Data Integration Project
Step 3: Cleaning & Showcase

Is Bigfoot an Alien?

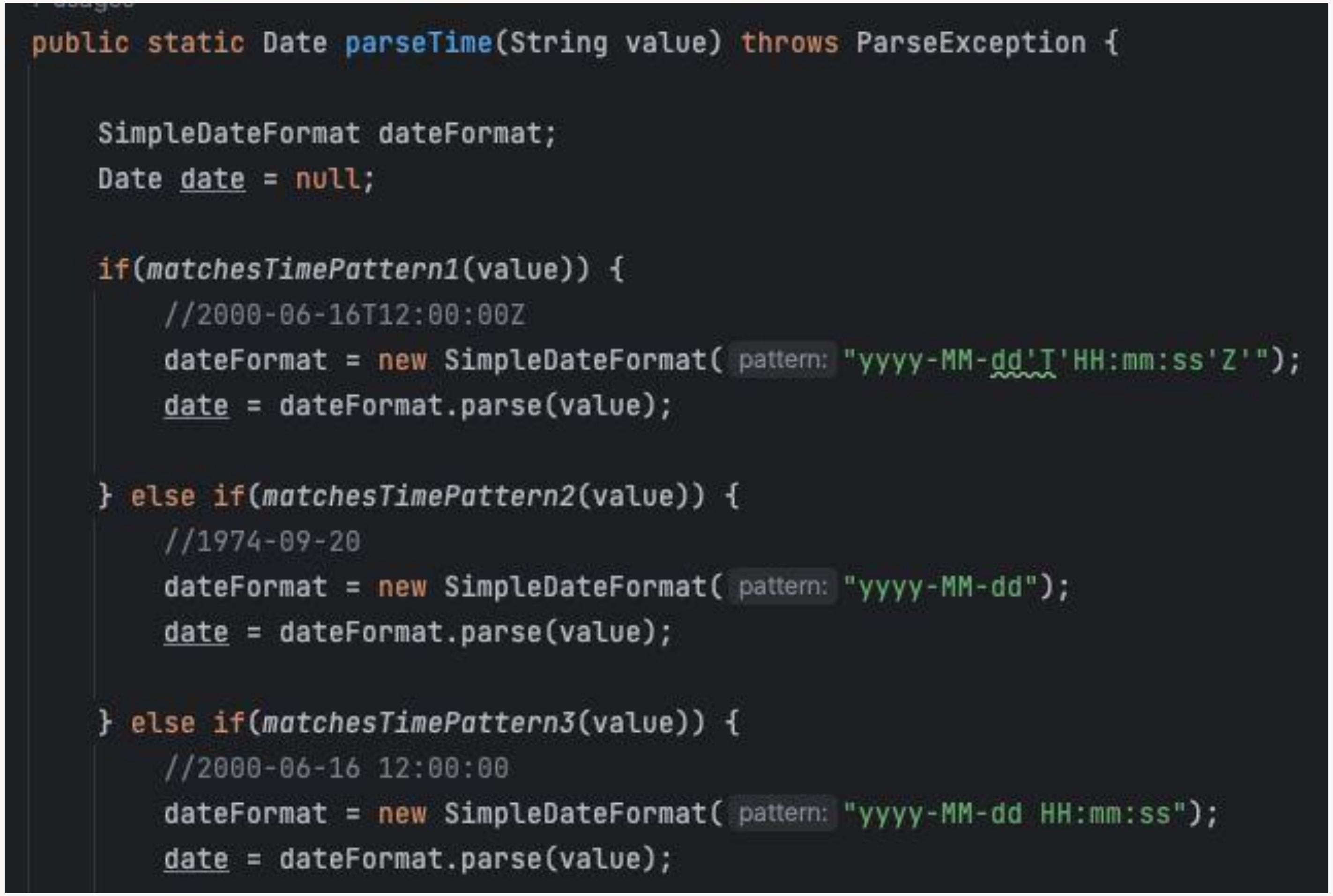
Julian Trösser

Zoe Chiying Lai

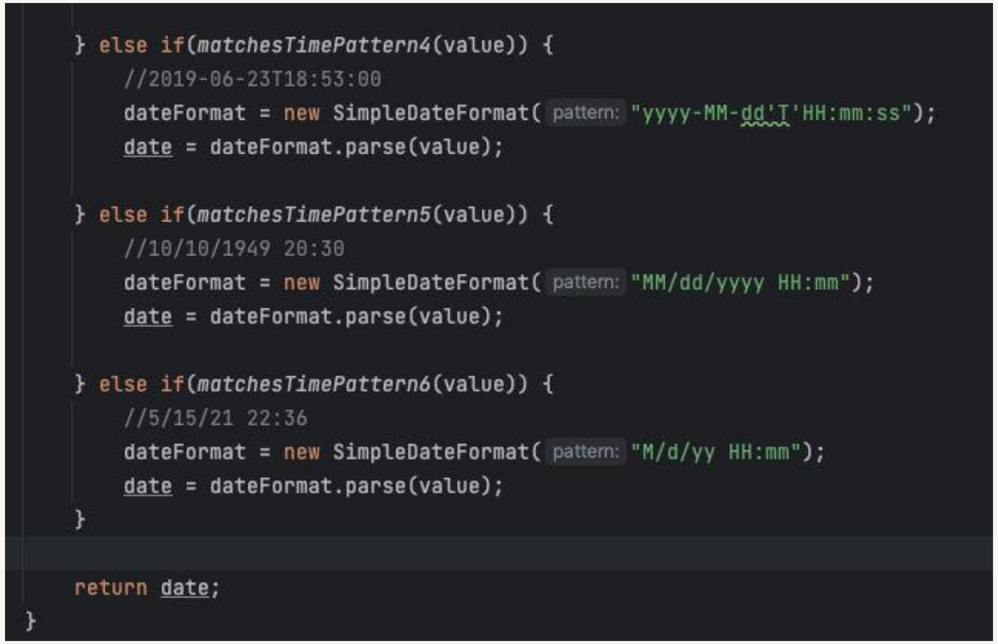
Data Standardization

- Data types are defined in SQL database
- For example:
  - TIMESTAMP for date
  - INTEGER for id
  - DECIMAL for temperature
  - VARCHAR for country
  - TEXT for description
- Fitted when writing from Java into SQL Database using setInt(), setDate() etc with Prepared Statement

