МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»

РТУ МИРЭА

Институт информационных технологий (ИИТ)
Кафедра прикладной математики (ПМ)

ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ

по дисциплине «Технологии и инструментарий анализа больших данных»

Практическая работа № 8

Студент группы *ИКБО-01-22 Прокопчук Роман Олегович* _____
(подпись)

Ассистент *Тетерин Николай Николаевич* _____
(подпись)

Отчёт представлен «__» декабря 2025 г.

Москва, 2025 г.

1 ЦЕЛЬ

Ознакомиться с возможностями работы Arenadata Cluster Manager совместно с бандлами для кластера и мониторинга.

2 ЗАДАНИЕ

1. Настроить виртуальную машину Ubuntu Server 22.04.5 LTS в системе виртуализации операционных систем, а также настроить Docker контейнеры PostgreSQL, Arenadata Cluster Manager совместно с бандлами для кластера и мониторинга (Prometheus, Grafana). **(1 балл)** В отчет прикрепить скриншот с интерфейса ADCM на странице Cluster->Overview где показано, что все сервисы подняты и работают в штатном режиме.
2. Создать базу данных ecommerce и развернуть public схему таблиц, на основе приложенного скрипта, для хранения данных транзакций в PostgreSQL посредством postgresql-client (DBeaver или любой другой админ баз данных). Для каждой таблицы сделать импорт данных посредством импорта из CSV-файла в следующем порядке **(1 балл):**
    - geolocation – geolocation.csv
    - leads_closed – leads_closed.csv
    - product_category_name_translation – product_category_name_translation.csv
    - sellers - sellers.csv
    - customers – customers.csv
    - leads_closed – leads_closed.csv
    - products – products.csv
    - orders – orders.csv
    - order_items – order_items.csv
    - order_reviews – order_reviews.csv
    - order_payments – order_payments.csv

    В отчет прикрепить визуализацию схемы таблиц, полученную в консоли командной строки или в виде диаграммы в админе баз данных (например DBeaver).
3. Написать запросы на языке SQL для базы данных ecommerce для получения данных по заданиям.

4. Построить дашборды в Grafana для визуализации данных из ecommerce на основе задач.

3 НАСТРОЙКА ВИРТУАЛЬНОЙ МАШИНЫ И БД

3.1 Настройка виртуальной машины

Интерфейс ADCM на странице Overview представлен на рисунке 1.

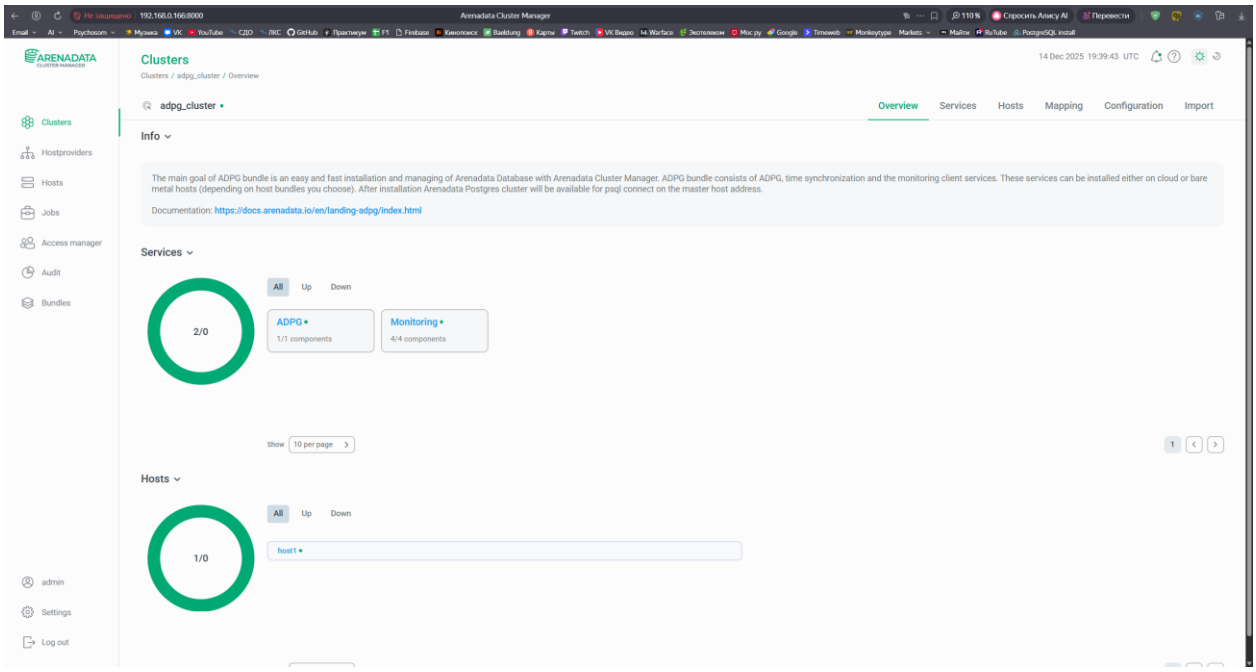

Рисунок 1 – Overview кластера в интерфейсе ADCM

3.2 Создание базы данных

Схема созданной базы данных представлена на рисунке 2.

