



MININGLAMP
明 略 数 据

MDP : Mininglamp Data Platform

产品白皮书

基于 Apache Hadoop 的大数据平台，专注于安全、可靠、易用、开放的企业级需求

明略数据

2015/10/1

MDP Introduction

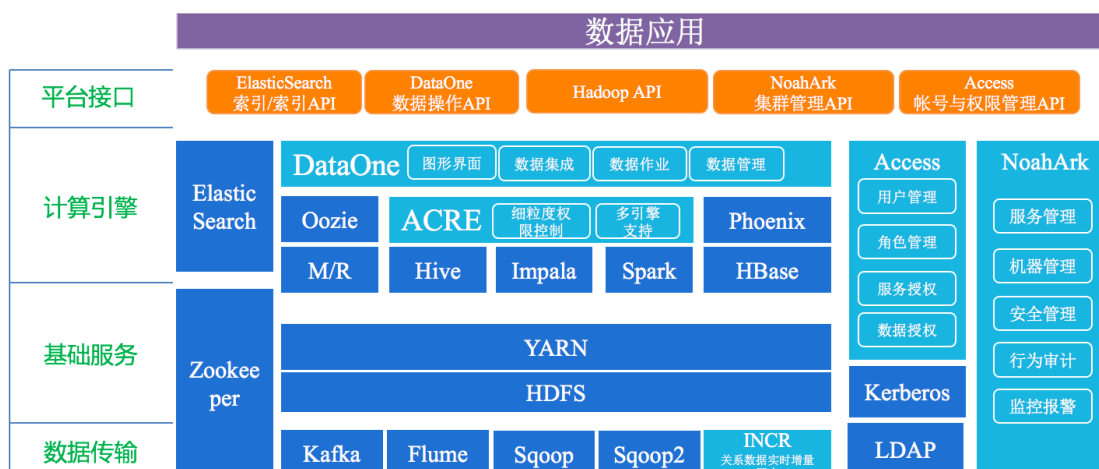


图 1 MDP 架构图

Mininglamp Data Platform (MDP)是明略数据基于 Apache Hadoop 研发的大数据平台。MDP 不仅提供海量数据存储和多种高性能计算框架，还为保护平台上的数据和服务提供了完整的安全保障体系、图形化的平台管理和作业、统一的身份认证和权限管理、细粒度的权限控制以及实时增量数据同步工具。

除了 Hadoop 组件外，MDP 集成了明略自有组件包含：

- 统一集群与服务管理 NoahArk
- 统一身份与授权管理组件 Access
- 统一数据作业平台 DataOne
- 细粒度的跨引擎 SQL 权限控制组件 ACRE
- 实时增量关系数据库同步工具 INCR

MDP 包含了十六种 Hadoop 社区最常用的组件，可以支撑多种计算类型的应用的混合负载，例如批处理应用、交互式查询、高频读写、全文检索、数据挖掘和实时流计算等多种计算

类型。各行各业可基于这些计算手段和方式进行上层应用的建设。

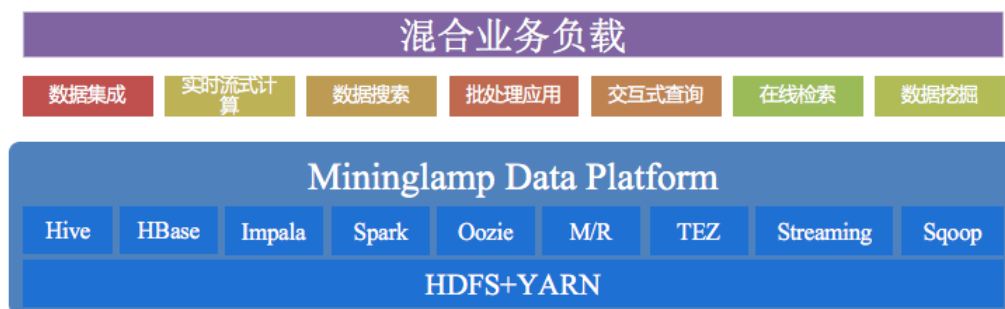


图 2 MDP 支持混合业务负载

统一集群与服务管理 NoahArk

NoahArk (诺亚方舟) 是明略自主研发的企业级Hadoop管理系统。NoahArk为MDP提供可视化的集群管理支持，帮助管理员提高运维效率、保证服务质量、优化集群性能、减少管理成本。NoahArk的主要功能包含仪表盘、机器管理、服务管理、调度器配置、安全审计、安全管理等功能模块。

NoahArk的特点：

➤ 极速部署

NoahArk系统针对hadoop生态系统中常用服务组合，提供简洁的整体安装方案，达到1小时内构建整个数据平台的急速部署。

➤ 精简操作

NoahArk的设计目标是在任何规模，都能简单和明确的管理企业数据平台；用户可以集中地部署和管理整个集群。 NoahArk重点优化集群管理的常用操作，力求一键式管理；并支持扩展非常用操作。

➤ 便捷扩展

NoahArk不仅能便捷横向扩展节点，也可以扩展符合标准的服务；使企业一站式管理集群上的全部服

务，包括hadoop生态系统中的组件和其他业务组件。

➤ 安全生产

NoahArk为大数据平台提供最高级别的保障，实时监控硬件、系统、数据平台、在线服务等各个部分，发现故障第一时间报警；同时，NoahArk还具备安全审计的功能，可对用户访问服务和数据的行为进行审计。

➤ 全面管理

NoahArk系统实现了所见即所得的效果，提供了统一的机器和服务的安装部署、实时监控、运维管理、日志汇总等功能，并提供可视化、易用性界面，满足运维管理人员针对集群机器、服务及组件的统一运维管理需求；统一集中式管理使得集群部署和运维操作更加快速、规范，监控报警更加实时有效；统一的日志管理方便更加快速定位、解决集群中存在的问题。