Data Formatting Example

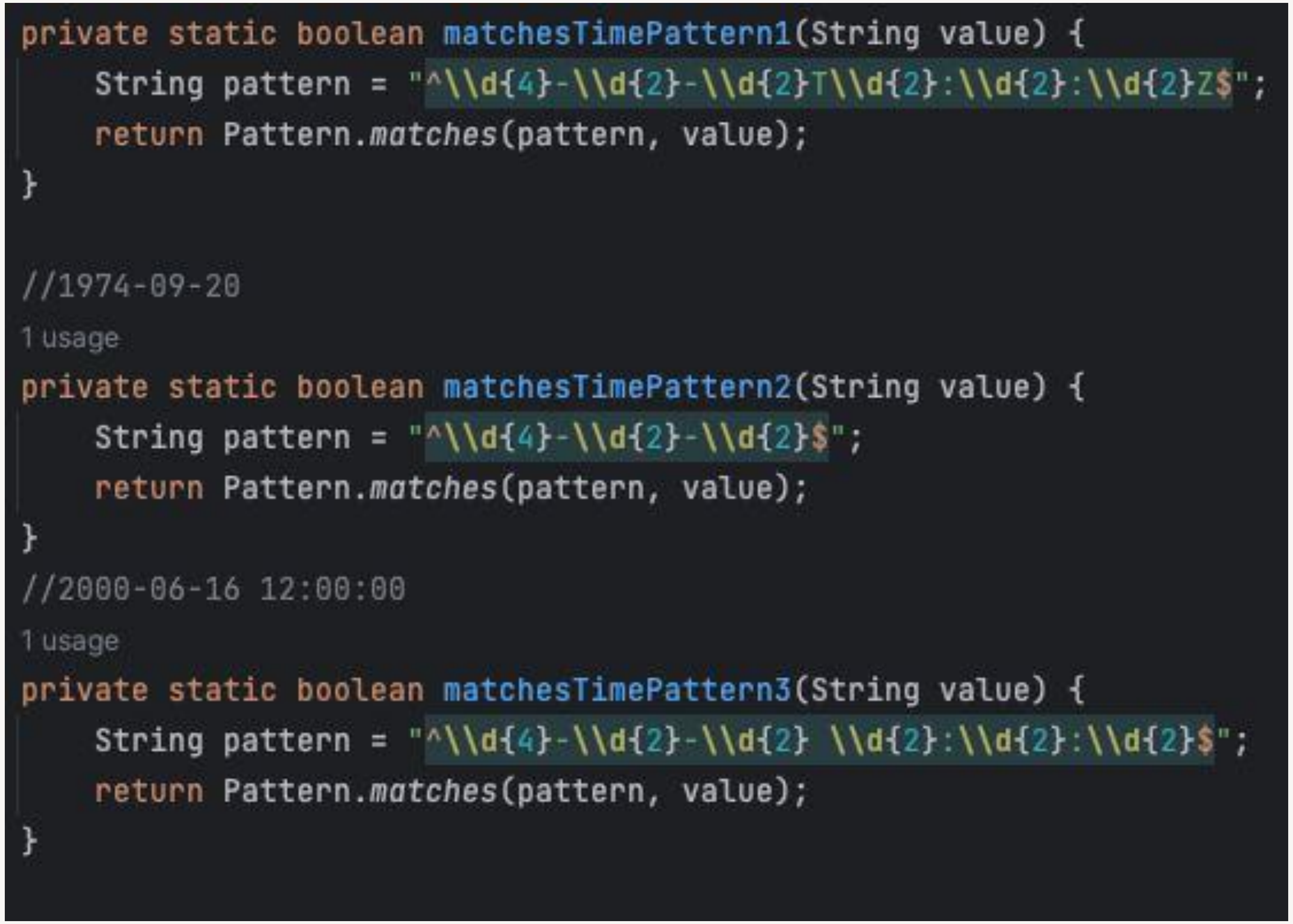


```java
public static Date parseTime(String value) throws ParseException {

    SimpleDateFormat dateFormat;
    Date date = null;

    if(matchesTimePattern1(value)) {
        //2000-06-16T12:00:00Z
        dateFormat = new SimpleDateFormat( pattern: "yyyy-MM-dd'T'HH:mm:ss'Z'");
        date = dateFormat.parse(value);

    } else if(matchesTimePattern2(value)) {
        //1974-09-20
        dateFormat = new SimpleDateFormat( pattern: "yyyy-MM-dd");
        date = dateFormat.parse(value);

    } else if(matchesTimePattern3(value)) {
        //2000-06-16 12:00:00
        dateFormat = new SimpleDateFormat( pattern: "yyyy-MM-dd HH:mm:ss");
        date = dateFormat.parse(value);
```

