Erasmus Mundus
master course in
Data Mining and Knowledge Management
a european master

Data WareHouse Proposal

Smart City:

Big Data Analytics using open data of Grand Lyon, France

DMKM1517

Monday 28th of March 2016

Objective

The objective of the Data Warehouse is to analyze the rating and visits (check ins) of interest points in Grand Lyon using the information available such as location, point and time.

User Needs

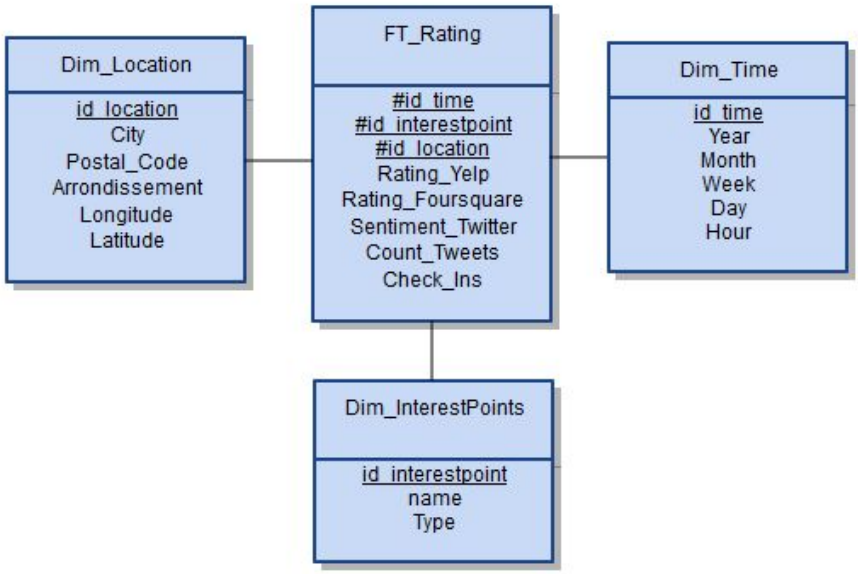
The visitors will be able to:

- explore the change in opinion of a specific point in time
- explore the aggregated opinion in a certain street, arrondissement, code postal or ville
- explore the check ins per place in time
- explore the aggregated check ins in a certain street, arrondissement, code postal or ville
- explore the opinion by source

Facts

Objective	Facts	Measures	Dimensions	Hierarchies
Analyze rating	Rating/ Time / Location / Point	Rating_Yelp Rating_Foursquare Sentiment_Twitter	Time Location Point	Hour - Day - Month - Year Latitude - Longitude - Street - Arrondissement - C.P. - Ville
Analyze Checkins	Checkins/ Time / Location / Point	Checkins	Time Location Point	Hour - Day - Month - Year Latitude - Longitude - Arrondissement - Postal_Code - Ville

Schema



Data Sources structure and content

interest_points

id	ID of the observation “Ex: 747606”
id_sitra1	Concatenation of *id* with *type* and *type_detail* “Ex: SITRA2_RES_747606”
type	Type of point of interest “Ex: RESTAURATION”
type_detail	Specific type of point of interest: “Ex: Restaurant traditionnel”
name	Name of the point of interest: “Ex: Brasserie les 3 brasseurs”
address	Address, specifically street. “Ex: 8 Chemin de Pontet et Crases”
postal_code	Postal code “Ex: 69130”
commune	Name of the area or district “Ex: Ecully”
telephone	Telephone number: “Ex: 0437460666”
email	Email address “Ex: info@ccc-lyon.com ”
website	Webpage “Ex: http://www.les3brasseurs.com“

facebook	Facebook page “Ex: http://www.facebook.com/pages/Havana-Caf%C3%A9/115023578571038”
open_hours	Opening hours “Ex: Ouvert tous les jours de 11h à 23h, sauf sam. jusqu'à minuit et dim. jusqu'à 22:30. ”
producteur	Segment of interest “Ex: Lyon Tourisme et Congrès ”
source_create_date	Creation in Grand Lyon database “Ex: 2015-12-23”
lsource_ast_update	Last update in Gran Lyon’s database “Ex: 2015-12-23”
in_use	Indicates if the interest point is shown in the map “Ex: false ”
coordinates_lat	Latitude coordinates “Ex: 45.7859839999999”
coordinates_long	Longitude coordinates “Ex: 4.774516”
sentiment	The sentiment evaluation has values from 1-5. “Ex: 4”

