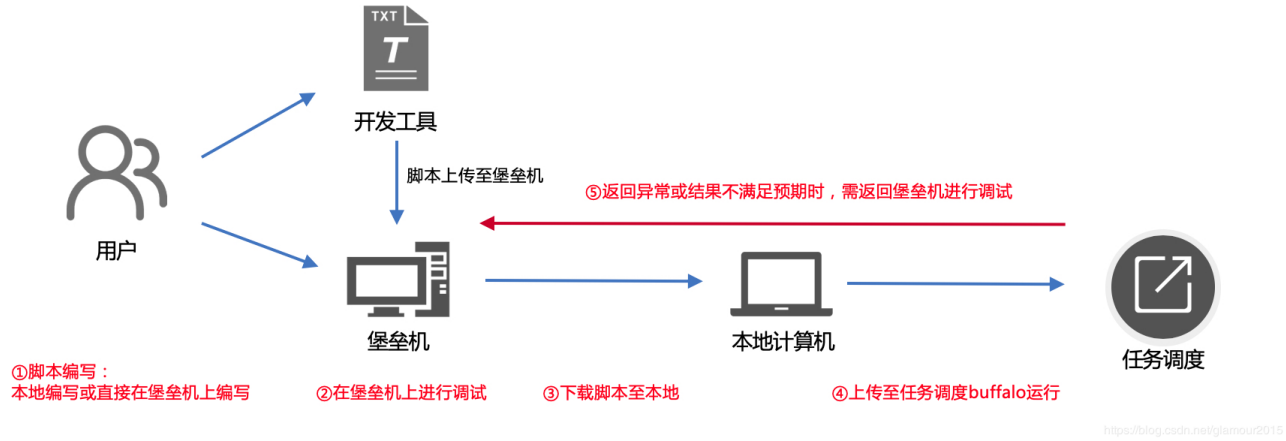
Zeppelin调研与数据开发平台

现状与用户痛点



1.开发、发布路径比较长
2.堡垒机开发、团队协调不够方便
3.堡垒机环境被调整后会造成脚本异常
4.表的管理操作缺少统一入口，表命名不规范，结构信息与负责人、所属项目、描述等信息缺乏。

Zeppline整体介绍

Apache Zeppelin 是一个兼具了 **Hadoop 大数据处理和 机器学习 / 深度学习算法交互式开发**的开源系统。

在大数据方面，Apache Zeppelin 是一个可以进行大数据可视化分析的交互式开发系统，可以承担数据接入、数据发现、数据分析、数据可视化、数据协作等任务，其前端提供丰富的可视化图形库，不限于SparkSQL，后端支持HBase、Flink 等大数据系统以插件扩展的方式，并支持Spark、Python、JDBC、Markdown、Shell 等各种常用Interpreter，这使得开发者可以方便地使用SQL 在 Zeppelin 中做数据开发。

在机器学习平台方面，Apache Zeppelin还可以完成机器学习的数据预处理、算法开发和调试、算法作业调度的工作，包括当前在各类任务中表现突出的深度学习算法，因为 Zeppelin 的最新的版本中增加了对TensorFlow、PyTorch等主流深度学习框架的支持，此外，Zeppelin将来还会提供算法的模型 Serving 服务、Workflow 工作流编排等新特性，使得 Zeppelin可以完全覆盖机器学习的全流程工作。

