Oozie应用开发

www.huawei.com

HUAWEI

目标

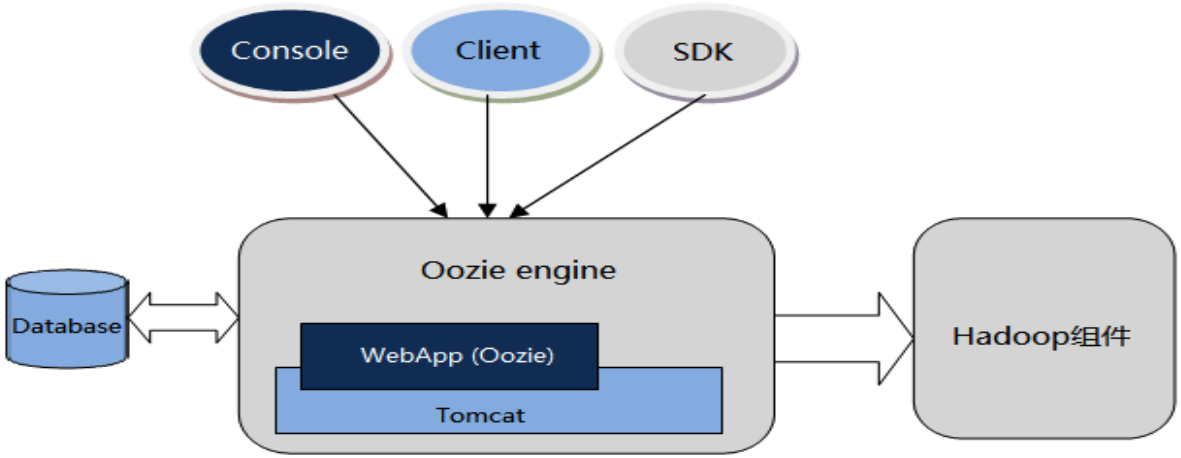
- 学完本课程后，您将能够:
  - 了解**Oozie**应用开发适用场景
  - 掌握**Oozie**应用开发
  - 熟悉并使用**Oozie**常用**API**

目录

Oozie简介

- **Oozie**是一个**Hadoop**作业的工作流调度管理系统。
- **Oozie**工作流（**Workflow**）是放置在控制依赖**DAG**（有向无环图）中的一组动作（**Action**）集合，控制依赖可确保后续操作在前面的操作已成功完成后才会启动。
- **Oozie**的协调作业（**Coordinator**）是通过时间（频率）和有效数据来触发当前的**Oozie**工作流。
- **Oozie**支持多种**Hadoop**作业（包括：**HDFS**，**MapReduce**，**Hive**，**Streaming MR**，**Loader**，**Spark**，**Distcp**）以及系统类作业（例如**Java**与**Shell**）。

Oozie架构回顾



Console	提供对**Oozie**流程的查看和监控功能。
Client	通过接口控制**workflow**流程：可以执行提交流程，启动流程，运行流程，终止流程，恢复流程等操作。**Hue**界面上的**Workflow**与**JobDesign**就属于**client**范畴
SDK	软件开发工具包**SDK**（**SoftwareDevelopmentKit**）是被软件工程师用于为特定的软件包、软件框架、硬件平台、操作系统等建立应用软件的开发工具的集合。
Database	**PG**数据库，用于存储作业信息。
WebApp（Oozie）	**webApp**（**Oozie**）即**Oozie server**，可以用内置的**Tomcat**容器，也可以用外部的，记录的信息比如日志等放在**PG**数据库中。
Tomcat	**Tomcat**服务器是免费的开放源代码的**Web**应用服务器。
HaDoop组件	底层执行**Oozie**编排流程的各个组件，包括**MapReduce**、**Hive**、**Spark**等。

