Механизм работы со слоями данных

Данная статья призвана утвердить как стандарт обработки данных в команде ДУАД, статья собирательный образ всех A best practice в промышленности данных

Рекомендация: 2 бутылки пива для того чтобы просто прочитать эту ахинею, 4 бутылки - уже будет понимание процесса, после 6 и пачки сигарет - возникнет дзен, и вы не только поймете сам процесс и сможете предложить мне более очевидные оптимизационные улучшения но также и риски

1 этап ИЗ ИСТОЧНИКА В S3 слой (RAW) AIRFLOW → S3 → PArquet

Стенд источника

Получение среза последних данных с источника к примеру 2х таблиц , основное требование к источнику - минимизация нагрузки снятия данных

Таблица 1

```
CREATE TABLE table1 (
   id BIGINT PRIMARY KEY,
   time_val TEXT, -- поле "Время срез"
   data_val TEXT -- поле "какие то полезные данные"
);
```

id	Время	Данные
1	Прошлое	Sensor-A: 42.7°C
2	Вчера	...
3	Сегодня	...

```
INSERT INTO table1 (id, time_val, data_val) VALUES
(1, '2025-10-21 10:15:00', 'Sensor-A: 42.7°C'),
(2, '2025-10-22 09:48:00', 'Sensor-B: 39.1°C'),
(3, '2025-10-23 14:02:00', 'Sensor-C: 41.9°C'),
(4, '2025-10-24 08:57:00', 'Sensor-A: 43.2°C'),
(5, '2025-10-24 13:33:00', 'Sensor-B: 40.6°C'),
(6, '2025-10-24 16:44:00', 'Sensor-C: 42.1°C');
```

Таблица 2

```
CREATE TABLE table2 (
   time_val TEXT, -- поле "Время / срез"
   data_val TEXT, -- поле "Данные"
   table1_id BIGINT -- ссылка на table1.id
);
```

Время	Данные	id первой таблицы
Прошлое	Event: Calibration Sensor-A OK	1
Вчера	.....	2
Сегодня	.....	3

```
INSERT INTO table2 (time_val, data_val, table1_id) VALUES
('2025-10-21 11:00:00', 'Event: Calibration Sensor-A OK', 1),
('2025-10-22 10:10:00', 'Event: Sensor-B maintenance', 2),
('2025-10-23 14:45:00', 'Event: Sensor-C alert threshold', 3),
('2025-10-24 09:15:00', 'Event: Sensor-A auto-restart', 4),
('2025-10-24 13:45:00', 'Event: Sensor-B power drop', 5),
('2025-10-24 17:00:00', 'Event: Sensor-C recovery', 6);
```

⚠ По итогу первое определение "СЫРЫХ" данных для RAW слоя и является Silver по определению (Архитектурно уточнить Афанасьев Олег)

- `table1` моделирует сырые данные с источников.
- `table2` связывает события с конкретными записями первой таблицы, что также является сырым данным

Research Байболов Данияр

Spark: Parquet