➤ 丰富API

NoahArk系统提供了丰富的API接口，允许运维管理人员以及其它有需求的业务部门通过对API的调用即可方便的进行集群的资源管理和监控报警等信息的获取

统一身份与授权管理 Access

Access 是 MDP 中负责账号和权限管理的组件。它主要由以下特点：

➤ 基于LDAP的独立身份系统

Access使用LDAP存储平台上的用户身份信息，并支持与企业原有的用户体系对接。

➤ 基于Kerberos的安全身份认证

Access集成了Kerberos作为MDP的安全身份认证体系，支持单点登录。

➤ 功能级别的服务授权管理

Access支持HDFS、YARN、HBase、Oozie等关键组件的服务授权功能，能够细粒度的限制用户使用服务的不同功能。

➤ 细粒度的数据授权（通过ACRE）

Access基于ACRE提供了MDP上细粒度的数据授权功能，能够对平台上的数据访问提供行访问权限控制以及列访问权限控制。

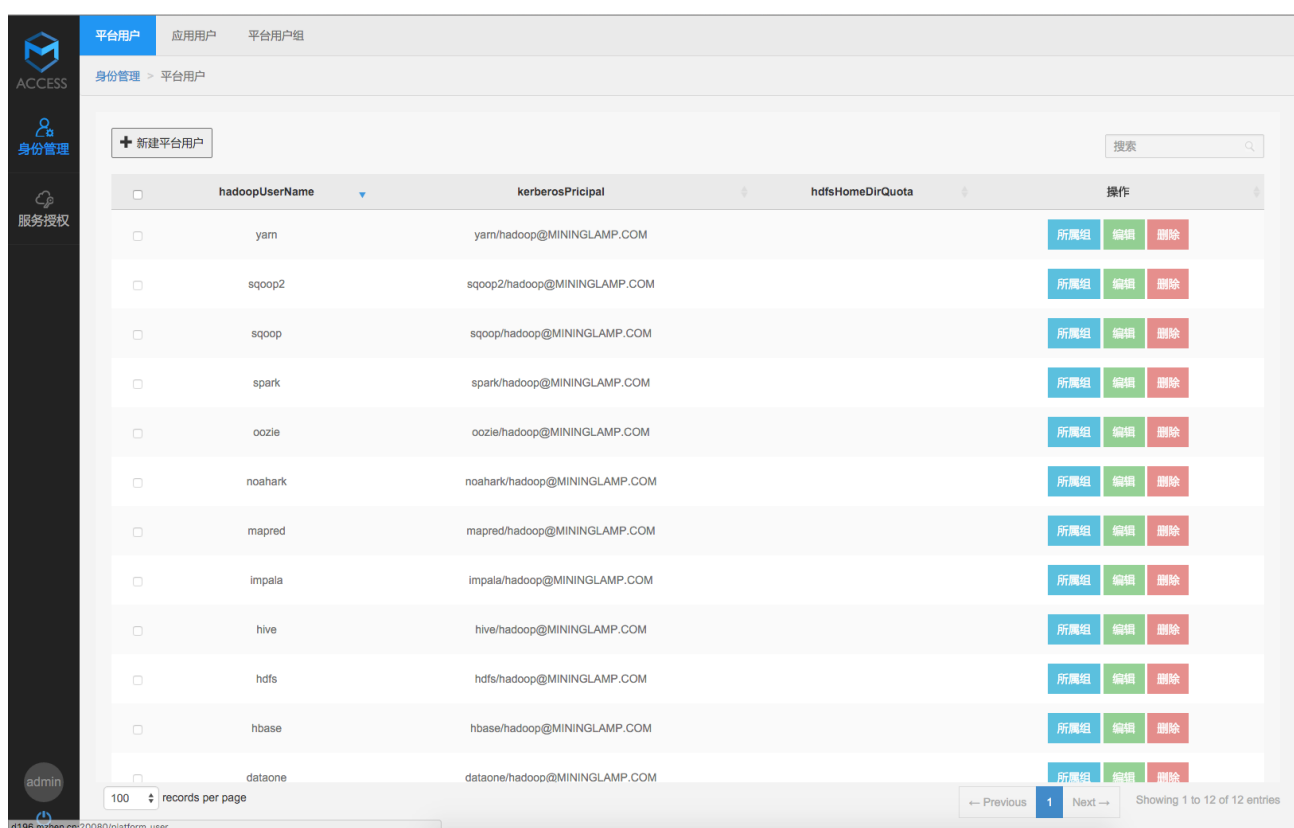


图 3 Access 平台用户管理

统一数据作业平台 DataOne

DataOne,是明略数据结合自己多年的大数据技术使用经验，考虑企业传统IT技术人才的

北京明略软件系统有限公司
MiningLamp Software System Co., Ltd.

A. 北京市昌平区中东路 398 号中煤建设大厦 1 号楼 4 层 邮政编码：102218
F4, 1#, Zhongmei Construction Group Plaza, Zhongdong Road, Changping District, Beijing
T. 010-8423389 F. 010-56842040 H. www.mininglamp.com

技术特点，为降低大数据实施成本，加速大数据落地，简化用户操作，优化用户体验而开发的统一大数据存储和处理的作业平台。DataOne是基于明略Hadoop大数据平台的全链路数据集成、存储管理、处理分析的大数据作业平台。其中主要包含了数据集成、数据处理、数据作业、数据查询等模块，通过全界面的方式提供给用户使用。

