Разработка расширяемой системы обработки и анализа разнородных данных

Автор: Шабанов В.С., гр. Б15.11-мм
Научный руководитель: ст. преп. Луцив Д. В
Рецензент: инженер-программист ООО "Девяносто Один" Карпов Д.Н.

Ситуационный центр Ленинградской области

- Данные поступают из различных источников: система 112, ручная загрузка из различных файлов и баз данных. Не больше нескольких раз в день
- Анализ данных в системе происходит вручную многими аналитиками одновременно
- Скорость анализа данных важнее скорости загрузки
- Аналитические запросы часто содержат функции агрегации и фильтрации

Постановка задачи

Цель работы - разработка новой масштабируемой аналитической системы

Задачи:

- Провести обзор существующих решений
- Собрать требования к системе
- Разработать архитектуру системы анализа данных
- Реализовать систему загрузки, обработки и анализа данных пользователя

Существующие решения

- Tableau
  + Возможности одновременной работы
- QlikView
  + Автоматически обнаруживает связи между таблицами
- Power BI
  - Только для Windows

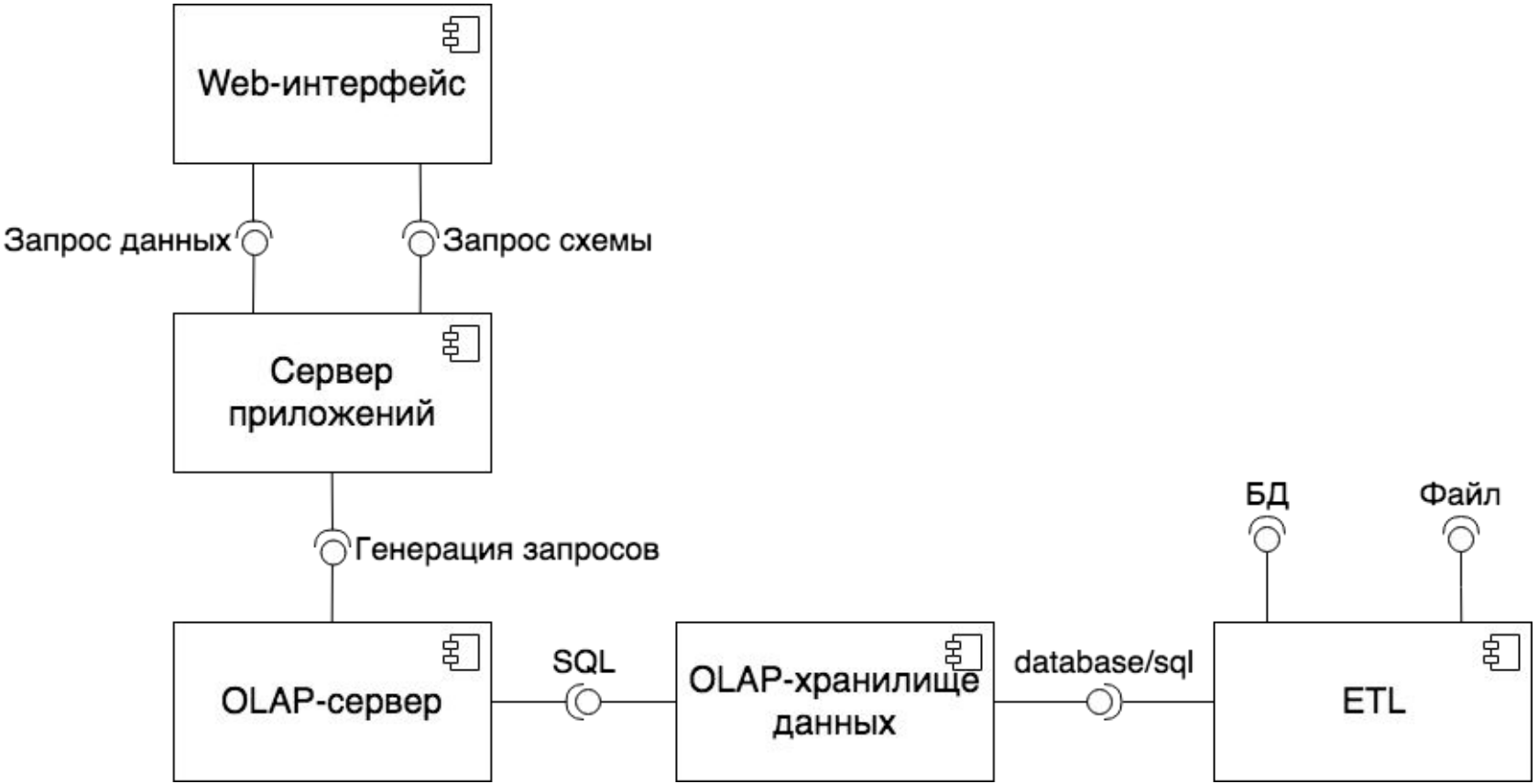
Общие проблемы:

