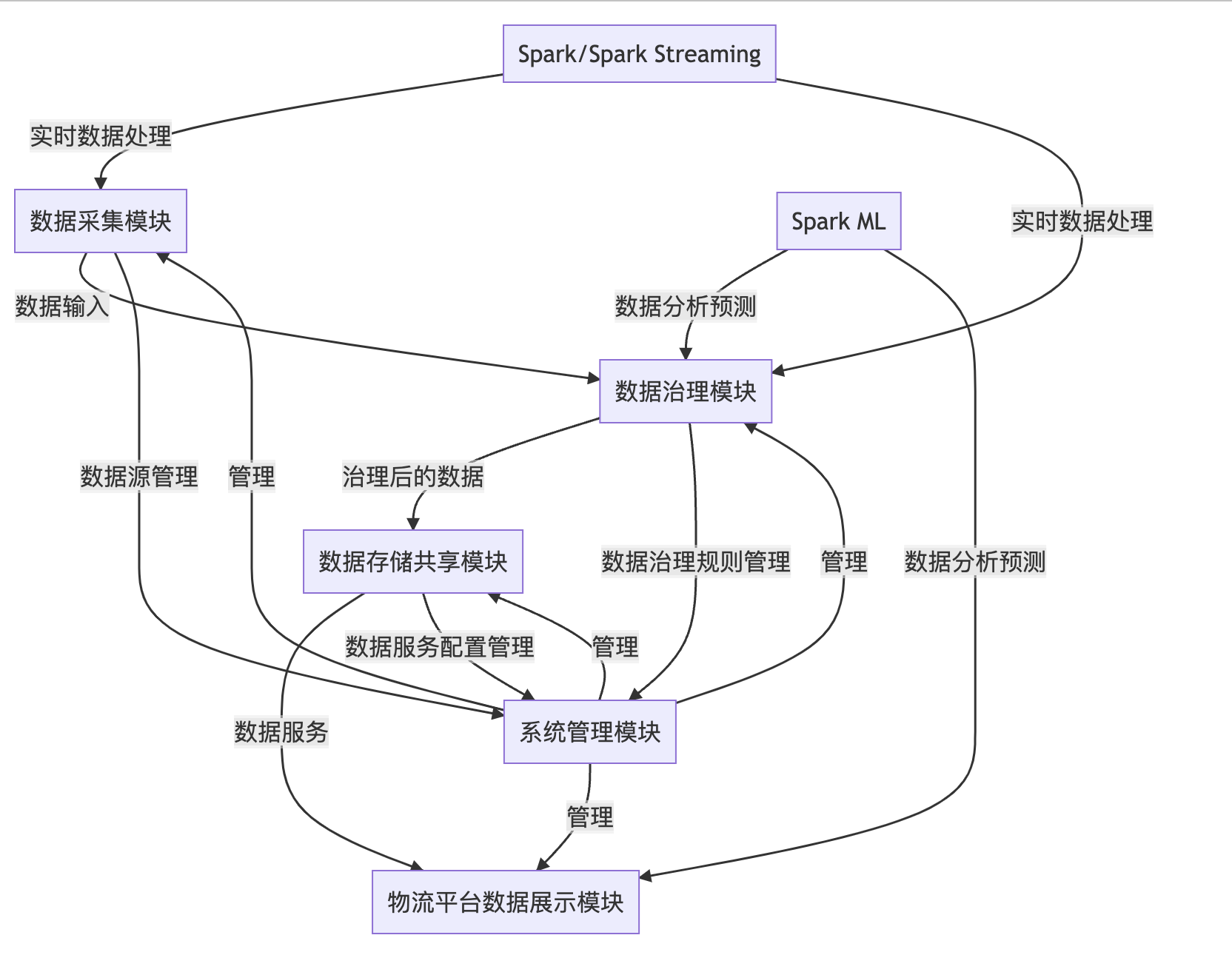
基于云原生大数据平台的数字化中台物流信息分析系统架构方案

实验方案：基于云原生大数据平台的物流信息分析系统



环境搭建准备： 使用达梦云原生大数据平台，将所需的开源软件，如Hadoop、Spark、Tomcat、Nginx等配置并运行在Kubernetes容器中。同时，安装并配置达梦数据库DM8和DMETL等国产软件(M8数数据库）等，我们主要是基于Spark/Flask/Python3技术栈的哈：

Spark 集成到数据采集模块和数据治理模块，用来处理大量的数据和进行复杂的数据治理任务。Spark Streaming 可用于处理实时的数据流，而 Spark ML 可用于数据分析和预测。数据采集模块：在数据采集模块中，Spark Streaming 可以被用来处理实时的数据流。例如，如果你的数据源是一个实时产生数据的API或者一个不断更新的数据库，你可以使用 Spark Streaming 来读取这些数据，然后将它们发送到数据治理模块。你也可以使用 Spark 的 batch processing 功能来处理批量的数据，例如从一个文件或者数据库中读取大量的数据。