DataOne 其突出的极大特点是：

1. 使用大数据平台连接传统数据平台成本低，速度快
2. 完成数据的全生命周期管理
3. 操作容易方便

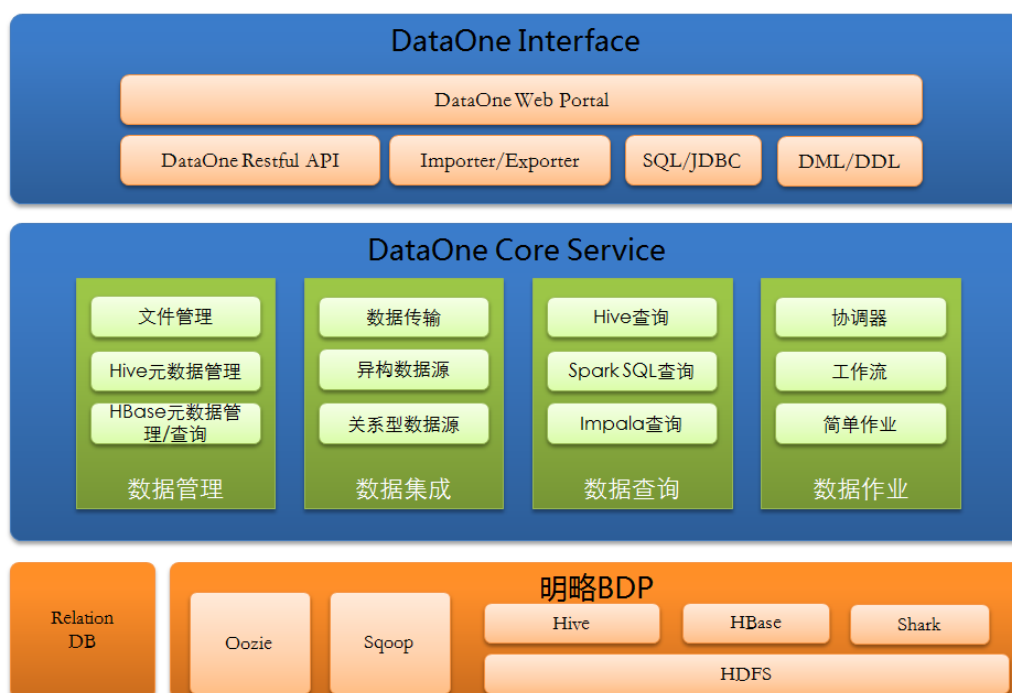


图 4 DataOne 产品结构图

细粒度的跨引擎 SQL 权限控制 ACRE

ACRE 是 MDP 独有的在 Hive、Impala、Spark SQL 上支持行列 (Cell) 级别访问控制的数据权限管理组件。

ACRE 有以下特点：

➤ **支持基于角色的授权模型**

ACRE中支持建立角色，且ACRE中的权限是赋予在角色上的。管理员可将角色授予不同的用户组，则用户组中的用户就有了角色上所赋予的权限。

➤ **支持Hive、Impala、Spark的统一授权管理**

ACRE是平台级的跨引擎的数据权限管理组件，MDP上的Hive、Impala、Spark SQL三种引擎，均可以通过ACRE进行权限控制。

➤ **支持行列（Cell）基本访问控制**

ACRE提供了细粒度的权限控制机制，管理员可以给角色赋予行权限和列权限，以控制用户在不同表内部的数据访问范围和访问方式。

➤ **基于访问过滤的实现无需修改原数据库/表结构**

ACRE提供的是插件式的访问控制机制，ACRE通过对访问SQL进行检查和过滤来发生作用，并不会对数据进行任何的修改。

实时增量关系数据库同步 INCR

INCR 独创性的搭建了传统关系型数据库到大数据平台的高速公路，让所有传统应用软件所产生的数据流实时、增量的同步到大数据平台，参与数据关联和实现真正意义上的实时数据分析。