- Отсутствие приближенных вычислений
- Узкий выбор аггрегационных функций. (corr, quantile, exp)
- Недостаточно гибкая функциональность фильтров (есть HAVING, но нет WHERE)

Требования

- Подсистема интеграции, информационного взаимодействия и загрузки данных
  - Форматы данных
  - Определение типов
  - Анализ загруженных данных
- Подсистема многомерной аналитической обработки информации
  - Агрегационные функции
  - Функции интерполяции
- Подсистема визуализации аналитических панелей с модулем детализации

Архитектура



Столбцовые СУБД

- Данные хранятся по столбцам, а не в строках
- Быстрее считать данные из отдельных столбцов, но медленнее из отдельных строк

Выбор хранилища данных

	Vertica	ClickHouse	InfiniDB	MonetDB	Hive
COUNT(*)	0.044	0.137	1.972	0.029	2.157
WHERE	0.124	0.094	0.801	2.557	58.777
SUM	0.250	0.442	86.484	87.650	47.942
UNIQ	0.927	0.610	28.658	14.286	61.381
AVG, GROUP BY	4.119	1.101	8.221	275.857	103.410

Выбор OLAP-сервера

- Mondrian
  - \+ Кроссплатформенность
  - \+ Подстраивается под разные хранилища данных
  - \- Нужно генерировать MDX запросы
  - \- Сложно обновить схему
- Druid
  - \+ Готовый веб интерфейс
  - \- Не подстраивается под другие хранилища данных

