

Loader 技术原理

www.huawei.com





目标

- 学完本课程后，您将能够：
 - 熟悉**Loader**是什么
 - 熟悉**Loader**能干什么
 - 熟悉**Loader**在**FusionInsight**产品的位置
 - 掌握**Loader**的系统架构
 - 掌握**Loader**的主要特性
 - 掌握如何管理**Loader**作业
 - 掌握如何监控**Loader**作业



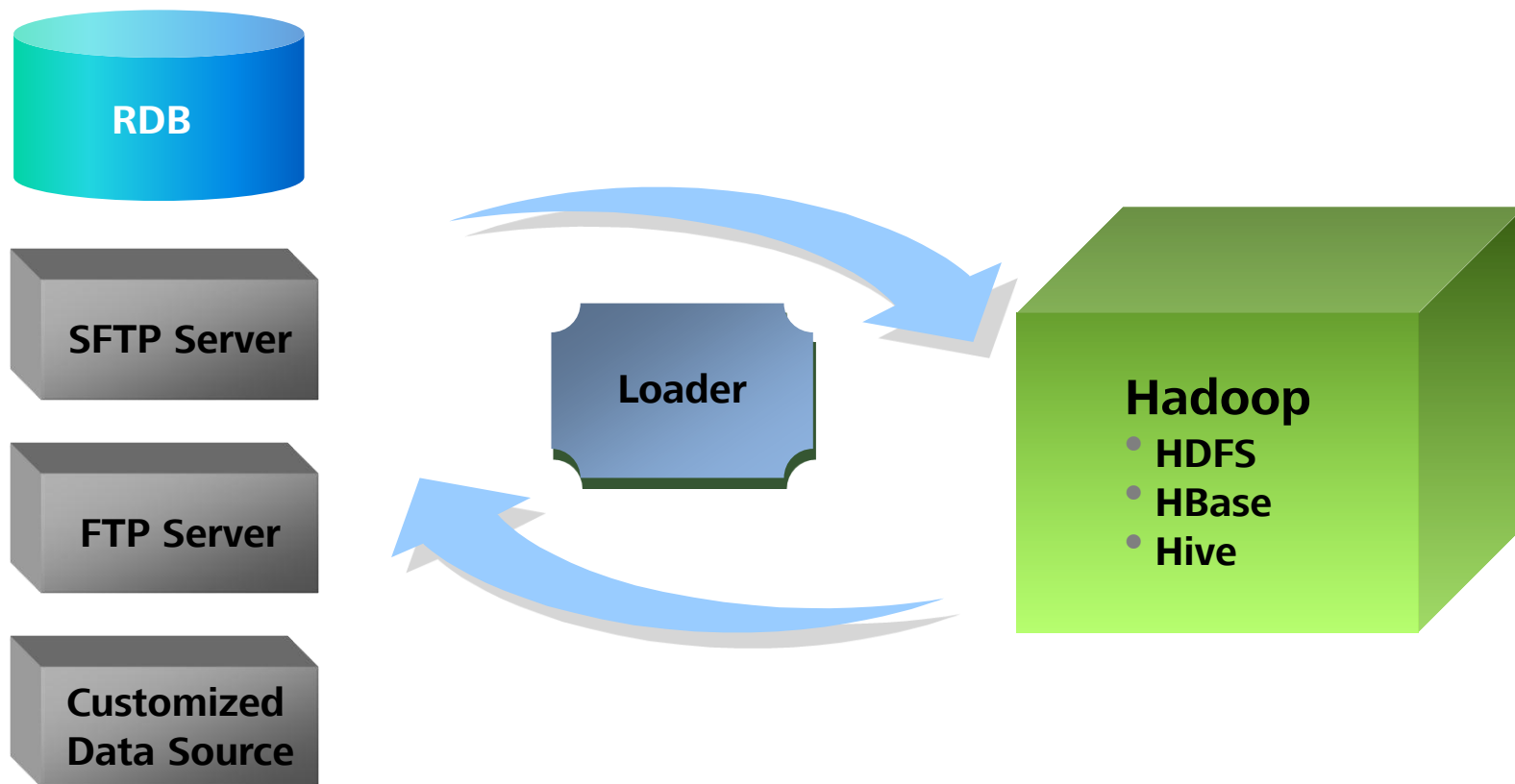