Oozie应用场景

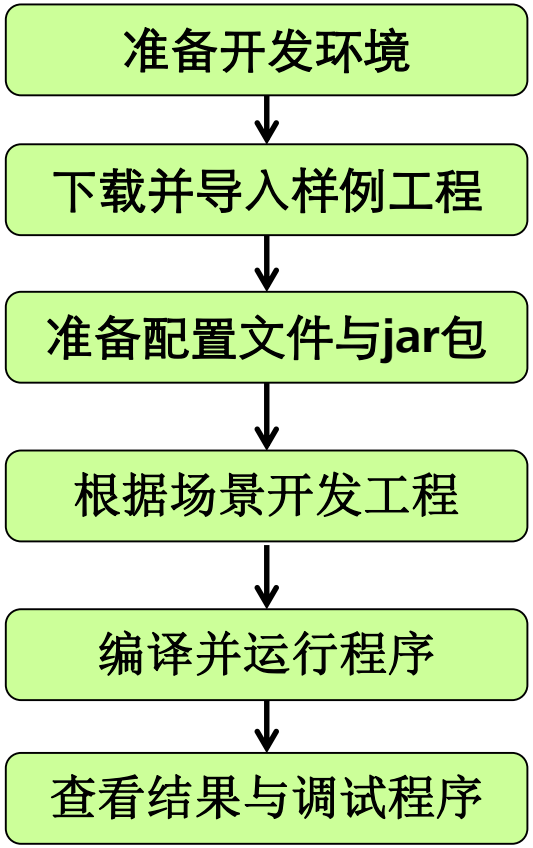
Oozie是一个工作流调度引擎，对各种类型的**Hadoop**作业进行编排与调度。

主要应用于以下几种场景：

- 编排与管理逻辑复杂的多种类型的**Hadoop**作业，按照指定的顺序协同运行，例如**ETL**任务。
- 基于时间(频率)触发工作流，例如每天/小时执行一次的重复任务或者指定时间执行的任务。
- 基于数据有效性触发工作流，当依赖的**HDFS**数据产生之后才触发下一步动作，可用于数据管道处理。
- 实时监控与管理集群的工作流，快速定位问题；掌握集群的资源使用情况，并根据需要调整工作流的调度，最大化利用集群资源。

目录

Oozie应用开发流程



准备开发环境

- 准备开发环境

准备项	说明
操作系统	**Windows**系统，推荐**Windows 7**以上版本。
安装**JDK**	开发环境的基本配置。版本要求：**1.7**或者**1.8**，推荐**1.8**。
安装和配置**Eclipse**	用于开发**Oozie**应用程序的工具。
网络	确保客户端与**Oozie**服务主机在网络上互通。

下载并导入样例工程

- 下载并导入**Oozie**样例工程
  - 安装**Oozie**客户端，拷贝样例工程到**windows**开发机器。
  - 在**FusionInsight Manager**页面新建用户，用于登录与操作。

    用户需要**Oozie**的普通用户权限，**HDFS**访问权限以及**Yarn**的队列提交权限。
  - 下载用户的认证凭据文件。
  - 配置认证凭据文件到**Oozie**客户端样例工程。
  - 导入样例工程到**Eclipse**开发环境。

准备工作流需要的配置文件与Jar包

- 开发工作流配置文件**workflow.xml(coordinator.xml**是对工作流进行调度，**bundle.xml**是对一组**coordinator**进行管理)与**job.properties**，配置示例参考后续章节或产品文档的**Oozie**应用开发指南章节。
- 如果有实现代码需要开发对应的**jar**包，例如**Java Action**；如果是**hive**则需要开发**sql**文件。
- 上传工作流依赖的配置文件、**jar**包（包括依赖的**jar**包）、**sql**文件等到**HDFS**，上传的路径取决于**workflow.xml**中的**Oozie.wf.application.path**配置的路径。

Job.properties配置

- 流程的属性定义文件，定义了流程运行期间使用的外部参数键值对，例如在**workflow.xml**中配置为**${nameNode}**，**job.properties**中设置**nameNode**真正的值，运行时替换到**workflow.xml**。

参数	含义
nameNode	**HDFS NameNode**集群地址
jobTracker	**MapReduce ResourceManager** 地址
queueName	提交到**Yarn**的队列名
dataLoadRoot	流程任务所在目录名
Oozie.coord.application.path	**Coordinator**流程任务在**HDFS**上的存放路径
start	定时流程任务启动时间
end	定时流程任务终止时间
workflowAppUri	**Workflow**流程任务在**HDFS**上的存放路径

workflow.xml配置

- 描述了一个完整业务的工作流程定义文件，一般由一个**start**节点、一个**end**节点和多个实现具体业务的**action**节点组成。

参数	含义
name	流程文件名
start	流程开始节点
end	流程结束节点
action	实现具体业务动作的节点（可以是多个）

coordinator.xml配置

- 周期性执行**workflow**类型任务的流程定义文件

参数	含义
frequency	流程定时执行的时间间隔
start	定时流程任务启动时间
end	定时流程任务终止时间
workflowAppUri	**Workflow**流程任务在**HDFS**上的存放路径
jobTracker	**MapReduce ResourceManager**地址
queueName	任务处理时使用的**Mapreduce**队列名
nameNode	**HDFS NameNode**地址

