CSI 4142 demo

how to build a data mart, basic OLAP, machine learning

Outline

1. Datamart creation
2. Data cleaning
3. Surrogate key pipeline
4. OLAP queries
5. Some dashboard options
6. Weka - connecting to your DB

Data used: UFO sightings from assignment 1

High level schematic plan

- Identify starting and ending points
  - CSV files ➞ PostgreSQL datamart
- Label known data sources
  - CSV file from A1
- Include placeholders for sources yet to be determined
  - currently unknown CSV file listing cities, states, countries across the world
- Label targets
  - Shapes, Locations, Dates, Facts
- Include notes about known problems
  - messy location data, 24+ hour long sightings, case sensitive shape names, generate additional attributes, check ref. int., role playing date dimension

Data cleaning

- Extract date and time of event
- Downcase shape name
- Missing values
  - discard when date
  - use ‘unspecified’ for shape name
- Duration
  - 23% under 10s …45% under 120s …80% under 15min … 96% under 1hr
    - (you could remove sightings > 5hrs)
- Partial match city, state to a CSV of worldwide locations
- ...

Creating surrogate keys

Ensure row uniqueness in dimension tables
Keys auto-increment from 1

In database *(not recommended)*

- dim_t = select distinct *col* from staging_t
- fact_table = insert values (select id *, select id *, select id *, select duration)

* from dim where staging_t.value = dim_t.value

In a script

Use data structure (DS) with fast insertion and searching (hashmap, dictionary, set...)

While iterating over your CSV rows, keep track of values seen, and use DS.num_elements as the ID when adding elements to the DS

Data-mart creation

1. Create database and tables
   create table *table_name* (*column_name1 datatype1, ...*)

2. Inserting data into your tables
   - import CSV files
     - use staging tables and SQL
     - script outputs 1 pre-processed CSV file per table
   - script: when done cleaning, insert in batches (with RI off to speed up inserting)

OLAP query examples

dice, slice, rolling ⬆, drilling ⬇, iceberg🚢

- Determining {measures} per {dimension attributes}
  - total number of sightings of circle-based shapes in Ontario during 2013, and so on.
- Contrasting {measures} per {dimension attributes} when compared to {dimension attributes}
  - number of sightings in Ontario of circle-based shapes, during 2014, with the number of sightings in Quebec of sphere-based, during 2014.
- Popular sightings
  - determine the 5 cities in the USA with the most sightings on Fri-Sun of sphere shapes
- Trends of {measures} over {dimension attributes}
  - determine trends of sighting numbers of light-based shapes over the years and countries in North America
  - determine trends of sighting numbers of all shapes over the years

Some dashboard options

- Build your own
- Tableau Desktop (drag-n-drop OLAP queries, dashboards)
- Klipfolio, Qlik (dashboards)
  - need ngrok (to *'securely'* expose your local postgres to the internet), or use uOttawa postgres

Machine learning / data mining with WEKA [how to connect]

1. Export a SQL command as a CSV
2. Open WEKA > Tools > Arff viewer
3. File > open (format=csv) find your file
4. File > save as .arff file
5. Weka > Explorer > open file

1. Download jdbc Postgres driver into app & DatabaseUtils.prop file into home directory (and modify as needed)
2. Open Weka > Explorer > open DB
3. URL = jdbc:postgresql://localhost:5432/DB_name
4. Hit the button with the lightning bolt
5. Query = SELECT mineable_attributes FROM fact INNER JOIN dimension1 d ON d.id = fact.dim1_id INNER JOIN ...

This method is harder unless you can skip step 1⬆

Some free tools (for students)

- postgreSQL server
  - web0.site.uottawa.ca:15432 …uozone login …uozone username for dbname)
  - Postgres.app, Postico (for Mac)
- Tableau Desktop (data exploration, OLAP queries, dashboards)
- Klipfolio, Qlik (dashboards)
  - need ngrok (to *'securely'* expose your local postgres to the internet), or use uOttawa postgres
- Weka, R, RapidMiner, SK Learn, etc

Good options for data cleaning and/or building your own dashboard:

Python, JavaScript, Ruby, SQL, + many others


THANK YOU!
ANY QUESTIONS?


wat


IM LIKE
SO CONFUSED MAN