OTUS

OLAP CH Superset

Курс СУБД

otus.ru

Проверить, идет ли запись

Меня хорошо видно && слышно?

Ставим “+”, если все хорошо
“-”, если есть проблемы

Тема вебинара

Apache superset: Построение отчетности

Ржевский Михаил

Преподаватель курсов OTUS
Базы данных
PostgreSQL для администраторов баз данных и разработчиков
PostgreSQL Cloud Solutions
MS SQL Server Developer

Опыт: 15 в сфере IT, более 10 лет в качестве преподавателя
1C, Green plum, MS SQL Server, MySQL, C# , HTML, XML, CSS, Javascript, JQuery, Unit Tests
сертифицированный разработчик Dynamics AX и Dynamics CRM

Правила вебинара

Активно
участвуем

Off-topic обсуждаем
в учебной группе

Задаем вопрос
в чат или голосом

Вопросы вижу в чате,
могу ответить не сразу

Цели вебинара

К концу занятия вы сможете

1. Понимать, что такое аналитическая отчетность ;
2. Зачем нужно собирать куб в Clickhouse;
3. Создать панель мониторинга в Superset.

Смысл

Что такое аналитическая отчетность.

1. Строится на отдельной СУБД или кубе.
2. Предшествует процесс сбора и обрабоки данных.
3. В общем случае BI отображает и агрегирует имеющиеся в хранилище данных цифры в удобной табличной или графической форме, подготавливая их для проведения анализа.

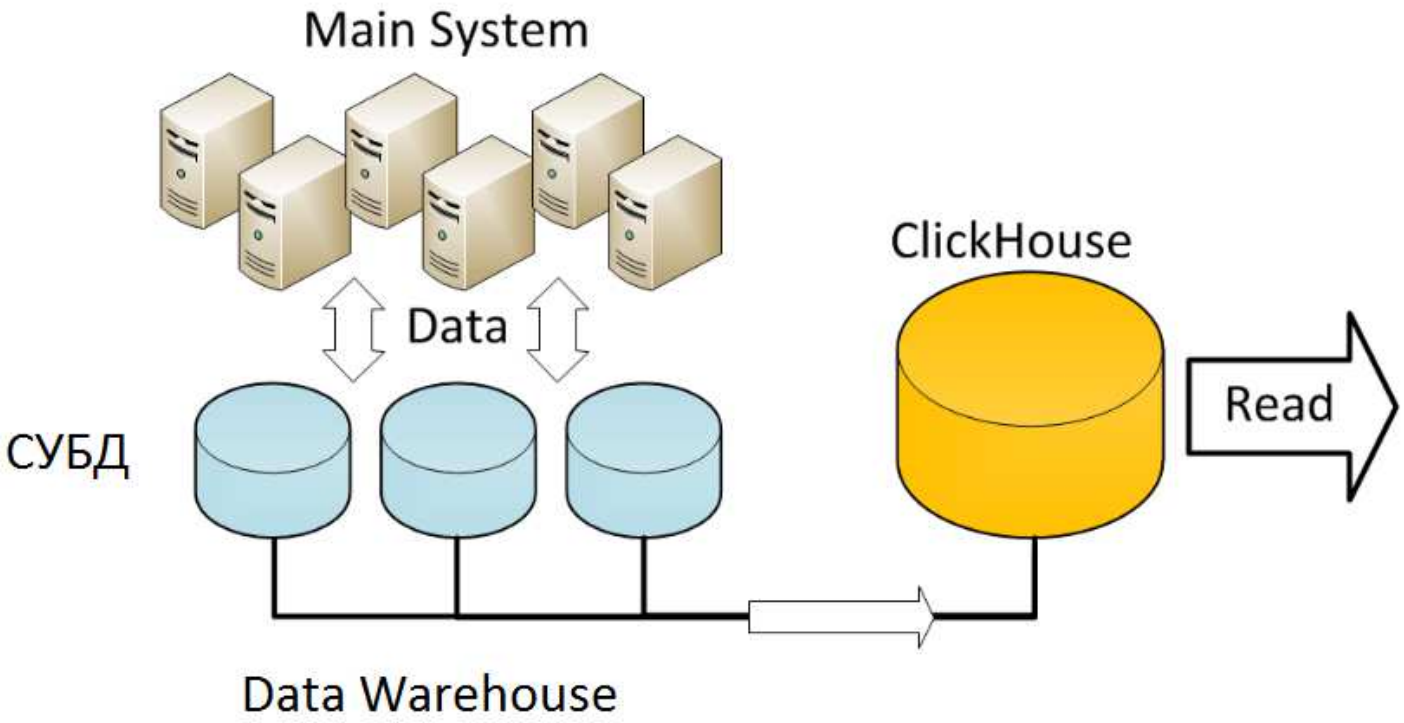
Выбор СУБД

Почему Clickhouse.

