大数据任务编排工具

文档状态：*草稿*

山东浪潮齐鲁软件产业股份有限公司

|  |  |  |  |
| --- | --- | --- | --- |
| 修订历史 | | | |
| 版本号 | 修改人 | 修改日期 | 摘 要 |
| V1.0 |  |  |  |
|  |  |  |  |
|  |  |  |  |

摘 要

摘要内容

关键字：

目录

[1 背景 1](#_Toc451188952)

[2 用户需求 2](#_Toc451188953)

[3 功能需求分析 3](#_Toc451188954)

[3.1 现有技术 3](#_Toc451188955)

[3.2 功能实现 4](#_Toc451188956)

[4 工作安排 6](#_Toc451188957)

[5 功能列表 7](#_Toc451188958)

[6 非功能需求分析 8](#_Toc451188959)

大数据任务编排工具需求分析

# 背景

随着大数据与云计算的发展。大数据的应用越来越广泛，各个领域利用大数据技术进行需求分析，发展规划以及规避风险等，各公司也迅速推出了自己的大数据产品，如HP的Vertica，Teradata的Teradata Aster，谷歌的BigQuery，浪潮的IDAP等。在大数据快速发展与组件化的潮流下，由于大数据开发的复杂性与多样性，开发过程中任务编排成了其面临的一大挑战。

为了提高开发效率与代码的复用性，提出了组件化得概念，组件化的出现解决了代码复用性低等问题，在项目应用中根据实际项目需求，只需要对已有组件进行编排即可完成特定功能，实现快速开发。

# 用户需求

为了满足客户与事业部需求，紧跟大数据发展步伐，推出了IDAP，且已完成 IDAP中部分组件的开发。现需要提供相应的开发工具，能够在可视化界面下对各组件以拖拽方式进行编排完成特定功能任务，同时可以监控各任务状态以及各任务实例运行状态与结果。

以旅游项目的应用场景为例：

* 从OTA网站上爬去酒店的数据(可以考虑定时、立即执行网络爬去数据的任务)
* 用flume将从OTA上的数据写到Hive数据表中
* 执行Hive SQL进行场景分析
* Sqoop将Hive SQL分析的数据抽取到关系型数据库中