twitter_feed

id	ID given by Twitter API “709,133,138,308,558,851”
username	twitterUsername
tweet	Déjeuner fort sympathique dans l'excellent Kitchen Café #restaurant #Lyon formule dejeuner22€
sentiment	The sentiment rate calculated on the scale of 1 to 5, where 1 - strong negative 2 - weak negative 3 - neutral 4 - weak positive 5 - strong positive
retweet	Retweet number
lat	Latitude, if available 45.7631569999999996
long	Longitude, if available 4.85960500000000017
language	Language of the tweets for the purpose of sentiment analysis
time	The timestamp

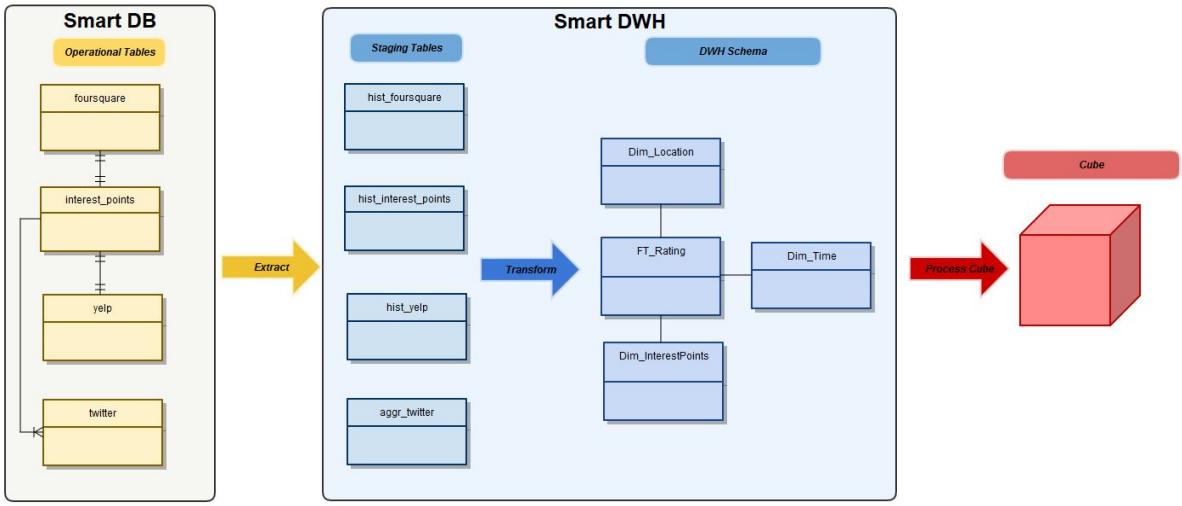
forsquare

idd	Foreign key to the respective interest point “149,370”
name	The name of the interest point “Parc de la Tête d'Or”
checkinscount	Number of checkins for the given point of interest
tipscount	Number of “tips” - comments on Forsquare for the given point of interest
userscount	Number of users who rated the interest point
rating	Rating on a scale of 0 to 10, 10 being the best.

yelp

idd	Foreign key to the respective interest point “149,370”
name	The name of the interest point “Parc de la Tête d'Or”
rating	Rating on a scale of 0 to 5, 5 being the best.
lat	Latitude, if available 45.7631569999999996
long	Longitude, if available 4.85960500000000017
image_url	The URL of the image representing the corresponding interest point, if available
phone	Telephone number
review_count	Number of users who rated the interest point

ETL



From the operational tables, a extraction process is applied, which creates the staging tables: history tables (for the interest points, foursquare and yelp data) and an aggregate table (for the twitter information: counts and sentiments).
This staging tables are now transformed into the DWH Schema that will contain three dimensions (Location, Time and Interest Points) and its respective Fact Table.
From the fact table, using a ROLAP approach we query and construct the cube that will allow to perform the correspondent analysis over the data.