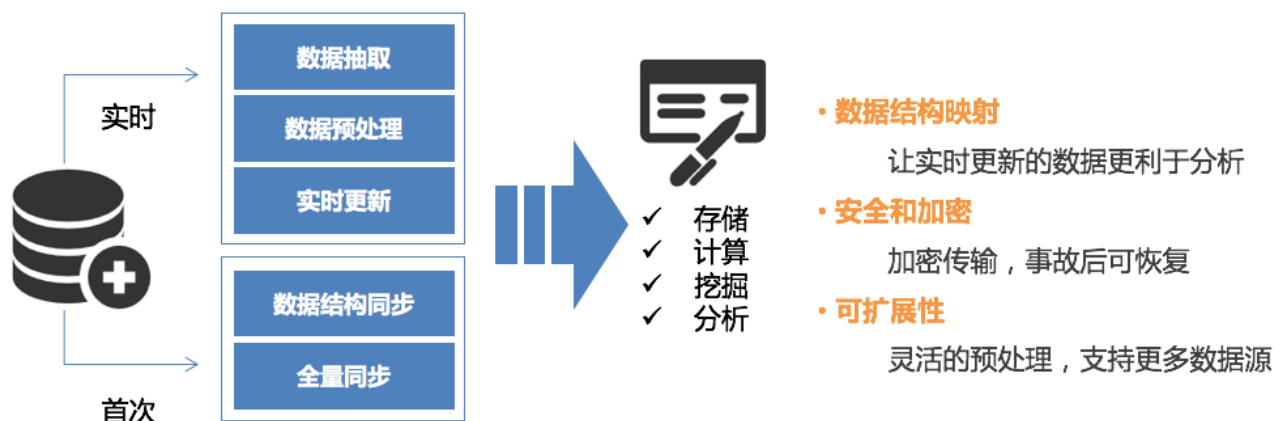


图 5 INCR 的特点

INCR 支持全量同步、实时增量同步、监控与报警、高性能等特点。同时，INCR 也具有非常好的可扩展性，可以通过横向扩展来支持多个数据源的同步。

MDP 之安全性

安全是 MDP 最关注的企业级需求，MDP 提供了网络安全、主机安全、服务安全、数据安全四层安全机制来保障整个平台的安全。

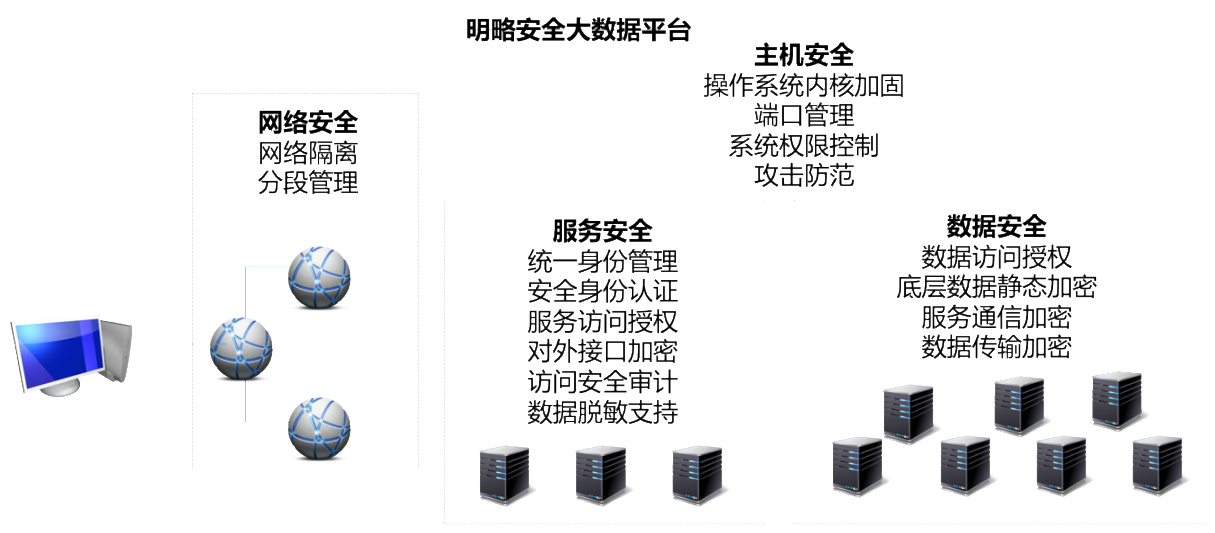


图 6 MDP 的四层安全保障

MDP 之可靠性

MDP 中的关键服务均进行了主从热备机制，这些服务包含了 HDFS Namenode, Yarn ResourceManager, HBase Master, HiveServer2, Oozie 等。另外，MDP 独有的组件，如 NoahArk, DataOne 等均提供了双活热备机制，来保障平台的正常运行。