1	Столбцовая СУБД.
2	Сжатие данных.
3	Хранение данных на диске
4	Параллельная обработка запроса на многих процессорных ядрах
5	Распределённая обработка запроса на многих серверах
6	Поддержка SQL
7	Наличие индекса

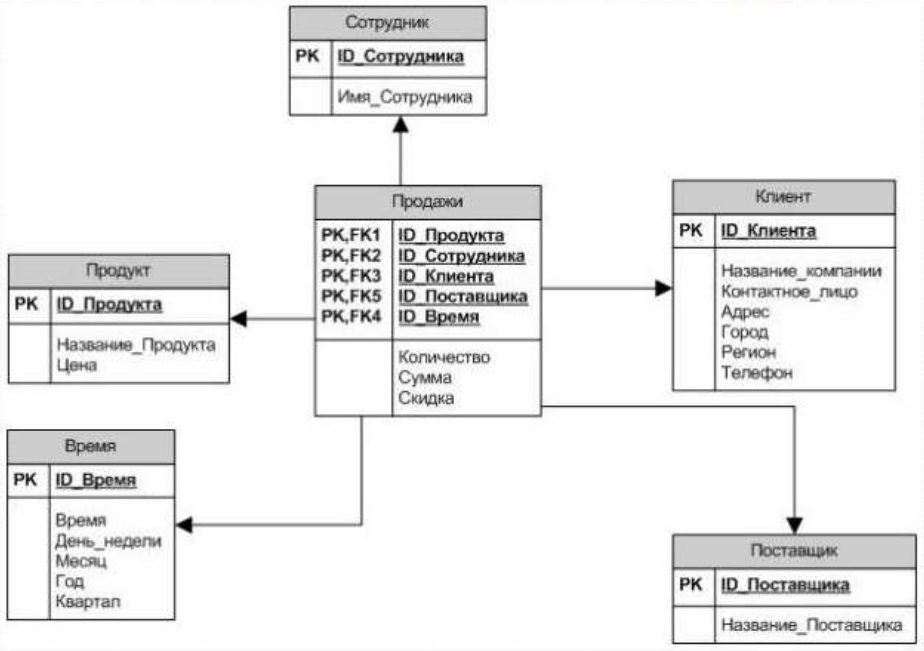
https://clickhouse.com/docs/ru/introduction/distinctive-features?ysclid=lm9ghv3xsi164866171

Построение аналитической отчетности



Проектирование OLAP

ROLAP – схема «звезда»



В центре – таблица фактов, по краям – таблицы измерений

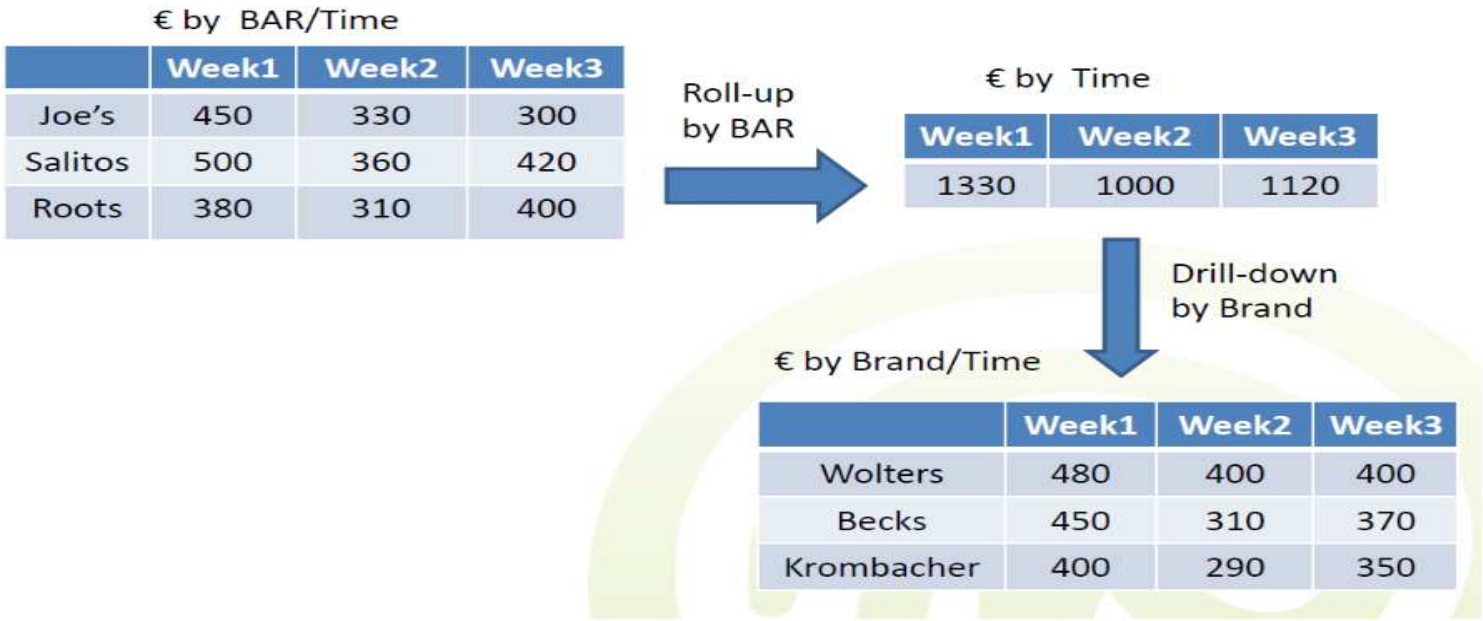
https://clickhouse.com/docs/en/getting-started/example-datasets/star-schema

Проектирование OLAP

ROLAP – схема «снежинка»

Основные операции OLAP

- Roll up: агрерация данных: по иерархии(-ям) до полного исключения измерения.
- Drill down: детализация: от обощенных данных к более детальным, от верхних уровней измерений – к нижним, детализация данных по дополнительным измерениям.

