# 概述

# 技术方案

## 概述

### 产品采用架构，兼容多服务器

产品采用的分布式架构，支持X86，ARM服务器。

答：完全支持。系统采用分布式架构，兼容X86和ARM服务器，确保在高并发、大数据量场景下仍能保持良好的性能和稳定性。

白鲸开源是ApacheDolphinScheduler和ApacheSeaTunnel原班人马成立的开源商业公司，提供功能更强多、稳定性更强的商业版本解决用户调度，数据开发、数据同步和ETL的问题，目前支持192种数据库的ETL与数据开发，全面替换Informatica与Talend等工具相应功能，在中信证券、中信建投、中国银行、中国人保、中国人寿、旺旺集团等多个行业头部企业都有成功商业版和实施替换案例。



WhaleStudio是一款通过拖拽式的实现调度和ETL功能，支持快速融合在企业/合作伙伴的系统当中，感兴趣直接发邮件给service@whaleops.com，或者扫描以下二维码：



项目测试

* 工作流编排： 允许用户以编程方式定义工作流，将一系列任务组织成有向无环图（DAG），并指定任务之间的依赖关系和顺序。
* 调度和执行： 它可以按照预定的时间表和依赖关系调度任务，并执行这些任务。
* 监控和告警： 提供了用户友好的用户界面，用于监控工作流的运行状态、任务的执行情况和日志信息。同时，它还支持告警功能，可以在任务失败或达到特定状态时发送通知。
* 任务编写和定制：系统提供了丰富的内置操作符，涵盖了常见的数据处理、连接器和工作流程操作。
* 数据源支持MySQL、Oracle等结构化和半非结构化数据源，支持信创数据源等。
* 数据转换和清洗： 对数据进行实时转换、清洗和处理，确保数据质量和一致性。
* 可视化和监控： 提供用户友好的界面和监控工具，帮助用户监视数据流和任务的运行情况，以及识别潜在的问题或瓶颈。

### 基础、工作流、任务实例定义

基础功能，工作流定义，任务实例定义

答：完全支持。系统具备基础功能和工作流定义的管理能力，用户可以根据实际需求进行工作流的创建、编辑和部署。同时，系统支持任务实例的定义和管理，可以实现对任务实例的创建、启动、暂停、恢复和终止等操作。

本系统架构为分布式多集群部署，支持进行分组管理，如下图：



* 水平扩展性：本系统支持水平扩展，允许通过增加调度器、执行器和工作节点来提高处理能力。此特性使用户能够根据业务需求动态扩展资源，从而满足不同规模和负载的要求。
* 高可用性配置：为了实现高可用性，系统提供了多种配置选项，如通过负载均衡器（例如 NGINX 或 HAProxy）进行流量分发，配置多个元数据库实例以避免单点故障，并支持多节点部署的调度器和执行器，保障系统的稳定运行。

支持集群部署与故障转移，从而实现不停机扩展和缩减系统服务：



* Edge Node： Edge Node 是集群的管理节点，负责管理整个集群。它提供 Web UI 和 API 来管理、配置和监控集群中的各个 Collector 节点。
* Collector Node： Collector Node 是实际执行数据处理任务的节点。它们在集群中分布，执行数据流水线并处理数据。Collector 节点通过与 Edge Node 通信，获取配置信息和任务调度。

。

### 数据源支持，三大数据库

数据源支持，mysql，oracle，postgresql

答：完全支持。系统支持多种数据库类型的数据源接入，包括但不限于MySQL、Oracle和PostgreSQL等。系统提供数据库配置管理功能，支持数据源的添加、配置、修改和删除等操作，确保数据源的安全可靠连接。

本系统通过多管理节点部署，实现了工作流和数据同步功能，并具备节点故障转移能力，从而支持不停机的系统扩展与缩减。多管理节点架构确保在任意节点发生故障时，其他节点能够无缝接管任务，保障任务调度的连续性和稳定性。多节点并行调度不仅显著提高了系统的处理效率，还增强了系统整体的高可用性和可靠性：

* 先进的架构：工具将采用现代化的架构设计，确保高效稳定的数据处理和调度能力。
* 原生分布式部署：支持在分布式环境中进行部署，以适应大规模数据处理需求。



本项目预计采购的数据调度工具将满足以下高稳定性和成熟度要求：

* 系统架构稳定性：工具将保持系统架构的稳定性，确保在长期运行中的高可靠性。
* 成熟技术采用：采用经过市场验证的成熟技术，以满足不同层次用户的业务需求。
* 高压力场景考验：经过高压力场景的严格测试，验证其在复杂和高负载环境下的表现。

|  |  |
| --- | --- |
| 指标项 | 实际测试指标 |
| 故障发现 | 1s |
| 自动切换 | 1s |
| 可用率（年） | >=99.9999% |