平台服务

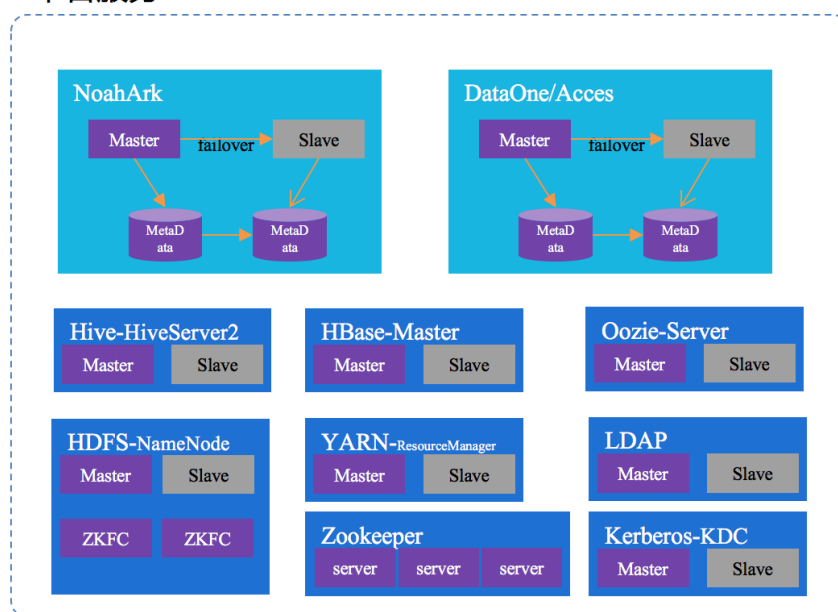


图 7 MDP 的服务热备机制

平台上各个服务的关键元数据，MDP 也结合数据库和存储设备的 raid 机制等进行了热备和冷备的多重备份机制。

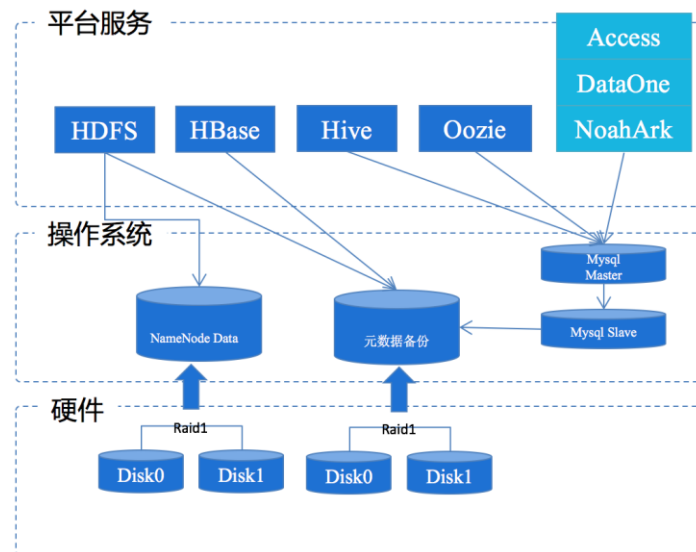


图 8 MDP 服务元数据备份机制

MDP 之易用性

MDP NoahArk 提供了全图形化的界面来支持平台运维工作。

