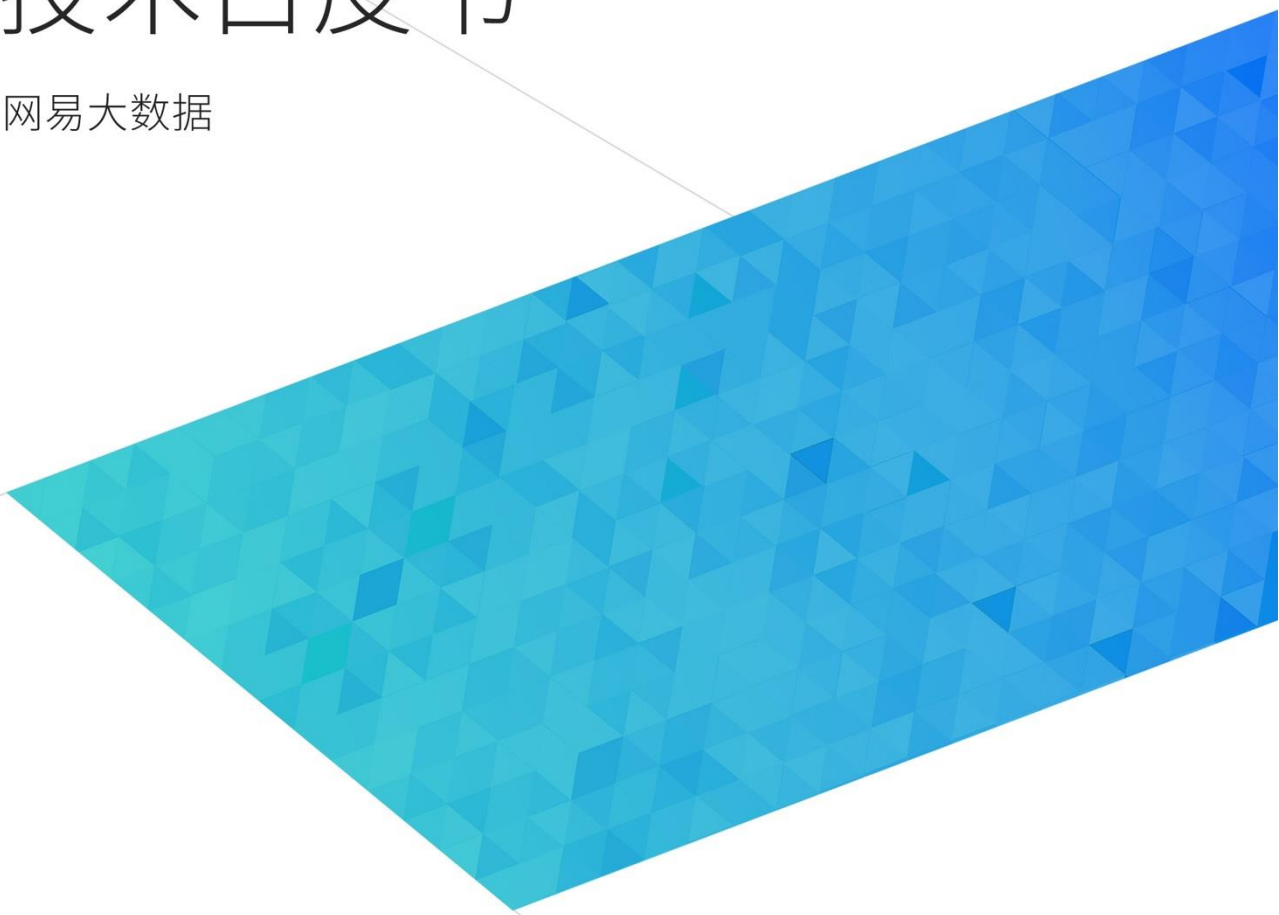
Mammut

网易猛犸大数据平台
技术白皮书

网易大数据



网易大数据
Netease Big Data

Mammut

目录

1. 猛犸大数据平台

猛犸大数据平台，网易大数据实践经验积累，一站式大数据应用开发和数据管理平台。猛犸大数据平台主要分为大数据开发套件和 Hadoop 发行版两部分。

猛犸大数据总体架构图

大数据开发套件主要包含数据开发、任务运维、自助分析、数据管理、项目管理及多租户管理等。大数据开发套件将数据开发、数据分析、数据 ETL 等数据科学工作通过工作流的方式有效地串联起来，提高了数据开发工程师和数据分析工程师的工作效率。

