大数据基础平台NDH
宣传手册

筑牢企业数字底座

网易数帆

2022

PREFACE

一

前言

网易数帆是网易旗下ToB企业服务品牌，定位于数字化转型技术与服务提供商，依托网易20余年互联网技术积累，推出三大数字生产力模型，帮助企业发展软件生产力、数据生产力、智慧生产力，沉淀企业数字资产，为企业数字化转型提质增效。目前，网易数帆已服务工商银行、兴业银行、华泰证券、东北证券、中信证券、平安产险、格力、OPPO、一汽解放、吉利集团、华能集团、南方电网、浙江电信、德邦快递、九州通、名创优品、科沃斯、温氏集团等300余家行业头部企业。

网易数帆旗下大数据产品线基于十多年数据技术积淀，以全面的技术及产品服务企业“看数”、“管数”、“用数”等业务场景，盘活企业数据生产力，助力企业人人用数据，时时用数据，推动企业数据生产力跃迁，全面释放数据价值。

多年来，网易数帆基于网易集团内外部业务场景在大数据基础技术领域有过诸多探索实践，沉淀了许多实用的技术产品（例如Kyuubi、EasyEagle），为用户提供了更加便利的SQL开发体验和更加稳定的任务运行保障。这些探索让网易数帆意识到，一个大数据基础平台首先需要提供高效稳定的批/流计算、交互式查询等数据仓库建设能力，同时保持架构开放、具有平滑演进能力的特性才能更好地满足客户长期长远的数据业务需求。在这样的技术理念和经验指导下，网易数帆推出大数据基础平台NDH，强调专家级、全过程的服务，不仅核心代码自主可控更通过组件功能增强实现多样化的数据存储分析。