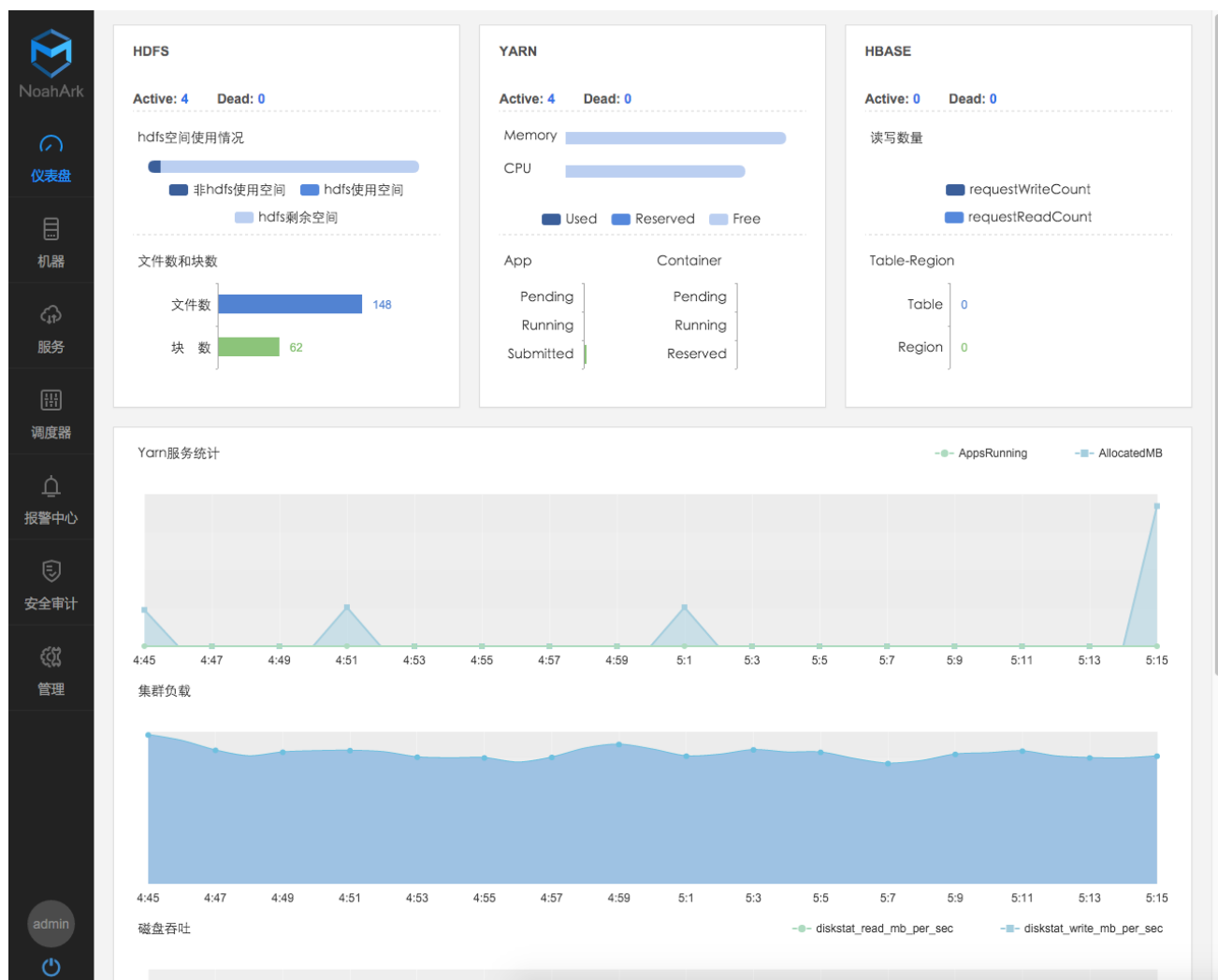


图 9 NoahArk 仪表盘

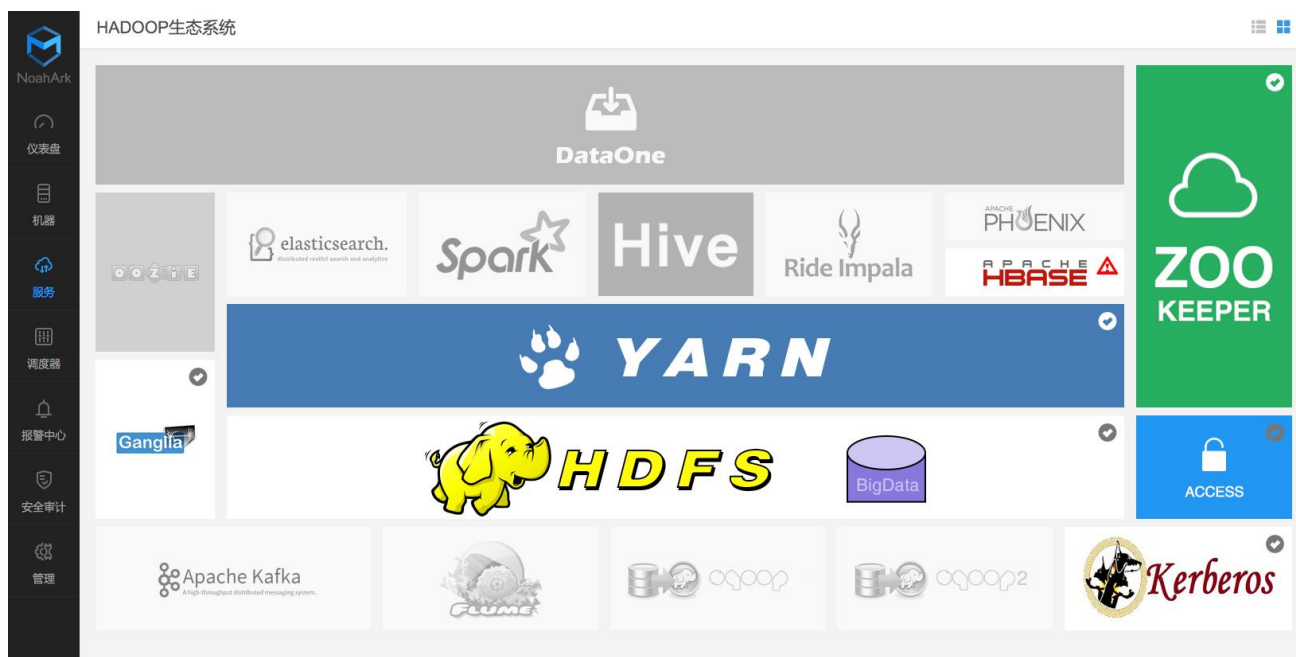


图 10 NoahArk 服务管理

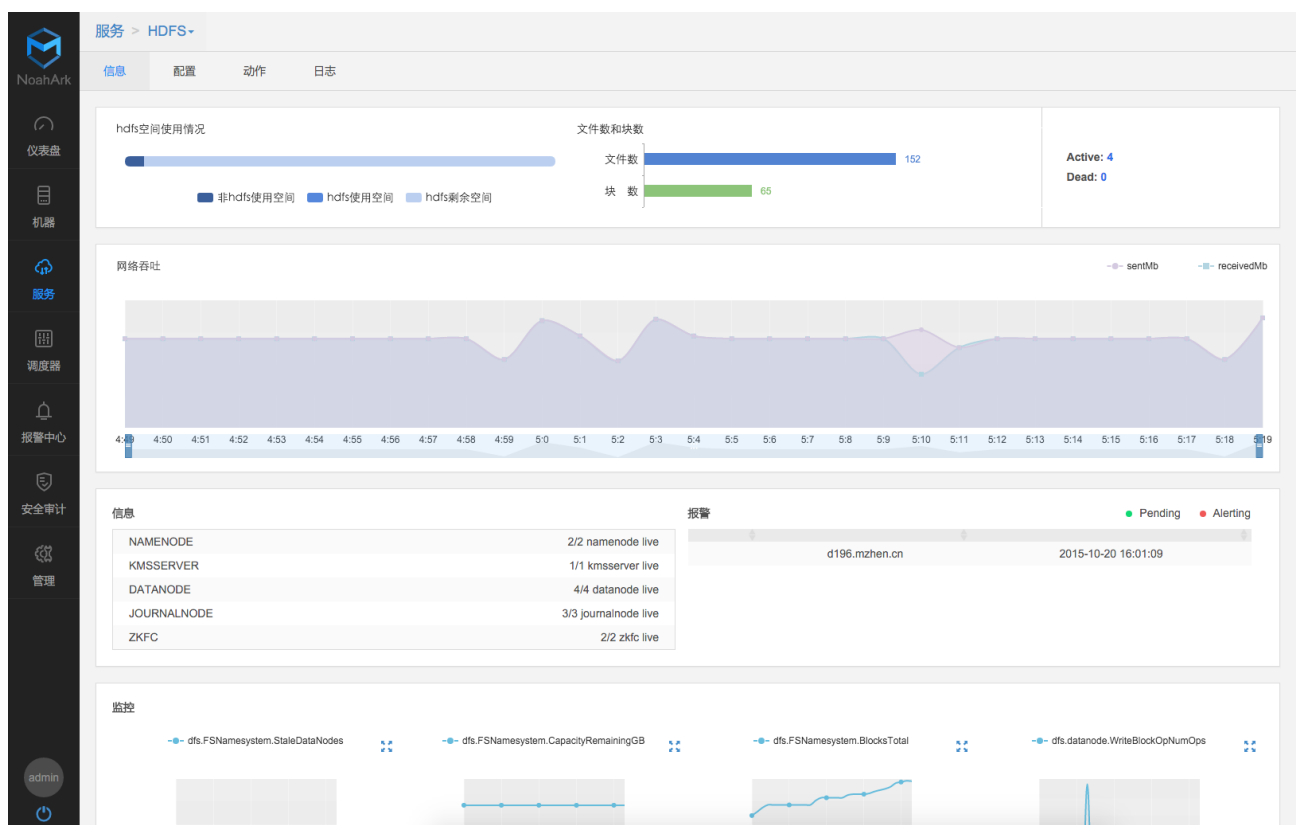
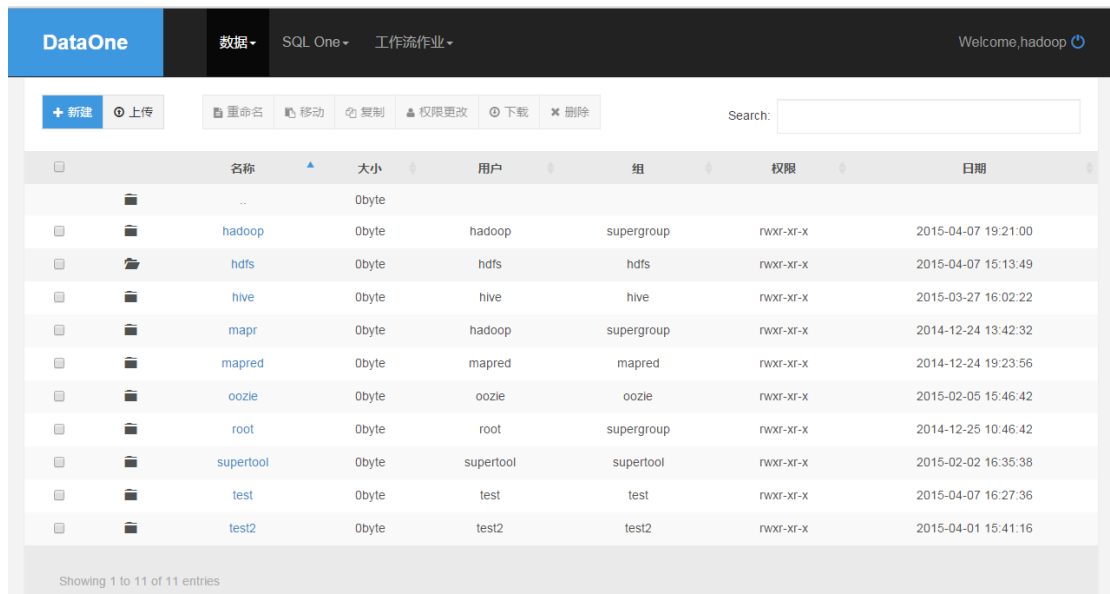