目录

1. Loader简介
2. Loader作业管理

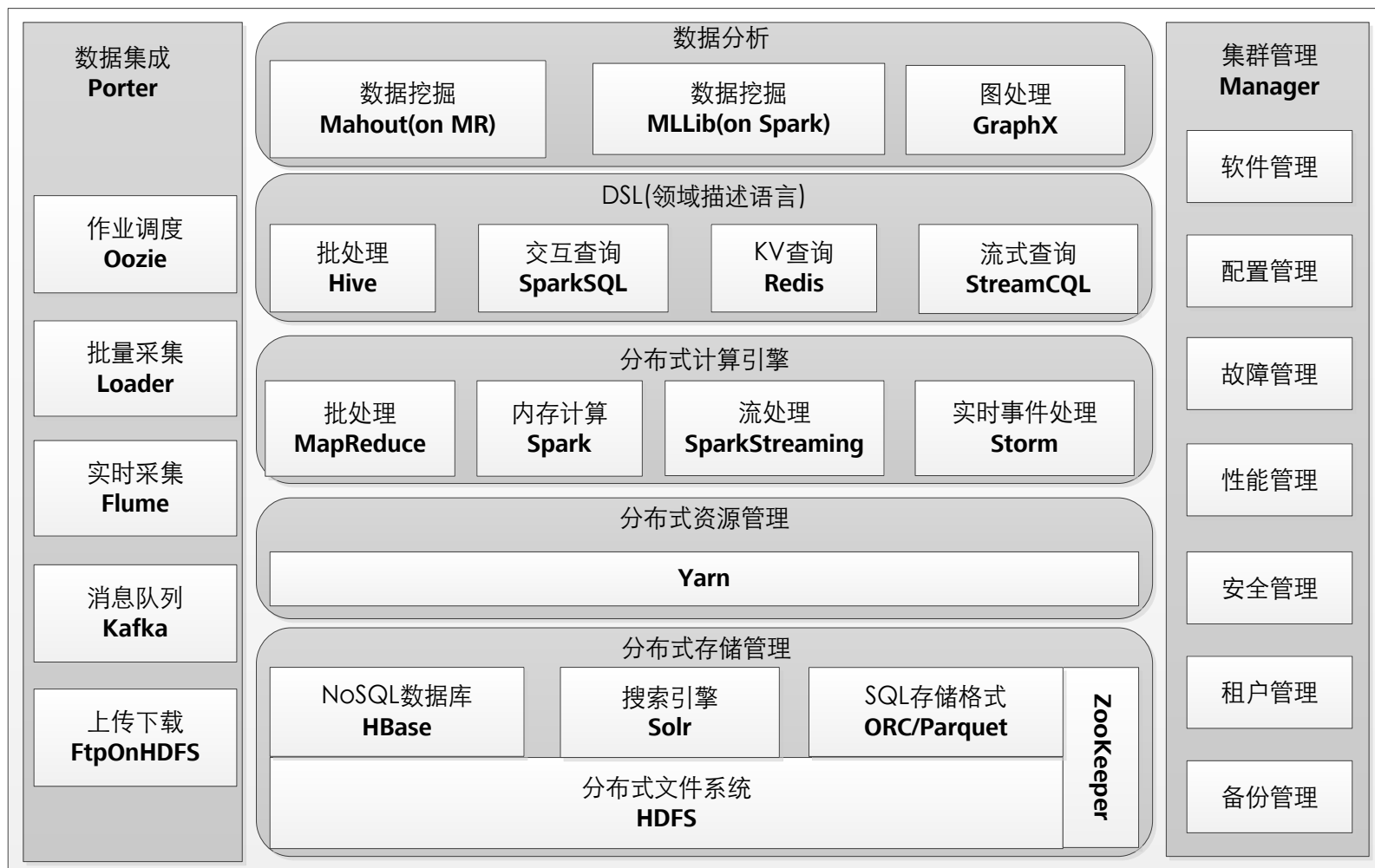
什么是Loader

Loader是实现**FusionInsight HD**与关系型数据库、文件系统之间交换数据和文件的数据加载工具。基于开源**Sqoop**研发，做了大量优化和扩展。提供可视化向导式的作业配置管理界面；提供定时调度任务，周期性执行**Loader**作业；在界面中可指定多种不同的数据源、配置数据的清洗和转换步骤、配置集群存储系统等。