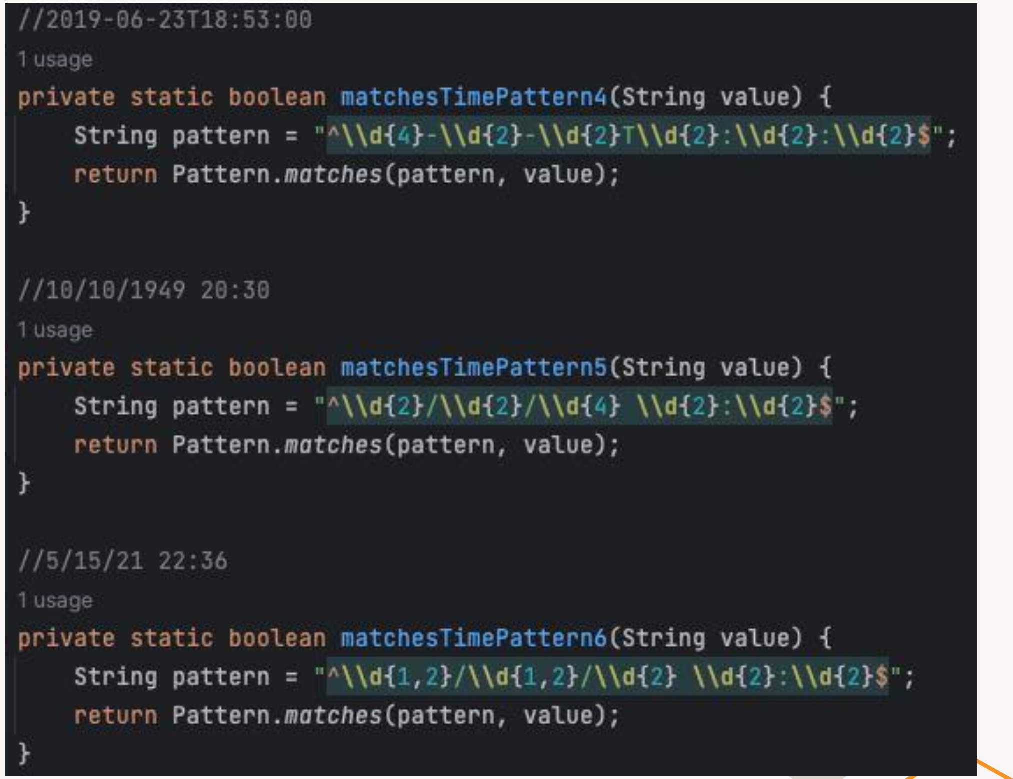


```java
    } else if(matchesTimePattern4(value)) {
        //2019-06-23T18:53:00
        dateFormat = new SimpleDateFormat( pattern: "yyyy-MM-dd'T'HH:mm:ss");
        date = dateFormat.parse(value);

    } else if(matchesTimePattern5(value)) {
        //10/10/1949 20:30
        dateFormat = new SimpleDateFormat( pattern: "MM/dd/yyyy HH:mm");
        date = dateFormat.parse(value);

    } else if(matchesTimePattern6(value)) {
        //5/15/21 22:36
        dateFormat = new SimpleDateFormat( pattern: "M/d/yy HH:mm");
        date = dateFormat.parse(value);
    }

    return date;
}
```

