Задание 2 «Источники данных»

1. Описать источники данных тип, формат или структуры данных, пакетные или потоковые, возможный способ доступа и регламент загрузки данных.

Источник	Описание	Тип данных	Формат	Загрузка	Способ доступа	Регламент загрузки
CRM-система банка	Данные о клиентах, продажах, каналах подключения	Транзакционные	Таблицы в СУБД (PostgreSQL/Oracle): –fact_sales –dim_client –dim_product	Пакетная	JDBC-подключение к DWH	Ежедневно, в 07:00
Система интернет-банка	Подключения через личный кабинет, поведение в системе	Транзакционные	JSON-логи, таблицы DWH	Пакетная	ETL-выгрузка через API или напрямую в DWH	Ежедневно, в 07:00
Рекламные платформы: Yandex.Direct	Данные о кликах, показах, конверсиях, UTM-метках	Внешние маркетинговые данные	CSV/JSON через API или SFTP: –clicks.csv –conversions.json	Пакетная	REST API (OAuth), SFTP	Ежедневно, в 07:00
CRM партнёров	Данные о клиентах, привлечённых через партнёрские площадки	Транзакционные	CSV/JSON по SFTP, API	Пакетная	SFTP, REST API	Ежедневно, в 7:00
Web-аналитика: Яндекс.Метрика	Поведение на сайте, цели, конверсии	Поведенческие	JSON через API, выгрузки в BigQuery	Пакетная / Потоковая	API с токеном	Ежедневно, в 07:00
Система UTM-маркировки	Данные о метках (источник, кампания, медиа)	Справочная информация	Таблицы в DWH:dim_utm_source,dim_campaign	Пакетная	ETL из рекламных платформ	Каждый понедельник в 7:00 / по запросу

2. Описать метаданные для Data Lineage

CRM-система банка

[CRM_BANK]
↓
Таблицы:
- fact_sales (sales_id, client_id, product_id, activation_dttm)
- dim_client (client_id, client_name, passport)
- dim_product (product_id, product_name)
↓
ETL-процесс:
- Выгрузка через JDBC
- Фильтр: только активные подключения
- Приведение client_id к единому формату
↓
Промежуточная таблица: stg_crm_sales
↓
Обогащение:
- JOIN с dim_client → получение ФИО, паспорта
- JOIN с dim_product → получение названия продукта
↓
Витрина: dm_marketing_attribution
↓
Поля витрины:
- client_id
- product_id
- conversion_dttm = activation_dttm
- client_name, passport

Регламент: Ежедневно в 07:00
Владелец данных: Владелец "Клиент", Владелец "Продукт"
Частота обновления: 1 раз в день
Качество: Проверка на NULL в `client_id`, `activation_dttm`

Система интернет-банка

[IB_SYSTEM]
↓
Источники:
- Логи сессий (JSON): user_id, session_start, utm_tags
- Таблица подключений: ib_activations (user_id, product_id, event_time)

↓
ETL-процесс:
- Парсинг JSON: извлечение utm_source, utm_medium, utm_campaign
- Маппинг user_id → client_id (через таблицу соответствия)
- Очистка UTM-меток: приведение к нижнему регистру, удаление пробелов
↓
Промежуточная таблица: stg_ib_journey
↓
Обогащение:
- Добавление marketing_source по utm_source
- Определение is_from_ad = 1, если utm_source не пустой
↓
Витрина: dm_marketing_attribution
↓
Поля витрины:
- client_id
- marketing_source
- click_dttm = session_start
- is_from_ad

Регламент: Ежедневно в 07:00
Владелец данных: Владелец `IB_SYSTEM`
Качество: Проверка на валидность UTM-меток, наличие `user_id`

Yandex.Direct

[Yandex.Direct API]
↓
Метод: reports.get (через OAuth)
- Поля: CampaignName, UtmSource, Clicks, Cost, Date, ClickID
↓
ETL-процесс:
- Запрос с фильтром за предыдущий день
- Приведение UtmSource к единому виду:
→ "yandex" вместо "YANDEX", "yandex_direct"
- Сопоставление ClickID с клиентом (через куки/email)
↓
Промежуточная таблица: stg_yandex_clicks
↓
Обогащение:
- Добавление channel_type = "Paid Search"
- Группировка кампаний по категориям (brand, non-brand)
↓

Витрина: dm_marketing_attribution
↓
Поля витрины:
- marketing_source = "yandex"
- campaign_name
- click_dttm
- cost_per_click
- attribution_type = "Last Click"

Регламент: Ежедневно в 07:00
Владелец данных: Маркетолог по контекстной рекламе
Качество: Проверка на дубли по `ClickID`, валидность даты