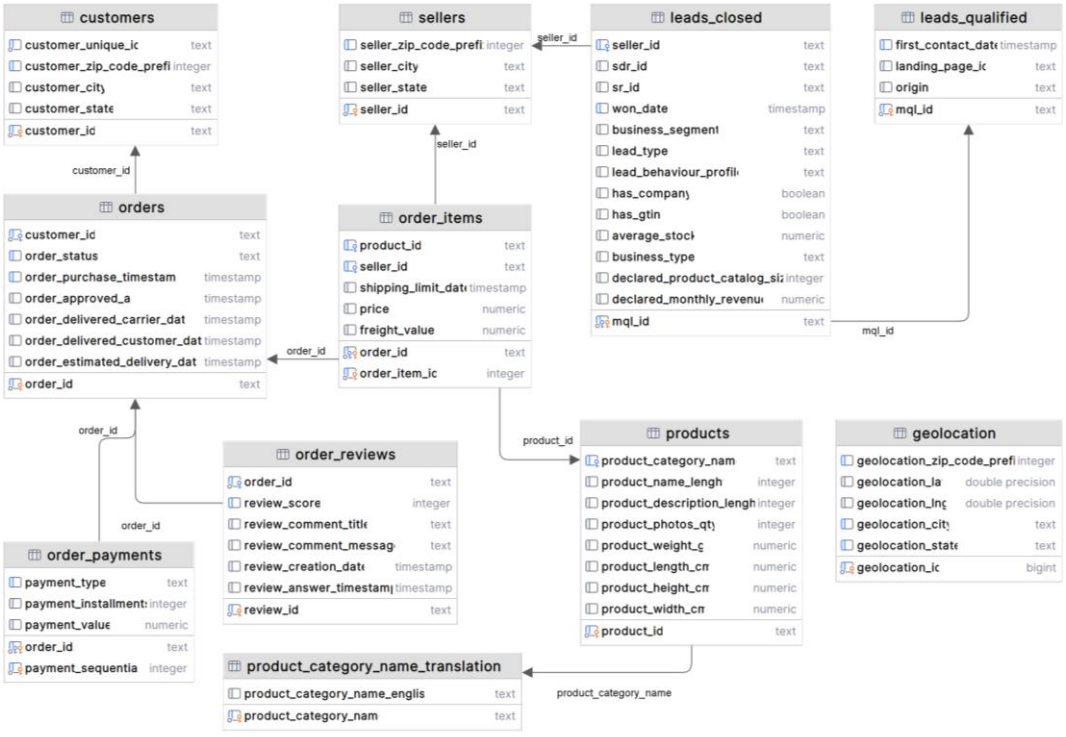

Рисунок 2 – Схема таблиц созданной БД

4 ВЫПОЛНЕНИЕ ЗАДАНИЙ

4.3 SQL запросы

4.3.1 ТОП-10 категорий по выручке, с переводом

Нужно определить десять товарных категорий с наибольшим оборотом продаж. Для каждой категории требуется посчитать сумму цен всех проданных позиций (это и есть выручка/GMV категории), количество уникальных заказов, в которых встречались товары этой категории, и общее число товарных позиций. Дополнительно к исходному названию категории нужно вывести её англоязычный перевод из словаря переводов. Итоговый набор данных должен быть отсортирован по выручке по убыванию и ограничен первыми десятью строками.

Разработанный для данного задания SQL-запрос представлен в листинге 1. Результат его работы – в таблице 1.

Листинг 1 – SQL-запрос для поиска топ-10 категорий по выручке

```
SELECT t.product_category_name_english AS category_name_en,
       SUM(oi.price)                   AS revenue,
       COUNT(DISTINCT oi.order_id)     AS unique_orders_count,
       COUNT(oi.product_id)            AS total_items_count
FROM order_items oi
         JOIN products p ON oi.product_id = p.product_id
         JOIN product_category_name_translation t ON
p.product_category_name = t.product_category_name
GROUP BY t.product_category_name_english
ORDER BY revenue DESC
LIMIT 10;
```

Таблица 1 – Топ-10 категорий по выручке

category_name_en	revenue	unique_orders_count	total_items_count
health_beauty	1258681.34	8836	9670
watches_gifts	1205005.68	5624	5991
bed_bath_table	1036988.68	9417	11115
sports_leisure	988048.97	7720	8641
computers_accessories	911954.32	6689	7827
furniture_decor	729762.49	6449	8334
cool_stuff	635290.85	3632	3796
housewares	632248.66	5884	6964
auto	592720.11	3897	4235
garden_tools	485256.46	3518	4347

4.3.2 Платежное поведение

Нужно сравнить используемые способы оплаты. По каждому типу оплаты требуется вывести количество зарегистрированных платежей, среднее число рассрочек (если применимо), средний размер платежа и долю заказов, в которых встречается данный способ оплаты.

