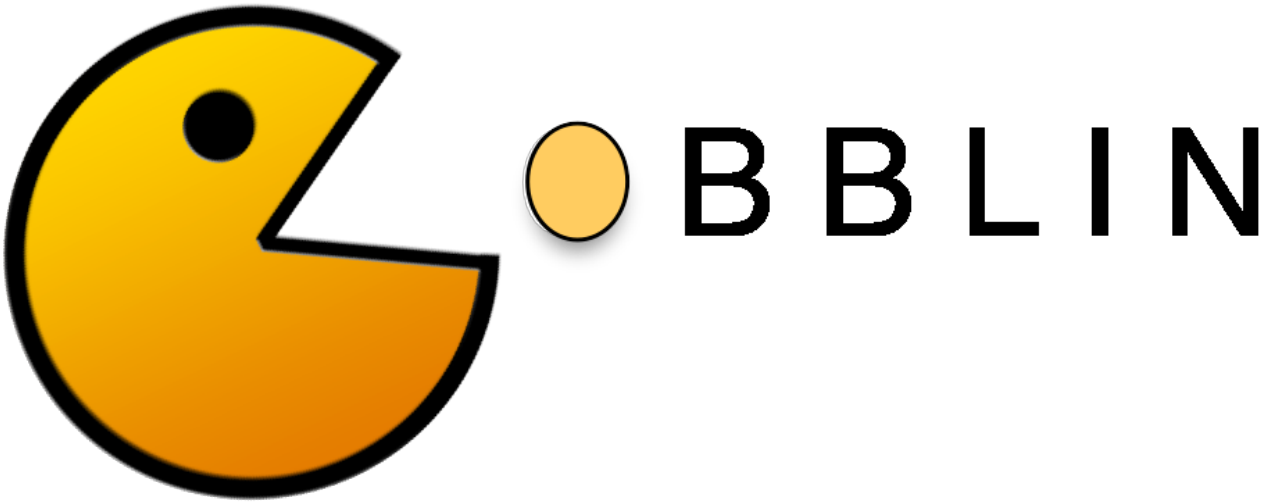
Apache® Gobblin™：开源分布式大数据集成框架

Apache Gobblin 是一个用于流数据和批处理数据生态系统的分布式大数据集成框架。可以简化大数据集成里面的常见问题，比如数据摄取、复制、组织以及生命周期管理等。该项目2014年起源于 LinkedIn，2015年开源，2017年2月进入 Apache 孵化器，2021年02月16日正式毕业成为 Apache 顶级项目。



如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：过往记忆大数据

多年来，LinkedIn
的数据基础架构团队构建了自定义的数据摄取解决方案，用于将不同的数据引入到 Hadoop 生态系统。最终，LinkedIn 运行了 15
种类型的摄取管道，这给数据质量、元数据管理、开发和操作带来了重大挑战。

上面这个问题促使 LinkedIn 构建了 Gobblin。 Gobblin
是一种通用数据摄取框架，用于从各种数据源（例如数据库、REST API、FTP/SFTP 服务器、文件管理器等）中提取、转换和加载大量数据到 Hadoop。Gobblin 处理所有数据摄取 ETL 所需的常见例行任务，包括作业/任务调度、任务分区、错误处理、状态管理、数据质量检查、数据发布等。Gobblin
在同一执行框架中摄取来自不同数据源的数据，并在同一个地方管理不同来源的元数据。 结合其他特性，例如自动扩展、容错性、数据质量保证、可扩展性和处理数据模型演化的能力，使 Gobblin 成为一个易于使用、自我服务且高效的数据摄取框架。

Gobblin
是围绕可扩展性的思想构建的，即用户可以轻松添加新适配器或扩展现有适配器以使用新数据。

Gobblin 的架构体现了这一思想：

Gobblin 作业建立在一组 constructs
上（由上图中的浅绿色框表示），它们以某种方式协同工作并完成数据提取工作。所有的 constructs 都可以通过作业配置插入，并且可以通过添加新的或扩展现有的实现来扩展。

一个 Gobblin
作业由一组任务组成，每个任务对应一个要完成的工作单元，负责提取一部分数据。Gobblin 作业的任务由 Gobblin 运行时（Gobblin
runtime）（由上图中的橙色框表示）根据选择的部署设置（由上图中的红色框表示）执行。

Gobblin 运行时（Gobblin runtime）负责在选择的部署设置上运行用户定义的 Gobblin 作业。它处理常见的任务，包括作业和任务调度、错误处理和任务重试、资源协商和管理、状态管理、数据质量检查、数据发布等。

Gobblin 目前支持两种部署模式：单节点的 Standalone 模式和 Hadoop 集群的 Hadoop MapReduce 模式。 当然，这部分还在扩展。

Gobblin 的运行和操作由一些组件和实用程序（由上图中的蓝色框表示）支持，它们处理重要的事情，例如元数据管理、状态管理、指标收集和报告以及监控。

关于 Apache® Gobblin™ 的更多介绍可以到 https://gobblin.apache.org/ 查看。