打造领先数据生产力
着力各行业实践深耕

NO.1

产品技术实力居于国内第一梯队

屡次

获评Gartner数据分析代表厂商、数据中台领域标杆厂商、Cloud ABI领域标杆厂商

200+

头部客户项目经验丰富，多行业两百余家客户成熟验证

100+

获信通院大数据产品能力评测等100余项权威荣誉

40+

拥有大数据技术授权专利40余项

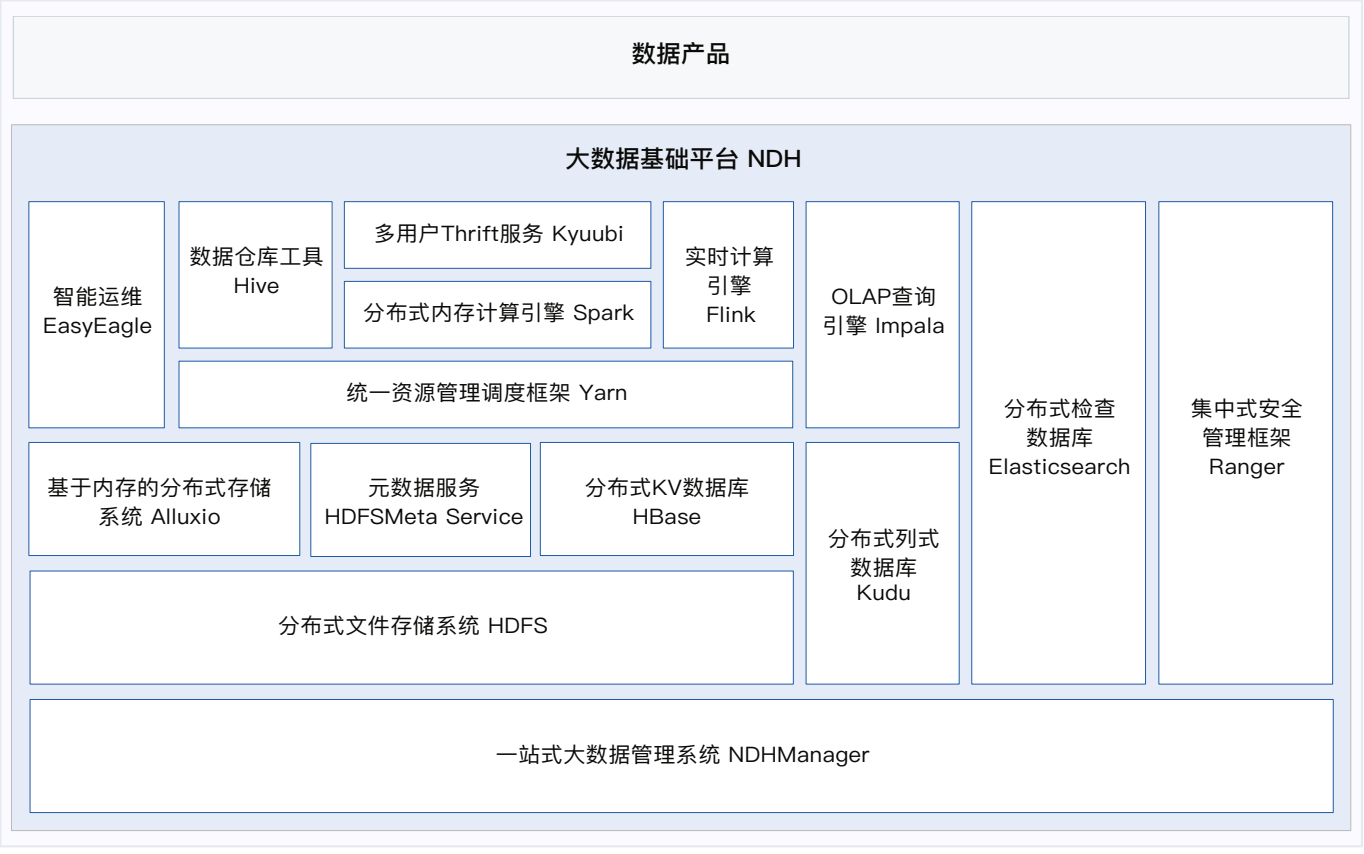
目录

大数据基础平台NDH

NDH（NetEase Data Hub）是网易数帆推出的企业级大数据基础平台。该产品基于最新开源技术打造，内置多种存储计算引擎，并对包括Hadoop、Spark、Impala等在内的多个核心组件做了功能及性能增强，新增EasyEagle组件实现智能运维和任务治理，支持企业级安全管控。

结合网易数帆在大数据领域多年的沉淀积累，大数据基础平台NDH不仅支持对核心代码的完全掌控，并且适配信创软硬件生态，可以更自主化、便捷化、智能化地提升企业数据运维效率，在加强数据安全保障的同时，筑牢企业数字底座。

产品全景



核心优势

自主可控	在开源技术的基础上，NDH分别对HDFS、Hive、Spark、Impala等多个大数据领域核心的组件进行了技术增强与安全保障，开发团队拥有强大的自主研发能力，每年为Apache社区贡献超过300+ Commits，还在Spark之上自研了企业级数据湖探索平台Kyuubi并贡献给Apache社区进行项目孵化。其中的OLAP开发团队是国内为数不多的专注于Apache Impala的团队，团队结合网易内部多个集群实践进行了大量的功能增强。NDH同时适配信创软硬件生态，可兼容国产数据库、华为鲲鹏等系列芯片、麒麟操作系统(v10)所有组件的部署应用，支持在华为云、阿里云、腾讯云等场景下云化部署，满足企业全信创类需求。
智能运维	EasyEagle为自主研发的任务异常诊断平台，汇集了网易数帆大数据多年大型平台运维经验，可以提供任务级别的全链路异常诊断分析、耗时优化建议以及资源优化建议。通过任务全链路标记以及丰富的故障诊断专家经验，可以帮助客户精准高效地定位任务运行失败原因，任务各阶段耗时分析以及优化建议。另外，EasyEagle还提供了Yarn队列资源监控以及任务级别的资源监控，可以帮助业务高效发现资源浪费严重的任务，并给出相关的优化建议，降低集群资源使用压力。

核心优势

安全稳定	NDH集成了包括Ranger、Kerberos、LDAP等整套安全管理模块，集中管理大数据核心组件审计日志，支持行列权限、数据脱敏等功能。另外平台基于网易数帆10多年数据管理经验，对底层组件进行了系列优化和增强。例如，HDFS上增强回收站功能，在实际生产线上多次有效防止了类似Drop Table后数据误删发生；Impala实现了基于虚拟数仓的隔离，支持对于同一集群中的不同节点进行分组，不同workload的业务配置不同的分组，避免业务之间相互影响，保障了平台以及任务的稳定性。
开发便捷	通过企业级数据湖探索平台Kyuubi ，企业可以像使用HiveServer2一样开发SparkSQL。Spark作为整个大数据计算领域最流行的计算框架，相比原先常用的Hive，在计算性能和资源利用方面有很大的提升，但大部分用户很难改变Hive模式下的使用习惯。NDH支持用户保留原有习惯和模式，基于Spark计算引擎上构建的SQL查询引擎，支持多租户隔离等特性，更好地实现分析计算。
平滑迁移	基于网易集团内部多年的成熟实践经验，网易数帆结合企业实际情况制定一整套平滑迁移的落地方案，满足在对业务不影响或影响时间较短(< 10 min)的情况下，完成整体平台迁移切换到NDH。迁移方案充分考虑组件兼容性、迁移可靠性，提供完整配套的迁移工具。