Под долей заказов понимается отношение числа уникальных заказов, где зафиксирован хотя бы один платёж данного типа, к общему числу уникальных заказов, по которым известны записи об оплате.

Следует учитывать, что у одного заказа может быть несколько платежей. Доля считается на уровне «есть/нет» для типа оплаты в заказе.

Разработанный для данного задания SQL-запрос представлен в листинге 2. Результат его работы – в таблице 2.

Листинг 2 – SQL-запрос для определения платежного поведения покупателей

```
WITH total_unique_orders AS (SELECT COUNT(DISTINCT order_id) as total_cnt
                             FROM order_payments)
SELECT payment_type,
       COUNT(*)                                AS payments_count,
       ROUND(AVG(payment_installments), 2) AS avg_installments,
       ROUND(AVG(payment_value), 2)        AS avg_payment_value,
       ROUND(
               (COUNT(DISTINCT order_id)::numeric / (SELECT total_cnt FROM
total_unique_orders)) * 100,
               2
       )                                       AS order_share_percent
FROM order_payments
GROUP BY payment_type
ORDER BY payments_count DESC;
```

Таблица 2 – Метрики по типу оплаты

payment_type	payments_count	avg_installments	avg_payment_value	order_share_percent
credit_card	76795	3.51	163.32	76.94
boleto	19784	1	145.03	19.9
voucher	5775	1	65.7	3.89
debit_card	1529	1	142.57	1.54
not_defined	3	1	0	0

4.3.3 Зависимость оценки от скорости доставки

Нужно изучить, как длительность доставки коррелирует с оценками в отзывах. Для каждого заказа с отзывом и известной фактической датой

доставки рассчитывается длительность доставки в днях как разница между датой доставки клиенту и датой покупки.

Далее каждый заказ попадает в один из трёх интервалов длительности:

– до пяти дней включительно,

– от шести до десяти дней включительно,

– более десяти дней.

Для каждого интервала требуется посчитать количество заказов и среднюю оценку из отзывов. Итог должен содержать три строки — по одному агрегату на интервал — в логическом порядке от самого быстрого интервала к самому длительному.

Разработанный для данного задания SQL-запрос представлен в листинге 3. Результат его работы – в таблице 3.

Листинг 3 – SQL-запрос для определения зависимости оценки от доставки

```
WITH delivery_data AS (SELECT o.order_id,
                              CASE
                                  WHEN EXTRACT(DAY FROM
(o.order_delivered_customer_date - o.order_purchase_timestamp)) <= 5
                                      THEN '0-5 days'
                                  WHEN EXTRACT(DAY FROM
(o.order_delivered_customer_date - o.order_purchase_timestamp)) BETWEEN 6
AND 10
                                      THEN '6-10 days'
                                  ELSE '> 10 days'
                                  END AS delivery_interval
                       FROM orders o
                       WHERE o.order_delivered_customer_date IS NOT NULL
                         AND o.order_purchase_timestamp IS NOT NULL)
SELECT dd.delivery_interval,
       COUNT(DISTINCT dd.order_id)   AS orders_count,
       ROUND(AVG(r.review_score), 2) AS avg_review_score
FROM delivery_data dd
         JOIN order_reviews r ON dd.order_id = r.order_id
GROUP BY dd.delivery_interval
ORDER BY CASE dd.delivery_interval
             WHEN '0-5 days' THEN 1
             WHEN '6-10 days' THEN 2
             ELSE 3
             END;
```

Таблица 3 – Зависимость оценки покупателя от времени доставки

delivery_interval	**orders_count**	**avg_review_score**
0-5 days	18993	4.43
6-10 days	32378	4.35
> 10 days	43409	3.9

4.3.4 Эффективность продавцов

Нужно определить десять товарных категорий с наибольшим оборотом продаж. Для каждой категории требуется посчитать сумму цен всех проданных позиций (это и есть выручка/GMV категории), количество уникальных заказов, в которых встречались товары этой категории, и общее число товарных позиций. Дополнительно к исходному названию категории нужно вывести её англоязычный перевод из словаря переводов. Итоговый набор данных должен быть отсортирован по выручке по убыванию и ограничен первыми десятью строками.

Разработанный для данного задания SQL-запрос представлен в листинге 4. Результат его работы – в таблице 4.

