循了从上层数据可视化应用普及，到数据处理技术升级，再到数据存储和管理的结构性演变，最终，数据基础设施能力的提升又反过来推动上层应用的技术革新。具体表现为，「数据可视化」在2016和2021年经历了两次热力跃迁，「搜索与分析」和「流处理」 在2019年热力跃迁，「交互式分析」和 「DataOps」从2018年和2021年经历了两次热力跃迁，「数据湖」在2020年热力跃迁。



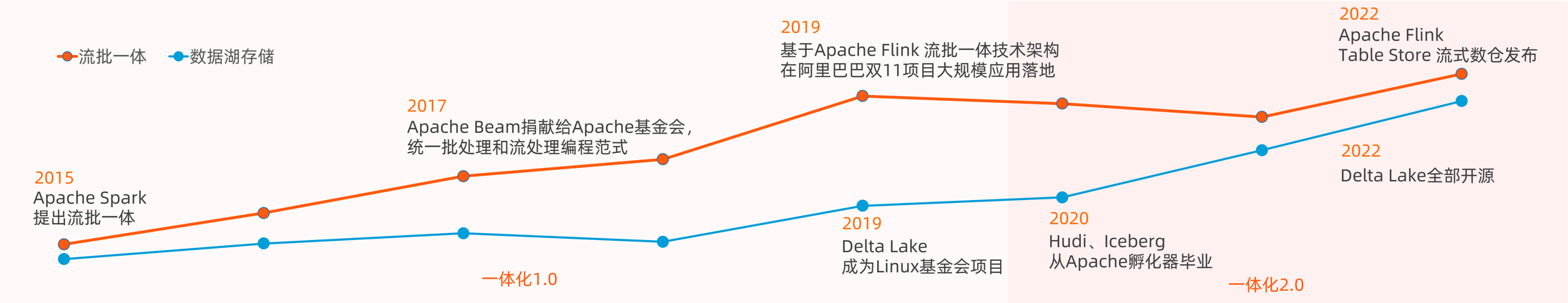


热点领域	热力值 年均复合增长率	代表项目
Hadoop体系	1%	Hadoop、Spark、Hive、HBase
搜索与分析	14%	Elasticsearch、Lucene、Solr、Opensearch
流处理	19%	Flink、Beam、Kafka、Pulsar
数据可视化	24%	Superset、Kibana、Grafana、Metabase
交互式分析	25%	ClickHouse、Presto、StarRocks、Doris
DataOps	27%	Airbyte、Airflow、Dolphinscheduler、Atlas
数据湖	34%	Iceberg、Hudi、Delta Lake、Alluxio

热力趋势二：一体化演进迈入2.0时代

从计算一体化到存储一体化

在对热力变迁数据的观察中，我们发现，从2015年开始，计算部分率先进入「一体化」演进历程，其中的典型代表「流批一体」在2019年出现第一个热力峰值。以数据湖存储为代表的存储一体化从2019年起进入了一个新的发展阶段，并在2021年前后进入了开发迭代的热力高速增长期，在此期间， 涌现了Delta Lake、 Iceberg和Hudi等热点项目。



热力变迁背后是用户使用痛点的转移

多元化技术的蓬勃发展，在一定程度上增加了开源生态体系的复杂性，系统架构也存在性能瓶颈，且扩展能力有限。业界需要统一、融合的大数据系统，能够将多种计算模式有机地融合在一起，易于扩展，能够支持新的模式，降低开源软件的开发、运维复杂度。

以「流批一体」为例，这种计算融合技术最早提出于 2015 年，它的初衷是让开发人员能够使用同一套接口实现大数据的流计算和批计算，进而保证处理过程与结果的一致性。使用统一的计算框架，用户可以不用区分实时和离线计算的场景，减少用户的学习成本，减少开发和维护两套框架的运维成本。流批一体技术演进过程中的几个关键时间节点，2015年Spark提出流批一体，到2019年基于Apache Flink在阿里巴巴双11项目中大规模落地流批一体应用，再到2022年Flink Table Store 流式数仓发布，每一次重大技术更迭，都会牵引大量开发者关注和参与，促使流处理领域热力值显著提升。

