# Apache Falcon概述

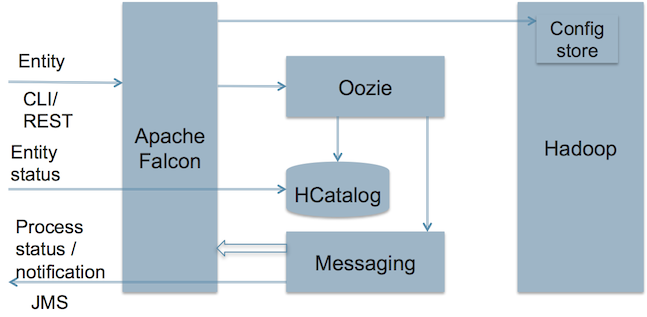
在数据处理过程中，由于种种原因，如数据类型不一致，各个部分数据设计不统一，或者不同数据源存储地址不同，造成各种问题，例如**数据孤岛、数据不一致**等，为了解决这些问题，需要采用大数据治理技术。

大数据治理包括数据生命周期管理、数据处理以及数据审计与追踪。其中**数据生命周期管理贯穿了对数据的采集、处理、备份保留以及清洗删除等过程**。Hadoop生态环境也有相关组件来处理与之对应的问题，像sqoop实现对数据采集抽取，MapReduce负责数据处理备份等。但是他们就像机器的零件，虽然能处理数据治理的问题，然而却**缺少有效的手段去整合调度这些零件使其成为一台秩序运作的机器**。为了解决这一问题，开发了Oozie这一Hadoop调度工具。但是Oozie在设置工作流调度时，更多的是面向Hadoop底层的设置，需要用户了解Hadoop的基础框架，同时生成的工作流十分繁琐。

## Falcon技术概况

Apache Falcon作为Hadoop集群数据处理和数据生命周期管理系统框架，通过数据生命周期管理及处理方案，解决**Hadoop的数据复制、业务连续以及血统追踪**等难题。**Falcon主要是对数据生命周期集中管理，促进了数据快速复制来实现业务连续性和灾难恢复，并通过实体沿袭追踪和审计日志收集为审计和合规性提供基础**，方便用户设定数据管理以及处理方案，并将其提交到Hadoop集群调度执行。

Falcon是一个建立在Hadoop上的数据集及其处理流程的管理平台，Falcon本质上通过标准工作流引擎将用户的数据集及其流程配置转换成一系列重复的活动，而本身不做任何繁琐的工作。所有功能以及工作流状态管理需求都委托给工作流调度器进行调度。



## Falcon发展概况

社区发展不活跃，hadoop版本支持到hadoop2.7，现阶段已退役，迁移至Apache Attic。

## Falcon技术优势

Falcon 通过更高层次的抽象，简化了数据处理管道的开发和管理，通过开箱即用的数据管理服务，在数据处理应用程序的开发过程中省略了复杂的编码，同时也简化了数据移动，灾难恢复和数据复制等工作流的配置和编排。

Falcon 架构也可以作用于其它Hadoop组件，例如Pig/HDFS/Oozie等。Falcon通过提供一个定义-部署-管理数据管道的框架来实现这种简化的管理，作为开源的数据生命周期管理项目，Apache Falcon能够提供以下服务：

(1) 建立各种数据之间的关系和处理Hadoop环境下的元素

(2) 数据集管理服务，例如数据保留/跨集群复制以及数据归档等

(3) 方便进行新工作流/管道上传，支持后期数据处理和Retry政策

(4) 集成了元数据库以及数据仓库，例如Hive/HCatalog集成

(5) 为终端用户提供基于可用性数据集组

(6) 支持在局部或全局聚合的本地处理来进行案例使用

(7) 获取数据集和处理程序的血统

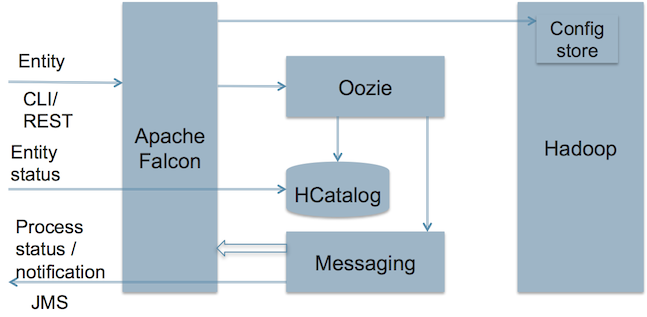
总的来说，Apache Falcon实现的企业级数据治理需求主要体现在3个方面，