Листинг 4 – SQL-запрос для определения самых эффективных продавцов

```
SELECT s.seller_id,
       COUNT(DISTINCT oi.order_id)          AS unique_orders,
       COUNT(DISTINCT c.customer_unique_id) AS unique_customers,
       COUNT(oi.product_id)                 AS items_sold,
       SUM(oi.price)                        AS total_revenue,
       ROUND(AVG(oi.freight_value), 2)      AS avg_freight_cost
FROM sellers s
         JOIN order_items oi ON s.seller_id = oi.seller_id
         JOIN orders o ON oi.order_id = o.order_id
         JOIN customers c ON o.customer_id = c.customer_id
GROUP BY s.seller_id
ORDER BY unique_orders DESC
LIMIT 10;
```

Таблица 4 – Топ-10 самых эффективных продавцов

seller_id	unique_orders	unique_customers	items_sold	total_revenue	avg_freight_cost
6560211a19b47992c3666cc44a7e94c0	1854	1824	2033	123304.83	13.75
4a3ca9315b744ce9f8e9374361493884	1806	1792	1987	200472.92	17.65
cc419e0650a3c5ba77189a1882b7556a	1706	1657	1775	104288.42	14.46
1f50f920176fa81dab994f9023523100	1404	1388	1931	106939.21	18.21
da8622b14eb17ae2831f4ac5b9dab84a	1314	1275	1551	160236.57	16.09
955fee9216a65b617aa5c0531780ce60	1287	1282	1499	135171.7	16.97
7a67c85e85bb2ce8582c35f2203ad736	1160	1155	1171	141745.53	17.85
ea8482cd71df3c1969d7b9473ff13abc	1146	1133	1203	37177.52	14.58
4869f7a5dfa277a7dca6462dcf3b52b2	1132	1124	1156	229472.63	17.45
3d871de0142ce09b7081e2b9d1733cb1	1080	1074	1147	94914.2	19.56

4.4 Визуализации в Grafana

4.4.1 Динамика числа заказов

Нужно показать динамику общего числа заказов по месяцам. Для каждого месяца в хронологическом порядке считается количество уникальных заказов, оформленных в этот месяц. Полученный временной ряд должен быть представлен линией, позволяющей увидеть рост, падения или сезонность количества заказов.

Созданная визуализация представлена на рисунке 3.

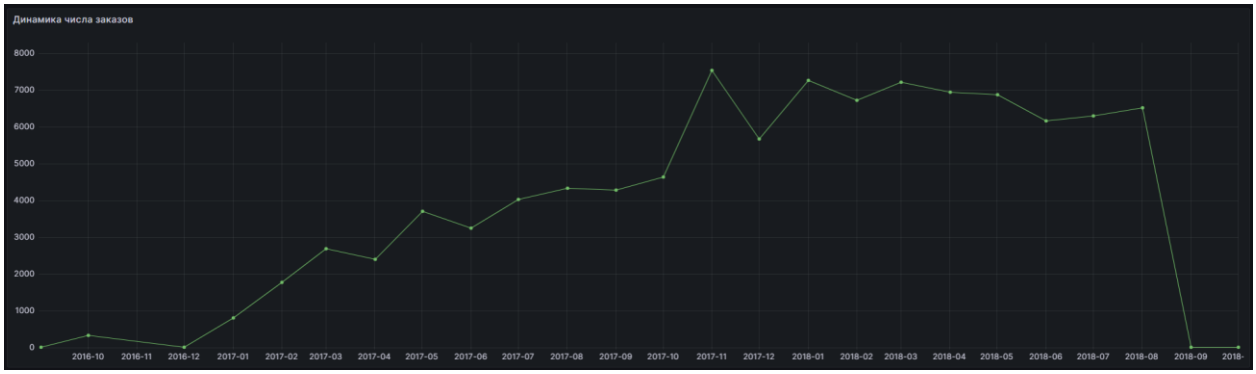

Рисунок 3 – Визуализация динамики числа заказов

График демонстрирует устойчивый тренд роста количества заказов начиная с января 2017 года. Ярко выраженный пик наблюдается в ноябре 2017 года, что, вероятнее всего, связано с сезонными распродажами («Черная пятница»). После пика следует небольшая коррекция, но общий уровень заказов в 2018 году остается стабильно выше показателей начала 2017 года. Резкий спад в конце графика (сентябрь-октябрь 2018) указывает на неполноту данных за последние месяцы выборки.

4.4.2 Количество заказов по оценкам

Нужно построить распределение количества заказов по оценкам клиентов. Для этого все отзывы группируются по выставленной оценке, и для каждой категории оценки подсчитывается количество заказов, получивших такую оценку. Результат следует отобразить в виде отдельных вертикальных столбцов, показывающих частоту каждой оценки.

Созданная визуализация представлена на рисунке 4.