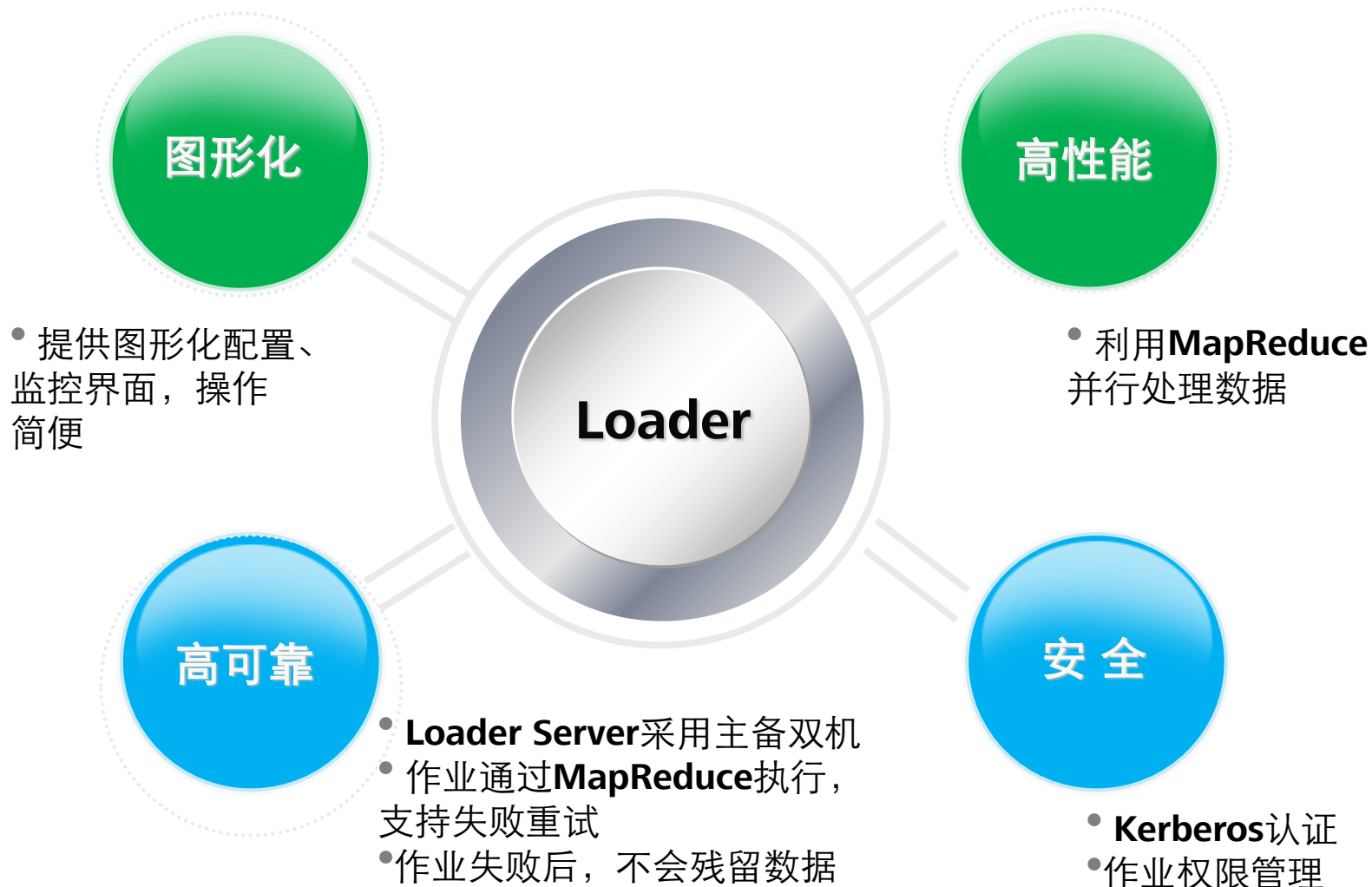
Loader的应用场景



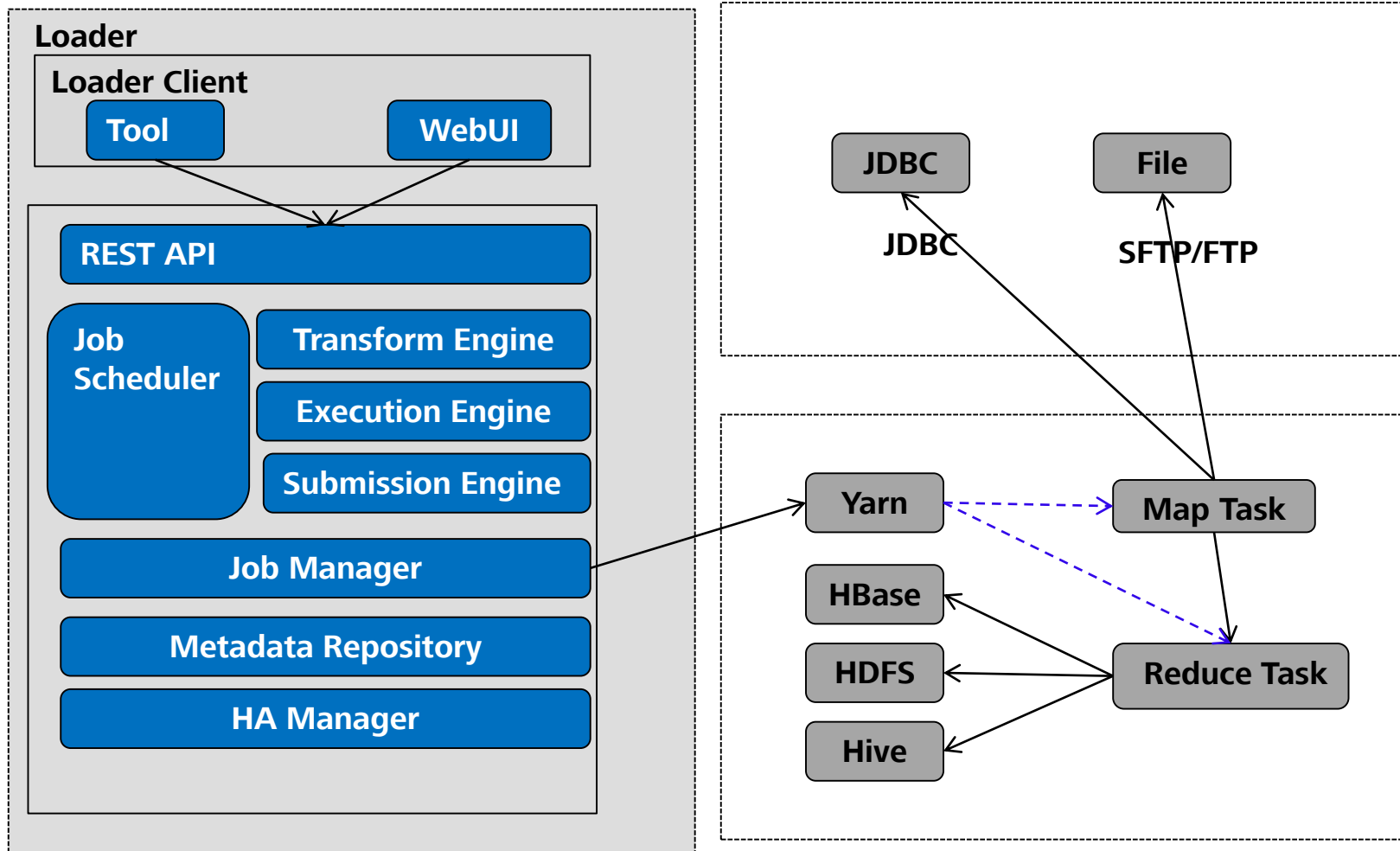
Loader在FusionInsight产品的位置



Loader特点



Loader模块架构



Loader模块架构 – 模块说明

名称	描述
Loader Client	Loader 的客户端，包括 WebUI 和 CLI 两种交互界面。
Loader Server	Loader 的服务端，主要功能包括：处理客户端请求，管理连接器和元数据，提交 MapReduce 作业和监控 MapReduce 作业状态等。
REST API	实现 RESTful (HTTP+JSON) 接口，处理来自客户端的请求
Job Scheduler	简单的作业调度模块，支持周期性的执行 Loader 作业。
Transform Engine	数据转换处理引擎，支持字段合并、字符串剪切、字符串反序等。
Execution Engine	Loader 作业执行引擎，包含 MapReduce 作业的详细处理逻辑。
Submission Engine	Loader 作业提交引擎，支持将作业提交给 MapReduce 执行。
Job Manager	管理 Loader 作业，包括创建作业、查询作业、更新作业、删除作业、激活作业、去激活作业、启动作业、停止作业。
Metadata Repository	元数据仓库，存储和管理 Loader 的连接器、转换步骤、作业等数据。