产品特色

1.新增组件

Kyuubi：支持企业可以像使用HiveServer2一样开发SparkSQL	
统一接口	接口简单易用，并兼容传统大数据任务，方便用户历史任务迁移
分布式计算服务	服务支持无限水平拓展，提供高客户端并发能力
高可用特性	服务支持平滑迁移，滚动升级
多租户特性	有效的进行计算资源的隔离和共享，提高集群资源的利用率；实现数据和元数据的隔离，保障数据安全
EasyEagle：提升资源利用率，提高问题诊断效率	
运行资源监控	支持集群维度、队列维度以及任务维度的CPU\内存资源监控
任务资源治理	帮助平台管理员以及用户分别了解队列、任务的实际资源使用率，指导用户对资源浪费严重的大任务进行合理资源优化，提升集群资源利用率
任务问题诊断	旨在提供快速的错误分析能力，尽量避免让用户去查看大量的原始日志信息，从而加快任务排错
任务性能优化	系统针对给定任务提供详细的性能问题类别以及优化建议
HDFSMetaService：应用于数据资产服务，指导数据治理	
准实时元数据查询服务	准实时解析HDFS元数据并基于此提供在线查询服务，业务可以查询指定路径的元数据信息、目录结构信息，还可以通过指标过滤查询，比如获取大小超过1T的目录集合等
HDFS元数据仓库服务	HDFS元仓服务可以统计文件增长最快的目录等，用于指导数据治理

产品特色

2.组件增强

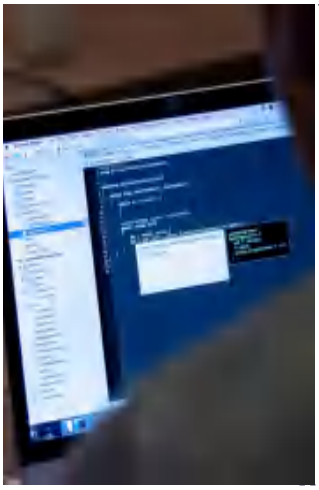
Impala：Apache社区开源组件，以性能优秀著称的数仓查询引擎。网易数帆在开源Impala基础上进行了大量易用性、高可用性以及稳定性的功能增强	
元数据自动同步增强	通过订阅Hive元数据服务（Hive Metastore, HMS）的DDL变更日志来自动刷新Catalog中缓存的表、分区和文件级元数据。通过变更日志批量获取、合并和元数据异步加载等方式提升元数据同步性能。
虚拟数仓服务	提供虚拟数仓功能，用于物理资源隔离和算力水平扩展，支持虚拟数仓间资源的分时复用；支持通过Hive JDBC或用户会话进行虚拟数仓选择
查询管理系统	能够收集保存Impala集群所有的查询信息，自动计算表的统计信息，自动取消慢查询等。基于历史查询建立标准化报表系统，从用户、查询数量、查询性能、扫描数据量、资源使用、查询时间分布、排队耗时等多维度展示集群状态
多表物化视图服务	支持通过控制台创建单表聚合或多表Join的物化视图，在源表数据更新后支持自动更新物化视图数据。支持对用户SQL进行透明改写使其命中物化视图
本地缓存服务增强	增强了社区版Impala的DataCache能力，提供了异步CacheFill（缓存读未命中时异步填充缓存数据）、缓存白名单（支持动态设置分区粒度的数据缓存区间）、Footer专用缓存（Parquet/ORC文件的Footer块用于进一步加速Scan性能）、缓存信息持久化（在impalad重启后可继续命中已缓存的数据块）等功能
用户权限管理增强	支持使用提交SQL的用户账号来进行数据的读写操作，避免社区版Impala使用同一账号操作数据导致权限不一致。增强了社区版Impala的数据脱敏功能，支持中文数据脱敏
Coordinator服务高可用和负载均衡特性	支持通过Zookeeper进行Coordinator节点的高可用配置，支持配置多个Coordinator用于高并发场景下的负载均衡。
Hive：基于Hadoop的一个数据仓库工具，是大数据数据仓库的事实标准	
Metastore服务高可用和负载均衡特性	解决社区Metastore服务单节点在高负载场景下无法横向扩展进而导致服务过载异常风险
HiveServer2服务高可用和负载均衡特性	解决社区HiveServer2服务单节点在高负载场景下无法横向扩展进而导致服务过载异常风险

