KORE DWH

PIMON 2024

ТТИИ

09.24

Описание продукта

Набор методик и инструментов для автоматизации построения хранилища данных.

Архитектурный дизайн системы

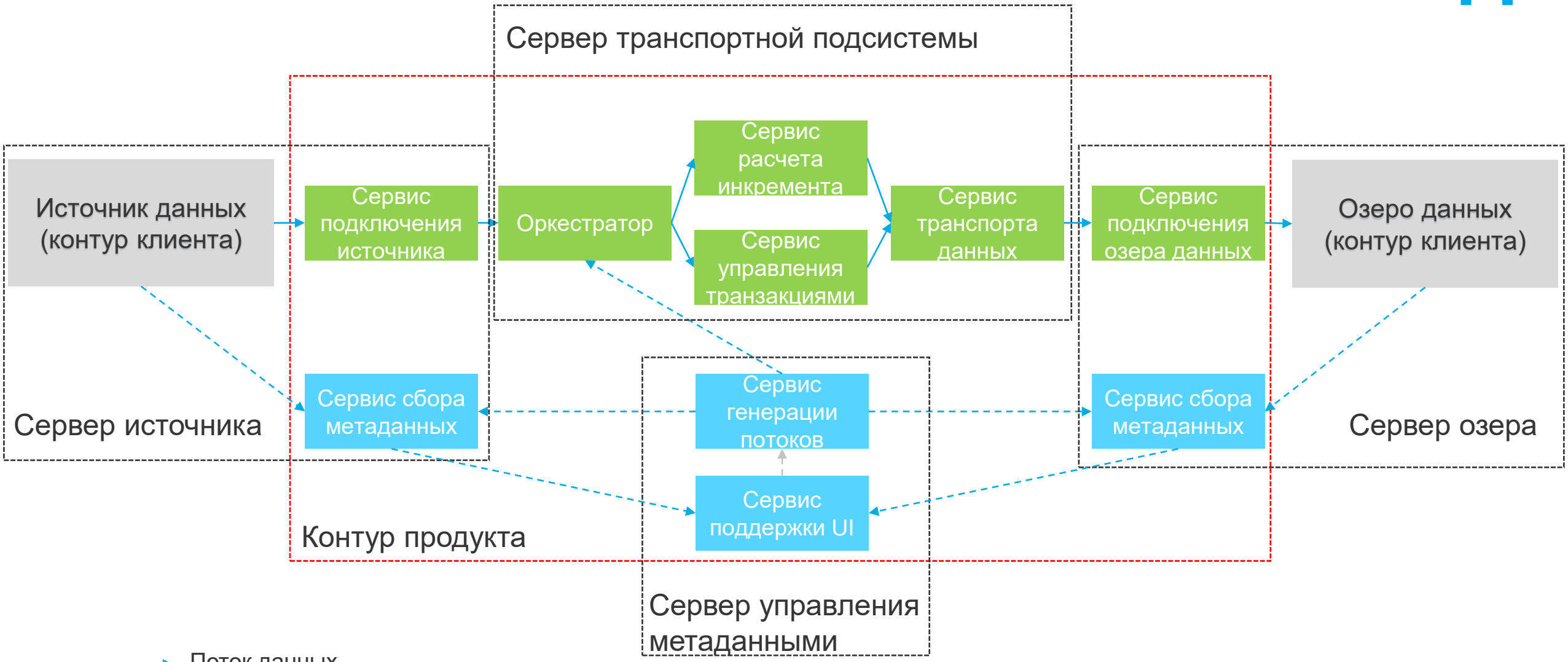
Стандарты моделирования проектирования и разработки

Готовые инструменты и технологии

Преимущества готового решения

- Базис для создания децентрализованной платформы данных
- Повышение гибкости и скорости для внесения изменений
- Увеличение управляемости нагрузкой на источники и получатели данных
- Устранение «бутылочных горлышек» в процессе обеспечения данными потребителей
- Объединение усилий всех участников процесса предоставления данных

Архитектура. Потоки данных



3

Функционал

Организация и выполнение интеграционных потоков

Работа с ресурсами

Логическая сущность системы, с которой выполняет работу поток. Каждая система представляется набором ресурсов, для управления разрабатывается провайдер, реализующий набор API по спецификации

Работа с транзакциями

Утилита для создания и управления распределенными транзакциями загрузки данных. Менеджер транзакций взаимодействует со всеми провайдерами ресурсов, участвующих в транзакции, по выделенным API