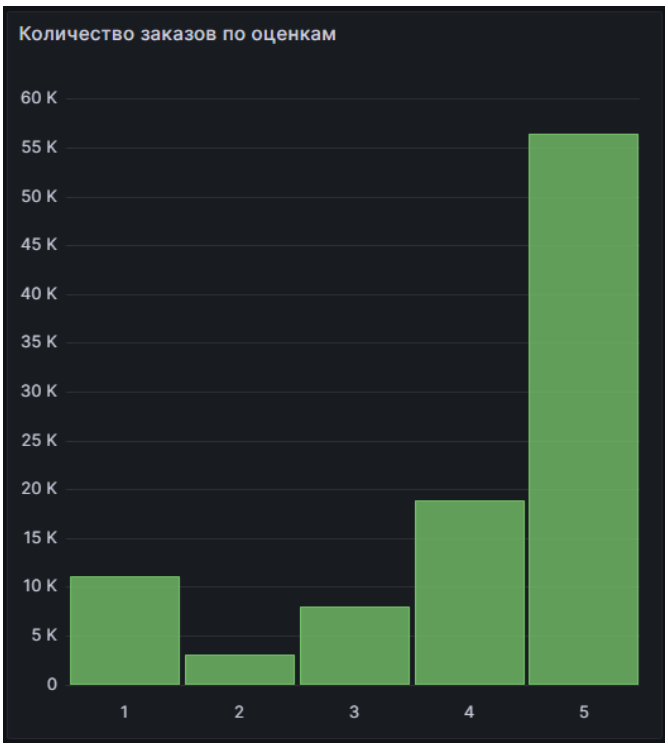

Рисунок 4 – Визуализация кол-ва заказов по оценкам

Распределение оценок показывает высокую удовлетворенность клиентов: подавляющее большинство заказов (более 55 000) получили максимальную оценку «5». Оценка «4» находится на втором месте (~19 000). Однако стоит отметить, что количество крайне негативных оценок «1» (около 10 000) существенно превышает количество нейтральных оценок «2» и «3». Это говорит о полярности мнений: клиенты либо очень довольны, либо сталкиваются с серьезными проблемами (например, дефектом товара).

4.4.3 Распределение способов оплаты

Необходимо показать распределение способов оплаты заказов. Для этого подсчитывается количество уникальных заказов, оплаченных каждым способом. Итоговые доли по каждому методу оплаты отображаются на круговой диаграмме, где сегменты представляют разные способы оплаты, а их размер соответствует доле.

Созданная визуализация представлена на рисунке 5.

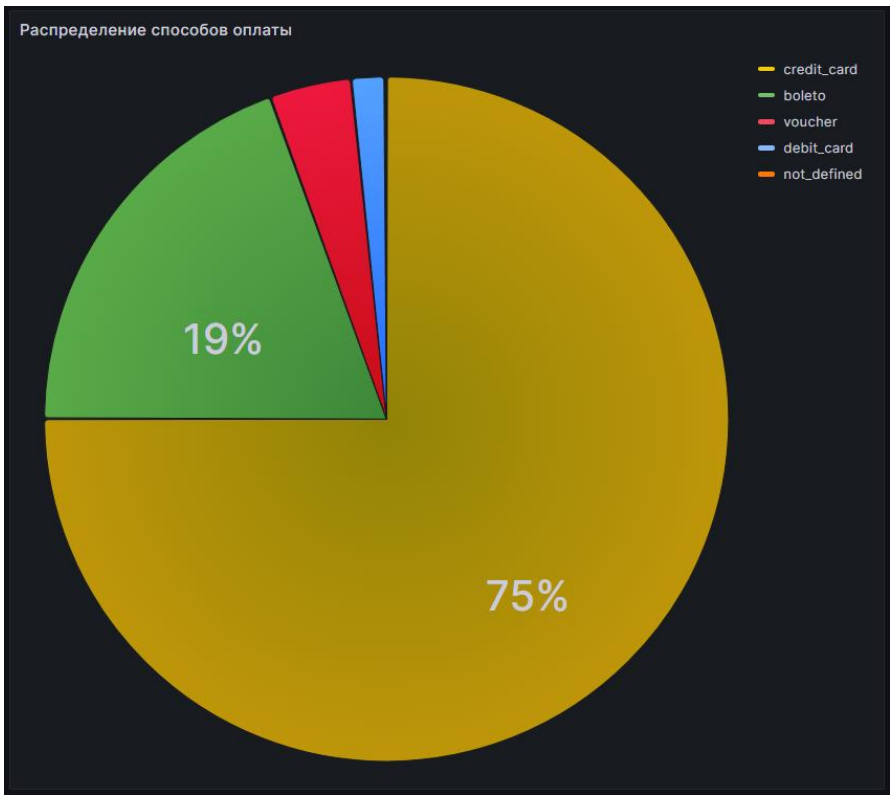

Рисунок 5 – Визуализация распределения способов оплаты

Анализ способов оплаты показывает абсолютное доминирование кредитных карт (credit_card), которыми оплачивается 75% всех заказов. Вторым по популярности методом является банковский перевод boleto (19%). Ваучеры (voucher) и дебетовые карты (debit_card) занимают малую долю, суммарно составляя около 6%.

