Задание 3 "Качество данных"

Проверки на пригодность данных

Проверка качества	Название	Описание	Источник данных	Требование	Метод проверки
Полнота	Наличие `client_id`	Каждая запись должна содержать идентификатор клиента	`IB_SYSTEM`, `CRM_BANK`	`client_id IS NOT NULL`	`SELECT COUNT(*) FROM raw_ib WHERE client_id IS NULL`
Полнота	Наличие реквизитов клиента	Если нет `client_id`, должны быть ФИО, паспорт, дата рождения	`PARTNER_API_*`, `PARTNER_FILE_*`	При `client_id IS NULL`: все поля заполнены	`SELECT COUNT(*) FROM raw_partner WHERE client_id IS NULL AND (first_name IS NULL OR passport IS NULL OR birth_date IS NULL)`
Полнота	Наличие `product_id`	Каждое подключение должно быть связано с продуктом	Все источники	`product_id IS NOT NULL`	`SELECT COUNT(*) FROM raw_source WHERE product_id IS NULL`
Точность	Корректность значения `channel_cd`	Поле `channel_cd` должно быть одним из: `IB`, `BRANCH`, `PARTNER`	Все источники	`channel_cd IN ('IB', 'BRANCH', 'PARTNER')`	`SELECT COUNT(*) FROM raw_source WHERE channel_cd NOT IN ('IB', 'BRANCH', 'PARTNER')`
Точность	Формат и валидность `activation_dttm`	Дата активации должна быть в правильном формате и ≥ 01.01.2024	Все источники	`activation_dttm >= '2024-01-01'` и `IS TIMESTAMP`	`SELECT COUNT(*) FROM raw_source WHERE activation_dttm < '2024-01-01' OR activation_dttm IS NOT A VALID TIMESTAMP`
Точность	Статус подключения = `'active'`	Только активные подключения учитываются	Все источники	`status = 'active'`	`SELECT COUNT(*) FROM raw_source WHERE status != 'active'`

Уникальность	Отсутствие дубликатов	Не должно быть одинаковых записей (по `client_id,product_id,activation_dttm`)	Все источники	Нет дублей	`SELECT client_id, product_id, COUNT(*) FROM raw_source GROUP BY client_id, product_id, activation_dttm HAVING COUNT(*) > 1`
Полнота	Наличие всех обязательных полей	Проверка, что все ключевые поля присутствуют в схеме	Все источники	Должны быть: `client_id` (или реквизиты), `product_id`, `channel_cd`, `activation_dttm`, `status`	`DESCRIBE raw_source` или проверка при загрузке

Проверки на регулярной основе (ежедневно)

Наличие `client_id` (для `IB_SYSTEM`, `CRM_BANK`)	Гарантия идентификации клиента
Наличие реквизитов клиента (для `PARTNER_*`)	Возможность мэтчинга без `client_id`
Корректность `channel_cd`	Предотвращение искажения каналов
Проверка диапазона дат (`activation_dttm ≥ 2024-01-01`)	Соответствие периоду анализа
Проверка статуса (`status = 'active'`)	Учёт только активных подключений
Проверка на дубликаты	Исключение завышения метрик

Документ описания метрик

Код метрики	Описание метрики	Слой DWH	Единица исчисления	Нижний порог хорошего качества	Верхний порог хорошего качества	Система источник	Бизнес владелец	SQL запрос
missing_client_id_count	Количество записей без client_id	raw (IB, CRM)	count	0	0.1	IB_SYSTEM, CRM_BANK	Владелец данных "Клиент"	SELECT COUNT(*) FROM raw_ib WHERE client_id IS NULL
missing_requisites_count	Количество записей без реквизитов для мэтчинга	raw (PARTNER)	count	0	0.1	PARTNER_API_*, PARTNER_FILE_*	Владелец данных "Клиент"	SELECT COUNT(*) FROM raw_partner WHERE client_id IS NULL AND (first_name IS NULL OR passport IS NULL OR birth_date IS NULL)
invalid_channel_cd_count	Количество записей с некорректным channel_cd	raw	count	0	0.1	Все источники	Владелец данных "Канал"	SELECT COUNT(*) FROM raw_source WHERE channel_cd NOT IN ('IB', 'BRANCH', 'PARTNER')
out_of_range_date_count	Количество записей с activation_dttm < 2024-01-01	raw	count	0	0.1	Все источники	Владелец данных "Продукт"	SELECT COUNT(*) FROM raw_source WHERE activation_dttm < '2024-01-01'
inactive_status_count	Количество подключений с status != 'active'	raw	count	0	0.1	Все источники	Владелец данных "Продукт"	SELECT COUNT(*) FROM raw_source WHERE status != 'active'
duplicate_count	Количество дублирующихся	raw	count	0	0.1	Все источники	ETL-разработчик	SELECT COUNT(*) FROM (SELECT client_id, product_id,

	подключений							`activation_dttm FROM raw_source GROUP BY client_id, product_id, activation_dttm HAVING COUNT(*) > 1) AS dups`
`total_records_loaded`	Общее количество загруженных записей	raw	count	100	∞	Все источники	ETL-разработчик	`SELECT COUNT(*) FROM raw_source`

Шаблон для информирования об инциденте качества данных

(этим цветом выделены примеры заполнения шаблона)

Дата и время инцидента:[ГГГГ-ММ-ДД НН:ММ]
Источник данных: PARTNER_API_X
Тип инцидента: Пропущенные поля
Описание проблемы: В выгрузке от партнёра за 15.04.2024 отсутствует `client_id` в 18% записей. Также обнаружены значения `channel_cd = 'partner_site'`, что не соответствует стандарту.

Последствия:
- Невозможно идентифицировать 18% клиентов
- Некорректное распределение подключений по каналам
- Риск искажения доли партнёрских каналов в отчёте

Решение/Рекомендации:
1. Немедленно уведомить владельца источника (`Представитель партнёрской интеграции`).
2. Временно использовать мэтчинг по ФИО+паспорт+дата рождения.
3. Заблокировать загрузку данных с некорректным `channel_cd`.
4. Запросить исправление формата выгрузки от партнёра.

Ответственный: [ФИО, должность]
Статус: [Новый / В работе / Решено / Закрыто]
Срок исправления:** [ДД.ММ.ГГГГ]