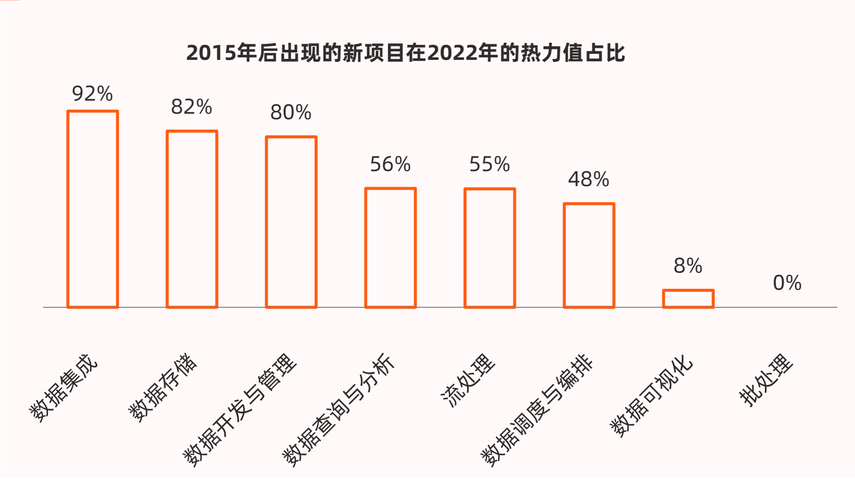
开发者在初尝了计算一体化带来的技术红利之后，开始在其他技术领域进行一体化的尝试。而另一方面，为多种不同的计算模型管理多套不同的存储已经成为了一个新的痛点。开发者深刻体会到传统数仓的难以逾越的缺陷，比如数据更新较为昂贵，缺乏跨数据源的高效联邦查询等。从2019年开始，数据湖存储解决方案 Delta Lake出现，以及后续的 Iceberg 和 Hudi等，都致力于解决存储一体化问题。

热力趋势三：云原生大规模重构开源技术栈

发轫于云端的技术重构

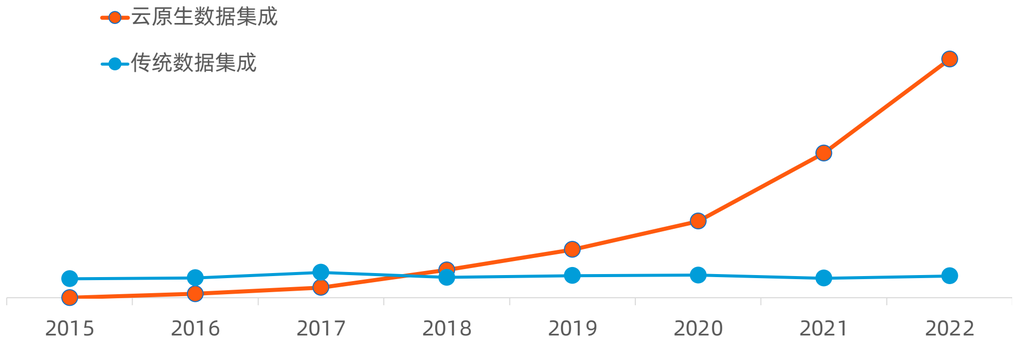
过去几年，数据源和数据存储正逐步迁移到云端，更多元化的计算负载也运行到了云端，计算与存储分离已成为大数据平台的标准架构。越来越多的开发者在云端开发中，对开源大数据项目进行云原生改造适配。云原生作为技术创新的实验场，改变了大量开源大数据技术的走向。

2015年后出现的新项目，无一例外地在云原生方向进行了积极的技术布局。Pulsar、DolphinScheduler、JuiceFS、Celeborn、Arctic等诞生于云原生时代的开源项目如雨后春笋般破土成长。这些新项目在2022年的热力值占比已经达到51%，其中，「数据集成」、「数据存储」、「数据开发与管理」等领域都发生了非常大的项目更迭，新项目热力值占比已经超过了80%。从2020年开始， Spark、 Kafka、Flink等主流项目陆续正式支持 Kubernetes。云原生推动的开源技术栈大重构正在进行时。