## 产品架构

### 产品采用架构，兼容多服务器

产品采用的分布式架构，支持X86，ARM服务器。

答：完全支持。系统采用分布式架构，兼容X86和ARM服务器，确保在高并发、大数据量场景下仍能保持良好的性能和稳定性。

本系统架构为分布式多集群部署，支持进行分组管理，如下图：



* 水平扩展性：本系统支持水平扩展，允许通过增加调度器、执行器和工作节点来提高处理能力。此特性使用户能够根据业务需求动态扩展资源，从而满足不同规模和负载的要求。
* 高可用性配置：为了实现高可用性，系统提供了多种配置选项，如通过负载均衡器（例如 NGINX 或 HAProxy）进行流量分发，配置多个元数据库实例以避免单点故障，并支持多节点部署的调度器和执行器，保障系统的稳定运行。

支持集群部署与故障转移，从而实现不停机扩展和缩减系统服务：



* Edge Node： Edge Node 是集群的管理节点，负责管理整个集群。它提供 Web UI 和 API 来管理、配置和监控集群中的各个 Collector 节点。
* Collector Node： Collector Node 是实际执行数据处理任务的节点。它们在集群中分布，执行数据流水线并处理数据。Collector 节点通过与 Edge Node 通信，获取配置信息和任务调度。

。

### 支持主流操作系统

支持centos，redhat等主流操作系统。

答：完全支持。系统支持在centos、redhat等主流操作系统上的运行和部署，确保系统的稳定性和兼容性。

\*\*大数据平台产品功能介绍\*\*  
  
\*\*一、系统概述\*\*  
  
本大数据平台是一款专为处理海量数据而设计的高性能解决方案，能够轻松应对各种规模的数据处理需求，为企业提供强大的数据分析与挖掘能力。  
  
\*\*二、支持主流操作系统\*\*  
  
本产品经过精心设计与优化，支持多种主流操作系统，包括但不限于CentOS、Red Hat等，确保在各种环境下都能稳定运行，为用户提供可靠的服务。  
  
\*\*三、核心功能介绍\*\*  
  
1. \*\*数据存储与管理\*\*  
  
本大数据平台提供强大的数据存储能力，能够处理结构化和非结构化数据，实现数据的统一管理。通过分布式文件系统，有效扩展存储能力，满足企业日益增长的数据存储需求。  
  
2. \*\*高性能数据处理\*\*  
  
平台支持多种数据处理框架，具备高效的数据处理能力，可对海量数据进行实时分析和处理，加速业务决策过程。  
  
3. \*\*实时数据流处理\*\*  
  
通过实时数据流处理技术，平台能够处理高速流动的数据，确保数据的实时性，适用于物联网、金融交易等场景。  
  
4. \*\*数据挖掘与分析\*\*  
  
提供丰富的数据分析工具与算法库，支持数据挖掘、机器学习等高级分析功能，帮助企业从海量数据中挖掘价值，发现业务机会。  
  
5. \*\*数据可视化\*\*  
  
平台提供直观的数据可视化功能，以图表、报告等多种形式呈现数据分析结果，帮助用户更好地理解数据，做出更明智的决策。  
  
6. \*\*数据安全与隐私保护\*\*  
  
平台注重数据安全和隐私保护，提供数据加密、访问控制、审计追踪等功能，确保数据的安全性和完整性。  
  
7. \*\*弹性扩展与容错性\*\*  
  
本大数据平台具备出色的弹性扩展能力，可根据业务需求灵活扩展资源。同时，平台具有高度的容错性，能够在节点故障时自动进行任务迁移，保证服务的连续性。  
  
8. \*\*集成与开放性\*\*  
  
平台支持多种数据源的接入，能够与其他系统无缝集成，同时具备开放性API接口，方便二次开发与定制。  
  
9. \*\*智能监控与日志管理\*\*  
  
提供智能监控功能，实时监控平台运行状态，自动预警潜在问题。同时，完善的日志管理功能，方便用户追踪问题原因，优化系统性能。  
  
\*\*四、总结\*\*   
本大数据平台是一款具备高性能、高可用性、高扩展性的数据处理解决方案，支持多种主流操作系统，为企业提供一站式的数据处理服务。通过强大的数据处理能力、丰富的分析工具和安全保障机制，帮助企业从数据中挖掘价值，推动业务创新与发展。