Data Formatting Example



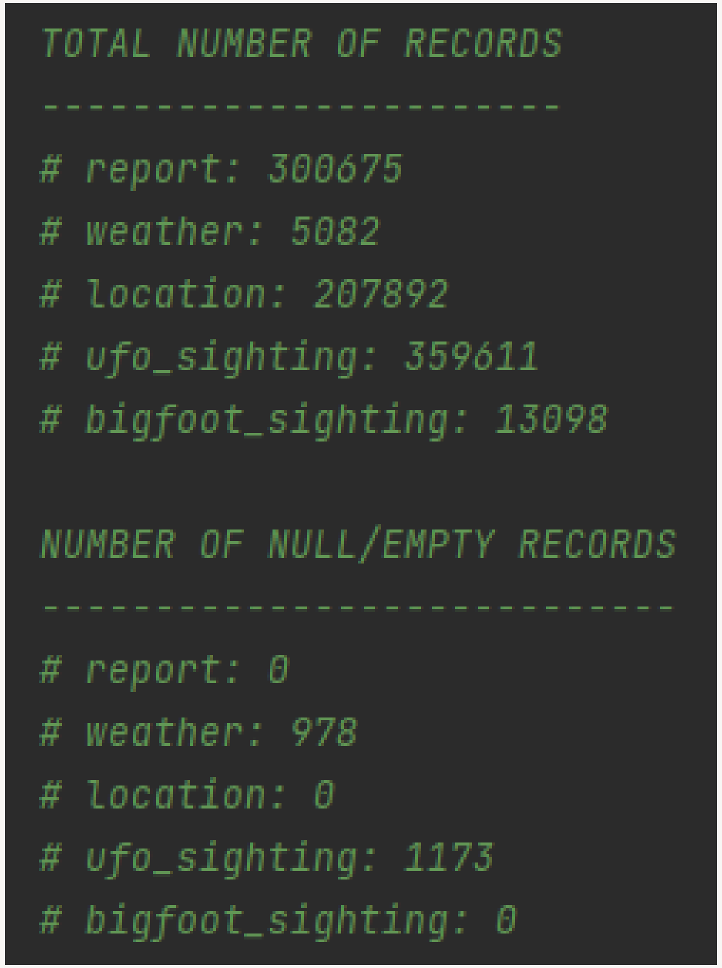
```
private static boolean matchesTimePattern1(String value) {
    String pattern = "^\\d{4}-\\d{2}-\\d{2}T\\d{2}:\\d{2}:\\d{2}Z$";
    return Pattern.matches(pattern, value);
}

//1974-09-20
1 usage
private static boolean matchesTimePattern2(String value) {
    String pattern = "^\\d{4}-\\d{2}-\\d{2}$";
    return Pattern.matches(pattern, value);
}
//2000-06-16 12:00:00
1 usage
private static boolean matchesTimePattern3(String value) {
    String pattern = "^\\d{4}-\\d{2}-\\d{2} \\d{2}:\\d{2}:\\d{2}$";
    return Pattern.matches(pattern, value);
}
```



```
//2019-06-23T18:53:00
1 usage
private static boolean matchesTimePattern4(String value) {
    String pattern = "^\\d{4}-\\d{2}-\\d{2}T\\d{2}:\\d{2}:\\d{2}$";
    return Pattern.matches(pattern, value);
}

//10/10/1949 20:30
1 usage
private static boolean matchesTimePattern5(String value) {
    String pattern = "^\\d{2}/\\d{2}/\\d{4} \\d{2}:\\d{2}$";
    return Pattern.matches(pattern, value);
}

//5/15/21 22:36
1 usage
private static boolean matchesTimePattern6(String value) {
    String pattern = "^\\d{1,2}/\\d{1,2}/\\d{2} \\d{2}:\\d{2}$";
    return Pattern.matches(pattern, value);
}
```