HA Manager	管理 Loader Server 进程的主备状态， Loader Server 包含 2 个节点，以主备方式部署。



目录

1. Loader简介
2. Loader作业管理

Loader的服务状态界面

点击“服务管理”，选择Loader，进入“Loader服务状态”界面。

The screenshot shows the FusionInsight Manager web interface. The top navigation bar includes icons for System Overview, Service Management (selected), Host Management, Alarm Management, Audit Management, Tenant Management, and System Settings. Below this, a breadcrumb trail shows 'Service' > 'Loader Service Status'. The main content area has tabs for 'Service Status' (selected), 'Instances', 'Service Configuration', and 'Resource Contribution Ranking'. A toolbar offers actions like 'Start Service', 'Stop Service', 'Download Client', and 'More Operations'. The 'Loader Overview' section displays a real-time status table with rows for Health Status (Good), Configuration Status (Synchronized), Version (sqoop-1.99.3), and Loader WebUI (LoaderServer(备) and LoaderServer(主)). Below this is a table for 'Operation Status and Health Status' showing LoaderServer instances in a 'Started' state with 'Good' health.

Loader 概述		实时
健康状态	✓ 良好	
配置状态	✓ 已同步	
版本	sqoop-1.99.3	
Loader WebUI	LoaderServer(备) LoaderServer(主)	

操作状态和健康状态		
角色	操作状态	健康状态
LoaderServer	✓ 2 已启动	✓ 2 良好

Loader作业管理界面

在“Loader服务状态”界面，点击“LoaderServer(主)”，进入Loader 作业管理界面。



Loader作业管理界面 – 作业

作业用来描述将数据从数据源经过抽取、转换和加载至目的端的过程。它包括数据源位置及数据源属性、从源数据到目标数据的转换规则、目标端属性。

Loader提供了诸多功能，用于管理与作业相关的操作。包括创建作业、导入作业、导出作业、迁移作业分组、批量删除作业、启动作业、停止作业、查看作业历史记录、复制作业和删除指定作业等功能。