产品特色

2.组件增强 (续)

Spark：大数据领域最流行的计算引擎之一，网易数帆拥有国内为数不多的内核研发团队	
Z-Order支持	支持通过Z-Order算法对多维数据进行分区排序，使得数据保持优秀的分布聚集，从而在查询侧提高数据的Data Skipping比例
自适应查询优化	自适应查询优化(Adaptive Query Execution, AQE) 是 Spark 3.0 版本引入的重大特性之一，可以在运行时动态的优化用户的SQL执行计划，很大程度上提高了 Spark 作业的性能和稳定性。AQE包含动态分区合并、Join数据倾斜自动优化、动态Join策略选择等多个子特性。网易内部版本在社区版本的基础上对AQE功能进行了增强
权限管控	基于Ranger实现用户读写库、表、列的权限认证
日志审计	所有用户操作行为都会记录到审计日志中，方便操作回溯以及统计分析
HDFS：大数据基础平台的存储基座，数据安全性至关重要	
公共回收站	由于HDFS的delete API接口会直接删除数据，容易引起数据误删无法恢复的问题，实际生产线和项目中曾发生过多次高危事件。针对此类场景特别新增公共回收站功能，可以将数据继续保留一段时间，保障误删数据一键恢复
删除黑名单机制	用户可以设置部分目录为"删除黑名单"，设置为"删除黑名单"的目录不允许被删除

应用场景



离线ETL计算

NDH提供了Kyuubi\Spark3作为主要的离线ETL计算引擎，帮助客户进行离线任务查询分析。Spark作为整个大数据计算领域最流行的计算框架，相比原先常用的Hive，在计算性能和资源利用方面有很大的提升。但大部分用户很难改变Hive模式下的使用习惯，NDH通过提供企业级数据湖探索平台Kyuubi ，帮助企业可以像使用HiveServer2一样开发SparkSQL。此外，NDH提供了EasyEagle服务，可以提供任务级别的全链路异常诊断分析、耗时优化建议以及资源优化建议，帮助企业快速分析定位任务异常原因、优化任务执行效率以及降低任务运行所需资源。

交互式查询服务

NDH主推Impala作为主要的交互式查询引擎，并在原生引擎的基础上做了大量的功能增强，比如建立基于虚拟数仓的隔离环境，支持对于同一集群中的不同节点进行分组，不同workload的业务配置不同的分组，避免业务之间相互影响。

实时更新计算

NDH使用Flink、HDFS\Kudu以及Impala作为主要的实时更新计算组件，数据通过Flink实时更新写入Kudu中，再通过Impala与HDFS中的数据进行联合查询分析。

迁移服务

· 结合业务场景制定整体迁移方案，提供原厂迁移服务。
· 整个迁移过程基本无需平台整体停服，迁移成本低、所需资源可控、风险可控。
· 无缝对接网易数帆数据中台产品，提质增效。

预备期

· 双方沟通确定迁移方案
· 评估迁移所需资源
· 准备服务器等硬件设施

迁移期

· 搭建新NDH集群及功能测试
· 用户权限迁移
· 离线任务迁移
· 元数据/数据迁移
· 迁移期两边平台并行双跑

运行期

· 所有数据及任务流迁移及验证完毕
· 将CDH集群下线扩容至NDH集群
· 生产业务完全切换到NDH集群中

数帆资质（部分）

大数据技术认证资质

工业信息安全发展研究中心评测鉴定所测试
信通院商务智能工具基础能力评测
工信部信创适配测试认证
浪潮技术兼容性测试认证
长江计算兼容性测试认证

管理体系相关资质

ISO 27001 信息安全管理体系认证
ISO 20000 信息技术服务管理体系认证
ISO 9001 质量管理体系认证
CSA STAR Certification 2013 服务管理体系认证
CMMI（三级）认证