需要支持的任务组件：

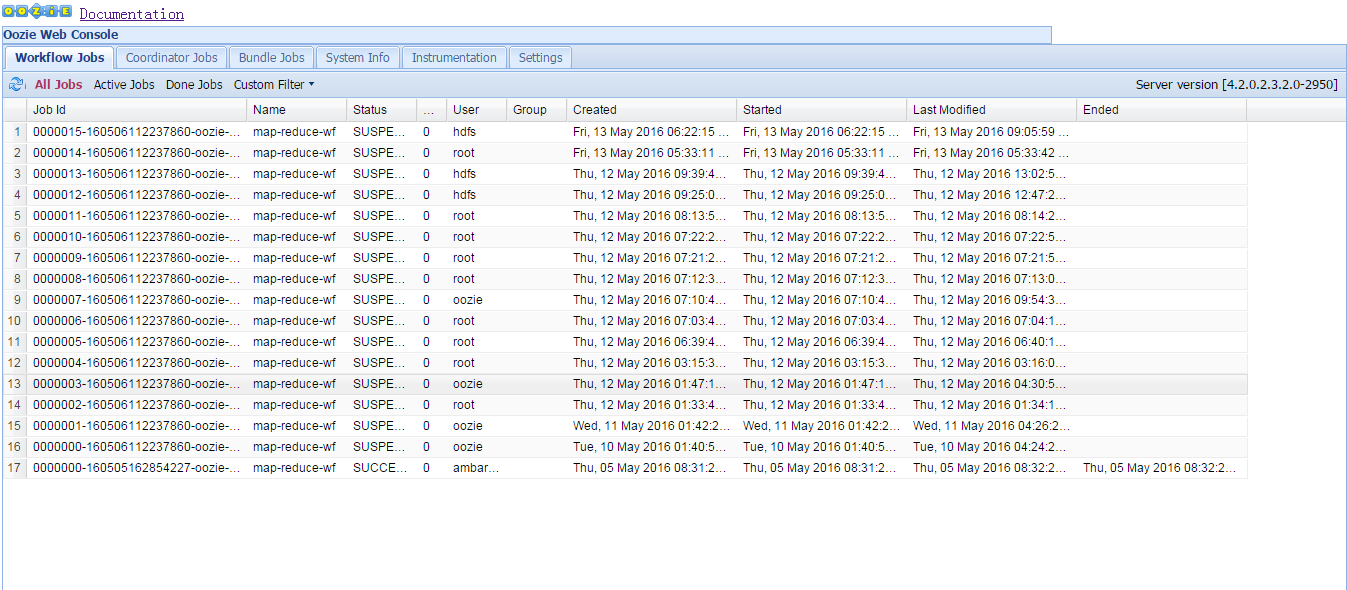
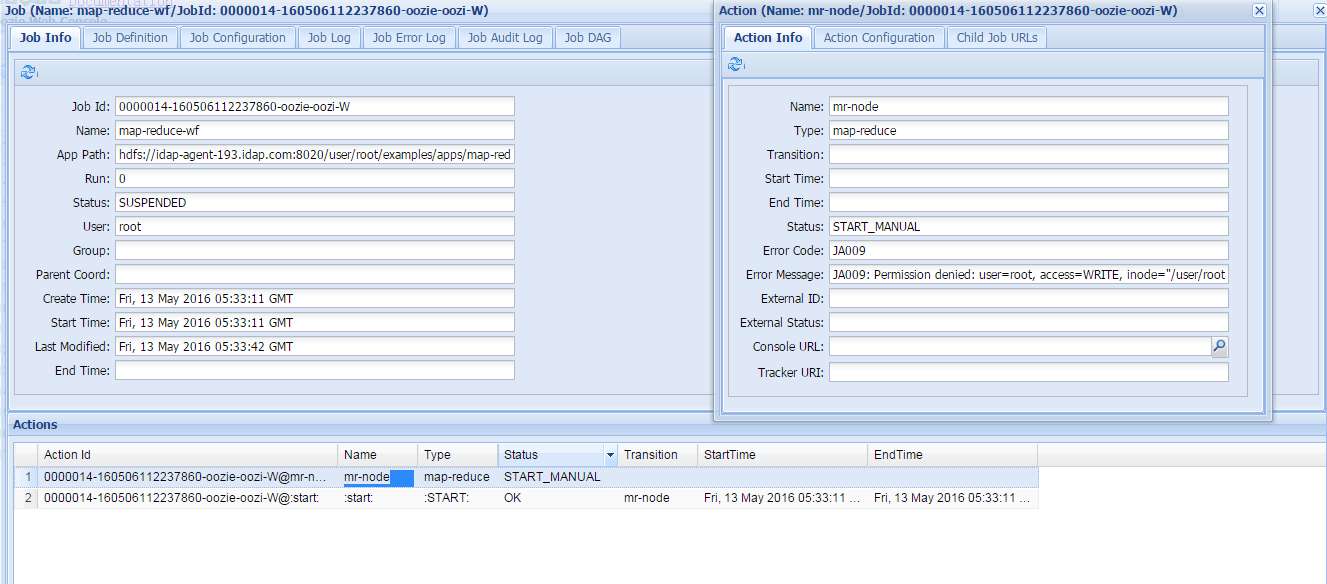
* Hive
* HBASE
* Sqoop
* Flume
* HDFS
* RocketMQ
* Spar
* Flume

# 功能需求分析

## 现有技术

1、Oozie

Ooize是一种用于支持hadoop任务编排的Java Web应用程序。主要包括：数据流引擎，任务管理等内容。数据流引擎本质上是一个以任务编排为目的,可以解析xml与运行Java代码的Java web应用程序，任务管理主要实现了对任务运行状态的跟踪，任务相关参数配置展示等功能。

Oozie作为一种成熟的大数据任务编排工具，还表现为支持多种任务节点：

* start 开始节点
* end 结束节点
* failed 失败节点
* kill 杀死节点
* decision 决策节点
* fork 分支节点
* join 会签节点