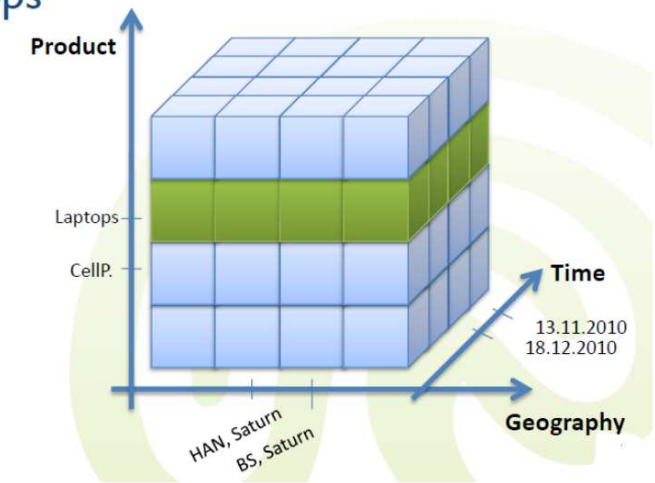


Основные операции OLAP

- Slice and dice: проекции и выборки – выборка нужных —ломтей кубика.

Slice

- Amounts to equality select condition
- WHERE clause in SQL
  - E.g., slice Laptops



Основные операции OLAP

- Pivot (rotate): вращение куба, визуализация, выборка и ориентация одно-, двух-, трехмерных срезов для визуального анализа
- drill across:кросс-детализация (условно – смена кубов при drilldown)
- drill through:переход с самого нижнего уровня детализации OLAPкуба, к фактам из выбранной ячейки (из исходной реляционной таблицы)

Pivot

- Pivoting on City and Day

	Mon	Tue	Wed	Thu	Fri	Sat	San	SubTotal
Hamburg	60	60	0	140	0	880	0	1140
Hannover	550	0	0	0	100	0	0	650
Braunschweig	540	300	0	0	0	0	50	890
SubTotal	1150	360	0	140	100	880	50	2680

	Hamb..	Han.	Bra..	SubTotal
Mon	60	550	540	1150
Tue	60	0	300	360
Wed	0	0	0	0
Thu	140	0	0	140
Fri	0	100	0	100
Sat	880	0	0	880
San	0	0	50	50
SubTotal	1140	650	890	2680

Вопросы?

Ставим “+”,
если вопросы есть

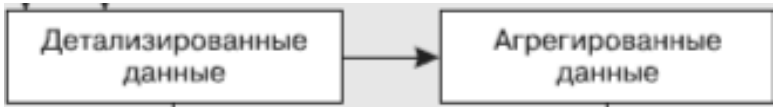
Ставим “–”,
если вопросов нет

Подготовка витрины для аналитики

Зачем?

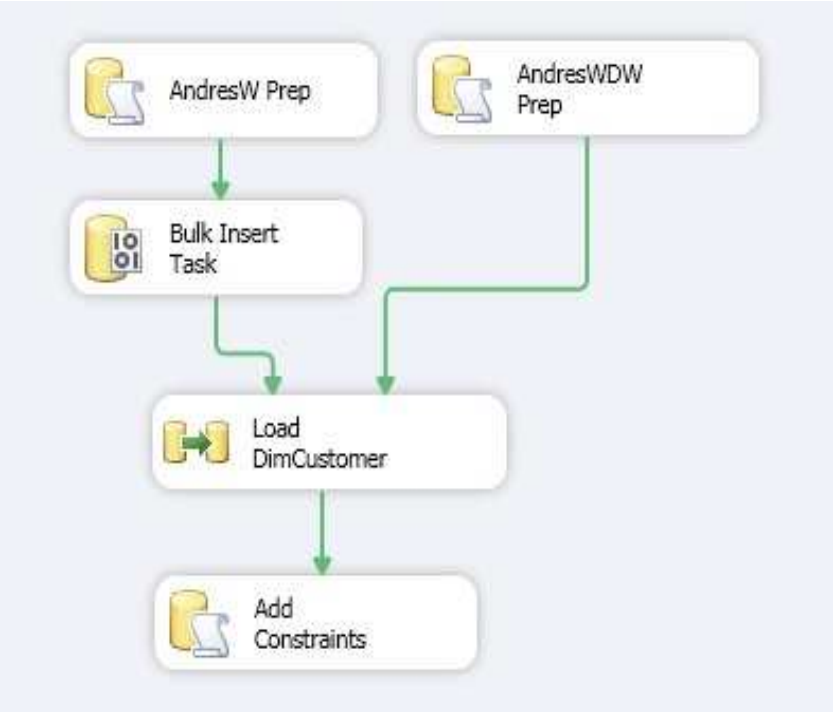
Задачи решаемые подготовкой данных. ETL. DWH

