Data Cleansing with SQL

Key Techniques for Effective Data Cleansing

- Identifying Duplicates
- Handling NULL Values
- Standardizing Data
- Correcting Data Formats
- Removing Unwanted Characters
- Outlier Detection
- Data Type Consistency
- Checking Referential Integrity

@Surjya Sabat

Identifying Duplicates

Using GROUP BY and HAVING

Syntex

```
SELECT column_name,
COUNT(*)
FROM table_name
GROUP BY column_name
HAVING COUNT(*) > 1
```

Removing Duplicates

Syntex

```
DELETE FROM table_name
WHERE id NOT IN (
SELECT MIN(id)
FROM table_name
GROUP BY column_name);
```

Handling NULL Values

Finding NULL Values

Syntex

```
SELECT * FROM table_name
WHERE column_name IS NULL;
```

Replacing NULLs

Syntex

```
UPDATE table_name
SET column_name =
'default_value'
WHERE column_name IS NULL;
```

Standardizing Data

Trimming Whitespace

Syntex

```
UPDATE table_name
SET column_name = TRIM(column_name);
```

Converting Case

Syntex

```
UPDATE table_name
SET column_name =
UPPER(column_name);
```

Correcting Data Formats

Fixing Dates

Syntex

```
UPDATE table_name
SET date_column = TO_DATE(date_column, 'YYYY-MM-DD')
WHERE date_column IS NOT NULL;
```

Formatting Numbers

Syntex

```
UPDATE table_name
SET number_column = ROUND(number_column, 2);
```

Removing Unwanted Characters

Removing Special Characters

Syntex

```
UPDATE table_name
SET column_name = REGEXP_REPLACE(column_name, '[^a-zA-Z0-9 ]', '')
```

Outlier Detection

Finding Outliers

Syntex

```
SELECT column_name
FROM table_name
WHERE column_name > (SELECT AVG(column_name) + 3 * STDDEV(column_name) FROM table_name)
   OR column_name < (SELECT AVG(column_name) - 3 * STDDEV(column_name) FROM table_name);
```

Data Type Consistency

Converting Data Types

Syntex

ALTER TABLE table_name

ALTER COLUMN column_name TYPE new_data_type USING column_name::new_data_type;

Checking Referential Integrity

Finding Orphan Records

Syntex

```
SELECT *
FROM child_table
WHERE foreign_key NOT IN (SELECT primary_key FROM parent_table);
```

Data cleansing with SQL ensures data accuracy and consistency by handling duplicates, NULLs, format corrections, and more. It automates detection and correction processes, enhancing data integrity for better analysis and decision-making. A vital tool for data quality.

Surjya Sabat

SQL | Python | Power BI | MSBI | VBA | Process Improvement