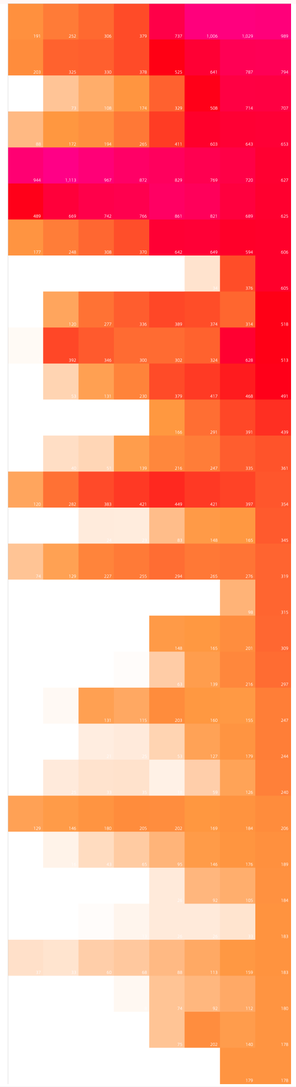
「数据集成」率先完成重构

随着云端多样化数据收集需求的爆发，以及下游数据分析逻辑的变化，数据集成从“劳动密集型”ETL工具演进到灵活高效易用的“数据加工流水线”。传统数据集成工具Flume、 Camel处于平稳维护状态， Sqoop已于2021年从Apache软件基金会退役。与云原生结合更紧密的Airbyte、Flink CDC、SeaTunnel等项目飞速发展。在热力趋势中可以看到，云原生数据集成在2018年超越了传统数据集成，从2019年开始，这一演进历程加速，热力值逐年翻倍。不少新孵化的项目热力值年均复合增长率超过100%，增长势头强劲。



项目名称	热力值 年均复合增长率	热力图谱 （2016～2022）	项目生命周期
Airbyte	325%		3年
Flink-CDC	159%		3年
SeaTunnel	119%		6年
InLong	111%		3年
ChunJun	72%		5年
Dbt-core	56%		7年
Debezium	52%		7年
DataX	12%		5年

开源大数据项目热力TOP30

排序	项目名称	技术领域	2022年热力值	热力图谱（2015～2022）
1	Kibana	数据可视化	989.40	
2	Grafana	数据可视化	793.55	
3	ClickHouse	数据查询与分析	707.42	
4	Airflow	数据调度与编排	653.00	
5	Spark	批处理/流处理	627.24	
6	Elasticsearch	数据查询与分析	624.52	
7	Flink	流处理	606.42	
8	Airbyte	数据集成	604.81	
9	Beam	流处理	517.67	
10	Superset	数据可视化	513.44	
11	Arrow	数据开发与管理	491.36	
12	Trino	数据查询与分析	439.23	
13	Pulsar	流处理	360.69	
14	Kafka	流处理	353.56	
15	Doris	数据查询与分析	344.59	
16	Metabase	数据可视化	318.99	
17	StarRocks	数据查询与分析	315.08	
18	DolphinScheduler	数据调度与编排	309.15	
19	Iceberg	数据存储	297.29	
20	RocketMQ	流处理	246.88	
21	Hudi	数据存储	243.76	
22	Datahub	数据开发与管理	240.37	
23	Hadoop	批处理	205.62	
24	Debezium	数据集成	189.15	
25	Duckdb	数据查询与分析	184.43	
26	SeaTunnel	数据集成	183.40	
27	Pinot	数据查询与分析	183.38	
28	Dagster	数据调度与编排	179.98	
29	Prefect	数据调度与编排	178.47	
30	OpenSearch	数据查询与分析	178.40	

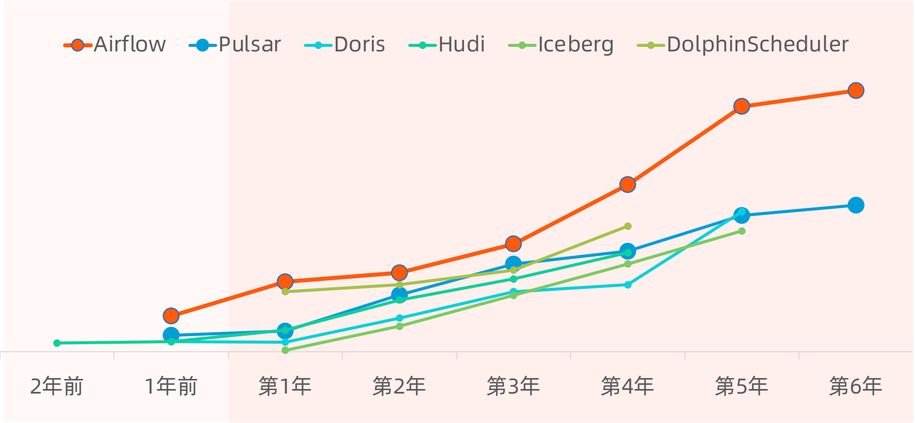
TOP项目热力跃迁逻辑研究

解决用户痛点是核心竞争力

每个项目都需要解决用户在某个细分场景的痛点，反过来，每个细分场景的用户问题都会有少数几个项目解决得最好。入围本次报告的102个项目，在细分领域分布上并不均衡。但TOP30项目的细分领域却均匀分布，每个领域3～5个项目。用户痛点并非一成不变，在前面章节，我们已经描述过技术趋势演变带来的项目热力变迁。我们观察到了无数新老交替，也观察到了一批优秀开源项目的与时俱进，成为热力趋势中的“常青树”。如Spark在2014年以Spark SQL代替Shark，2016年发布Structured Streaming，推动着大数据技术向前发展。又如，Flink围绕实时处理的核心需求，陆续延展出数据集成（Flink CDC）、数据分析（Flink SQL）、机器学习（Flink ML）、规则引擎（Flink CEP）、动态表存储（Flink Table Store）等多种场景能力。