图 11 NoahArk 服务页面

MDP DataOne 提供了全图形化的界面来支持平台上的各种数据作业操作。

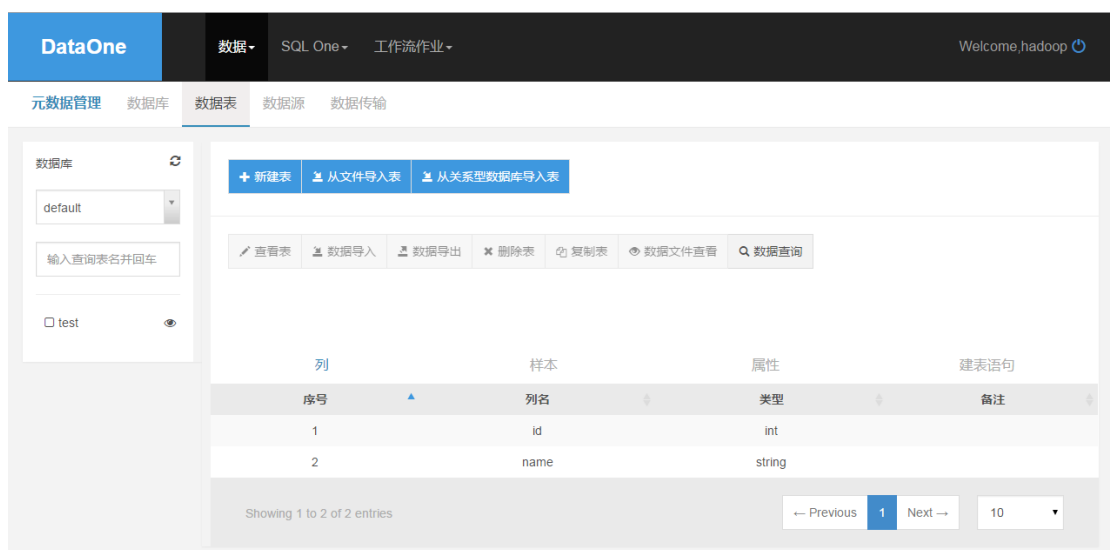
Dataone 数据管理包含了基础的文件管理(HDFS)、结构化数据的数据管理(Hive)、非结构化数据的数据管理(HBase)等功能。



The screenshot shows the DataOne file management interface. It features a top navigation bar with 'DataOne', '数据' (Data), 'SQL One', and '工作流作业' (Workflow Jobs). A search bar is located on the right. Below the navigation bar, there are tabs for '+ 新建' (New), '上传' (Upload), and a set of action buttons: '重命名' (Rename), '移动' (Move), '复制' (Copy), '权限更改' (Change Permissions), '下载' (Download), and '删除' (Delete). The main area displays a table of files and directories.

	名称	大小	用户	组	权限	日期
	..	0byte				
	hadoop	0byte	hadoop	supergroup	rwxf-xf-x	2015-04-07 19:21:00
	hdfs	0byte	hdfs	hdfs	rwxf-xf-x	2015-04-07 15:13:49
	hive	0byte	hive	hive	rwxf-xf-x	2015-03-27 16:02:22
	mapr	0byte	hadoop	supergroup	rwxf-xf-x	2014-12-24 13:42:32
	mapred	0byte	mapred	mapred	rwxf-xf-x	2014-12-24 19:23:56
	oozie	0byte	oozie	oozie	rwxf-xf-x	2015-02-05 15:46:42
	root	0byte	root	supergroup	rwxf-xf-x	2014-12-25 10:46:42
	supertool	0byte	supertool	supertool	rwxf-xf-x	2015-02-02 16:35:38
	test	0byte	test	test	rwxf-xf-x	2015-04-07 16:27:36
	test2	0byte	test2	test2	rwxf-xf-x	2015-04-01 15:41:16

Showing 1 to 11 of 11 entries



The screenshot shows the DataOne table management interface. It features a top navigation bar with 'DataOne', '数据' (Data), 'SQL One', and '工作流作业' (Workflow Jobs). Below the navigation bar, there are tabs for '元数据管理' (Metadata Management), '数据库' (Database), '数据表' (Data Table), '数据源' (Data Source), and '数据传输' (Data Transfer). The main area displays a table of tables.

列	样本	属性	建表语句
序号	列名	类型	备注
1	id	int	
2	name	string	

Showing 1 to 2 of 2 entries

DataOne 数据集成能够定义关系型数据源和异构数据源，支持建立数据传输任务，将数据源中的数据导入到 DataOne 中。数据集成支持自定义异构数据集成脚本，异构数据经处理后存储为结构化数据。DataOne 结合明略实时数据增量传输工具 INCR，可实现与关系型数据源的实时同步。

北京明略软件系统有限公司
 MiningLamp Software System Co., Ltd.

A. 北京市昌平区中东路 398 号中煤建设大厦 1 号楼 4 层 邮政编码：102218
 F4, 1#, Zhongmei Construction Group Plaza, Zhongdong Road, Changping District, Beijing
 T. 010-8423389 F. 010-56842040 H. www.mininglamp.com

DataOne 数据 SQL One 工作流作业 Welcome, hadoop

元数据管理 数据库 数据表 数据源 数据传输

+ 新建数据传输作业 × 删除数据传输作业 修改数据传输作业 复制数据传输作业

数据库源作业 异构数据源作业

序号	名称	属性	执行次数	创建时间	最后执行时间	当前状态	操作
29	hds_import_xls	import	19	2014-12-15 1...	2015-01-05 2...	DELETED	运行
30	hds_import_csv	import	9	2014-12-15 1...	2015-03-17 1...	DELETED	运行
41	KK_KIOX	import	3	2015-01-05 1...	2015-01-05 1...	DELETED	运行
38	adwawd	export	1	2014-12-29 1...	2014-12-29 1...	DELETED	运行
42	v_gnlk	import	2	2015-01-05 1...	2015-01-05 1...	DELETED	运行
22	rds_import_h...	import	7	2014-12-15 1...	2015-02-02 1...	DELETED	运行
31	hds_import_script	import	7	2014-12-15 2...	2014-12-16 1...	DELETED	运行
44	useinfo	import	2	2015-01-05 1...	2015-01-05 1...	DELETED	运行
45	persion	import	2	2015-01-05 1...	2015-01-05 1...	DELETED	运行
46	BTV2	import	19	2015-01-13 1...	2015-01-13 1...	DELETED	运行
47	ss	import	4	2015-01-13 1...	2015-01-19 1...	DELETED	运行

