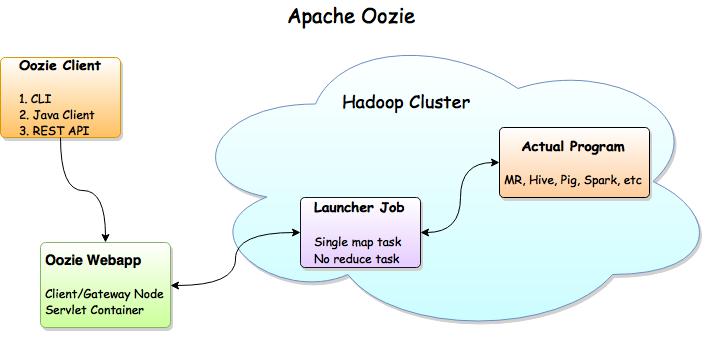
Apache Oozie

Apache Oozie是用于Hadoop平台的工作流调度引擎，部署到JServlet容器中的Java应用，提供对Hadoop MR和Pig Jobs的任务调度与协调。在Oozie中，工作流是一系列操作的有向无环图，任务之间的操作的输入依赖于前一个任务的输出，只有前一个操作完成后才开始第二个。Oozie工作流通过hPDL（XML的流程定义语言）定义，通过远程系统启动任务，当任务完成后，远程系统会进行回调来通知任务已经结束，然后再开始进行下一各操作。

Oozie的系统架构图如下所示:



在Oozie中有几个重要的概念：

1. WorkFlow

工作流（以DAG的方式串接起来），任务开始和结束的过程以及工作流Job的执行路径，并提供一种机制来控制一种机制来控制工作流执行路径（比如Decision/Fork以及Join节点等）。 Oozie定义了一种基于XML的hPDL(Hadoop Process Definition Language)来描述workflow的DAG，在Workflow中定义了：

* 控制流节点（Control Flow Nodes）
* 动作节点（Active Nodes）

其中，控制流节点定义了流程的开始和结束（start/end），以及控制流程的执行路径（Execution Path），如decision、fork、join等。而动作节点包括Hadoop的任务、SSH、HTTP等。

示例如下：



其中fork/join是成对出现，表示工作流的并发执行，最后汇聚到一个Node。从Oozie的工作流调度机制可以看出，Oozie是没有能力表达复杂的DAG，比如嵌套的依赖关系。此外Oozie工作流可以参数化，比如在工作流中定义使用像${inputDir}之类的变量，然后通过job.properties配置对应参数，在启动的时候，将参数传入工作流。

1. Coordinator

多个WorkFlow可以组成一个Coordinator，可以把前几个WorkFlow的输出作为后一个WorkFlow的输入，也可以额定义WorkFlow的触发条件来定时触发，其定义格式如下：

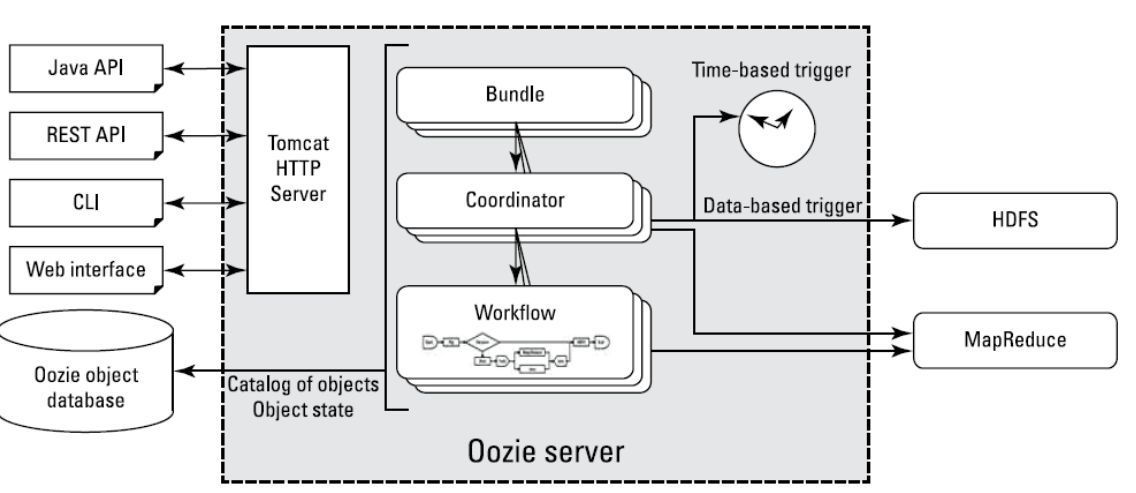


1. Bundle

控制一个或者多个Coordinator应用，其写法如下所示：



Oozie的细节架构如下所示：



Oozie简介：

http://www.infoq.com/cn/articles/introductionOozie

http://crazyadmins.com/tag/oozie-tutorials-hortonworks/

Oozie概念：

https://www.cnblogs.com/smartloli/p/4758181.html

https://www.cnblogs.com/ilinuxer/p/6804339.html

https://blog.csdn.net/oracle8090/article/details/54599068

<http://oozie.apache.org/docs/4.0.0/DG_Examples.html>

https://www.ibm.com/developerworks/cn/data/library/bd-hadoopoozie/

https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems

http://www.cnblogs.com/juncaoit/p/6122187.html