- **Консолидация** — комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему. Данные на предприятии расположены в различных источниках самых разнообразных форматов и типов — в отдельных файлах офисных документов (Excel, Word, обычных текстовых файлах), в учетных системах ERP
- **Обогащение** — процесс дополнения данных некоторой информацией, позволяющей повысить эффективность решения аналитических задач.
- **Очистка** данных от «мусора»
- **Агрегация**. Расчет показателей



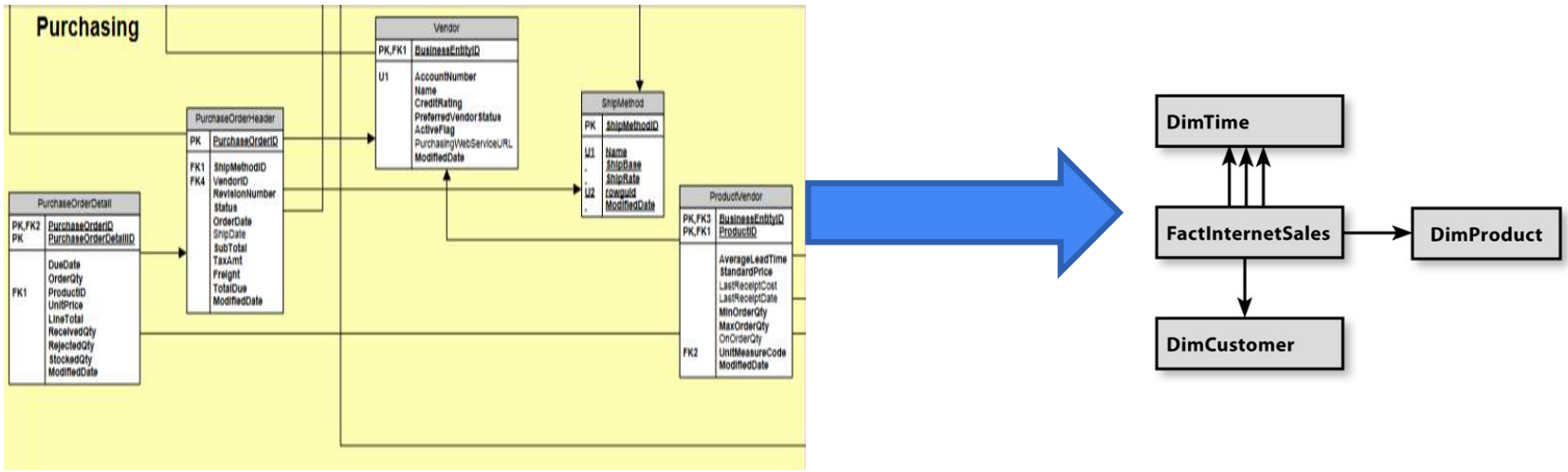
Пример подготовки данных для аналитики. ETL. SSIS

- Создание промежуточной области (пустого хранилища данных OLAP), содержащей только сценарии схемы базы данных.
- Загрузка данных из другого источника (плоский файл) во временную базу данных OLTP (задача массовой вставки).
- Добавление дополнительных столбцов в измерения для SCD, если они не существуют, для отслеживания исторических значений.
- Отключение одного ограничения в хранилище данных OLAP и в конце повторное добавление этих ограничений в сценарии.
- Представление потока управления SSIS



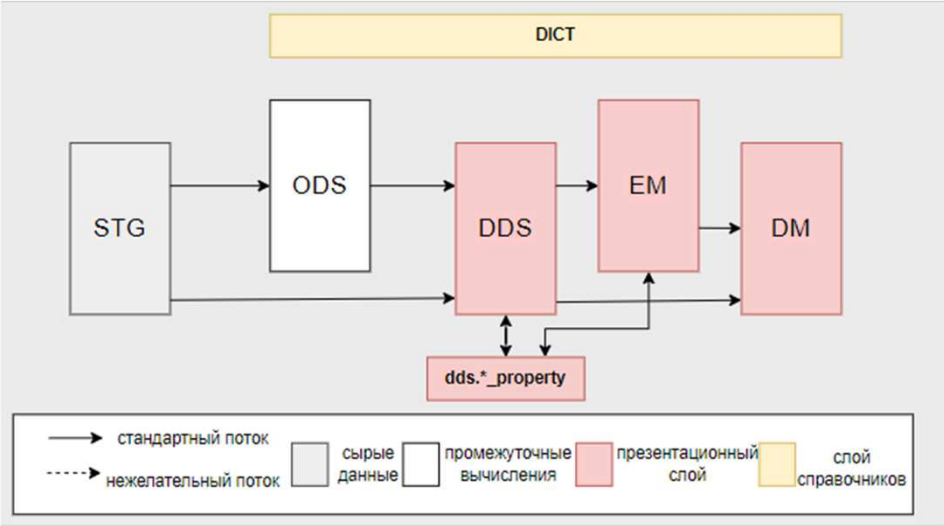
https://github.com/andrescastillol/ETL-DesignSolution/blob/master/Scripts/DimCustomerQuery.sql