掌握开源社区运作的方法论

对于新开源项目，进入基金会孵化器能够帮助项目快速成长，Airflow、Pulsar等项目进入孵化器后的热力趋势验证了这一点。欧美开源运作发展较为成熟，除了加入基金会，也有不少独立存在的优秀开源项目，如Elasticsearch、ClickHouse等。这是开源发展到一定阶段的产物，背后有一批开源经验丰富的人才在不同项目间流动。无论是哪一种方式，这些TOP项目背后的开源社区运作模式都能够通过基金会、人才流动或者文化传播沉淀为方法论，传承到下一个有潜力的项目。



开源项目进入基金会孵化器前后的热力趋势

持续关注开发者体验

在社区起步阶段，找到种子用户非常关键，这一阶段项目需要快速迭代满足他们的需求。而在社区发展趋于成熟时，则更需要关注大众开发者的产品体验。无论处于什么阶段，都需要保持良好的开发者体验，如Issue、邮件咨询等社区互动行为，保证及时反馈SLA。对于诞生于国内的开源项目，拥有良好体验的英文项目文档，是做好国际化的先决条件。接受本地开发者的文化和沟通习惯，用他们喜欢的方式发展社区。

商业化对于开源社区发展是双刃剑

热力TOP30中有超过9成的项目背后存在商业化公司运作。开源与商业化可以并存，并且能够相互促进，这已经成为业界共识。但我们也在研究中发现，当前能够做到商业化与开源社区平衡发展的项目并不多。这里存在几种不同类型：第一类，在长期经营的开源生态上已经建立起强大“护城河”，商业化相对克制和保持节奏。另一类，因为不得已的原因而更改开源策略，开源社区发展受到一定影响，以此换取商业回报。第三类，也是最多的一类，商业化已经启动，同时开源社区也处于快速发展阶段，商业化软件开发模式在一定程度上改变了“集市”类型的开源软件开发模式，开源的“速度”变得更快。

我们认为，开源背后的商业化更多体现为良性的促进作用。在某个时间段出现商业化和开源之间的排异现象，市场和社区都会自动消化和调整，最终回归到稳定状态。

致谢

联合发起

开放原子开源基金会 OPENATOM FOUNDATION

X-lab 开放实验室

阿里巴巴开源 Alibaba Open Source

战略合作

OSCHINA

InfoQ 极客传媒

阿里云开发者社区 ALIBABA CLOUD DEVELOPER COMMUNITY

社区合作

CSDN

DataFun.

segmentfault 思否

开源社 kaiyuanshe

专家顾问（按照姓氏拼音为序）

代立冬　Apache Member、Apache DolphinScheduler PMC Chair
金耀辉　白玉兰开源开放研究院执行院长、上海交通大学教授
李　钰　Apache Member、Apache Flink & Apache HBase PMC Member
刘京娟　开放原子开源基金会副秘书长
王　峰　阿里巴巴开源委员会大数据AI领域主席、Apache Flink 中文社区发起人
王青兰　开放群岛开源社区委员会法律合规组组长
王一鹏　InfoQ 总编
翟　佳　Apache Pulsar & Apache BookKeeper PMC Member
郭　炜　Apache Member、Apache SeaTunnel(incubating) 导师、ClickHouse中文社区发起人
李　潇　Apache Spark PMC Member
刘　冬　开源中国创始人，Gitee (码云) 创始人 & CTO
秦江杰　Apache Flink & Kafka PMC Member
王晶昱　阿里巴巴开源办公室秘书长
王　伟　X-lab开放实验室负责人、华东师范大学研究员、博士生导师
于邦旭　CSDN高级副总裁
周　晓　阿里云智能大数据AI运营总经理

报告贡献者（按照姓氏拼音为序）

蔡芳芳　InfoQ 主编
郭雪雯　开放原子开源基金会专家
李　萌　开源中国社区负责人
林日华　开源中国主编
聂励峰　Apache SeaTunnel PPMC、 Apache DolphinScheduler Committer
涂　南　阿里巴巴开源办公室运营专家
王荷舒　开放原子开源基金会专家
郭　皓　开放原子开源基金会专家
李　博　开放原子开源基金会专家
李　雪　开放原子开源基金会专家
刘晓清　阿里云开发者社区专家
是　溪　阿里云开源大数据运营专家
王殿进　StreamNative社区运营负责人
赵生宇　X-lab实验室核心成员、同济大学计算机在读博士

报告合作与反馈