Cleaning

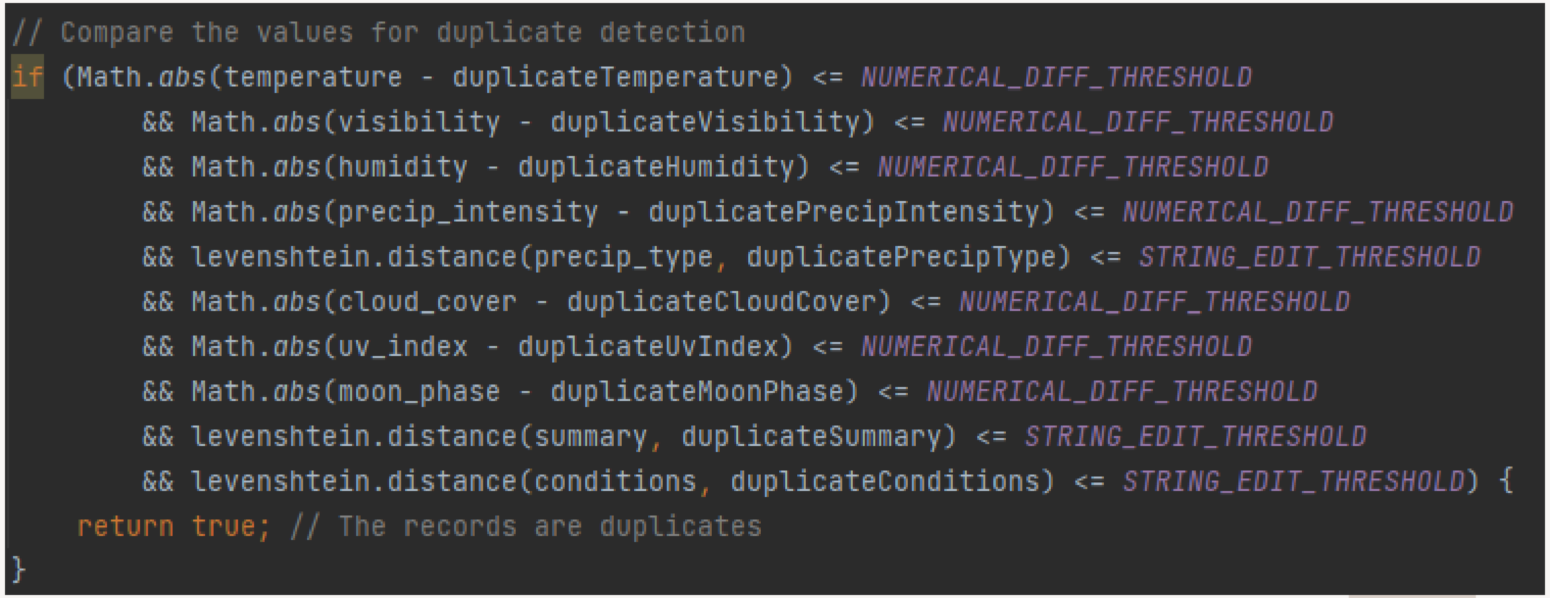
- Many weather information missing
- Some empty records found
- Potential duplicates expected



```
TOTAL NUMBER OF RECORDS
-----------------------
# report: 300675
# weather: 5082
# location: 207892
# ufo_sighting: 359611
# bigfoot_sighting: 13098

NUMBER OF NULL/EMPTY RECORDS
----------------------------
# report: 0
# weather: 978
# location: 0
# ufo_sighting: 1173
# bigfoot_sighting: 0
```