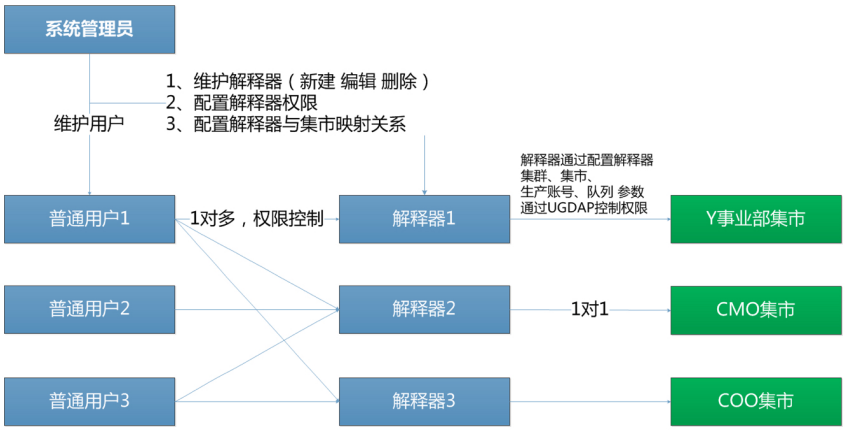
在平台部署和运维方面，Zeppelin还提供了单机 Docker、分布式、K8s、Yarn 四种系统运行模式，无论你是小规模的开发团队，还是 Hadoop 技术栈的大数据团队、K8s 技术栈的云计算团队，Zeppelin 都可以让数据科学团队