2．Hue

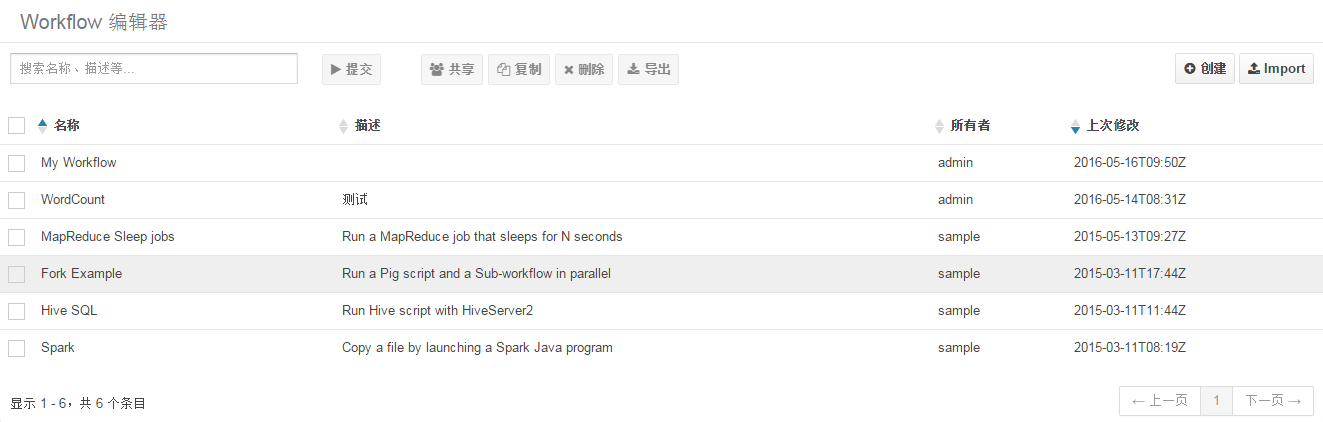
Hue是一个开源的Apache Hadoop UI系统，最早是由Cloudera Desktop演化而来，由Cloudera贡献给开源社区，它是基于Python Web框架Django实现的。通过使用Hue可以在浏览器端的Web控制台上与Hadoop集群进行交互来分析处理数据，例如操作HDFS上的数据，运行MapReduce Job等。Hue已经为开发者提供了任务管理，实例管理，拖拽界面。



## 功能实现

综合以上背景和需求，需要为IDAP平台提供一个可视化的，可在线编辑与拖拽功能的任务编排工具，能够帮助开发者在可视化界面下以拖拽方式快速开发出特定功能的大数据产品。并可以在任务编排过程中生成相关配置文件与执行文件，通过导出方式能够获取相关文件，保证文件的快速移植，同时需要对各任务实例的状态进行跟踪监测并进行展现，保证开发者掌握产品运行各阶段的状态与结果。综合分析大数据任务编排工具需要提供给以下功能：

