Дипломный проект

по профессии Инженер данных

Задание

Данные для выполнения дипломного задания - https://www.kaggle.com/aungpyaeap/supermarket-sales?select=supermarket_sales+-+Sheet1.csv

1. Вам необходимо разработать и задокументировать ETL-процессы заливки данных в хранилище, состоящее из слоёв:
   - NDS - нормализованное хранилище и DDS - схема звезда;
   - Data Quality - опционально, будет большим преимуществом в вашей работе;
2. на основании DDS построить в Табло дашборды

Рекомендации при выполнении работы:

1. ETL процессы можно делать:
   - с помощью Pentaho;
   - с помощью Python (pandas) + SQL;
2. датасет:
   - предложен вам в CSV формате выше;
   - сбор данных вы также можете сделать из сторонних API, это станет вашим преимуществом;
3. дополнительно вы можете сделать оркестровку с помощью Airflow;
4. опционально можно сделать отдельный слой метаданных в хранилище, а также дашборды на основании данных из этого слоя, где будет отображаться кол-во прогрузок и их статусы;

Результат:

- дашборды
- задокументированная схема хранилища данных
- документированная схема ETL-процессов

Введение

Для выполнения дипломного проекта был использован следующий технологический стек:

- Docker
- PostgreSQL
- DBeaver
- Apache Spark
- Scala
- Metabase
- Graphana

Анализ исходных данных

Предложенный датасет (https://www.kaggle.com/aungpyaeap/supermarket-sales?select=supermarket_sales+-+Sheet1.csv) содержит информацию по продажам в 3х филиалах супермаркета Мьянмы.

Атрибуты датасета:

- Invoice id: Номер счета, сгенерированный компьютером
- Branch: Филиал (доступно 3 филиала, которые идентифицированны как A, B и C).
- City: Местоположение филиала
- Customer type: Тип клиента (с картой магазина или без)
- Gender: Пол клиента
- Product line: Категория товара
- Unit price: цена каждого товара в $
- Quantity: Количество товаров, купленных клиентом
- Tax: 5% налог на покупки
- Total: Итоговая цена, включая налог
- Date: Дата продажи
- Time: Время продажи

- Payment: Тип оплаты
- COGS: Себестоимость проданных товаров
- Gross margin percentage: Валовая рентабельность в %
- Gross income: Валовой доход
- Rating: рейтинг пользователя

При начально анализе данных датасета было замечено, что значения COGS, Gross margin percentage, Gross income не соответствуют описанию.

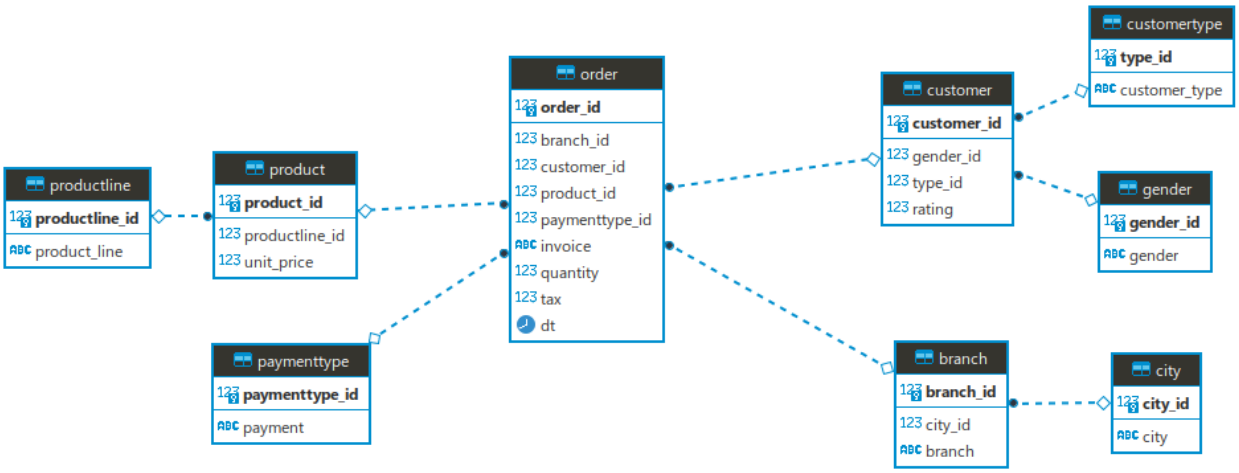
В столбце COGS находятся значения итоговой цены товара исключая налог, Unit price * Quantity