轻松的进行部署和使用 Zeppelin丰富的数据和算法的开发能力。

Zeppline目标用户

分布式计算、数据分析从业者、机器学习算法工程师

Zeppline应用方案



系统管理员

1. 配置系统；
2. 包括用户、解释器、运行环境等；

普通用户

1. 使用系统；
2. 创建、编辑、删除、运行笔记等；

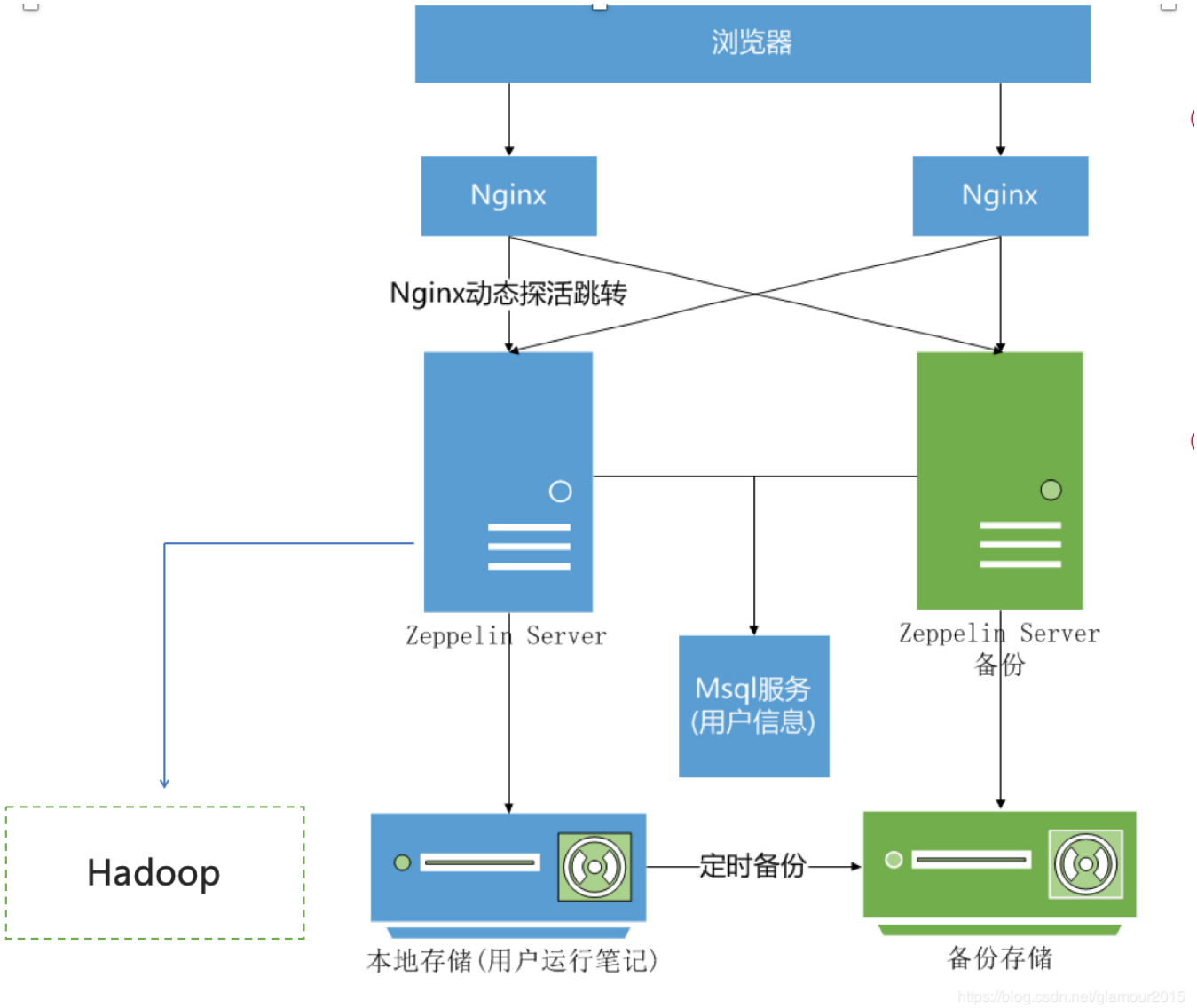
https://blog.csdn.net/glamour2015

运行资源
除主服务外，每个解释器启动一个JVM进程，硬件资源需求大，建议8C64G;
主服务停止后，解释器进程不会停止，需手动停止方可释放资源；

Zeppline特征

Zeppelin 集群模式（Cluster Mode）

1.由单节点提供服务，存在一定时间的单点故障,对服务可用性的要求可能无法满足；
2.定时备份数据，保证数据安全，备份时间间隔内的数据会丢失；
3.主服务失败时，启动备份节点提供服务。



集群模式下，我们可以同时启动多个Zeppelin Server(解释器)，基于Raft 算法选主（Master）、同步，共同对外提供服务。用户通过 Nginx 反向代理域名访问这些 Zeppelin 服务。同时，集群模式还提供了 Cluster 元数据管理的能力，集群中所有的 Zeppelin Server 的运行状况，以及所有的解释器进程，都会记录在元数据中，用户可以通过Nginx 配置访问不同的 Server，创建不同的解释器。解释器进程可以在集群中自动寻找资源最为富余的 Server 来运行，而当某个 Server 挂了且难以恢复，用户仍然可以通过元数据启动另外一个 Server，继续未完成的工作。Zeppelin 集群模式只需在参数中配置3个服务器的列表，并将其启动，即可自动组建 Zeppelin 集群，不需要借助 ZooKeeper。通过专门的集群管理页面，用户可以清晰看到集群中的服务器、解释器的数量和运行状态。

本机 Docker

无论是单机模式还是集群模式，用户都可以在本机 Docker 上创建解释器进程。通过集群模式+ Docker，用户不需要 Yarn 或者 Kubernetes，即可创建 Zeppelin 集群，提供高可用服务，核心功能和Zeppelin On Yarn/ Kubernetes 并无二致，而且部署和维护也很简单，无需复杂的网络配置。

Zeppelin On Yarn

Zeppelin 的解释器可以创建在 Yarn 的运行环境中，支持Yarn 2.7及以上的版本。Zeppelin 容器的维护需要模拟终端，Zeppelin 支持通过shell 命令进入 Docker 进行维护，如安装所需的 Python 库、修改环境变量等。

多 Hadoop 集群

Zeppelin 支持通过配置，即指定不同的 Hadoop / Spark Conf 文件，即可用一个 Zeppelin 集群，去连接所有的 Hadoop 集群，而无需为所有 Hadoop 集群分别创建多个 Zeppelin 服务，从而简化管理和维护的复杂度，同时保证服务的可靠性。

可视化调参

不同的机器学习框架有不同的参数配置，甚至不同的算法参数都不同，传统命令行的方式容易配置出错，Zeppelin 基于其前端可视化展示能力，将支持针对每个算法自行设置一个参数调整界面，和模型一起发布，模型使用者可以使用该可视化界面，根据需要动态地调整参数。