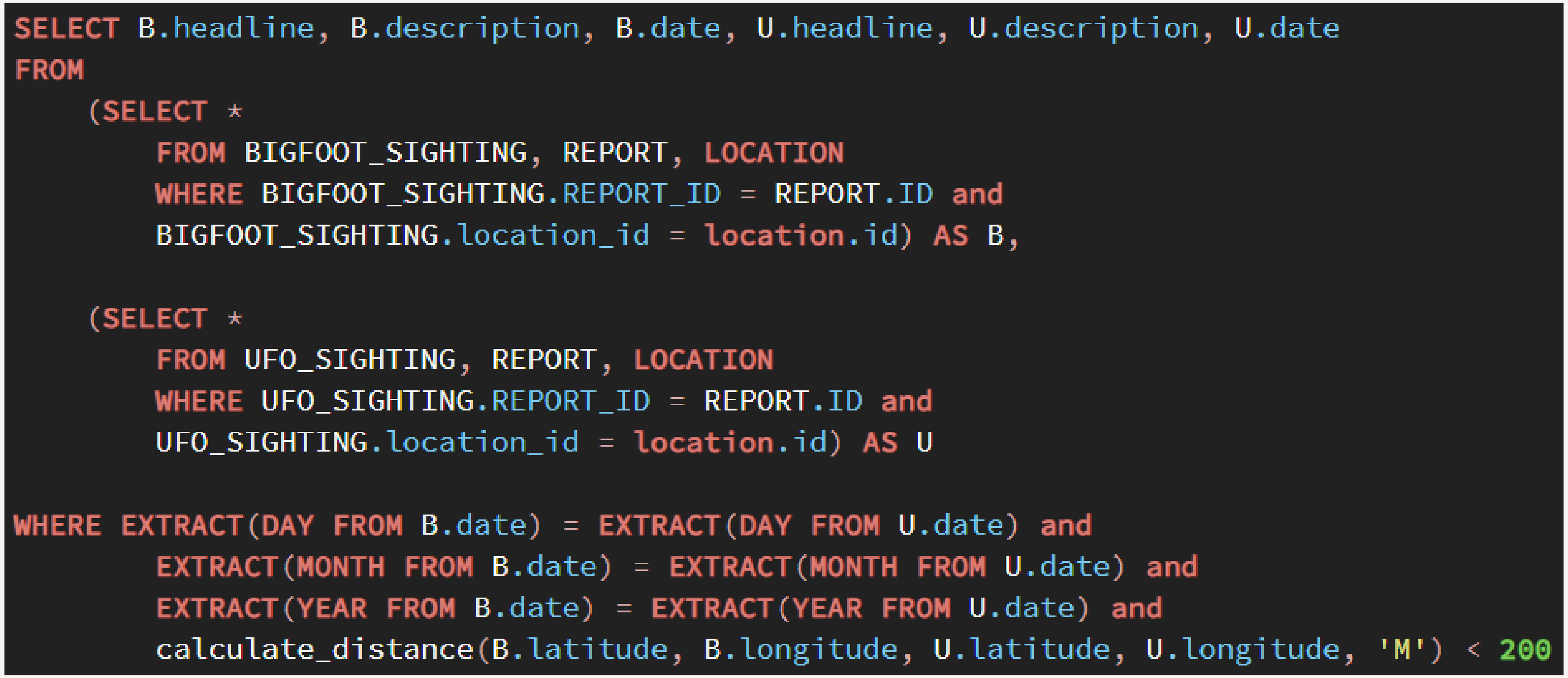
Cleaning Example: Weather

- Duplicates when:
    - All numerical fields have difference <= 5
    - All string fields have difference in Levenstein distance <= 2
- Tested with several thresholds
- No duplicates found



```
// Compare the values for duplicate detection
if (Math.abs(temperature - duplicateTemperature) <= NUMERICAL_DIFF_THRESHOLD
        && Math.abs(visibility - duplicateVisibility) <= NUMERICAL_DIFF_THRESHOLD
        && Math.abs(humidity - duplicateHumidity) <= NUMERICAL_DIFF_THRESHOLD
        && Math.abs(precip_intensity - duplicatePrecipIntensity) <= NUMERICAL_DIFF_THRESHOLD
        && levenshtein.distance(precip_type, duplicatePrecipType) <= STRING_EDIT_THRESHOLD
        && Math.abs(cloud_cover - duplicateCloudCover) <= NUMERICAL_DIFF_THRESHOLD
        && Math.abs(uv_index - duplicateUvIndex) <= NUMERICAL_DIFF_THRESHOLD
        && Math.abs(moon_phase - duplicateMoonPhase) <= NUMERICAL_DIFF_THRESHOLD
        && levenshtein.distance(summary, duplicateSummary) <= STRING_EDIT_THRESHOLD
        && levenshtein.distance(conditions, duplicateConditions) <= STRING_EDIT_THRESHOLD) {
    return true; // The records are duplicates
}
```

Showcase: Is Bigfoot an Alien?



```
SELECT B.headline, B.description, B.date, U.headline, U.description, U.date
FROM
    (SELECT *
        FROM BIGFOOT_SIGHTING, REPORT, LOCATION
        WHERE BIGFOOT_SIGHTING.REPORT_ID = REPORT.ID and
        BIGFOOT_SIGHTING.location_id = location.id) AS B,

    (SELECT *
        FROM UFO_SIGHTING, REPORT, LOCATION
        WHERE UFO_SIGHTING.REPORT_ID = REPORT.ID and
        UFO_SIGHTING.location_id = location.id) AS U

WHERE EXTRACT(DAY FROM B.date) = EXTRACT(DAY FROM U.date) and
        EXTRACT(MONTH FROM B.date) = EXTRACT(MONTH FROM U.date) and
        EXTRACT(YEAR FROM B.date) = EXTRACT(YEAR FROM U.date) and
        calculate_distance(B.latitude, B.longitude, U.latitude, U.longitude, 'M') < 200
```

- Query result:
  6629 records

- There are in total:
  - Bigfoot sightings: 13098
  - UFO sightings: 358438

Thank you