MapReduce Aciton配置示例

- **MapReduce Action**说明

参数	含义
name	**map-reduce action**的名称
job-tracker	**MapReduce ResourceManager**地址
name-node	**HDFS NameNode**地址
queueName	任务处理时使用的**MapReduce**队列名
mapred.mapper.class	**Mapper**类名
mapred.reducer.class	**Reducer**类名
mapred.input.dir	**MapReduce**处理数据的输入目录
mapred.output.dir	**MapReduce**处理后结果数据输出目录
mapred.map.tasks	**MapReduce map**任务个数

HDFS Aciton配置示例

- **HDFS Action**说明

参数	含义
name	**FS**活动的名称
delete	删除指定的文件和目录的标签
move	将文件从源目录移动到目标目录的标签
chmod	修改文件或目录权限的标签
path	当前文件路径
source	源文件路径
target	目标文件路径
permissions	权限字符串

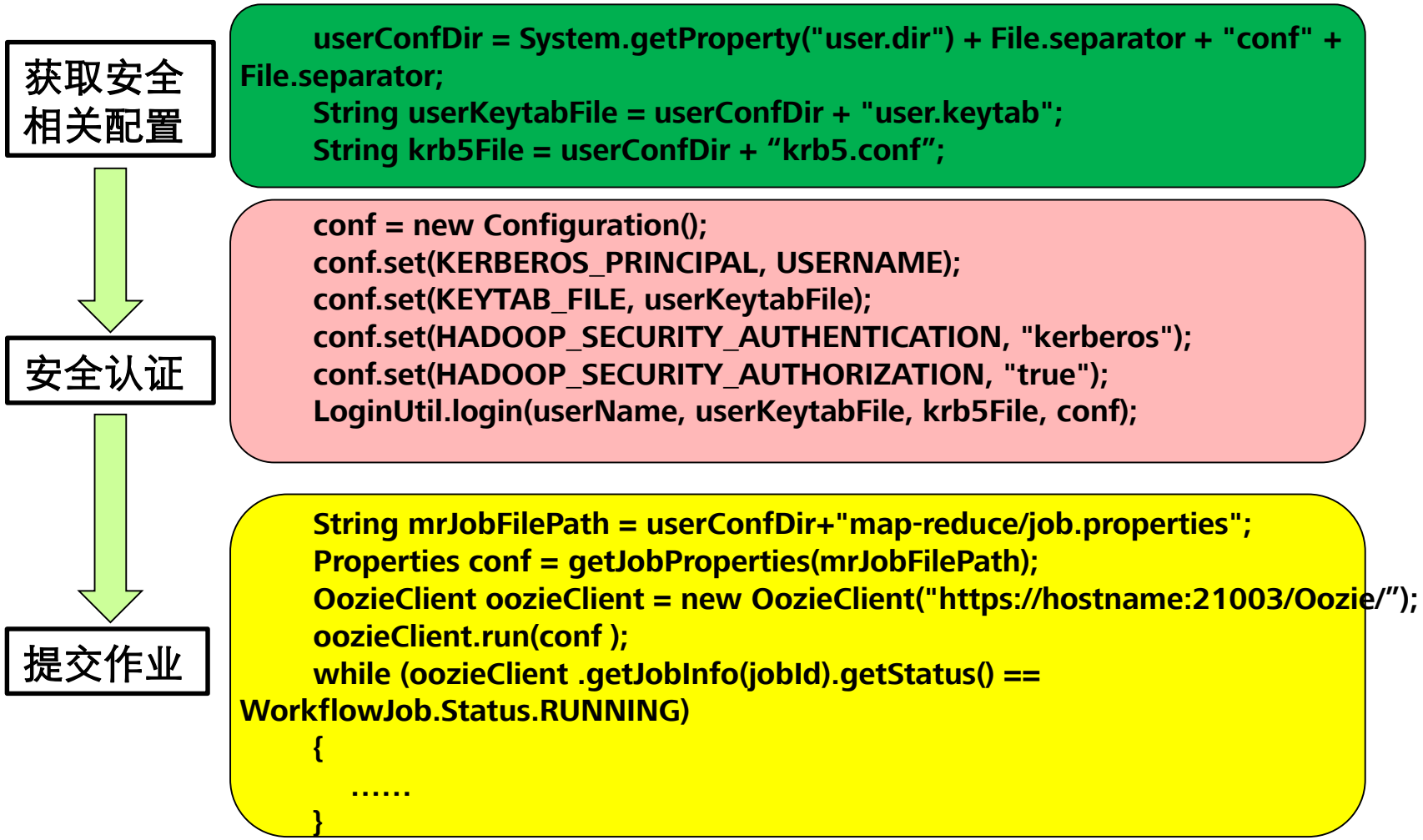
提交工作流

- **1. Oozie**提供如下几种方式对工作流进行操作，本文主要介绍的是**Java API**的方式提交作业，详情请参见产品文档应用开发指南章节。

  Shell命令；**Java API**；**Hue(UI)**； **Rest API**

- **2. Oozie**客户端提供了比较完整的**example**示例供用户参考，包括各种类型的**Action**，以及**Coordinator**以及**Bundle**的使用，以客户端安装目录为**/opt/client**为例，**example**具体目录为**/opt/client/Oozie/Oozie-client-4.2.0/examples/apps**，详情请参见产品文档的《应用开发指南》中**Oozie**相关章节。

JAVA API提交作业代码示例



编译并运行程序

1、Windows编译并运行样例工程

在开发环境**Eclipse**中，右击**OozieMain.java**，单击“**Run as > Java Application**”运行对应的应用程序工程。

2、Linux中运行样例工程

- 在**eclipse**中导出样例工程的**jar**包。
- 在**Linux**环境新建目录，例如“**/opt/test**”，并创建子目录“**lib**”和“**conf**”。将样例工程中“**lib**”的**Jar**包，以及导出的**Jar**包，上传到**Linux**的“**lib**”目录。将样例工程中“**conf**”的配置文件上传到**Linux**中“**conf**”目录。
- 在“**/opt/test**”根目录新建脚本“**run.sh**”(内容请参考产品文档**Oozie**应用开发指南)，切换到“**/opt/test**”，执行**sh run.sh** 命令运行**Jar**包。