В столбце Gross income находятся значения налога на покупку, Total - (Unit price * Quantity)

В столбце Gross margin percentage значения 100*(Total - (Unit price * Quantity)) / Total

Нормализованное хранилище (NDS)

ER-диаграмма



Описание БД

таблица nds.city

содержит данные по городам

Столбец	Тип	Модификаторы	Описание
city_id	int	NOT NULL	Суррогатный ключ
city	varchar(250)	NOT NULL	Местоположение филиала

таблица nds.gender

содержит данные по половой принадлежности клиентов

Столбец	Тип	Модификаторы	Описание
gender_id	int	NOT NULL	Суррогатный ключ
gender	varchar(250)	NOT NULL	Пол

таблица nds.customertype

содержит данные по типам клиентов

Столбец	Тип	Модификаторы	Описание
customertype_id	int	NOT NULL	суррогатный ключ
customertype	varchar(250)	NOT NULL	тип клиента

таблица nds.productline

содержит данные по категориям товаров

Столбец	Тип	Модификаторы	Описание
productline_id	int	NOT NULL	суррогатный ключ
productline	varchar(250)	NOT NULL	категория товара

таблица nds.paymenttype

содержит данные по типам платежей

Столбец	Тип	Модификаторы	Описание
paymenttype_id	int	NOT NULL	суррогатный ключ
paymenttype	varchar(250)	NOT NULL	метод оплаты

таблица nds.branch

содержит данные о филиалах

Столбец	Тип	Модификаторы	Описание
branch_id	int	NOT NULL	суррогатный ключ
city_id	int	NOT NULL	внешний ключ
branch	varchar(250)	NOT NULL	филиал

таблица nds.product

содержит данные о товарах

Столбец	Тип	Модификаторы	Описание
product_id	int	NOT NULL	суррогатный ключ
productline_id	int	NOT NULL	внешний ключ
unitprice	decimal(20,2)	NOT NULL	цена за еденицу товара

таблица nds.customer

содержит данные о клиентах

Столбец	Тип	Модификаторы	Описание
customer_id	int	NOT NULL	суррогатный ключ
gender_id	int	NOT NULL	внешний ключ
customertype_id	int	NOT NULL	внешний ключ
rating	numeric(3,1)	NOT NULL	рейтинг