Было решено реализовать свой OLAP-сервер

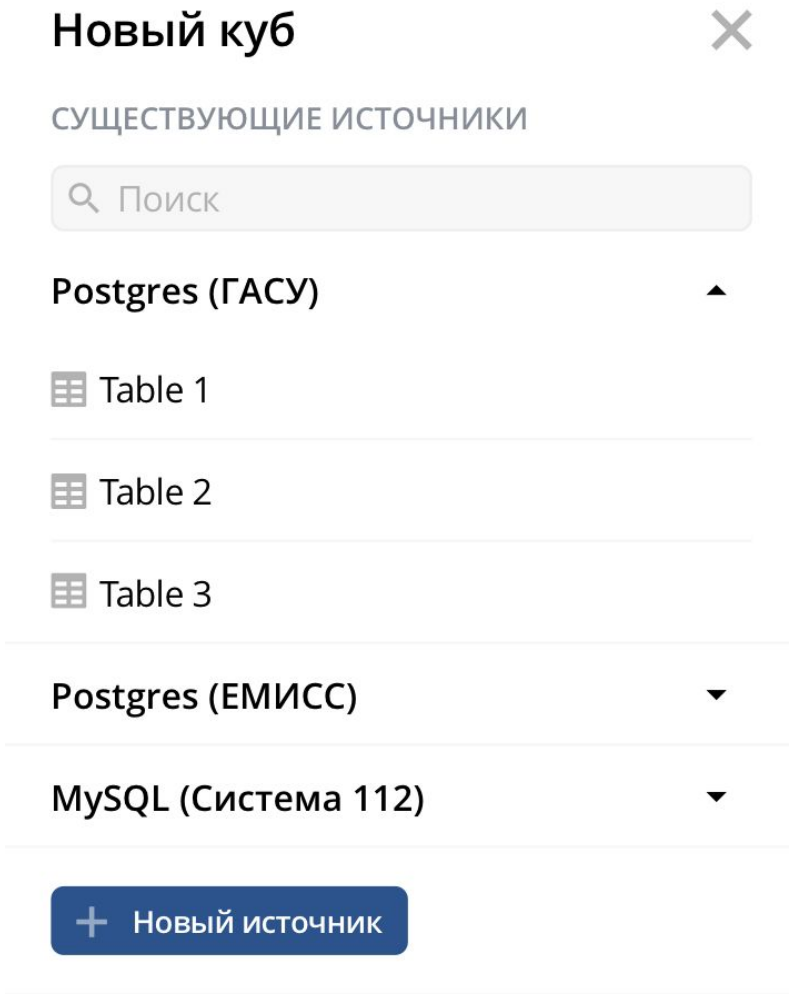
Технологии

ETL-сервис: Go

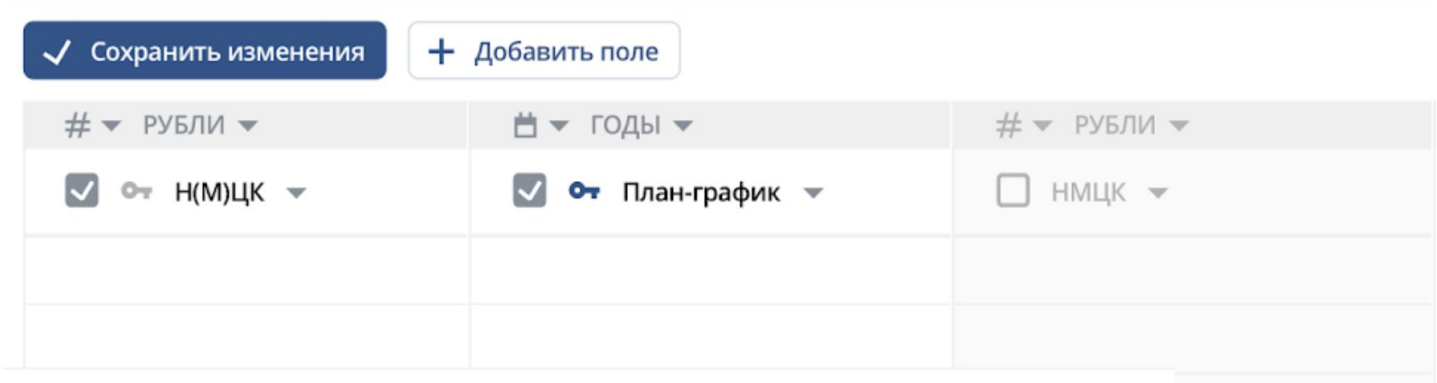
Сервер приложений и OLAP-сервер: Ruby on Rails

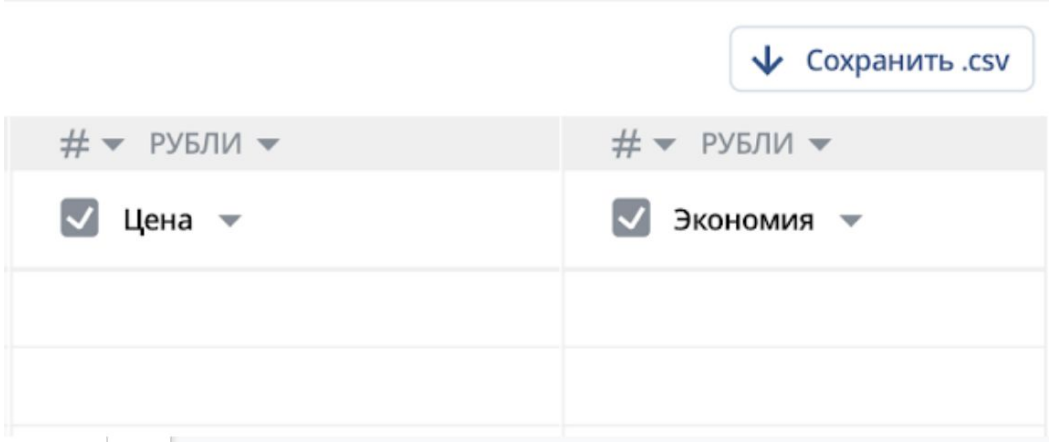
Веб-интерфейс: TypeScript + React

Интерфейс


Новый куб
СУЩЕСТВУЮЩИЕ ИСТОЧНИКИ
Поиск
Postgres (ГАСУ)
Table 1
Table 2
Table 3
Postgres (ЕМИСС)
MySQL (Система 112)
Новый источник

Интерфейс


Сохранить изменения
Добавить поле
РУБЛИ
ГОДЫ
РУБЛИ
Н(М)ЦК
План-график
НМЦК


Сохранить .csv
РУБЛИ
РУБЛИ
Цена
Экономия

Результаты

- Проанализированы основные решения, существующие на рынке: QlikView, Tableau, Power BI. Выделены общие черты интерфейса и сценарии знакомые пользователю
- Собраны требования к системе анализа данных. Определены СУБД и форматы файлов необходимые заказчику
- Создана модульная архитектура системы. В качестве хранилища выбран ClickHouse. OLAP-сервер было решено реализовывать с нуля
- Выполнена реализация модуля загрузки, обработки и анализа данных