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

Project Proposal

Smart City:

Big Data Analytics using open data of Grand Lyon, France

DMKM1517

Monday 29th of February 2016

Executive Summary

Objective

Using data provided by the open data platform of Grand Lyon, assign a score to each touristic site analysing live-stream and historic data coming from social networks.

Hypothesis

There exists a relation between livestream data sentiment and touristic places.

Goals

- To visualize real-time opinion of touristic sites in Grand Lyon
- To rank touristic sites based on their opinion ranking
- To consolidate opinion of touristic places and be able to visualize changes in time.

Solution

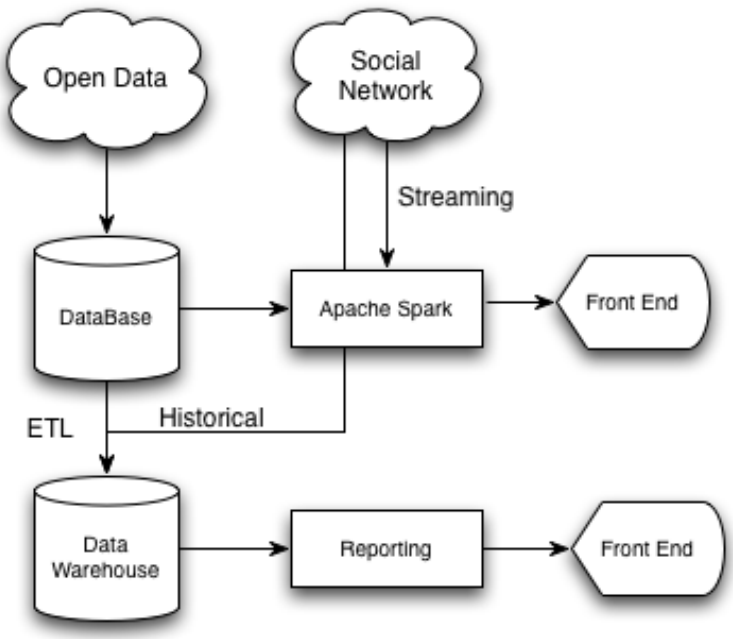
Implement, validate and deploy a platform to:

- Fetch, acquire and consolidate the data from the open data sources.
- Extract, transform and load the operational daily data to a Data Warehouse.
- Acquire, consolidate and process streaming data coming from social networks.
- Apply machine learning algorithms to determine the sentiment and/or opinion contained in social network and relate it to geospatial data of the sites.
- Visualize the analysis using web-based, front end technology

Project Roles

Business User	Project Sponsor	Project Manager	Business Intelligence Analyst	Data Engineer	Database Admin	Data Scientist
Katerina Bastrakova Fadila Bentayeb	Youakim Badr	Carlos López	Katerina Bastrakova Saul Garcia	Krishna Kalyan Jose Millan Rodney Ledesma	Jose Millan Rodney Ledesma	Carlos López Krishna Kalyan

Overall diagram and description of components



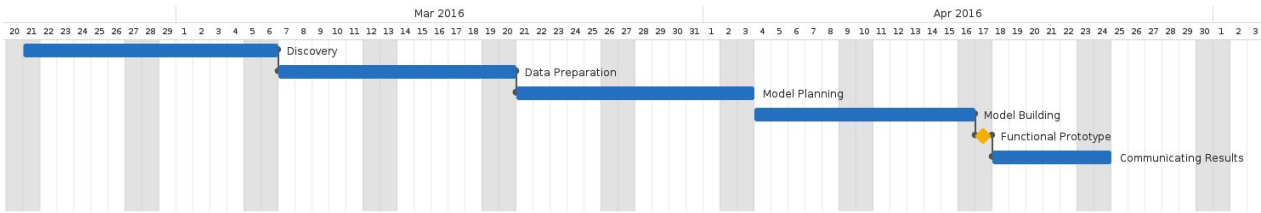
- The data sources includes several datasets coming from the open data repository of Grand Lyon. This data sources must be queried daily, looking for changes, transformed and loaded to the operational database.
- Also, social networks, such as twitter are to be queried, both in real time, and in regular basis to retrieve live and historical data.
- The operational database and the social networks must be loaded through ETL process to a Data Warehouse for further analysis.

- The analytics, such as geospatial recognition, sentiment analysis, opinion mining, topic modelling, etc, is to be carried out in real time using in-memory analytics platform Apache Spark.
- Also a custom reporting tool is to be implemented to query and summarize the Data Warehouse.
- Both tools will display their results in custom made web platforms.

Project Development

The data project is to be developed using EMC's Data Analytics Lifecycle methodology, and the code is to be developed using Agile Software Development practices.

An overview of the planning



After the first cycle, a new cycle is to be developed, with the possibility to add more functionalities on top of the existing prototype, producing the final product to be operationalized in production servers if any by June.

Deliverables

- Source code
- Documentation
- Written Report
- Test Cases