查看结果与调试程序

- 查看结果与调试程序
  - 查看**Oozie API**返回结果是否符合预期。
  - 可以通过**Hue**的**workflow**仪表板查看程序运行结果。
  - 可以通过**Hue**的文件浏览器查看是否在**HDFS**指定目录生成了期望的文件。
  - 通过**Oozie**自带的**Web UI**也可以查看作业执行结果。

目录

Oozie常用接口示例

- **Shell**常用接口清单

命令	参数	说明
Oozie job	-config <arg>	指定job配置文件（job.properties）路径
	-Oozie <arg>	指定Oozie server地址
	-run	运行job
	-start <arg>	启动指定的job
	-submit	提交job
	-kill <arg>	删除指定的job
	-suspend <arg>	暂停指定的job
	-resume <arg>	恢复指定的job
	-D <property=value>	给指定的属性赋值
Oozie admin	-Oozie <arg>	指定Oozie server地址
	-status	显示Oozie服务状态

Oozie常用接口示例

- **Java API**常用接口清单

方法	说明
public String run(Properties conf)	运行**job**
public void start(String jobId)	启动指定的**job**
public String submit(Properties conf)	提交**job**
public void kill(String jobId)	删除指定的**job**
public void suspend(String jobId)	暂停指定的**job**
public void resume(String jobId)	恢复指定的**job**
public WorkflowJob getJobInfo(String jobId)	获取**job**信息

总结

- **Oozie**简介以及应用场景介绍
- 应用开发流程介绍，包括准备环境，安全认证，开发，调试
- 配置文件与安全认证及提交任务代码示例
- 常用**API**介绍

思考题

1. 新建的**Oozie**开发用户提交**MR**工作流需要哪些权限?
2. **Java API**登录认证需要那几项必要信息?

习题

- 判断题

  1. 提交**Oozie**作业的新建用户不需要**HDFS**权限。 （**T or F)**
  2. **coordinator.xml**是工作流配置文件。 （**T or F)**
  3. 提交作业前需要先上传作业依赖的配置文件或**jar**包到**HDFS**（**T or F)**

- 多选题

  1. 有几种方式提交**Oozie**作业（ ）？

     A.Shell命令

     B.Java API

     C.Hue(UI)

     D.Rest API

Thank you

www.huawei.com