Hadoop 发行版涵盖了网易大数据所有底层平台组件，包括自研组件、基于开源改造的组件。丰富而全面的组件，提供完善的平台能力，使其能轻易地构建不同领域的解决方案，满足不同类型的业务需求。

敏捷易用

基于业务场景设计的用户操作界面提高了系统的易用性，结束了平台命令行运维的繁琐状态。数据开发工程师和数据分析师通过简单拖拽和表单填写即可完成数据科学相关工作。

成熟稳定

持续内部需求驱动帮助打磨平台，网易互联网各业务验证。同时，网易杭研院成熟的 QA 体系为猛犸大数据平台保驾护航。

安全可靠

猛犸平台提供多租户支持，不同租户之间相互隔离。底层使用 Kerberos 认证，实现了数据的安全性和隔离性。除了认证系统，利用 Ranger 实现了细粒度权限控制，保证了各个租户只能查看授权访问的库、表或字段。此外，平台提供审计功能，通过对用户平台行为的记录、分析和汇报，用来帮助事后生成合规报告、事故追根溯源，提高平台安全性。

2. 产品核心功能

2.1 猛犸大数据开发套件

猛犸大数据开发套件提供可视化界面，用户可以进行数据开发、任务运维、自助分析、数据管理及项目管理。大数据开发套件降低了大数据技术门槛，帮助企业快速落地大数据项目。

网易猛犸开发套件

数据开发

数据开发模块提供数据库传输、SQL、Spark、OLAP Cube、MapReduce 及 Script 各种类型任务的敏捷开发界面，任务开发者通过拖拽创建任务，方便地进行数据集成、数据 ETL、数据分析等数据科学工作。以数据库传输为例，用户只需将“数据库传输”组件拖拽到画布上并双击，通过下拉框选择和手动输入填写表单，快速完成数据传输的任务开发。

此外，企业还能根据自身业务场景按需进行任务调度管理，用户可以设置任务的执行顺序、优先级以及执行周期。针对任务失败的情况，设置重试次数、重试间隔及

报警规则。最后，任务产生的结果可以对接主流 BI 系统进行数据可视化分析，或者直接回流到线上系统支撑辅助线上业务。

任务运维

任务运维模块包含可视化的任务管理和实例运维。

任务管理：用户可以查看当前产品线任务列表及各个任务的状态、创建人、修改时间、最近执行时间及调度信息。针对单个任务，用户可以查看详情（包括修改历史、执行历史及执行计划）、编辑任务或补数据。补数据可以对任务执行发生在过去一段时间的调度。

实例运维：用户可以查看任务实例列表及各个实例的状态、运行方式、开始时间、结束时间、运行时长、计划执行时间及提交人信息。此外，用户可以按照不同的维度（开始时间、关键字、运行方式、状态及提交人）快速定位感兴趣的实例。针对单个实例，用户可以查看详情、日志或重跑。

自助分析

自助分析提供交互式数据分析的 Notebook。单个 Notebook 切分成不同段落，便于分析师使用多个段落同时进行交互式分析。除了交互式数据分析， 用户可以使用自助分析进行历史数据查询和自助取数。

数据管理

数据管理模块包括元数据管理、数据源管理、权限设置及权限查看。通过主题视图，企业可以实现数仓分层，用户可以根据主题快速定位感兴趣的表。

数据源管理提供登记关系型数据库数据源的入口。登记数据源后，数据开发工程师可以将数据源的数据集成到猛犸平台，并做进一步的操作如数据 ETL 和数据分析。目前支持的关系型数据库包括 MySQL、SQL Server、PostgreSQL、DB2 及 Oracle。除了数据源登记，项目管理员可以修改、删除数据源或测试数据源连通性。

通过权限管理，项目管理员可以按照角色进行细粒度权限控制，并且针对某个角色，授予库、表和列的不同权限（select、update、create、drop 和 alter 等）。此外，用户可以查看各个角色的授权情况。

项目管理

为了满足现代企业多部门多集群的需求，项目管理提供创建项目、管理项目成员管理审计项目活动等功能。针对单个项目，项目管理员可以进行管理集群用户、目录、队列及资源。

2.2 调度系统

用户可以通过调度系统灵活方便地配置和调度大数据 ETL 任务。支持 Sqoop、hive、Spark、HadoopMR、Script、Java 等类型的大数据任务，通过配置任务之间的依赖关系，可以灵活地组织任务流。支持任务流的定期调度、历史回溯调度、历史任务重跑等多种调度方式。支持跨任务流的任务依赖和任务的细粒度分配，并且所有服务节点都实现了高可用机制。