4.4.4 Топ товарных категорий по объему продаж

Задача заключается в том, чтобы вывести топ-10 товарных категорий по общему объёму продаж. Для каждой категории суммируется стоимость всех товарных позиций, после чего выбираются десять лидеров. Визуализация выполняется в виде горизонтальных столбцов (линеек), что позволяет удобно сравнивать категории между собой по длине.

Созданная визуализация представлена на рисунке 6.

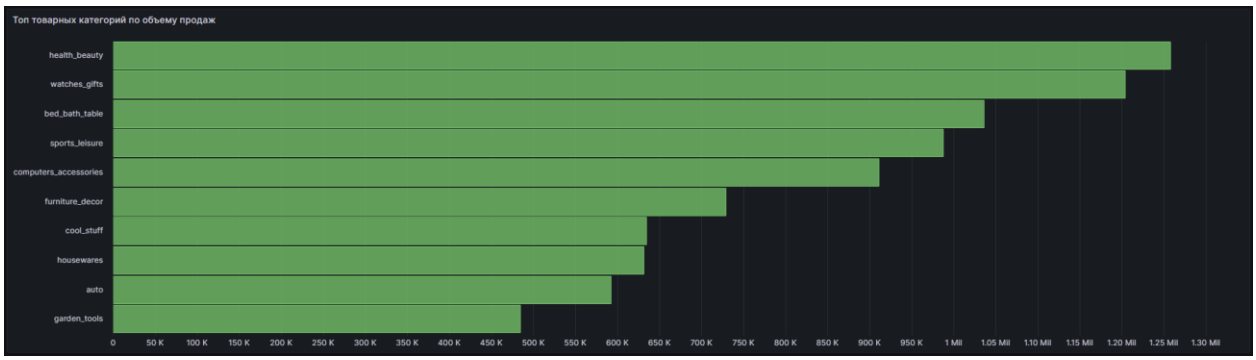

Рисунок 6 – Визуализация самых прибыльных категорий товаров

Лидирующей категорией по выручке является «Красота и здоровье» (health_beauty), объем продаж которой превышает 1.25 млн. Практически на том же уровне находятся категории «Часы и подарки» (watches_gifts) и «Постельное белье» (bed_bath_table). Плотная конкуренция в топ-3 говорит о широком ассортименте маркетплейса. Категории, связанные с домом и декором, занимают значительную часть топа.

4.4.5 Дополнительные визуализации

В качестве двух дополнительных визуализаций были выбраны показатели, связанные с географией бизнеса:

1. Штаты по кол-ву заказов (рисунок 7). Эта визуализация позволит оценивать развитие бизнеса в разных регионах.
2. Среднее время доставки по всем штатам (рисунок 8). Эта визуализация будет указывать на зоны для роста организации в сфере логистики в разных регионах.

Визуализация распределения заказов показывает, что распределение заказов сильно неравномерное. Абсолютным лидером является Сан-Паулу (SP) – на него приходится основная часть всех заказов. Далее с заметным отрывом следуют Рио-де-Жанейро (RJ) и Минас-Жерайс (MG). Эти три штата формируют ядро спроса.

В остальных регионах объём заказов существенно ниже, а в таких штатах, как Рорайма (RR), Амапа (AP) и Акри (AC), количество заказов минимально, что говорит о низкой концентрации клиентов.