Loader作业管理界面 – 作业转换规则

Loader提供了丰富的作业转换规则，能将数据按照不同的业务场景进行转换和清洗，转换成目标数据结构，实际应用中，如果不需要转换，可以不指定转换规则。

Loader提供了14种转换算子，描述如下：

- 长整型时间转换：实现长整型数值与日期类型的互换。
- 空值转换：将空值替换成指定值。
- 增加常量字段：生成常量字段。
- 随机值转换：生成随机数字段。
- 拼接转换：拼接已有字段，生成新字段。
- 分隔转换：将已有字段，按指定分隔符，分隔出新字段。
- 取模转换：对已有字段取模，生成新字段。
- 剪切字符串：通过指定起止位置，截取已有字符串类型的字段，生成新字段。

Loader作业管理界面 – 作业转换规则

- **EL操作转换**：指定算法，对字段值进行运算，目前支持的算法有：**md5sum**、**sha1sum**、**sha256sum**和**sha512sum**等。
- **字符串大小写转换**：对已有的字符串类型字段，切换大小写，生成新字段。
- **字符串逆序转换**：对已有的字符串类型字段，做逆序变换，生成新字段。
- **字符串空格清除转换**：对已有字符串类型字段，清除左右空格，生成新字段。
- **过滤行转换**：配置逻辑条件过滤掉含触发条件的行。
- **更新域**：当满足某些条件时，更新字段的值。