Из чего что должно получиться



https://blog.jpries.com/wp-content/uploads/2015/12/AdventureWorks2008_db_diagram.pdf

Методология прохождения данными слоев обработки



•STG — слой, содержащий сырые ext-таблицы. Отдельного ELT-потока в GP нет, таблицы только подключаются к Hadoop.
•ODS — слой для накопления истории из слоя STG по нужным атрибутам. Используется редко, так как историю сырых данных мы не копим.
•DDS — содержит детальные данные по основным сущностям.
•EM — содержит витрины с агрегированными показателями базовых сущностей: клиентский портфель, портфель HR, кредитные и депозитные портфели и т.д.
•DM — витрины с рассчитанными агрегатами, сложными расчетами атрибутов. Этот слой содержит общие витрины для всех департаментов. Только на его основе далее формируются отчеты. Также есть отдельные DM-слои под конкретные бизнес направления.
•DICT – слой справочников.

https://habr.com/ru/companies/rosbank/articles/678646/

Инструменты для построения витрины DWH. Trino

https://en.m.wikipedia.org/wiki/Trino_(SQL_query_engine)

https://github.com/trinodb/trino?ysclid=lmau6le9x528225103

https://hub.docker.com/r/trinodb/trino

Преимущества Trino:
1. Масштабируемость: Trino может эффективно работать с огромными объемами данных, разделенными по нескольким узлам. Он может горизонтально масштабироваться, добавляя новые узлы к кластеру и распределяя нагрузку для обработки запросов.
2. Высокая производительность с Big Data: Trino оптимизирован для выполнения аналитических запросов на больших данных. Он использует параллельную обработку запросов, что позволяет ускорить выполнение сложных запросов и улучшить общую производительность.
3. Гибкость и совместимость: Trino поддерживает стандарт SQL и может работать с различными источниками данных, такими как Hadoop HDFS, Amazon S3, Apache Kafka, MySQL, PostgreSQL и многими другими. Он также интегрируется с различными инструментами и платформами анализа данных.

Инструменты для построения витрины DWH. Trino

Недостатки Trino:
1. Сложность настройки: Настройка и управление Trino может быть сложным заданием, особенно для непрофессионалов. Он требует наличия квалифицированных специалистов для правильной настройки и оптимизации производительности.
2. Ограниченная поддержка административных функций: Trino сфокусирован на выполнении аналитических запросов и обработке данных, поэтому у него может быть ограниченная поддержка административных функций, таких как мониторинг, резервное копирование и восстановление данных. Вам может понадобиться дополнительные инструменты или настройки для этих задач.
3. Отсутствие встроенной системы управления ресурсами: Trino не имеет встроенной системы управления ресурсами или планировщика. Это означает, что вы должны использовать сторонние инструменты или настройки для эффективного распределения ресурсов между запросами и контроля за производительностью кластера.
4. Зависимость от сторонних инструментов и платформ: Trino интегрируется с различными инструментами и платформами анализа данных, но его функциональность может зависеть от этих сторонних компонентов. Это может создать сложности в управлении и обновлении всей экосистемы, особенно при использовании новых версий или дополнительных интеграций.
5. Не подходит для транзакционных операций: Trino не предназначен для выполнения транзакционных операций, таких как вставка, обновление и удаление данных. Если вам требуется обработка транзакций, вам следует рассмотреть другие системы, специализирующиеся на этой области.

Инструменты для построения витрины DWH. Airflow

https://ru.wikipedia.org/wiki/Apache_Airflow

https://habr.com/ru/articles/512386/

https://hub.docker.com/r/apache/airflow

Преимущества Airflow:

- небольшой, но полноценный инструментарий создания процессов обработки данных и управления ими – 3 вида операторов (сенсоры, обработчики и трансферы), расписание запусков для каждой цепочки задач, логгирование сбоев;
- графический веб-интерфейсдля создания конвейеров данных (data pipeline), который обеспечивает относительно низкий порог входа в технологию, позволяя работать с Airflow не только инженеру данных (Data Engineer), но и аналитику, разработчику, администратору и DevOps-инженеру.
- Расширяемый REST API, который относительно легко интегрировать Airflow в существующий ИТ-ландшафт
- Программный код на Python, который считается относительно простым языком для освоения и профессиональным
- Наличие собственного репозитория метаданных на базе библиотеки SqlAlchemy, где хранятся состояния задач, DAG’ов, глобальные переменные и пр.
- Масштабируемость за счет модульной архитектуры и очереди сообщений для неограниченного числа DAG’ов

Инструменты для построения витрины DWH. Airflow

https://ru.wikipedia.org/wiki/Apache_Airflow

https://habr.com/ru/articles/512386/

https://hub.docker.com/r/apache/airflow

Ограничения Airflow:

- наличие неявных зависимостей при установке, например, дополнительные пакеты типа greenlet, gevent, cryptography
- большие накладные расходы (временная задержка 5-10 секунд) на постановку DAG'ов в очередь и приоритизацию задач при запуске;
- необходимость наличия свободного слота в пуле задач и рабочего экземпляра планировщика.
- пост-фактум оповещения о сбоях в конвейере данных, в частности, в интерфейсе Airflow логи появятся только после того, как задание, к примеру, Spark-job, отработано. Поэтому следить в режиме онлайн, как выполняется data pipeline, приходится из других мест,