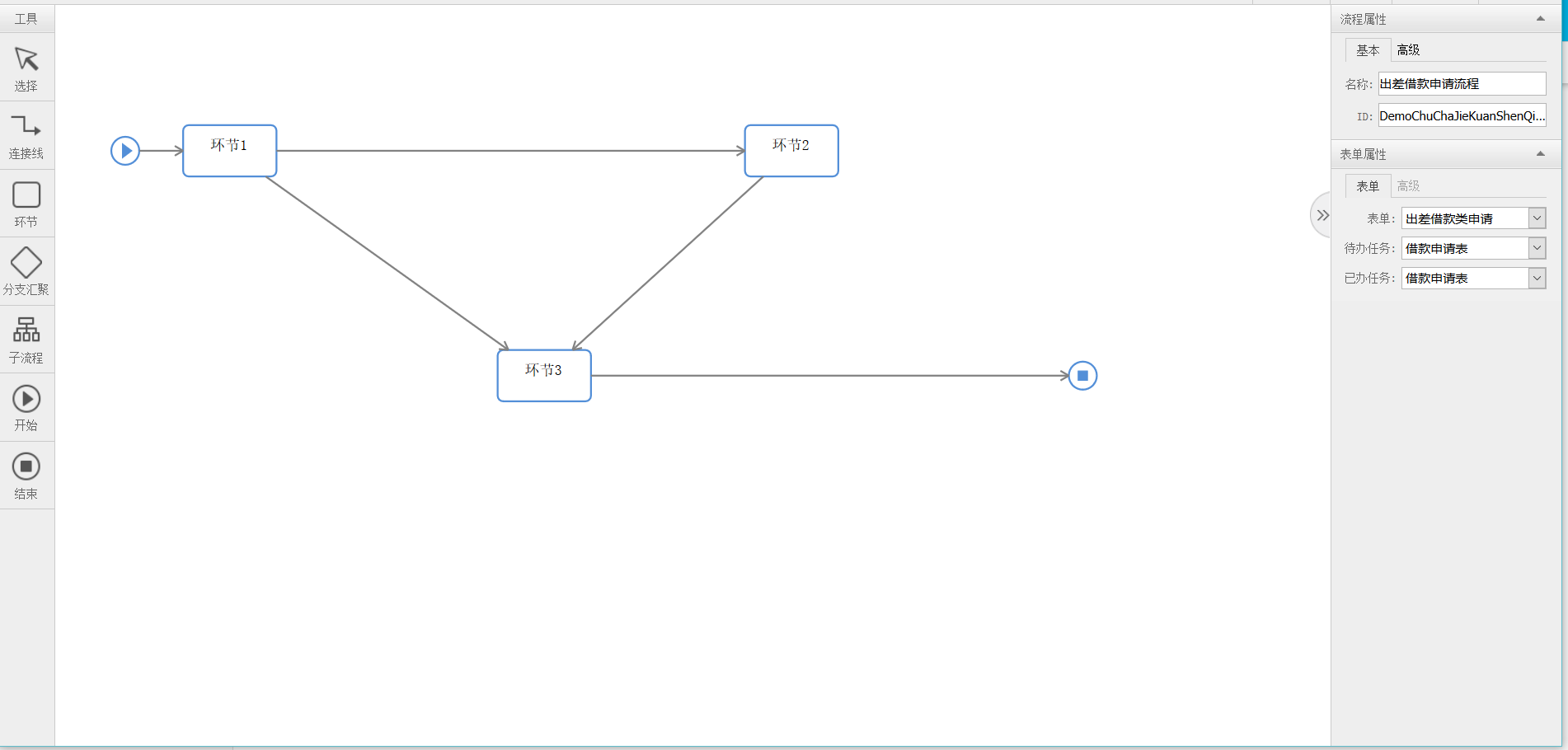
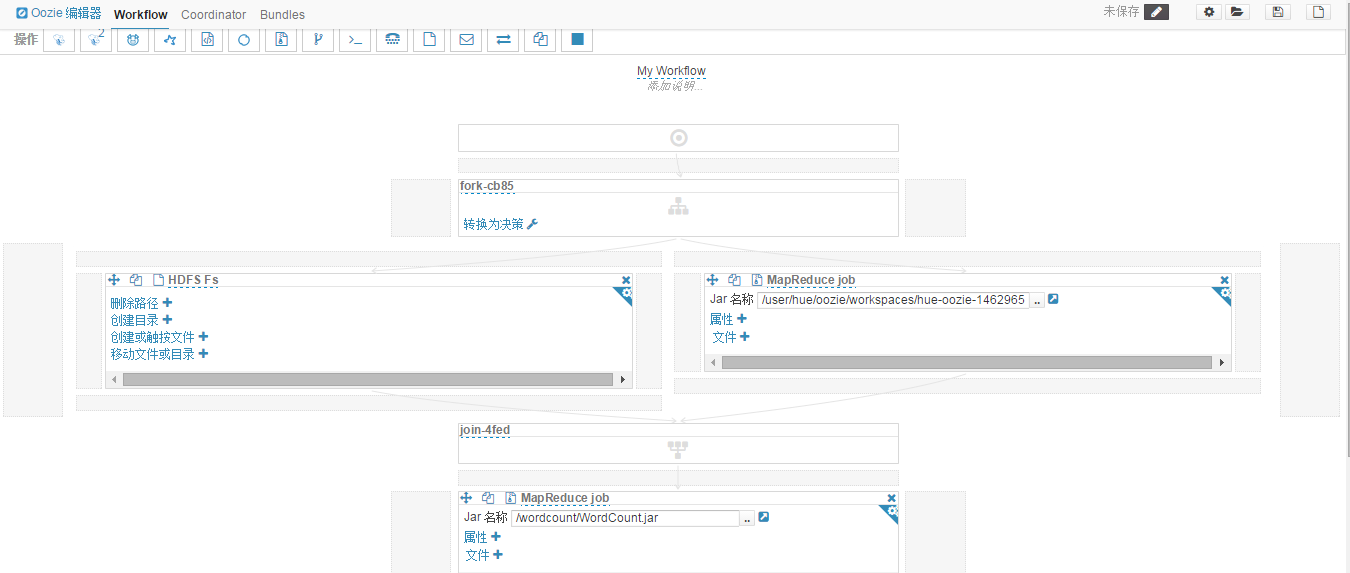
* 任务管理：主要实现对定义任务的管理。包括：添加，删除，复制，执行，流程展现，任务编辑以及流程文件的导入导出，参数配置展现与修改等操作，同时在任务管理界面也可对各组件进行操作。