在数据治理模块中，你可以使用 Spark 来进行数据处理和清洗。Spark 提供了很多内置的函数，可以用来进行数据转换、过滤、聚合等操作。此外，如果你的数据规模非常大，使用 Spark 可以帮助你提高处理速度。你也可以使用 Spark Streaming 来实时处理数据；在物流平台数据展示模块中，你可以使用 Spark ML 来进行数据分析和预测。例如，你可以使用 Spark ML 的机器学习算法来预测港口的吞吐量，或者分析不同类型货物的流向。你可以将这些分析结果保存到数据库中，然后在物流平台数据展示模块中展示出来。在数据采集模块中，使用 Spark Streaming 从实时数据源读取数据，或者使用 Spark 的 batch processing 功能从文件或数据库中读取批量数据。

具体流程：

将读取到的数据发送到数据治理模块。在这个模块中，使用 Spark 进行数据清洗和处理，例如进行数据转换、过滤、聚合等操作。

将处理后的数据保存到 DM8 数据库中，作为数据存储共享模块的一部分。

在物流平台数据展示模块中，使用 Spark ML 进行数据分析和预测。将分析和预测的结果保存到数据库中。

在物流平台数据展示模块中，从数据库中读取数据和分析结果，然后在前端页面中展示出来。

在系统管理模块中，管理员可以管理数据源、数据治理规则、数据服务配置等

总的系统可以这样拆分：

数据采集： 开发数据采集模块，采用Python(scrapy)/selenuim 技术栈 编写爬虫脚本，规定并采集从mysql数据库、hdfs存储、minio存储的数据。这些数据源的文件格式可以是txt、csv、xls等。此外，可以编写动态管理（创建，修改，删除）数据源的功能。

该模块负责制定数据采集规则并采集数据，从mysql数据库、hdfs存储和minio存储中采集数据，并处理txt、csv、xls等格式的数据。该模块还需要包含一个爬虫组件，用于抓取互联网上的公开数据。在数据采集过程中，需要确保能够处理各种异常和错误。

这个模块应该使用一个灵活的配置系统，以便动态地管理数据源的创建、修改和删除。这可能包括一个用户界面，允许用户输入数据源的详细信息，以及一个后端服务，将这些信息保存在数据库中。

数据治理： 开发数据治理模块，制定并按照治理规则对数据进行处理。例如，对身份证号进行合规性检查，检验物流信息中的客户是否在客户数据中存在等。同时，可以开发动态配置治理规则的功能，如数据聚合，数据去重，异常值检测，数据转换等。

该模块负责制定数据治理规则，并按照这些规则处理数据。这包括验证客户的身份证号码，确保物流信息中的客户存在于客户数据库中，并将治理完成的数据和异常数据分开存放。模块应该包含一个治理规则配置系统，允许动态地配置数据治理规则，如数据聚合，数据去重，异常值检测，数据转换等。这可能需要一个用户界面，允许用户创建和修改治理规则，以及一个后端服务，将这些规则保存在数据库中（M8)

数据存储与共享： 数据存储和共享模块将负责将治理后的数据存储在DM8数据库中，并提供数据服务，例如提供接口服务和共享数据库。此模块还可以开发动态配置接口服务、共享数据的功能。模块负责存储治理后的数据，并对外提供数据服务。这可能包括一个RESTful API(Flask Cros API)，允许其他系统获取数据，以及一个数据库接口，允许直接访问数据库。

该模块应该使用DM8数据库来存储数据，并需要一个配置系统，允许动态地配置接口服务和共享数据。这可能包括一个用户界面，允许用户创建和修改接口服务和共享数据的配置，以及一个后端服务，将这些配置保存在数据库中

数据分析与展示： 开发一个物流信息分析系统，该系统使用从数据存储共享模块获取的数据。系统将包含港口的吞吐量分析，港口不同类型货物吞吐趋势，港口货物吞吐同比环比，不同货物吞吐占比，不同货物流向分析，不同类型货物堆场流转周期分析等功能模块。此外，可以拓展开，进行港口效率、物流业务查询、集装箱跟踪等分析和查询。该模块负责展示数据，包括港口的吞吐量分析，港口不同类型货物的吞吐趋势，港口货物吞吐的同比环比，不同货物的吞吐占比，不同货物的流向分析，以及不同类型货物的堆场流转周期分析，模块应该使用现代的前端技术（例如Vue+boostrasp）来创建用户界面，并使用一种可视化库（例如D3或Chart.js）来展示数据。它还应该包含一个后端服务，该服务使用从数据存储共享模块提供的API或数据库接口获取数据。

系统管理模块

负责提供系统管理功能，包括用户管理，权限管理，系统配置等。它应该包含一个用户界面，允许管理员创建和管理用户，分配权限等。并且，它也应该包含一个后端服务，管理这些信息并保存在数据库中。

扩展分析： 结合其他数据集以充实数据分析，如网上公开的气象历史数据，分析气候条件对港口吞吐量，货物流入流出等的影响，并完成相关数据的展现。

项目管理： 使用敏捷开发方法进行项目管理，对整个项目进行迭代开发和优化。利用单元测试和集成测试保证软件质量，并使用持续集成/持续部署（CI/CD）工具提高开发效率。容器化和微服务架构：Docker，Kubernetes等；CI/CD：Jenkins，GitLab CI/CD等