|  |  |
| --- | --- |
| 需求 | 特征 |
| 数据生命周期的统一管理 | 统一定义和管理数据采集/处理以及输出管道  确保灾难准备和业务连续性  用于数据复制和保留的开箱即用策略  端到端监控数据管道 |
| 合规性和审计 | 可视化数据管道血统  跟踪数据管道审计日志  使用事物元数据标记数据 |
| 数据库复制和归档 | 跨本地和基于云存储目标的数据复制  通过文档和示例的数据血统支持  HDFS中的异构分层存储  集群中的数据热/冷存储层的定义 |

## Falcon架构

### 整体架构

Apache Falcon 提供了一个用于治理和编排[Hadoop](https://www.oschina.net/p/hadoop" \t "https://www.oschina.net/p/_blank)内部和周边数据流的数据处理框架。该框架为获取和处理数据集、复制与保留数据集、重新定向位于非Hadoop扩展中的数据集、维护审核跟踪与沿袭提供了关键性的管控框架。



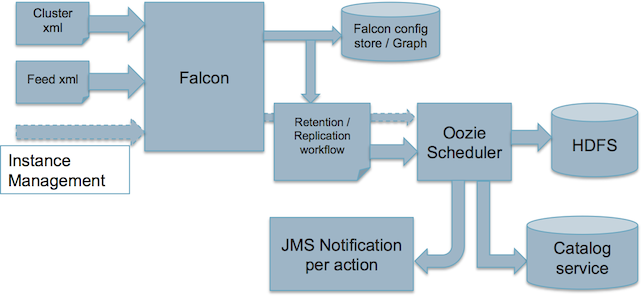
调度器

Falcon目前默认使用oozie作为调度器。

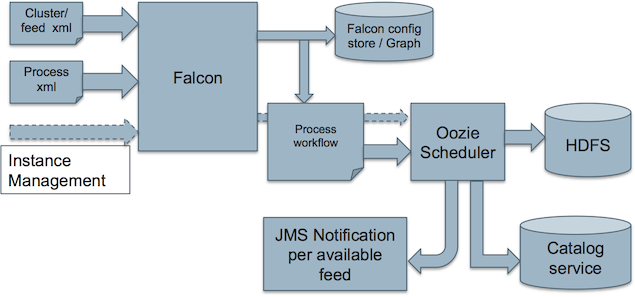
与oozie集成

**Falcon通过JMS (Java消息服务) 与oozie进行通信,控制oozie上对应的作业流,oozie也通过JMS给falcon报告作业运行情况**.

### Feed Schedule flow



### Process Schedule flow



## 总结

总结来说, Falcon是在调度器Oozie上封装了一层,**用户可以用Web UI上的wizard来配置数据流水线**, 数据生命周期管理非常方便. **它最大的优点就是增强了Oozie的易用性**, 对于业务逻辑比较复杂的系统, 用Falcon管理比起直接用Oozie会大大简化.

但是调研中发现, **Falcon Web UI上呈现的血缘关系只是以实体为中心, 并没有全局层面上整个data pipeline的血缘关系. 如果能够以pipeline为中心, 画出血缘关系图,提供zoom in功能,以及在图中把实体运行状态可视化, 将会是一个很有用的特性.**  
 **虽然被称为数据治理工具,但是它的功能只是集中在数据生命周期管理和流水线管理,要与元数据管理(Atlas),数据安全管理(Ranger)等组合,才能形成一个完整的数据治理解决方案**.

# Apache Falcon场景设计与实现

* 数据管道

**Oozie workflow**

* 结构化数据导入分布式文件系统

HDFS🡨🡪Mysql

* 结构化数据与数据仓库的交互

Hive🡨🡪HDFS🡨🡪Mysql

* 跨集群数据传输

**Hadoop的DistCp服务**

* 数据镜像
* 数据仓库中的数据操作

Hive