Работа с инкрементами

Используется для регистрации изменений в областях данных, с которыми работают провайдеры ресурсов. Менеджер дельты регистрирует все операции над ресурсом и создает версии ресурса, обслуживаемых конкретным провайдером

Стандартные преобразования

Большое количество подготовленных функций и операторов, решающих задачи преобразования данных, работы с различными алгоритмами версионирования и историчности данных, возможность добавлять собственные функции

4

Интеграция на основе метаданных

Основной принцип построения интеграции с использованием платформы KORE – максимальное использование автоматизировано обрабатываемых метаданных для описания входов, выходов, настроек и правил трансформации

KORE DWH интегрирован с инструментом управления метаданными, который выступает в роли пользовательского интерфейса

Цель – снизить количество ручных операций по поиску и описанию данных и сократить расстояние от поиска данных до тестирования потока

Технические метаданные используются для описания потенциальных объектов источников и получателей данных, а также для анализа графа прослеживаемости

Бизнес-метаданные используются для логических объектов платформы KORE и связанных описаний

Связь с инструментами проверки качества данных для сквозной информации по источникам данных и результате загрузки

Профилирование данных для контроля работы потока: количества обрабатываемых записей

Группы пользователей

Аналитик	Поиск источников	Создание запроса	Тестирование
Пользователь-потребитель данных	▪ Поиск в каталоге подходящих таблиц-источников, подключенных к KORE	▪ Формирование заявки на данных с описанием желаемых преобразований и фильтров. Согласование запроса	▪ Ожидание отработки потока перекладки данных. Анализ профиля и качества данных таблицы-результата

Разработчик	Анализ ТЗ	Настройка потока	Расширение по требованиям
Пользователь, разрабатывающий потоки данных хранилища	▪ Анализ бизнес постановки на загрузку и постобработку данных	▪ Выбор конфигурации рабочего и управляющего потока в соответствии с регламентами	▪ Разработка собственных функций обработки данных при функциональном дефиците базового функционала

Разработчик источника	Составление семантики	Определение источника	Настройка провайдера
Ответственный за разработку интеграции со стороны систем-поставщиков	▪ Анализ логической модели и требований семантического слоя	▪ Определение таблиц и сервисов, откуда можно брать данные, формирование правил подключения по требованиям ИБ и производительности	▪ Создание/загрузка ресурсов в соответствующем провайдере. Создание необходимых оповещений в инструменте управления метаданными

Преимущества

Классические ETL решения

Инструмент стандартизирует и автоматизирует интеграцию на уровне шага процесса передачи данных

Инструмент для разработчика, частично для системного аналитика

Область применения – только команда хранилища или озера данных

Метаданные – справочная информация, использование ограничено

Инструменты классического стека работы с данными, хорошо применимо для классических архитектур ХД

Платформа KORE DWH

Уровень стандартизации – тип потока перекладки данных. Платформа определяет и инкапсулирует все стандарты работы с потоком

Платформа содержит функционал для совместной работы всех групп пользователей.

Разделяет функции процесса работы с данными между командами ХД и источника

Метаданные – ядро системы, важный рабочий инструмент

Инструменты modern data stack, лучше применимо для распределенных архитектурных шаблонов, таких как data fabric, data mesh и т.д.

Технологический стек

Провайдер ресурсов

Обеспечивает взаимодействие с системой. Реализуется на базе компонентов **KORE.DWH (Python)**. Реализует проверки доступности и готовности источника и предоставляет необходимую информацию для подключения

Предоставление метрик

Обеспечивает на стороне источника и платформы расчет агрегированных значений, маскирование данных, семплирование перед отправкой данных в платформу. Реализуется на базе **Python** или **DBT**

Транспортный слой

Выполняет подготовленные потоки данных в необходимом режиме взаимодействия. Для реализации могут использоваться компоненты решений Заказчика (например **PXF**) или комбинация **NiFi** или **Airflow** + **Spark**

Шаблоны потоков и трансформаций

Обеспечивает хранение структурированных настроек и их трансформацию в конфигурации объектов транспортного слоя. Реализуется на базе **Airflow**, **DBT**, а также компонентами системы. Версионирование осуществляется в **Gitlab**

Управление метаданными

Обеспечивает ведение пользовательских настроек и просмотр метаданных подключённых систем. Единая точка пользовательского интерфейса. Реализуется на базе **React JS** + **OpenMetadada (Arenadata Catalog)**

Хранение наборов данных

Зона хранения подготовленных наборов данных и предоставления их пользователям-заказчикам на базе **S3**, **HDFS**, **GreenPlum**, возможно использование реляционной базы заказчика, совместимой с **DBT**