DataOne 数据查询中整合了多种支持 SQL 查询（结构化数据查询）的查询引擎，并提供统一的查询界面。

DataOne 数据 SQL One 工作流作业 Welcome, test2

Hive Editor 查询编辑器 我的查 Hive Spark SQL Impala

辅助 数据库 default 输入查询表名并回车

orc_l_o_incr...
orc_l_o_incr...
orc_l_o_incr...
orc_l_o_incr...
orc_l_o_incr...
orc_l_o_incr...
orc_l_o_incr...
orc_l_o_incr...
orc_l_o_incr...
orc_l_o_incr...

```
1 select * from t_trans_info_cache
```

执行 另存为... 解释 或创建一个 新查询 结果导出

结果	列	图表			
id	trans_at	settle_dt	loc_trans_dt_tm	mchnt_cd	pri_acct_no
1	1111	2013-01-01	0101000000	100000001348587	10000000000115917406
2	2222	2013-01-01	0101000000	100000002533180	100000000000456575036
3	3333	2013-01-01	0101000000	100000000211701	100000000000827549676
4	4444	2013-01-01	0101000000	100000001874274	100000000000534851160

结果导出

DataOne 数据作业包含了作业编辑、管理、提交以及任务管理的功能。其中作业分为简单作业、工作流、协调器(定时器)等几种不同的作业类型。简单作业就相当于一个个独立的程

序，工作流可以把多个作业串起来，协调器负责控制工作流的运行时机。当作业提交执行后，任务管理将监控作业运行的状况，以图形化的方式进行任务管理。

DataOne
数据
SQL One
工作流作业
Welcome, test2

编辑器
简单作业
工作流
协调器

+ 创建协调器
+ 导入协调器
修改协调器
删除协调器
复制协调器

ID	协调器名称	工作流名称	创建用户	频率	创建时间	上次修...	当前状态	详细信息	操作
88a6a65...	indata1	hive-test	hadoop	1minute	2014-10-...	2015-04-...	shared		运行
18	MMMM	UMSETL...	test2	1month	2015-03-...	2015-03-...	shared	ETL月计...	运行
134aa00...	copy_of_...	UMSETL...	test	1month	2015-01-...	2015-01-...	shared	ETL月计...	运行
412f29a2...	copy_de...	hive-test	hadoop	1minute	2015-01-...	2015-01-...	shared		运行
e697c59...	UMSETL...	UMSETL...	hadoop	1month	2014-12-...	2014-12-...	shared	ETL月计...	运行
bf2de652...	UMSETL...	UMSETL...	hadoop	1month	2014-12-...	2014-12-...	shared	ETL月计...	运行
3c290e9...	UMSETL...	UMSETL...	hadoop	1day	2014-12-...	2014-12-...	shared	ETL之宽...	运行
45dc795...	mytest	mymytest	hadoop	2minute	2014-12-...	2014-12-...	shared		运行
9d7c20b...	coord_fil...	godsaythi...	hadoop	2minute	2014-12-...	2014-12-...	shared		运行
9ed2ba5...	coord_w...	work_demo	hadoop	1day	2014-12-...	2014-12-...	shared		运行
619b7b0...	xxx	d196_my...	hadoop	1day	2014-11-...	2014-12-...	shared	1111	运行

DataOne
数据
SQL One
工作流作业
Welcome, test2

控制面板
任务
协调器

请输入查询的序号或名称

正在运行

序号	名称	持续时间	提交者	创建时间	开始时间	结束时间	最后修改时间	当前状态	操作
没有数据									

首页
上一页
下一页
每页 20

已完成

序号	名称	持续时间	提交者	创建时间	开始时间	结束时间	最后修改...	当前状态	操作
0000002-1...	xingye_de...	58s	test2	2015-04-0...	2015-04-0...	2015-04-0...	2015-04-0...	SUCCEEDED	重运行
0000001-1...	xingye_de...	54s	test2	2015-04-0...	2015-04-0...	2015-04-0...	2015-04-0...	SUCCEEDED	重运行
0000025-1...	xingye_de...	103693s	test2	2015-03-3...	2015-03-3...	2015-04-0...	2015-04-0...	KILLED	重运行

经过培训的技术人员，能够基于 DataOne 快速完成一系列复杂的数据集成、数据处理、数据分析等工作，大大加速大数据作业的速度。

MDP 之开放性

MDP 中的所有开源组件来源于 Apache 社区，不包含任何明略自有的黑盒定制代码，客户后期可自主选择通过社区或者其他途径升级和维护各个组件的版本，MDP 不通过定制化的黑盒代码绑定客户。

使用 MDP 这样的开源平台的好处有：

	商业开源的 MDP	商业闭源 Hadoop 平台
快速升级	纯开源组件，可随社区升级	由于自身引入大量代码，跟随社区脚步慢，难以做到快速升级
依赖绑定	不受平台供应商绑定，可自行根据社区进行平台与应用的开发、维护与升级	受供应商绑定，需依赖供应商提供新的组件版本进行平台升级和应用升级
知识产权	所有基于开源平台的开发代码均提供给甲方，业务实现技术手段对甲方透明	闭源部分不提供给甲方，以黑盒的方式面对客户
可持续性	Hadoop 生态系统异常繁荣，可预见其生命力将非常持久	平台生命力依赖供应商公司，一旦公司业务变化或其他问题将带来不可控的风险
开发维护支持	由产品供应商和开源社区双重保障，开发维护不依赖于供应商	需依赖于供应商提供开发维护支持，其闭源特性可能引入特定的开发接

口,性能优化和开发设计均需要原厂支持

MDP 之核心组件构成

MDP 的开源组件和独有组件列表：

Component	Name	Version
OpenSource	Zookeeper	3.4.5
	HDFS	2.6.0
	YARN	2.6.0
	Hive	1.2
	Hbase	1.0
	spark	1.5.1
	impala	2.0
	sqoop	1.4.4
	sqoop2	1.9.9
	Phoneix	4.4.0
	oozie	4.1.0
	flume	1.6.0



	kafka	2.10-0.8.2.1
	elasticsearch	1.5.2
	ganglia	3.7.1
MDP 独有		
	NoahArk	2.0
	DataOne	2.0
	Access	2.0
	Acre	2.0