* 实例管理：主要对运行，挂起，停止等状态下的实例进行管理。包括：停止，运行，状态展现，数据展现，日志展现等操作。



* 拖拽开发：主要是在可视化界面下以拖拽组件的方式，对任务进行编排开发，设置相关参数。同时对编排的任务进行流程展现，为对IDAP进行完全支持，拖拽开发需要支持Oozie全部动作节点。



* 在线可编辑：以拖拽方式编排的任务，最终生成运行相关文件，开发者可以在线编辑相关文件进行参数修改优化。
* 自定义组件：根据任务不同，可以添加，删除，修改各组件。
* 数据展示：用于展示任务执行各阶段生成的数据。
* 日志展示：用于展示任务执行过程中产生的日志信息，方便开发者对任务进行调试管理。

# 工作安排

主要分为三个阶段：分别为：技术验证阶段，控制管理，可视化。

技术验证阶段：数据流开发第一阶段，主要实现：通过手动编写相关xml文件与配置文件的方式，可以对IDAP相关组件进行任务编排，保证任务运行。主要完成工作包括：验证Oozie可用性，熟悉任务编排流程，可通过命令形式对任务进行启动，执行，停止等操作。

控制管理：数据流开发第二阶段，完成对任务管理，实例管理的开发，同时支持文件的导入、导出操作。主要实现：可以对任务进行添加，删除，编辑等操作，可以在任务管理页面以导入、导出文件的方式生成相关任务，所有任务以列表形式展现。可以对实例进行管理，所有运行实例以列表形式进行展示，可以查看实例参数配置。

可视化：实现所有功能，主要实现:任务管理，实例管理，同时支持任务拖拽，页面可编辑，参数配置自动生成等操作以及数据展现，日志展现等，对IDAP有高度支持，同时具有高稳定性与高可操作性。

# 功能列表

|  |  |  |
| --- | --- | --- |
| 功能 | 说明 | 优先级 |
| 任务管理 | 对编排任务进行管理与操作。 | 高 |
| 实例管理 | 对运行实例进行管理，查看运行状态等。 | 高 |
| 拖拽开发 | 可视化的组件拖拽式开发。 | 高 |
| 在线可编辑 | 可以对部分流程进行手动编辑优化。 | 高 |
| 自定义组件操作 | 可以根据需要手动添加删除部分组件。 | 高 |
| 运行流程展示 | 拖拽过程中展示整个设计执行流程。 | 高 |
| 日志展现 | 展示任务执行各阶段产生的日志信息。 | 低 |
| 数据展现 | 展示任务执行各阶段产生的数。 | 低 |

# 非功能需求分析

略。