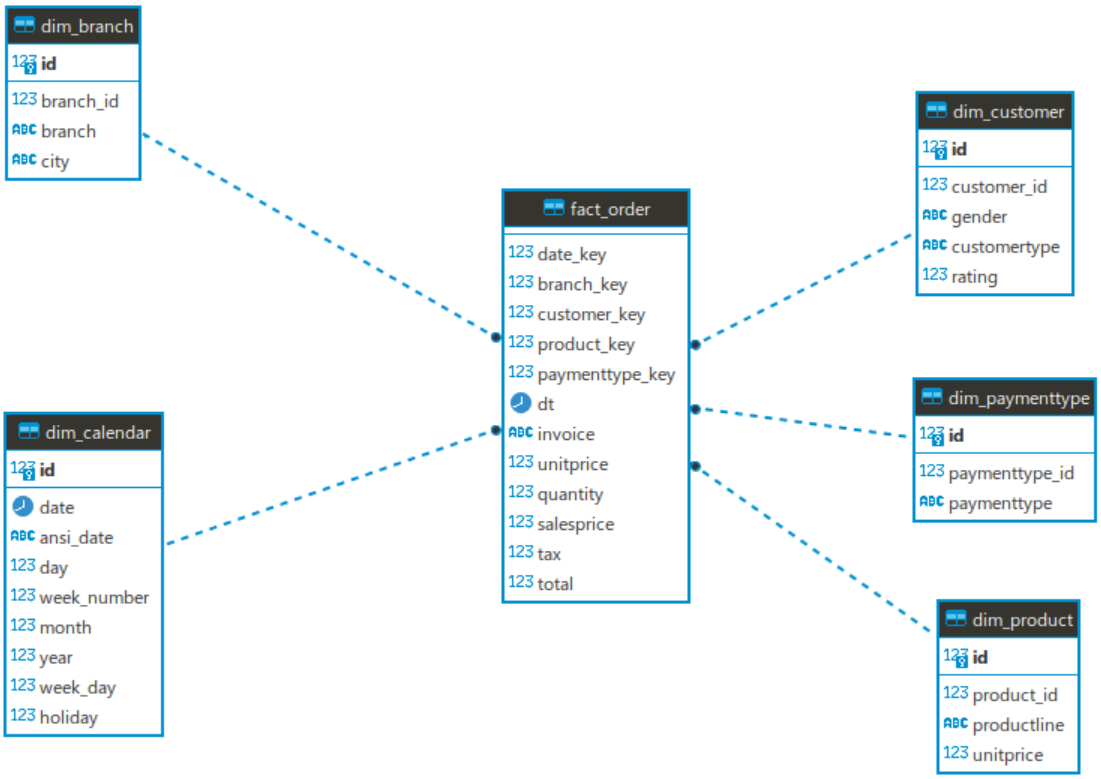
таблица nds.order

содержит данные о заказах

Столбец	Тип	Модификаторы	Описание
order_id	int	NOT NULL	суррогатный ключ
branch_id	int	NOT NULL	внешний ключ
customer_id	int	NOT NULL	внешний ключ
product_id	int	NOT NULL	внешний ключ
paymenttype_id	int	NOT NULL	внешний ключ
invoice	varchar(250)	NOT NULL	номер счета
quantity	int	NOT NULL	количество товаров
tax	decimal(20,4)	NOT NULL	налог на покупки
dt	timestamp	NOT NULL	дата и время продажи

Многомерное хранилище (DDS)

ER-диаграмма



Описание БД

таблица dds.dim_calendar

таблица измерения “дата”

Столбец	Тип	Модификаторы	Описание
id	int	NOT NULL	суррогатный ключ
dt	date	NOT NULL	дата
ansi_date	text	NOT NULL	дата в формате ansi
day	int	NOT NULL	номер дня в году
week_number	int	NOT NULL	номер недели в году
month	int	NOT NULL	номер месяца в году
year	int	NOT NULL	год
work_day	int	NOT NULL	рабочий день
weekend	int	NOT NULL	выходной день
holiday	int	NOT NULL	официальный праздник

таблица dds.dim_branch

таблица измерения “филиал”

Столбец	Тип	Модификаторы	Описание
id	serial	NOT NULL	суррогатный ключ
branch_id	date	NOT NULL	ключ из nds.branch
branch	varchar(250	NOT NULL	филиал
city	varchar(250	NOT NULL	местоположение филиала

таблица dds.dim_customer

таблица измерения “покупатель”

Столбец	Тип	Модификаторы	Описание
id	serial	NOT NULL	суррогатный ключ
customer_id	int	NOT NULL	ключ из nds.customer
gender	varchar(250)	NOT NULL	пол
customertype	varchar(250)	NOT NULL	тип клиента
rating	numeric(3,1)	NOT NULL	рейтинг

таблица dds.dim_product

таблица измерения “товар”

Столбец	Тип	Модификаторы	Описание
id	serial	NOT NULL	суррогатный ключ
productline_id	int	NOT NULL	ключ из nds.product
productline	varchar(250)	NOT NULL	категория товара
unitprice	decimal(20,2)	NOT NULL	цена за еденицу товара

таблица dds.dim_paymenttype

таблица измерения “тип платежа”

Столбец	Тип	Модификаторы	Описание
id	serial	NOT NULL	суррогатный ключ
paymenttype_id	int	NOT NULL	ключ из nds.paymenttype
paymenttype	varchar(250)	NOT NULL	метод оплаты

таблица dds.fact_order

таблица фактов “покупки”

Столбец	Тип	Модификаторы	Описание
date_key	int	NOT NULL	внешний ключ
branch_key	int	NOT NULL	внешний ключ
customer_key	int	NOT NULL	внешний ключ
product_key	int	NOT NULL	внешний ключ
paymenttype_key	int	NOT NULL	внешний ключ
dt	timestamp	NOT NULL	дата и время продажи
invoice	varchar(250)	NOT NULL	номер счета
unitprice	decimal(20,2)	NOT NULL	цена за еденицу товара
quantity	int	NOT NULL	количество товаров
salesprice	decimal(20,2	NOT NULL	сумма продажи без налога
tax	decimal(20,4)	NOT NULL	налог на покупки
total	decimal(20,4)	NOT NULL	сумма продажи включая налог

Rejected -таблицы

Таблицы rejected.customer, rejected.product, rejected.order хранят некачественные строки.

Описание процедуры ETL

1. Нормализованное хранилище

Считываем данные из файла supermarket_sales.csv. Удаляем столбцы, которые не будем использовать для нормализованного хранилища. Строки с NULL-значениями записываем в файл ndsnullvalues.csv. Оставшиеся записи распределяем по датафреймам, добавляя к каждой первичный ключ. Записываем датафреймы: каждый в свою таблицу БД.