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

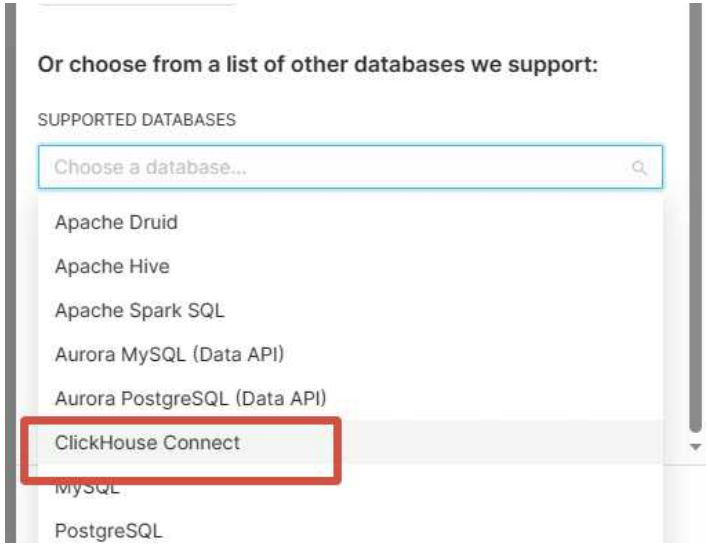
Визуализация данных Superset

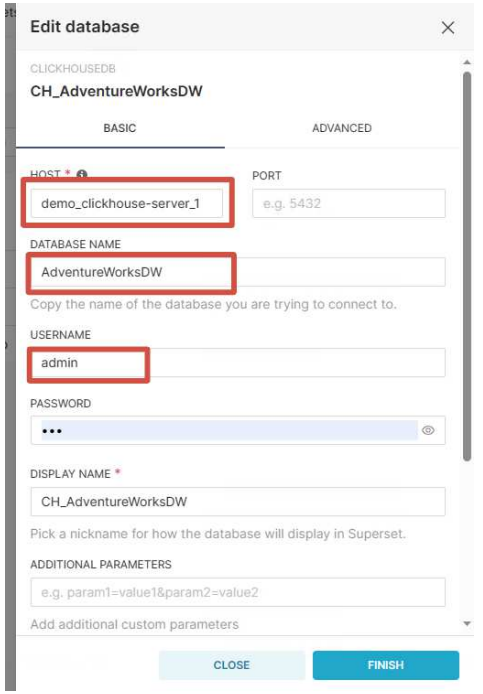
Apache Superset.Настройка для CH

- Установить драйвер перед **Settings->DataBase connection**

Superset использует драйвер clickhouse-connect для подключения к ClickHouse https://pypi.org/project/clickhouse-connect/ и устанавливается следующей командой : https://superset.apache.org/docs/databases/clickhouse/

pip install clickhouse-sqlalchemy - актуально для версии 2.0
pip install clickhouse-connect - актуально для версии 2.1.0