创建Loader作业 - 步骤1

- 一、配置作业的基本信息，包括名称、类型、数据源连接，所属分组，队列和优先级。

当没有需要的连接时，可以通过“连接”属性后的“添加”功能，创建连接。

作业 > 新建作业

1.基本信息

2.输入设置

3.转换

* 名称

* 类型

* 连接

组

* 队列

优先级

下一步

取消

sftpConnector

* 名称

Sftp服务器的IP	Sftp服务器端口	Sftp用户名	Sftp密码	操作
10.1.1.1	22	root	删除

添加

创建Loader作业 - 步骤2

- 二、配置作业的数据源属性。

作业 > 新建作业

1.基本信息 2.输入设置 3.转换 4.输出设置

* 输入路径

* 文件分割方式

过滤器类型

路径过滤器

* 文件过滤器

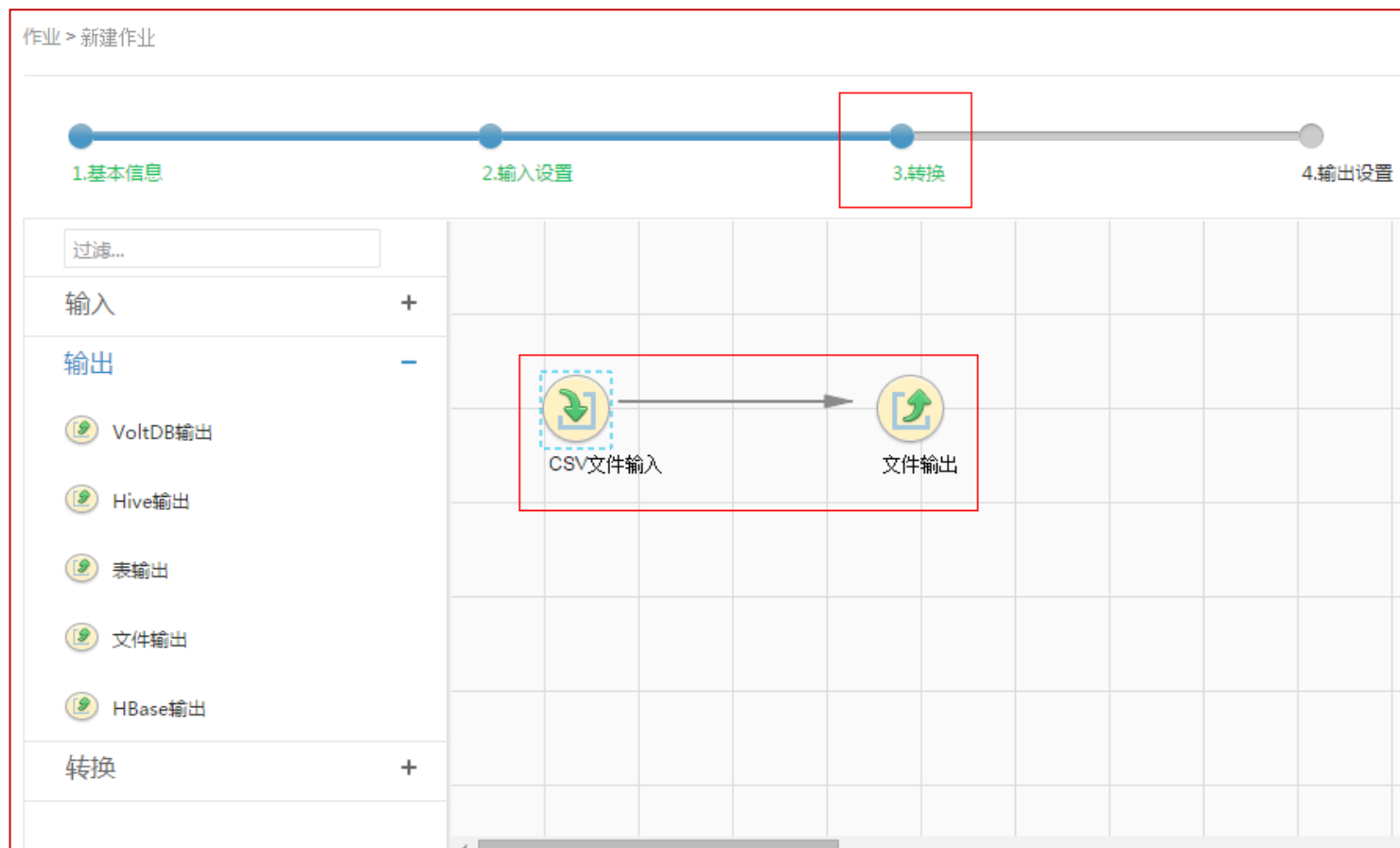
编码类型

后缀名

* 压缩 ☐ true ☒ false

创建Loader作业 - 步骤3

- 三、配置作业的数据转换规则。



创建Loader作业 - 步骤4

- 四、配置作业的目标端属性。

作业 > 新建作业



1.基本信息 2.输入设置 3.转换 4.输出设置

* 存储类型 HDFS ▼

* 文件类型 TEXT_FILE ▼

压缩格式 请选择... ▼

* 输出目录 /user/test

文件操作方式 OVERRIDE ▼

☒ Map数 ☐ Map数据块大小

* 个数 2

返回 保存 保存并运行 取消

监控作业执行状态

- 查看所有作业执行状态：
 - 1.进入**Loader**作业管理界面。
 - 2.界面会显示当前的所有作业和作业最后一次执行状态。
 - 3.选中一个作业，可以点击上方或右方“操作”一栏中的按钮执行相应的操作。

监控作业执行状态 – 作业历史记录

- 查看指定作业历史记录：
 - 1.选中一个作业，点击“操作”中的“历史记录”按钮，进入作业历史查看界面。
 - 2.该界面显示作业每次执行的开始时间、运行时间（秒）、状态、失败原因、行/文件读取数、行/文件 写入数、行/文件 跳过数、脏数据链接、**MapReduce**日志链接。

监控作业执行状态 – 脏数据查看

- 脏数据是指不符合**Loader**转换规则的数据，查看方式如下：
 - 在作业历史查看界面上，发现跳过记录数不为**0**时，点击“脏数据”按钮，进入该次作业执行产生的脏数据目录。
 - 脏数据存放在**HDFS**，每个**Map Task**处理的脏数据分别记录到相应文件

Browse Directory

/user/loader/etl_dirty_data_dir/2/1459942782531_0003

Go!




Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwx-----	admin	hadoop	0 B	Thu Apr 07 14:13:03 2016	0	0 B	2_m_000000
drwx-----	admin	hadoop	0 B	Thu Apr 07 14:13:14 2016	0	0 B	2_m_000001
drwx-----	admin	hadoop	0 B	Thu Apr 07 14:13:15 2016	0	0 B	FILE_ROW_CNT_INFO

Hadoop, 2015.