2. Многомерное хранилище (DDS)

dds.Dim_Branch - справочник филиалов

Считываем данные из таблиц nds.city и nds.branch, объединяем и записываем в таблицу dds.Dim_Branch.

Проверки значений полей на null не требуются, так как в исходных данных такое ограничение уже заложено.

dds.Dim_Customer - справочник покупателей

Считываем данные из таблиц nds.customer, nds.gender, nds.customertype. Объединяем данные и после проверки качества данных записываем в таблицу dds.Dim_Customer

Проверки качества данных:

rating	Значения должны быть в диапазоне от 1 до 10

Проверки значений полей на null не требуются, так как в исходных данных такое ограничение уже заложено.

Данные не прошедшие проверку записываются в таблицу rejected.customer

dds.Dim_Product - справочник товаров

Считываем данные из таблиц nds.product, nds.productline. Объединяем данные и после проверки качества данных записываем в таблицу dds.Dim_Product.

Проверки качества данных:

unitprice	Значение должно быть больше 0

Проверки значений полей на null не требуются, так как в исходных данных такое ограничение уже заложено.

Данные не прошедшие проверку записываются в таблицу rejected.product

dds.Dim_Paymenttype - справочник типов платежей

Считываем данные из таблицы nds.paymenttype и записываем в таблицу dds.Dim_Paymenttype.

Проверки значений полей на null не требуются, так как в исходных данных такое ограничение уже заложено.

dds.Fact_Order - покупки

Считываем данные из таблицы nds.order, присоединяем суррогатные ключи всех таблиц измерений. Считаем значения salesprice (сумма продажи без налога) и total (сумма продажи включая налог).

После проверки качества данных записываем в таблицу dds.Fact_Order.

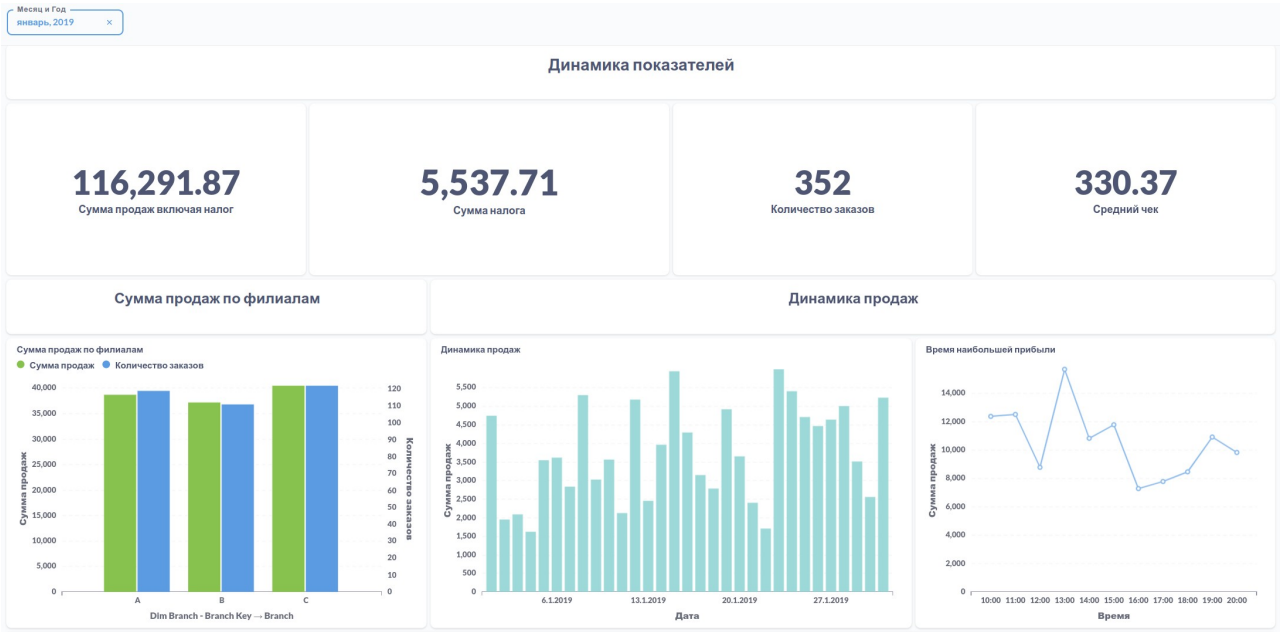
Проверки качества данных:

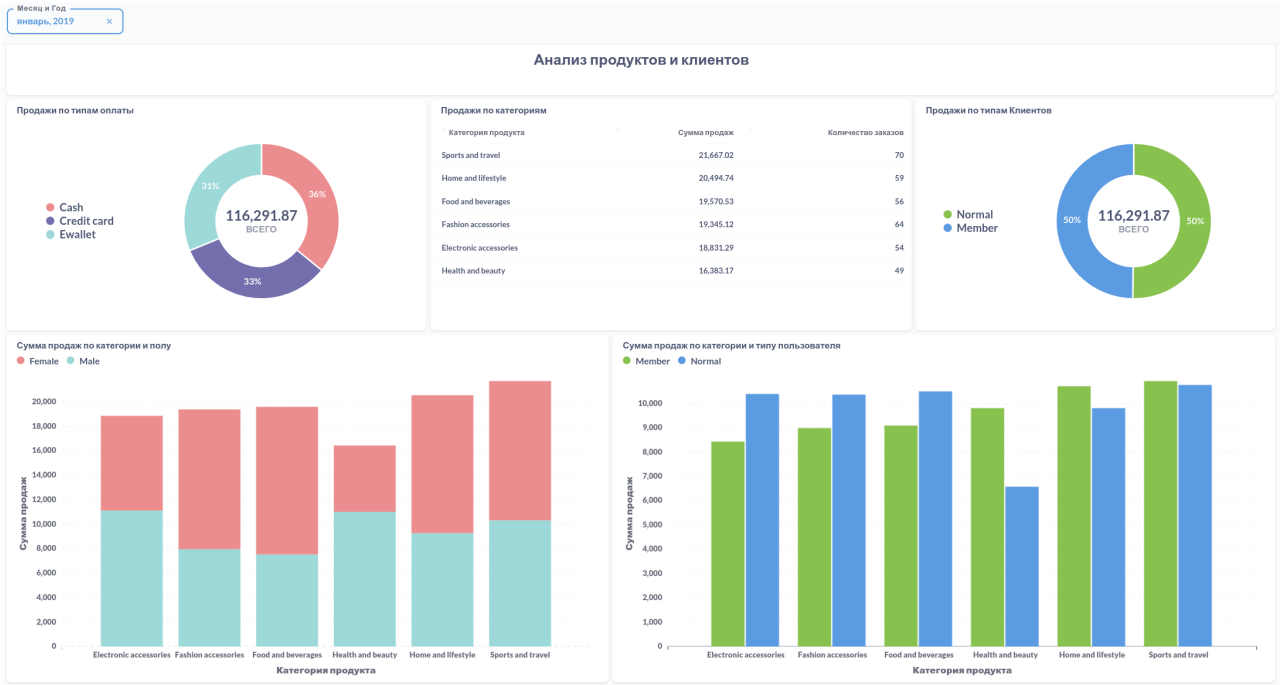
dt	не должно быть больше текущей даты
invoice	Значение должно быть в формате ###-##-####, где x — число от 0 до 9
quantity	Значение должно быть больше 0
tax	Значение должно быть больше 0 и не должно быть больше salesprice