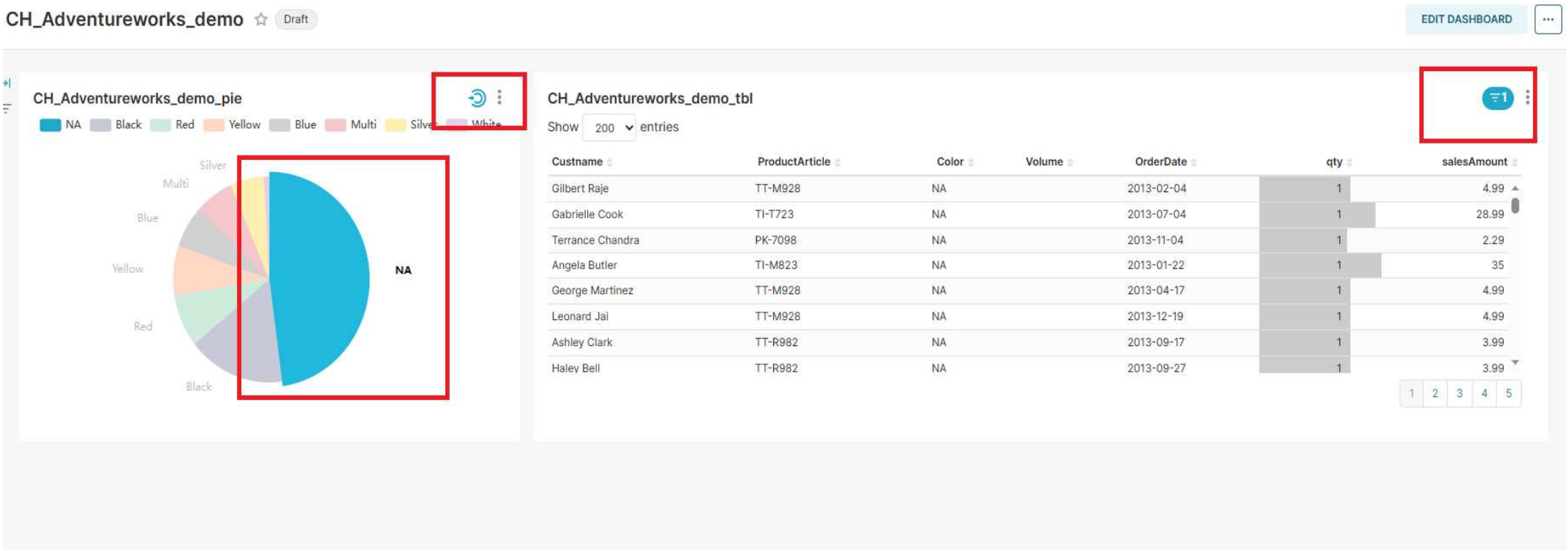


Apache Superset. Создание дашборда

1 Выполняем запрос, формируем dataset
2 Создаем chart (график)
3 Вставляем в dashboard

Apache Superset. Создание дашборда

Cross filter -> drill down



Apache Superset. Аналоги операций OLAP

Drill down позволяет детализировать агрегированный срез данных путем:

- Добавление нового фильтра без до настройки superset.
- Применение нового параметра для группировки без до настройки superset.

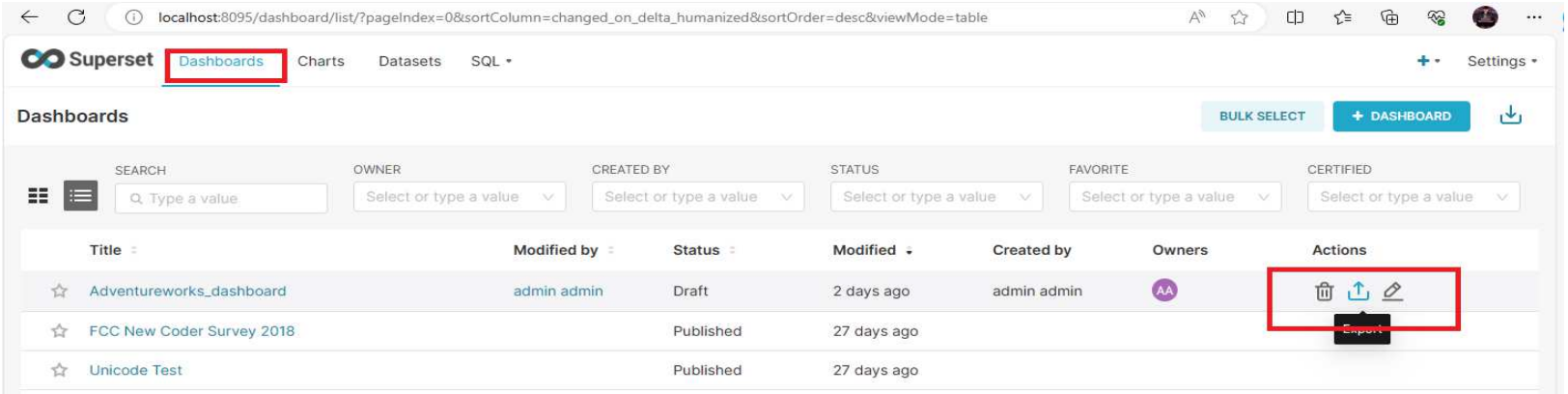
С до настройкой superset (флаг `DRILL_TO_DETAIL`) можно добавить контекстное меню.

Drill down поддерживает детализацию в *любом измерении*

https://preset.io/blog/drill-down-and-drill-by/

https://docs.preset.io/docs/drilling-to-chart-details

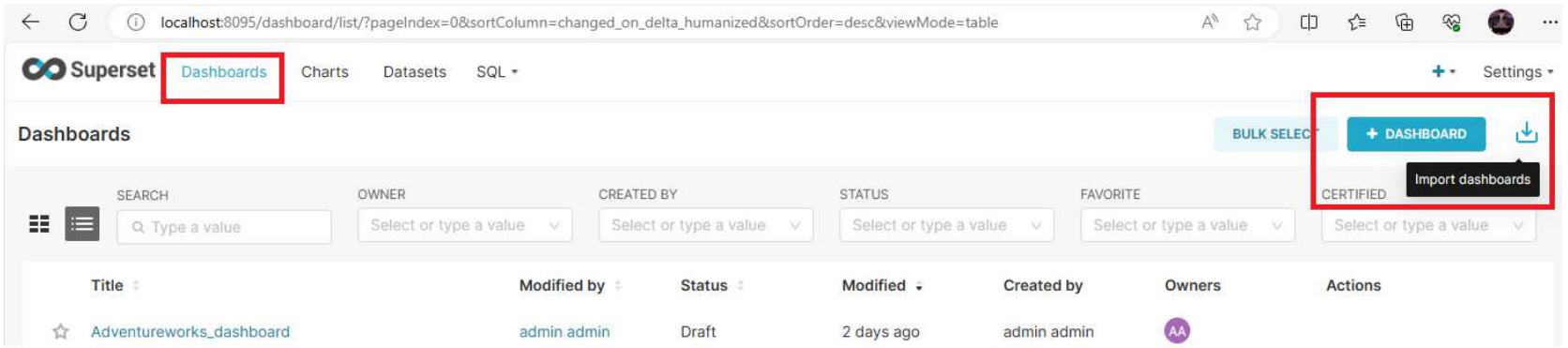
Apache Superset. Export




```
curl -H "Authorization: Bearer {API_KEY}" \
      -H "Content-Type: application/json" \
      -X GET {SUPERSET_URL}/api/v1/dashboard/{DASHBOARD_SLUG} \
       > {DASHBOARD_SLUG}.json
```

https://superset-bi.ru/apache-superset-api-export-and-import-dashboard/?ysclid=lm9fcpxc2z515126949