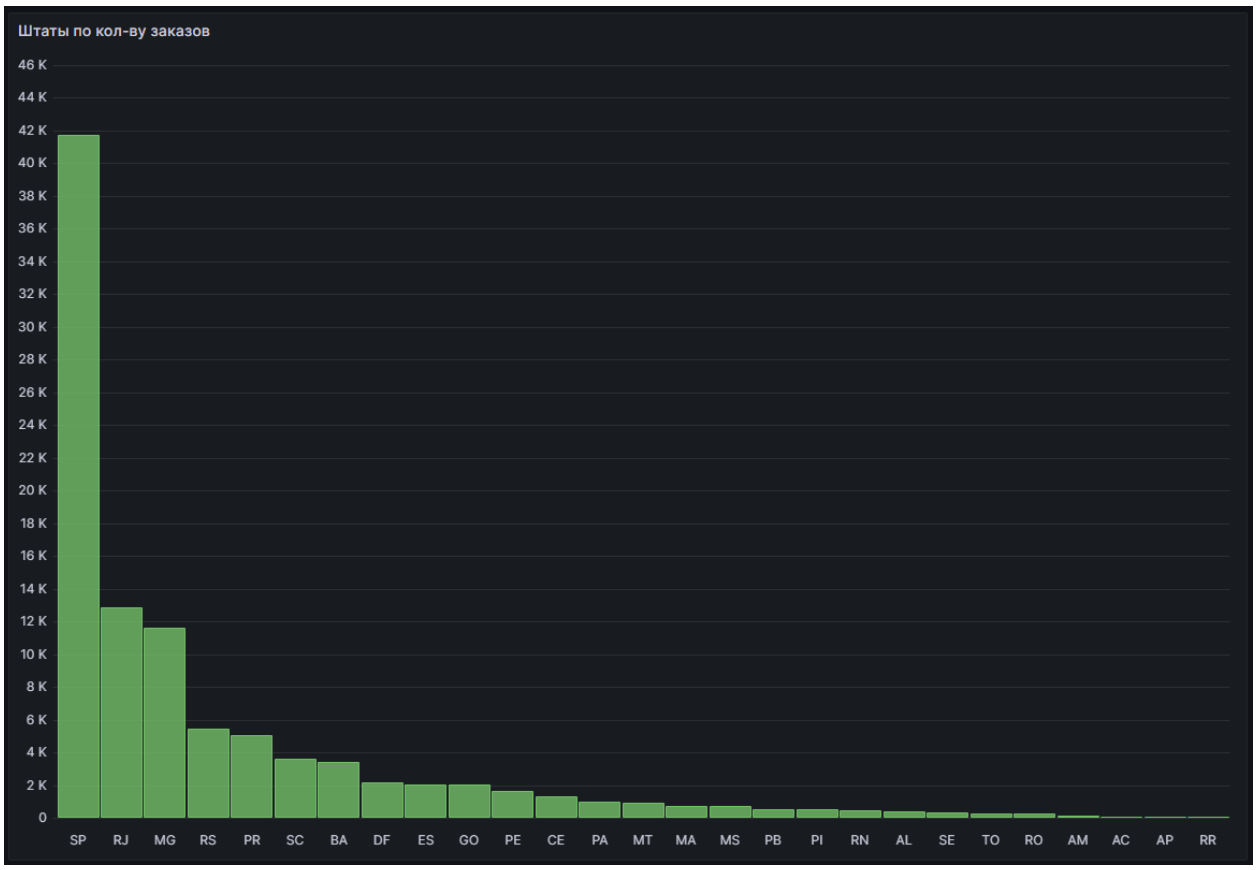

Рисунок 7 – Визуализация распределения заказов по штатам

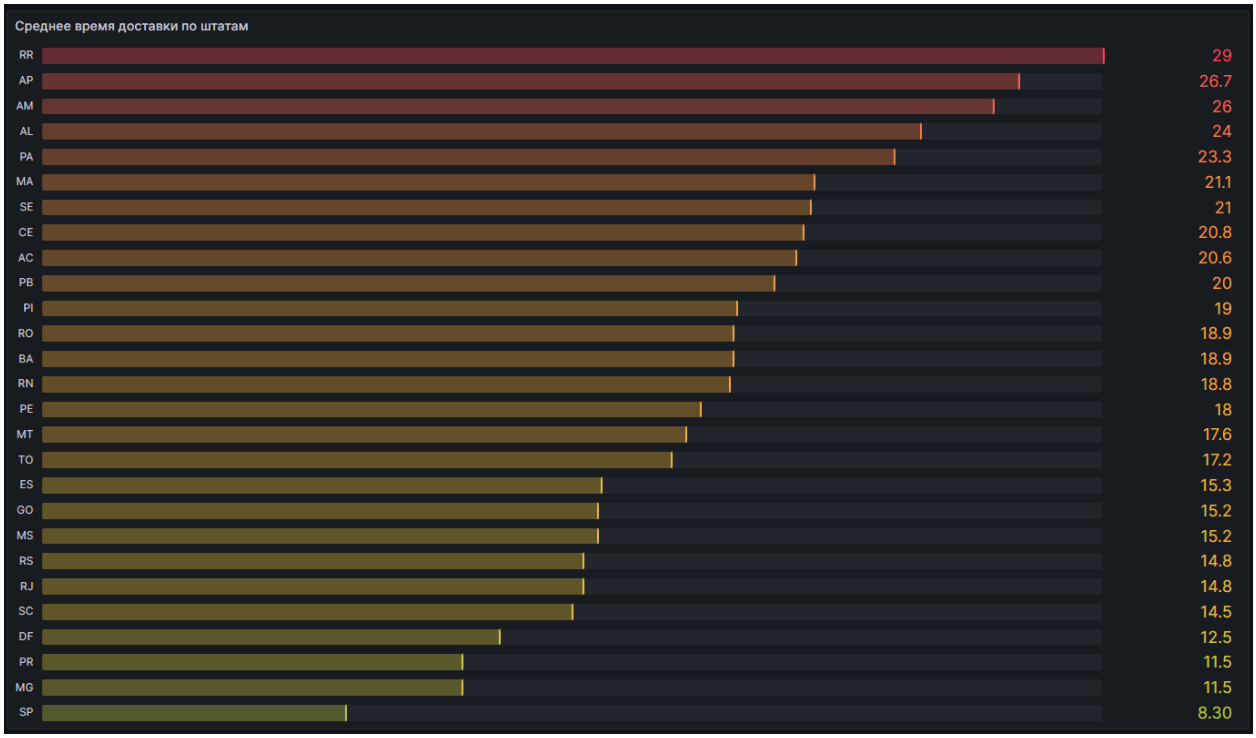

Рисунок 8 – Визуализация среднего времени доставки по штатам

Визуализация среднего времени доставки показывает, что наименьшее среднее время доставки наблюдается в штатах с самым высоким объёмом заказов – Сан-Паулу (SP), Минас-Жерайс (MG), Парана (PR) и Федеральный

округ – Бразилиа (DF) (примерно 8–12 дней). Это указывает на хорошо развитую логистическую инфраструктуру в регионах с высоким спросом.

