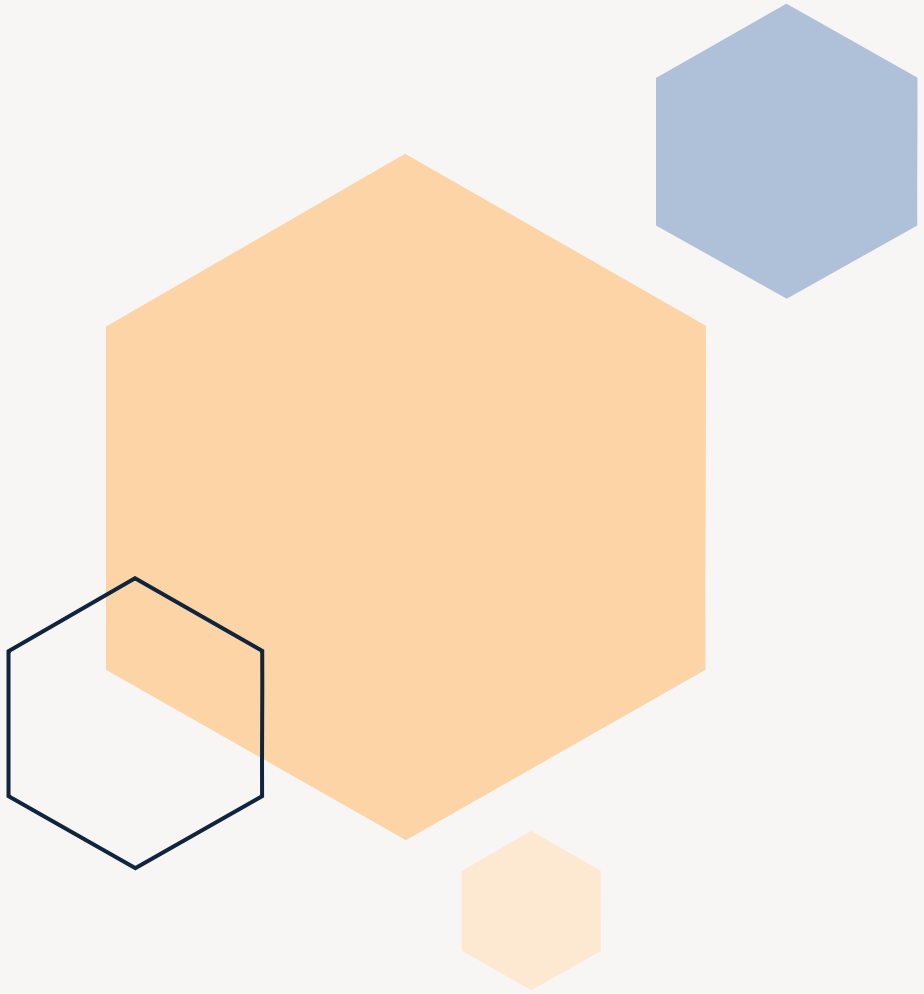
Data Integration Project
Step 3: Cleaning & Showcase

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

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
```

Thank you