CRM партнёров

[CRM партнёров]
↓
Форматы:
- CSV по SFTP: partner_clients.csv
- JSON через API: /api/v1/conversions
- Поля: external_id, full_name, passport, product, created_date, source_system
↓
ETL-процесс:
- Загрузка с SFTP (каталог: /inbound/partners/)
- Парсинг JSON из API
- Нормализация source_system → единый код: "PARTNER_AUTO", "PARTNER_INSURANCE"
- Мэтчинг по ФИО + паспорт → client_id
↓
Промежуточная таблица: stg_partner_conversions
↓
Обогащение:
- Добавление marketing_source = "partner_" || partner_type
- Пометка is_partner_lead = 1
↓
Витрина: dm_marketing_attribution
↓
Поля витрины:
- client_id
- marketing_source
- conversion_dttm = created_date
- is_partner_lead

Регламент: Ежедневно в 07:00
Владелец данных: Представитель партнёрской интеграции
Качество: Проверка на наличие паспортных данных, валидность `created_date`

Яндекс.Метрика

[Яндекс.Метрика API]
↓
Метод: stat.v1.data
- Поля: date, utm_source, utm_campaign, goal_id, visitor_id, hit_time
- Цель: "Подключение продукта" (goal_id = 101)
↓
ETL-процесс:
- Запрос за предыдущий день
- Фильтр: только цели с goal_id = 101
- Приведение utm_source к справочному виду
- Сопоставление visitor_id с client_id (через email из формы)
↓
Промежуточная таблица: stg_ym_conversions
↓
Обогащение:
- Добавление conversion_type = "Web Goal"
- Расчёт времени от первого визита до конверсии
↓
Витрина: dm_marketing_attribution
↓
Поля витрины:
- client_id
- marketing_source
- campaign_name
- conversion_dttm = hit_time
- conversion_path_length

Система UTM-маркировки

[UTM Справочники]
↓
Таблицы:
- dim_utm_source (utm_source_raw, marketing_source, channel_type)
- dim_campaign_category (campaign_name_pattern, category)
↓
[Примеры]

- "yandex_direct" → marketing_source = "yandex", channel_type = "Paid Search"
- "insta_ads" → marketing_source = "instagram", channel_type = "Social Media"
↓
ETL-процесс:
- Еженедельное обновление на основе новых кампаний
- Ручное утверждение новых значений маркетологом
↓
Справочная таблица: ref_marketing_sources
↓
Используется в ETL всех источников для нормализации полей
↓
Влияет на все записи в dm_marketing_attribution, где есть UTM

Регламент: Еженедельно / по запросу
Владелец данных: Маркетолог / Аналитик по цифровым каналам
Качество: Все `utm_source_raw` должны быть сопоставлены с `marketing_source`

3. Классифицировать используемые элементы данных по типам корпоративных данных (транзакционные, основные, НСИ)

`client_id`	Мастер-данные	Уникальный идентификатор клиента — центральная сущность
`client_name`	Мастер-данные	Атрибут клиента из справочника
`passport`	Мастер-данные	Реквизит клиента, используется для идентификации
`product_id`	Мастер-данные	Уникальный идентификатор продукта (кредит, карта и т.д.)
`product_name`	Мастер-данные	Название продукта — часть справочника продуктов

`sales_id`	Транзакционные данные	Идентификатор факта продажи — событие
`activation_dttm`	Транзакционные данные	Момент подключения продукта — событие
`click_id`	Транзакционные данные	Событие клика по рекламе
`click_dttm`	Транзакционные данные	Время клика — часть транзакции
`conversion_dttm`	Транзакционные данные	Момент конверсии — факт завершения цели
`goal_id`	Транзакционные данные	Идентификатор цели в веб-аналитике
`cost`/`cost_per_click`	Транзакционные данные	Финансовые показатели по рекламе — события расходов
`marketing_source`	НСИ	Справочник рекламных источников: `yandex`,`google`
`utm_source`	НСИ	Исходное значение UTM — нормализуется по справочнику
`utm_medium`	НСИ	Тип канала (cpc, banner, social) — классификатор

`campaign_name`	НСИ	Название кампании — управляется маркетингом
`campaign_category`	НСИ	Группировка кампаний (brand, non-brand) — справочная гранулярность
`channel_type`	НСИ	Тип канала (Paid Search, Social Media) — классификатор
`attribution_model`	НСИ	Модель атрибуции (Last Click, First Click) — правило расчёта
`product_status`	НСИ	Справочник статусов: "активен", "отменён", "в обработке"
`is_partner_lead`	НСИ	Признак, определяемый по справочнику источников