任务执行

调度系统支持几乎所有主流的大数据类型任务，对任务的执行进行了严格的权限控制和资源隔离，保证用户任务正常执行。用户可以灵活便捷地配置任务参数，系统可用性好。任务的执行采用独立进程执行的方式，任务插件的升级和扩展不会对系统使用有任何影响。

任务流执行控制

调度系统除了支持多种形式的调度方式以外，还支持多维度的精细化的调度参数的设置：支持多层级的任务流并发执行，内置多种任务异常处理策略，提供多种任务流执行状态的通知报警方案。

其他

除了支持调度任务的核心功能，调度系统还支持执行 sql 执行结果的预览和下载、任务执行日志的预览、保存和下载等提高用户使用体验的功能。


调度系统
调度系统
系统高可用
服务节点多镜像多副本高可用
任务插件化动态接入
结果下载
统一执行结果、日志保存/预览/下载
调度管理
定期调度
历史回溯调度
立即执行
历史任务重跑
任务执行
Script任务
Hadoop MR
Hive
Spark
Sqoop

调度系统

2.5 交互式分析查询

Impala 是基于 MPP 架构的新型查询系统，它提供比现有 SQL-on-Hadoop 引擎具有简易使用和快速查询的特点，支持标准的 ANSI SQL 语法；Impala 支持 Hive 元数据查询存储在多种存储系统上的数据。另外 Impala 具有较好的可扩展性，可以很好的与典型 BI 应用系统协同工作，对于即席查询(Ad-hoc 查询)需求无疑是首选工具。网易猛犸团队对社区版本做了以下改进提升：

用户权限隔离

开源版本的 Impala 只支持 impala 用户执行所有的数据访问操作，不同用户的操作会造成数据权限不一致，无法被其它查询引擎使用等问题，我们基于开源 Impala 版本添加支持用户权限隔离，实现用户数据的自治和不同引擎之间的共享。

基于 Zookeeper 高可用和负载均衡

Impala 典型的高可用方案是基于 HAproxy+Keepalived 实现，但是这种方案扩展性一般并且不能够和 Hive 兼容，我们由此开发了基于 Zookeeper 的高可用负载均衡方案，以此兼容 Hive 的使用方式。

集中式的查询审计和管理系统

每一个 Impalad 都可以作为 SQL 引擎提供服务 ,导致每一个节点保存了部分的查询详细信息，这样增加了用户的使用难度，由此我们开发了集中式的查询审计和管理系统，支持不同用户查看不同的 SQL 查询信息。

细粒度的权限控制

开源版本的 Impala 只支持 ALL/INSERT/SELECT 三种权限，无法做到诸如 CREATE/UPDATE/DROP 等细粒度的权限，我们对此进行修改以支持细粒度的权限控制，更好的保证了数据安全。

元数据同步

Impala 和 Hive 等 SQL 引擎共享元数据存在无法同步 DDL 操作的问题，我们基于现有的 Impala 架构增加了同步 DDL 操作的功能，实现元数据在不同 SQL 引擎之间实时的同步。

元数据权限集成 Ranger

社区版本 Impala 权限系统只能与 Apache Sentry 集成，我们针对这个问题实现了与 Apache Ranger 的集成，实现统一的元数据和数据管理。

兼容 Apache Hive 的客户端

Impala 虽然可以直接使用 Hive 的 URL 进行连接，但是仍然存在一些参数有所区别，因此对原有客户端进行封装以支持使用与 Hive 完全一致的 URL 访问 Impala。

2.3 数据安全

原生 Hadoop 在数据安全领域的限制较少，非常开放。但在实际业务中，尤其是涉及机密和敏感数据时，仅限授权用户访问就至关重要。同时访问是否合理等信息也需要系统记录下来，让管理员可以回溯，进一步保证数据安全。平台通过认证（Authentication）、授权（Authorization）、审计（Audit）三个方面来保证数据安全。