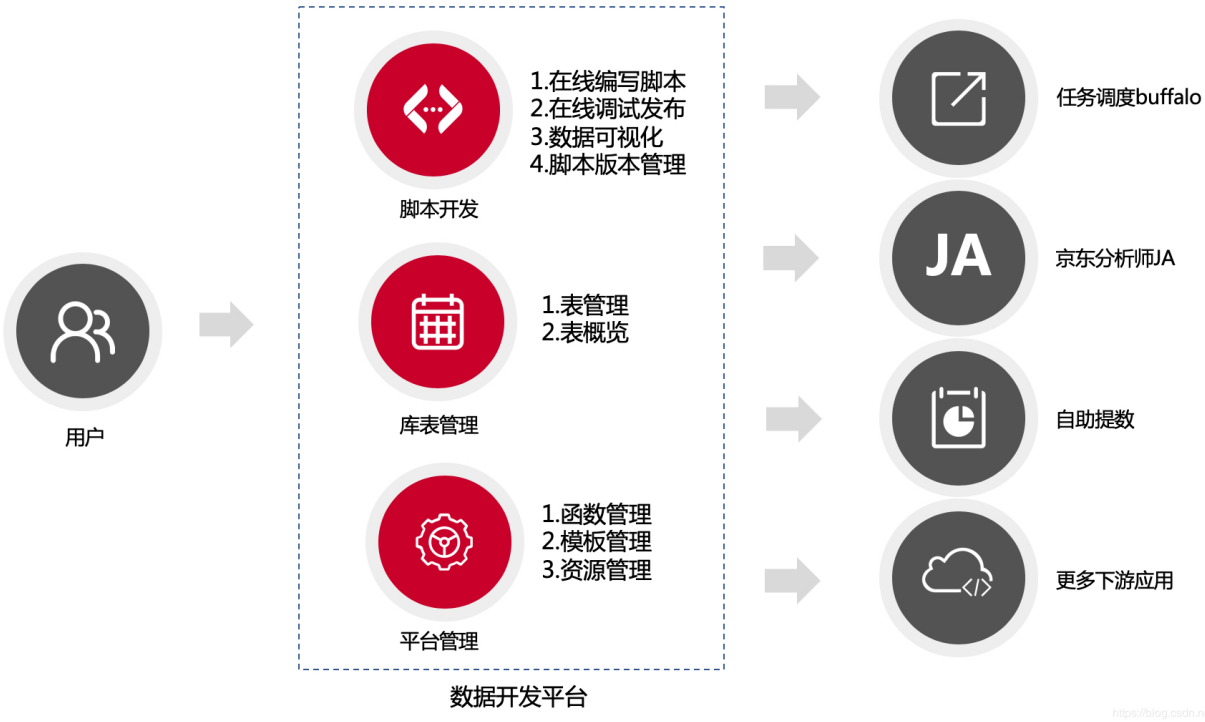
Zeppelin WorkFlow

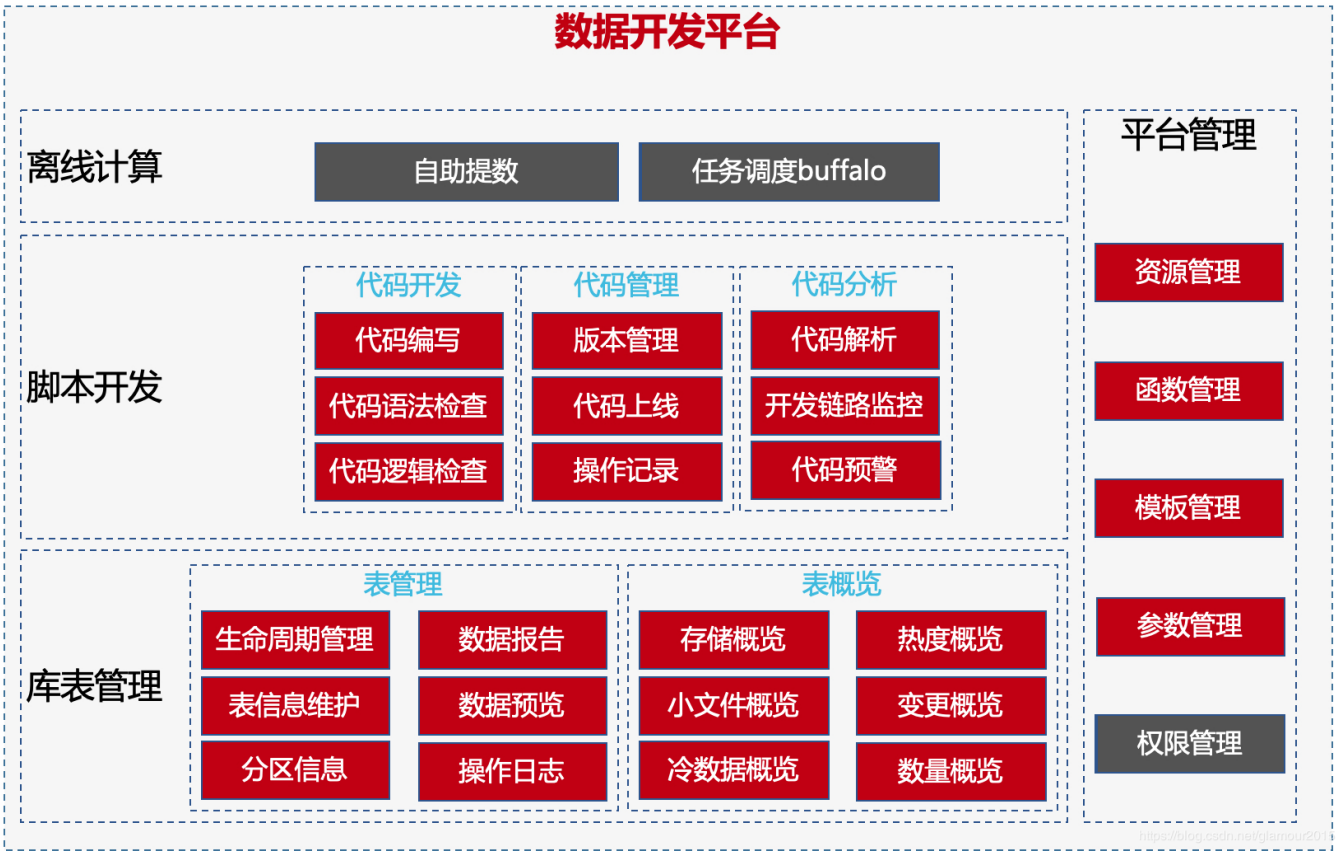
用户可以在按照 Zeppelin 提供的一种类似 Azkaban 的数据格式，编写 Node 之间的依赖，下方形成一个可视化的 WorkFlow 图，通过拖拽的方式可以编排整个工作流，设置每个节点的动作。结合参数的配置，用户可以编写一个复杂的 Zeppelin 工作流，在右边设置触发的条件，如按时间点、Rest 接口手动触发，或者按照周期性时间、数据变化来设置。

Zeppelin总结

Apache Zeppelin 覆盖机器学习全流程，让数据科学工作者能够以可视化的方式，方便地编写机器学习算法、调参和进行机器学习任务管理。针对大数据任务的特点，Zeppelin 也做了分布式的优化。同时，Zeppelin 还能与其他 Apache 大数据生态项目也能很好地集成，可以更好地满足不同团队的需求。

数据开发平台规划




数据开发平台
离线计算
自助提数
任务调度buffalo
脚本开发
代码开发
代码编写
代码语法检查
代码逻辑检查
代码管理
版本管理
代码上线
操作记录
代码分析
代码解析
开发链路监控
代码预警
库表管理
表管理
生命周期管理
数据报告
表信息维护
数据预览
分区信息
操作日志
表概览
存储概览
热度概览
小文件概览
变更概览
冷数据概览
数量概览
平台管理
资源管理
函数管理
模板管理
参数管理
权限管理