专业技术机构资质

工信部云计算服务能力标准首批试点单位
信通院云计算标准和开源推进委员会会员单位
信通院大数据技术标准推进委员会成员
信通院首批开源供应商
信通院可信开源合规计划首批正式成员
浙江省云计算和大数据省级企业研究院
大数据系统软件浙江省工程实验室
浙江省网易大数据重点企业研究院
浙江省增强显示与智能交互工程技术研究中心
CNCF 官方认可的 Kubernetes 服务提供商
云计算开源产业联盟企业数字化发展共建共享平台成员单位

云计算技术认证资质

工业信息安全发展研究中心评测鉴定所测试
工信部信创测试
信通院可信云服务网格先进级（最高级别）评估
信通院数字化可信服务能力认证
信通院可信云微服务评估先进级认证
信通院可信开源项目、OSCAR 开源社区/项目
信通院可信开源供应链风险管理能力评估
Kubernetes 一致性认证
华为鲲鹏技术兼容性测试认证
飞腾处理器平台兼容性认证
浪潮服务器产品兼容互认认证
H3C服务器产品兼容性互认认证

国家级组织资质

国家高新技术企业
国家规划布局内重点软件企业
国家企业技术中心
国家级博士后科研工作站

省级组织资质

浙江省重点企业研究院
浙江省企业技术中心
浙江省数字工厂标杆企业

荣誉资质（部分）

顶级分析机构评选认定

《Gartner 2022中国低代码应用平台竞争格局》低代码代表厂商
《Gartner 2022中国中国分析平台市场指南》数据分析代表厂商
《Gartner 2021中国 ICT 技术成熟度曲线》数据中台领域标杆厂商
《Gartner 2020中国 ICT 技术成熟度曲线》数据中台领域标杆厂商
《Gartner 2019中国 ICT 技术成熟度曲线》Cloud ABI领域标杆厂商

客户/合作伙伴评选认定

华泰证券2021年度金融科技最具创新成果奖
2021光合组织解决方案大赛优秀奖

工信部/信通院评选认定

信通院可信云计算最佳实践-云原生类-服务网格
信通院 OSCAR 尖峰开源用户奖
信通院 OSCAR 尖峰开源创新（二次开发）奖
信通院大数据'星河'案例-行业大数据应用优秀案例
云计算开源产业联盟云原生十大优秀案例
云原生产业联盟2022年度云原生技术创新领航者
云原生产业联盟2022年度云原生应用实践先锋

权威三方机构评选认定

亿欧2021全球产业数字化服务商TOP40
海比研究院"综合中台优秀企业奖获奖
2021长三角金融科技创新与应用大赛——
中国金融科技 · 最佳供应商

政府相关科技项目评选认定

教育部科技进步一等奖
浙江省大数据应用服务创新奖
2015-2016全球云计算大会云鼎奖“全球最佳实践奖”
第十届全球云计算大会云鼎奖“2021-2022年度优秀解决方案奖”
中国长三角数字经济大会“数字经济优秀案例企业”
2022CCF大数据与计算智能大赛优秀案例奖
2019CCF Top10大数据应用最佳实践案例
2021数博会“十佳大数据案例”
2021“科创中国”开源创新榜

专业媒体/社区评选认定

2021 年度 OSCHINA 优秀开源技术团队
infoQ2020最有价值技术团队
InfoQ2020最佳技术社区驱动力奖
infoQ 中国技术力量年度榜单
infoQ2021年度最具价值技术团队奖
infoQ2021年度最佳技术内容运营奖
思否 SegmentFault 中国技术品牌影响力企业
DTCC 第十一届中国数据库技术大会创新产品奖
CTDC 年度优秀微服务创新产品奖I
数字化观察网金数奖2021年度最具影响力数字化服务商
LowCode低码时代 2021“年度最佳低代码产品”

了解网易数帆更多大数据产品

产品宣传手册

- ○ 《 有数BI产品宣传手册 》
- ○ 《 数据开发治理平台 EasyData宣传手册 》
- ● 《 大数据基础平台NDH宣传手册 》

产品列表

有数BI

标签画像 EasyTag

消费者运营平台 EasyCDP

数据开发治理平台 EasyData

大数据基础平台NDH

订阅产品公众号及课程培训中心
获取最新产品动态

有数公众号　　有数学堂官网

DIGITAL
SAIL

数帆官微

数帆官网

2022年9月印刷