认证

认证是用户进入系统的第一道屏障。平台采用了 MIT 开发的 Kerberos 做用户级别的认证。Kerberos 的设计主要针对 client-server 模型，基于加密方法建立用户（和系统）识别自己的方法，对个人通信以安全的手段进行身份认证，用户和服务器都能验证对方的身份。

授权

平台提供基于角色和个人的访问控制。对 HDFS、Hive 等实现了统一的，细粒度的数据访问控制。从数据角度，可以查看当前何种角色/何人有何种权限。从角色/个人角度，可以查看对哪些数据有何种权限。

审计

平台为项目安全提供较直观的整体评估和事件跟踪，包括实时监测对系统敏感信息的访问和操作行为，根据规则设定报警并及时阻断违规操作，收集并记录行为，可检索所有记录，提供统计信息五个方面。

监控处理的信息包括用户动作，管理员动作两大类。用户动作，所有用户的登录信息，对数据、对资源、对服务的访问和操作等；管理员动作，管理员对项目、成员等做出的配置等。

2.4 平台运维与监控

Ambari 是大数据生态组件管理系统，包含了安装部署、配置管理、监控告警等组件与集群管理功能，并集成了所有网易大数据生态组件，包括自研组件 Mammut、DataStream、Sloth 等以及社区版本中并未集成的 Impala 等。网易猛犸团队对社区安装部署方式进行改进，提供富安装包模式，无需外网或者部署 Repo 仓库即可完成安装部署，使其更适用于企业环境的安装部署。丰富监控能力，让问题更显而易见。丰富告警能力，不止支持邮件，还支持易信告警。

2.5 数据可靠性

Hadoop 通过数千台机器组成大规模集群提供大数据能力，当集群规模变大以后，机器的各类型故障将变得频繁。例如：假设硬盘年故障率 3%，以 1000 台规模的集群计算，每台机器 12 块硬盘，则一年中将会有 360 块左右的磁盘故障，这对于数据可靠性来说是一个巨大的挑战。

HDFS 通过多方面的技术手段来保证数据可靠性。HDFS 通过把数据多副本保存到多机器来避免磁盘损坏导致数据丢失的风险；并通过自动恢复副本的能力，保证在磁盘损坏后维持集群中数据的副本数。同时 Hadoop 发行版通过 Ambari 进行集群管理，可以对每个节点从硬件、操作系统、进程状态到业务层面进行监控，及时发现各类异常状态，并及时产生告警，使得故障检测时间和修复时间大大缩短，从而保证集群稳定性与数据可靠性。

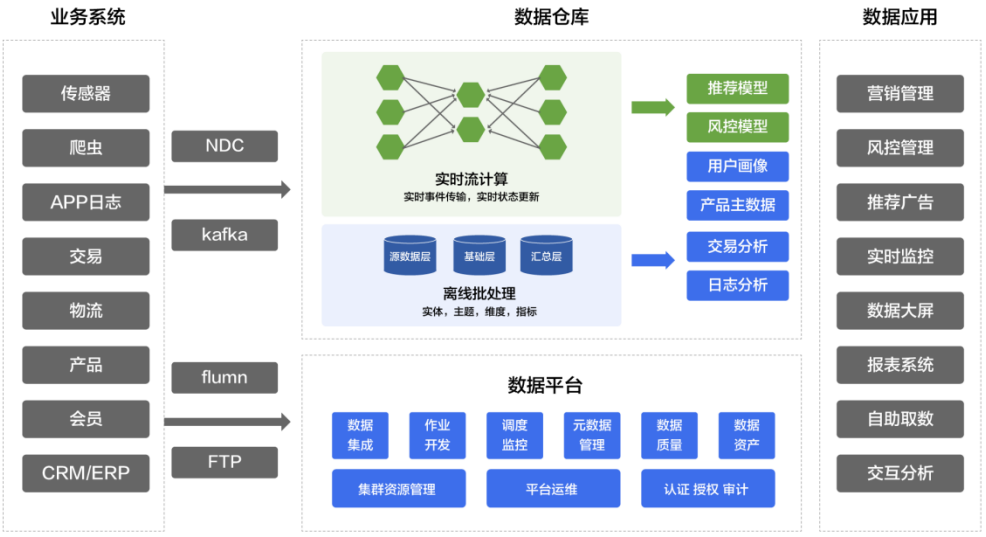
以磁盘故障举例：1000 个节点，每个节点 12 块盘，年故障率 3%，HDFS 副本数默认使用 3 个副本；根据网易大数据集群的运维实践，从磁盘故障、收到告警到完成换盘过程耗时 5 分钟左右；所以在 5 分钟之内同时坏掉 3 个磁盘导致 3 个副本全部失效的概率只有：0.000004%，系统的数据可靠性可达 99.99999%（>7 个 9）。