В то же время штаты с небольшим числом заказов, такие как Рорайма (RR), Амапа (AP), Амазонас (AM) и Алагоас (AL), характеризуются наибольшими сроками доставки (20–29 дней). В целом прослеживается чёткая зависимость: чем больше заказов в штате, тем быстрее осуществляется доставка.

Итоговый дашборд, получившийся в результате разработки визуализаций в Grafana представлен на рисунках 9-10.

Рисунок 9 – Итоговый дашборд (1/2)

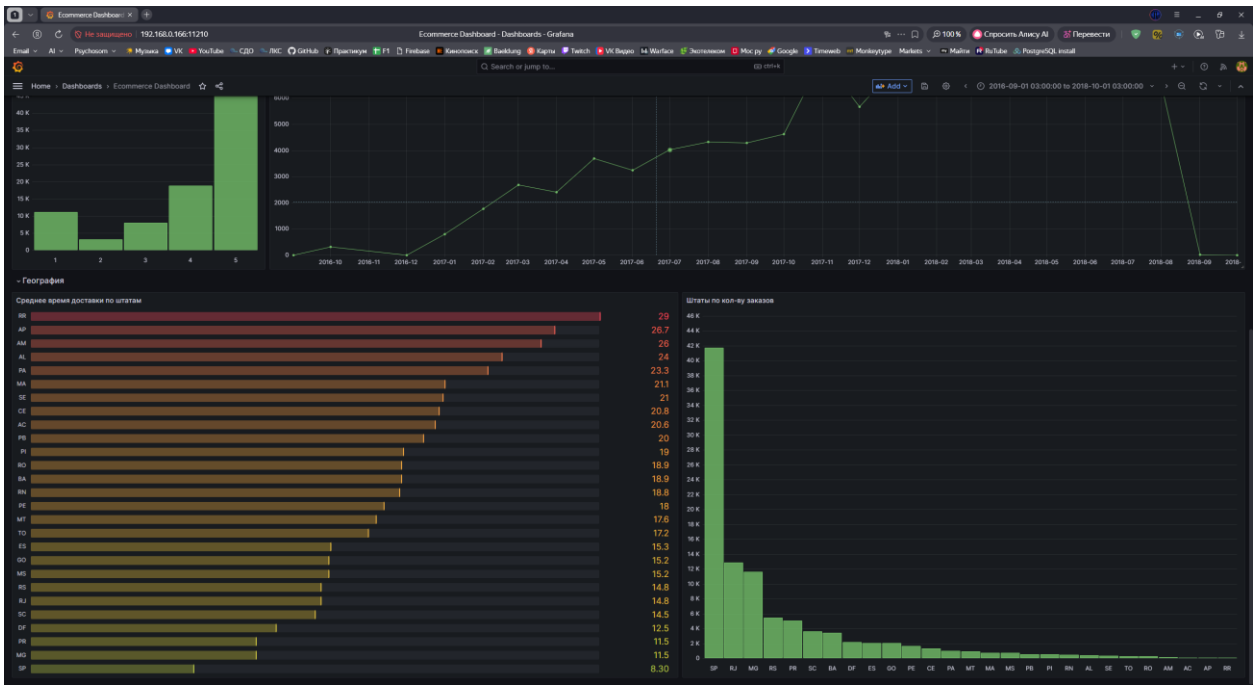

Рисунок 10 – Итоговый дашборд (2/2)

5 ВЫВОД

В ходе выполненной практической работы проведено ознакомление с возможностями работы Arenadata Cluster Manager совместно с бандлами для кластера и мониторинга.