监控作业执行状态 – MapReduce 日志

- 在作业历史查看界面，点击“日志”按钮，进入该次执行的**MapReduce**日志界面。



MapReduce Job job_1459942782531_0003

Logged in as: admin

Job Overview

Job Name: Loader: sftp-hdfs_1460009555574

User Name: admin

Queue: default

State: SUCCEEDED

Uberized: false

Submitted: Thu Apr 07 14:12:40 CST 2016

Started: Thu Apr 07 14:12:52 CST 2016

Finished: Thu Apr 07 14:13:15 CST 2016

Elapsed: 23sec

Diagnostics:

Average Map Time 15sec

ApplicationMaster




Attempt Number	Start Time	Node	Logs
1	Thu Apr 07 14:12:46 CST 2016	loader2:26010	logs

Task Type	Total	Complete
Map	2	2
Reduce	0	0

Attempt Type	Failed	Killed	Successful
Maps	0	0	2
Reduces	0	0	0

监控作业执行状态 – 作业失败告警

作业运行失败，会上报告警。

作业							
<div>新建作业 导入 导出 作业迁移 批量删除</div> <div>根据作业名称搜索 每10秒刷新</div>							
作业ID	名称	描述	开始时间	执行者	进度	状态	操作
2	sftp-hdfs	从 SFTP 导入 到 HDFS	2016-04-07 14:23:46	admin	0%	提交失败	  
10	总条数: 1 < 1 >						

告警管理 > 告警				
告警 事件				
 导出全部  清除告警				
<input type="checkbox"/>	告警ID	告警名称	告警级别	产生时间
<input type="checkbox"/>	23003	Loader任务执行失败	严重	2016-04-07 14:23:50
当前页: 1 总页数: 1 < > 跳转到 确定				
告警详情				
告警ID:	23003	告警名称:	Loader任务执行失败	
告警级别:	严重	产生时间:	2016-04-07 14:23:50	
清除时间:		清除类型:	未清除	
是否自动清除:	否	序列号:	11	
告警原因:	提交任务失败。		定位信息:	ServiceName=Loader;RoleName=LoaderServer;HostName=loa
附加信息:	Details=The sftp file does not exist. cause: /opt/temsdtpfile			

客户端脚本介绍

Loader除了提供图形化操作界面外，还提供了一套完整的**shell**脚本，通过这些脚本，可实现数据源的增删查改，作业的增删查改、启动作业、停止作业、查看作业状态，判断作业是否正在运行等功能。

脚本介绍如下：

- **lt-ctl**：简称作业控制工具，用于查询作业状态、启动作业、停止作业以及判断作业是否在运行中。
- **lt-ucj**：简称作业管理工具，用于查询、创建、修改和删除作业。
- **lt-ucc**：简称数据源管理工具，用于查询、创建、修改和删除数据源连接信息。



本章总结

- 描述**Loader**的主要功能
- 描述**Loader**的主要特性
- 描述**Loader**作业的管理
- 描述**Loader**作业的监控

思考题

- 判断题： **FusionInsight**的**Loader**仅支持从关系型数据库与**Hadoop**的**HDFS**和**HBase**之间的数据导入、导出。
- 判断题： **Loader**作业必须配置转换步骤。



习题

- (多选题) 以下说法正确的有()
 - A、 作业运行了一段时间后失败了, 不会残留原始文件
 - B、 脏数据是指不符合转换规则的数据
 - C、 **Loader**客户端脚本只能提交作业
 - D、 创建了一个人机账号, 就可以操作所有**Loader**作业
- (发散题/单选) 以下说法正确的是()
 - A、 **Loader**将作业提交到**MR**执行后, 如果**Loader**故障, 则此作业执行失败。
 - B、 **Loader**将作业提交到**MR**执行后, 如果某个**Mapper**执行失败, 能够自动进行重试。
 - C、 **Loader**作业执行失败, 将会残留数据, 需用户手动清除。
 - D、 **Loader**将作业执行到**MR**执行后, 在该作业执行完成前, 不能再提交其他作业。

Thank you

www.huawei.com