3. 基于猛犸的大数据应用建设方案

数据仓库建设方案

对于当下日益激励的市场环境，企业为提升市场竞争力，在生产制造过程，供应链，销售等经营过程收集数据，分析挖掘，用于过程精细化流程控制，大数据分析和挖掘方法为企业完成大数据落地提供了方法支持。企业管理系统如：ERP，CRM，CMS等，还是日渐完善的物联网数据，结合现代数据采集和传输技术，更容易被采集、传输并存储，结构化，半结构化，甚至视屏、音频等二进制数据的加工和利用，数据内容的种类更加丰富。传统的数据计算平台，无论容量，计算能力都难以跟上数据多样性和数据体量的增长速度。

猛犸大数据平台，依托开源社区 Hadoop 更好的适应现代数据应用场景，平台通过 Sqoop，Flume 等数据传输工具，将多样的数据形式从不同的数据源导入到平台，通过 NDC，Kafka 实现实时数据接入，在数据平台行进行统一存储，清洗，加工，集成，建模，将多种不同源的数据在平台上进行关联与集成，按数据层次组织划分 数据主题，建立维度，度量，指标等，丰富数据宽度，沉淀数据中间层。



猛犸平台能满足离线，准实时，实时等的多种数据应用场景，构建不同时间周期的数据应用，例如：流量日志实时监控，生产设备状态实时监控预警，风控实时预警等实时应用；又如：用户画像，用户标签，商品推荐，精准营销，交叉销售等离线数据分析和挖掘场景，平台提供友好的交互界面，降低交互式分析过程使用门槛，为业务分析团队数据探索和业务建模过程提供良好的平台和工具支持。

4. 技术规格

Impala 模糊查询指标

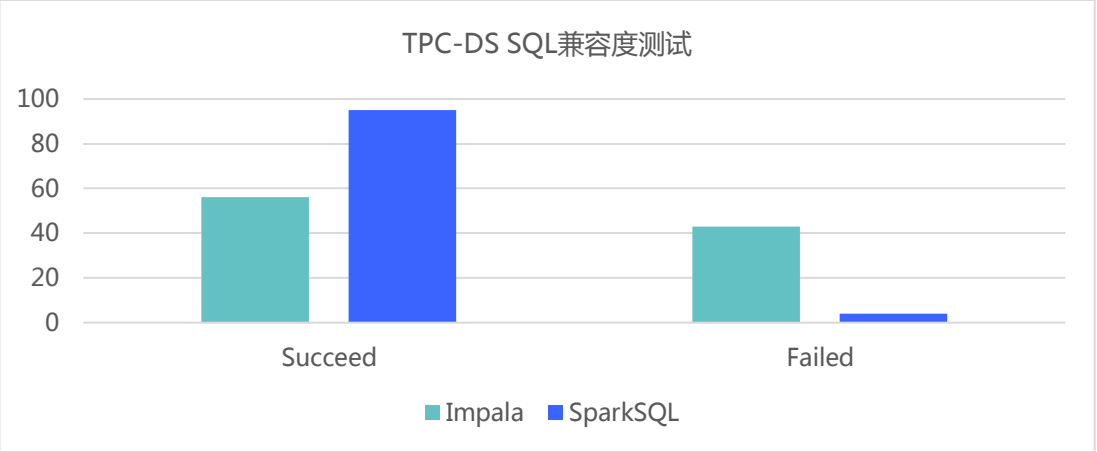
分类	指示	性能	说明
Impala 模糊查询	Impala 使用 like 进行指定字符串列模糊匹配查询性能 记录格式：<(id bigint)，(license string)> 匹配字段长度：9 Bytes 测试记录行数：28799781846	240ms	20 节点集群节点配置： CPU: 2 * E5-2630 内存：128G 磁盘：12 x 3.6T SATA