Apache Superset. Import




```
curl -H "Authorization: Bearer {API_KEY}" \
     -H "Content-Type: application/json" \
     -X POST {SUPERSET_URL}/api/v1/dashboard/import \
     -d @/{DASHBOARD_SLUG}.json
```

https://superset-bi.ru/apache-superset-api-export-and-import-dashboard/?ysclid=lm9fcpxc2z515126949

Apache Superset. Jinja динамические шаблоны и макросы

```
SELECT *
FROM tbl
WHERE
     dttm_col > '{{ from_dttm }}' AND dttm_col < '{{ to_dttm }}'
```

from_dttm: начальное значение даты и времени из выбранного диапазона (если не определено)

```
SELECT *
FROM jinja_username_demo
WHERE "username" = '{{ current_username() }}'
```

https://preset.io/blog/intro-jinja-templating-apache-superset/

https://superset-bi.ru/examples-of-using-jinja-templates-in-apache-superset/

Apache Superset. Настройка доступа

Roles

Managing Data Source Access

Customizing Permissions

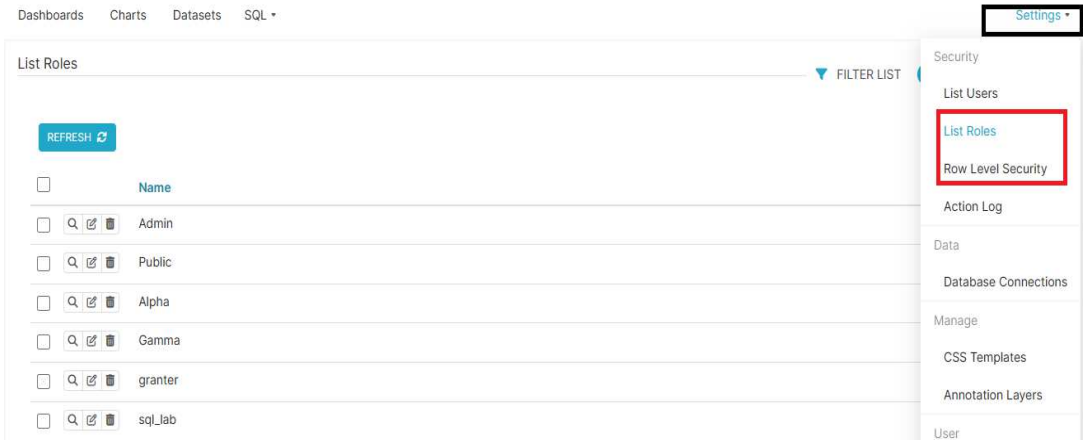
Restricting Access to a Subset of Data Sources

Row Level Security

Content Security Policy (CSP)

https://superset.apache.org/docs/security/

Apache Superset. Настройка доступа



- Admin – администраторы, имеют все возможные права, включая предоставление или отзыв прав других пользователей и изменение чужих фрагментов и информационных панелей;

- Alfa - пользователи имеют доступ ко всем источникам данных, могут их добавлять и изменять, но не могут предоставлять или отзывать доступ другим пользователям, ограничены в изменении объектов, которыми не владеют;

- Gamma - пользователи могут потреблять данные, поступающие из источников данных, к которым им предоставлен доступ через другую дополнительную роль.

https://bilab.ru/sozdanie-uzerov-s-ogranichennimi-pravami-v-apache-superset?ysclid=lm9fkh8nbs199296432

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

LIVE

Apache superset.
Демонстрация создания дашборда

Demo.

Создаем тестовую базу на Clickhouse

https://github.com/topics/adventureworksdw

Используем запрос

```
select cust.FullName AS Custname, prod.ProductAlternateKey AS ProductArticle,
prod.Color AS Color, prod.`Size` as Volume,
fact.OrderDate AS OrderDate,
SUM(fact.OrderQuantity) as qty, SUM(fact.SalesAmount) AS salesAmount
from AdventureWorksDW.FactInternetSales fact
LEFT SEMI JOIN AdventureWorksDW.DimCustomer cust using(CustomerKey)
LEFT SEMI JOIN AdventureWorksDW.DimProduct prod on prod.ProductKey =
fact.ProductKey
GROUP BY fact.CustomerKey, fact.ProductKey, fact.OrderDate,
cust.FullName, prod.ProductAlternateKey, prod.Color, prod.`Size`
```

Рефлексия

Ключевые тезисы занятия

Подведем итоги

1.
2.
3.
4.

Рефлексия

С какими впечатлениями уходите с вебинара?

Как будете применять на практике то,
что узнали на вебинаре?

Заполните, пожалуйста, опрос о занятии по ссылке в чате