СБЕР УНИВЕРСИТЕТ

Перезапуск Data Engineer

ETL

1. Извлечение (Extract)
2. Преобразование (Transform)
3. Загрузка (Load)

Нейминг

source (src) – источник данных

staging (stg) – промежуточный слой временного хранения

meta – слой мета-данных

target (tgt) – приемник данных

INSERT

Вставка 1 строки

1) INSERT INTO table (id, val, dt_update)

VALUES (5, ‘AA’, sysdate)

ID	VAL	DT_UPDATE
1	A	20.03.2021 06:07:38
2	B	20.03.2021 06:07:38
3	C	20.03.2021 06:07:38

2) INSERT INTO table

VALUES (5, ‘AA’, sysdate)

3) INSERT INTO table (id, val)

VALUES (5, ‘AA) --- в пропущенное поле dt_update будет вставлен NULL

INSERT

Вставка нескольких строк

```
INSERT INTO table (column1, column2, … column_n )
SELECT expression1, expression2, … expression_n
FROM source_table
[WHERE conditions];
```

Пример:

```
INSERT INTO src_increment (id, val)
select 6, 'A’ from dual
union all
select 8, 'B’ from dual
union all
select 9, 'C’ from dual
```

UPDATE

Обновление 1 таблицы

```
UPDATE table
SET column1 = expression1,
        column2 = expression2,
        …
        column_n = expression_n
[WHERE conditions];
```

Обновление 1 таблицы данными другой таблицы

```
UPDATE table1
SET column1 = (SELECT expression1
                      FROM table2
                      WHERE conditions
                      )
[WHERE conditions];
```

Инкрементальная загрузка

ID	VAL	DT_UPDATE
1	A	20.03.2021 06:07:38
2	B	20.03.2021 06:07:38
3	C	20.03.2021 06:07:38

Инкрементальная загрузка определяется тем, что загружает из базы данных только новые или измененные записи. Все прочие данные должны быть доступны.

ID	VAL	DT_UPDATE
1	A	20.03.2021 06:07:38
2	B	20.03.2021 06:07:38
3	C	20.03.2021 06:07:38
4	D	21.03.2021 16:15:04

Добавленная запись

Инкрементальная загрузка

ID	VAL	DT_UPDATE
1	A	20.03.2021 06:07:38
2	B	20.03.2021 06:07:38
3	YY	20.03.2021 06:07:38

Измененная запись

ID	VAL	DT_UPDATE
1	A	20.03.2021 06:07:38
2	XX	20.03.2021 06:07:38
3	C	20.03.2021 06:07:38
4	D	21.03.2021 16:15:04

Измененная запись

Инкрементальная загрузка

1. Выделить инкремент на источнике и положить его в stg-таблицу.
2. Вычислить в stg insert-записи и загрузить их в целевую таблицу.
3. Вычислить в stg update-записи и загрузить их в целевую таблицу.
4. Вычислить на источнике удаленные записи и удалить их из целевой таблицы.
5. Обновить мета-данные