HBase 性能指标

分类	指标项	规格	说明
HBase 性能指标	100%写入 :平均每节点写入记录数(每条记录 500 Bytes)，响应时间小于 20ms	39000 records/s	8 节点集群节点配置： CPU: 2 * E5-2440 内存： 96G 磁盘：12 x 3.6T SATA
	100%随机读 :平均每节点写入记录数(每条记录 500 Bytes)，响应时间小于 20ms	13000 records/s	
	顺序扫描 :平均每节点 scan 操作数(每条记录 500 Bytes)，响应时间小于 50ms	7000 ops	
	读写混合(1:1): 平均每节点操作记录数(每条记录 500 Bytes)，响应时间小于 20ms	25000 records/s	

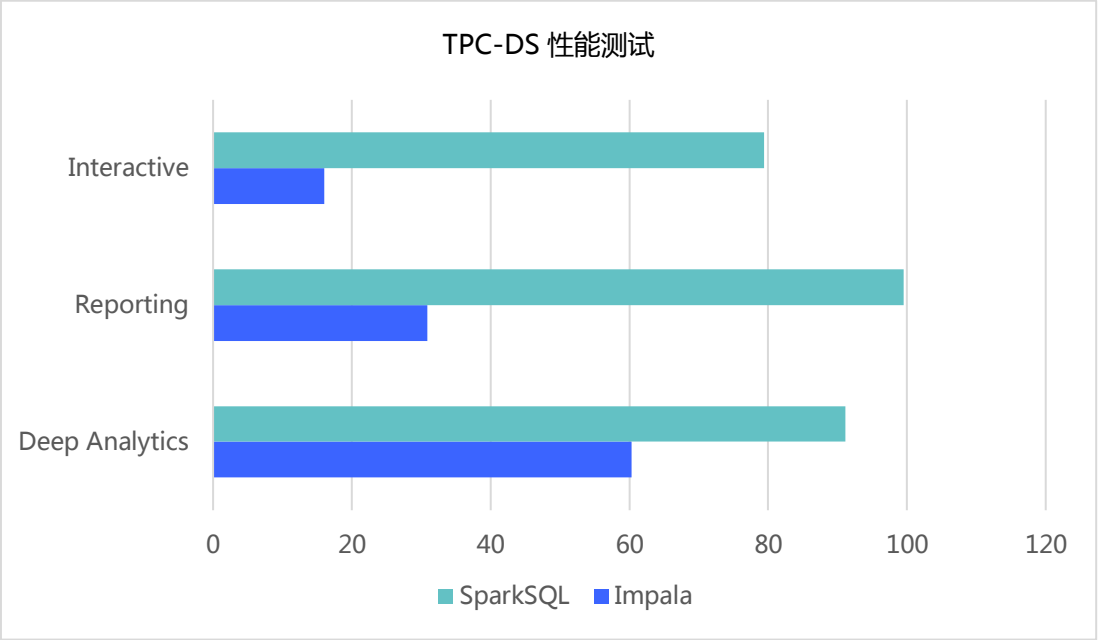
TPC-DS SQL 兼容度测试

分类	测试 SQL 集	Succeed	Failed
Impala	TPC-DS SQL99	56	43
SparkSQL		95	4



测试类型	Impala 性能(平均)	SparkSQL 性能(平均)	性能提升倍数
Interactive	16s	79.4s	5.0
Reporting	30.9s	99.5s	3.2
Deep Analytics	60.3s	91.1s	1.5

TPC-DS 性能测试



备注： 数据集由 TPC-DS 自带的工具生成，数据集的大小通过参数 scale-factor=10240，数据集大约 10T

5. 组件版本

组件名称	版本	说明	说明
Hadoop	2.7.3	-	-
Zookeeper	3.4.6	-	-
Kerberos	1.10.1	-	-
Ambari	2.4.2	-	-
Sqoop	1.4.6	改造	-
Hive	1.2.1	改造	-
Impala	2.8.0	改造	Cloudera 版本，非 Apache 版本
Spark	2.1.1	改造	-
HBase	1.1.2	-	-
Azkaban	3.0.0	改造	-
Kafka	0.9.0.1	-	-
Ranger	0.5.4	改造	
Hadoop-Meta	4.0	自研	代理创建 Kerberos、LDAP 用户，设置 Yarn 队列配置等
LDAP	2.4.40	-	-

关于我们

网易猛犸团队由网易的各领域技术产品专家组成，各成员已在各自领域积累了十年的卓越技术和宝贵经验，强强联合只为打造国内最优秀的一站式大数据开发和数据管理平台。

网易大数据

电话：0571-89852485

邮箱：bigdata-bd@hz.netease.com

地址：浙江省杭州市滨江区网商路599号