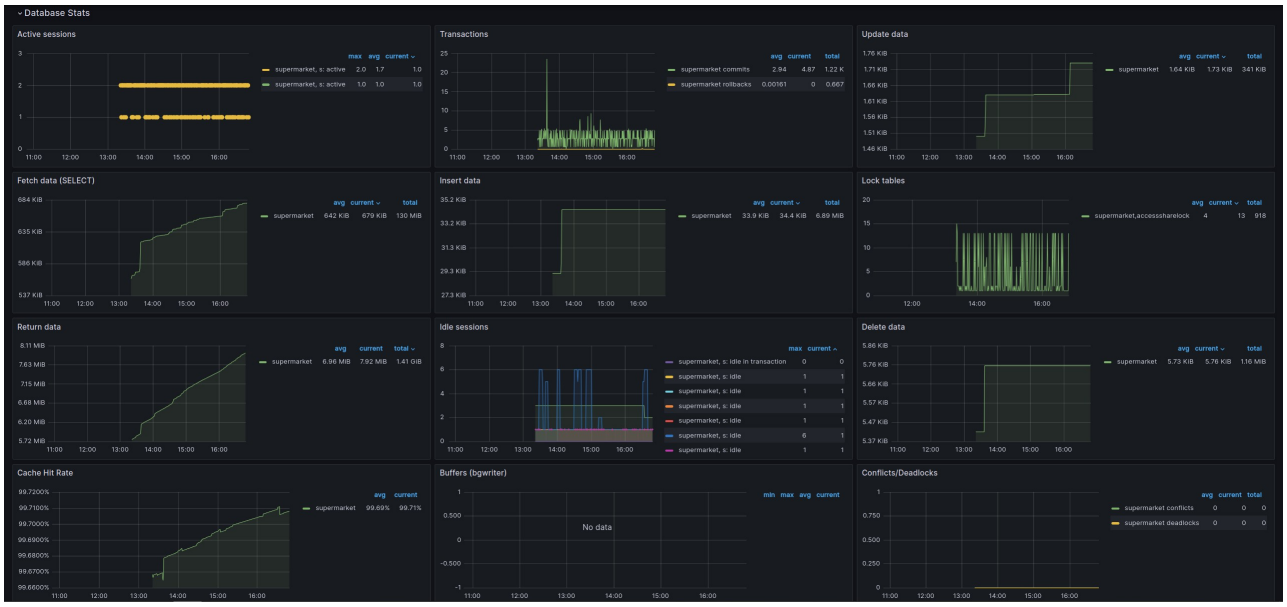
Проверки значений полей на null не требуются, так как в исходных данных такое ограничение уже заложено.

Данные не прошедшие проверку записываются в таблицу rejected.order

Дашборды


Месяц и Год
январь, 2019
Динамика показателей
116,291.87
Сумма продаж включая налог
5,537.71
Сумма налога
352
Количество заказов
330.37
Средний чек
Сумма продаж по филиалам
Динамика продаж
Сумма продаж по филиалам
Сумма продаж
Количество заказов
Сумма продаж
Количество заказов
A
B
C
Dim Branch - Branch Key → Branch
Динамика продаж
Сумма продаж
6.1.2019
13.1.2019
20.1.2019
27.1.2019
Дата
Время наибольшей прибыли
Сумма продаж
Время


Месяц и Год
январь, 2019
Анализ продуктов и клиентов
Продажи по типам оплаты
Cash
Credit card
Ewallet
36%
33%
31%
116,291.87
ВСЕГО
Продажи по категориям
Категория продукта
Сумма продаж
Количество заказов
Sports and travel
21,667.02
70
Home and lifestyle
20,494.74
59
Food and beverages
19,570.53
56
Fashion accessories
19,345.12
64
Electronic accessories
18,831.29
54
Health and beauty
16,383.17
49
Продажи по типам Клиентов
Normal
Member
50%
50%
116,291.87
ВСЕГО
Сумма продаж по категории и полу
Female
Male
Сумма продаж
Electronic accessories
Fashion accessories
Food and beverages
Health and beauty
Home and lifestyle
Sports and travel
Категория продукта
Сумма продаж по категории и типу пользователя
Member
Normal
Сумма продаж
Electronic accessories
Fashion accessories
Food and beverages
Health and beauty
Home and lifestyle
Sports and travel
Категория продукта


Database Stats
Active sessions
max avg current
supermarket, s: active 2.0 1.7 1.0
supermarket, s: active 1.0 1.0 1.0
Transactions
avg current total
supermarket commits 2.94 4.87 1.22 K
supermarket rollbacks 0.00161 0 0.667
Update data
avg current total
supermarket 1.64 KiB 1.73 KiB 341 KiB
Fetch data (SELECT)
avg current total
supermarket 642 KiB 679 KiB 130 MiB
Insert data
avg current total
supermarket 33.9 KiB 34.4 KiB 6.89 MiB
Lock tables
avg current total
supermarket,accesssharelock 4 13 918
Return data
avg current total
supermarket 6.96 MiB 7.92 MiB 1.41 GiB
Idle sessions
max current
supermarket, s: idle in transaction 0 0
supermarket, s: idle 1 1
supermarket, s: idle 1 1
supermarket, s: idle 1 1
supermarket, s: idle 1 1
supermarket, s: idle 6 1
supermarket, s: idle 1 1
Delete data
avg current total
supermarket 5.73 KiB 5.76 KiB 1.16 MiB
Cache Hit Rate
avg current
supermarket 99.69% 99.71%
Buffers (bgwriter)
min max avg current
No data
Conflicts/Deadlocks
avg current total
supermarket conflicts 0 0 0
supermarket deadlocks 0 0 0