**National Research University**

**Higher School of Economics**

Faculty: Business Informatics

Department: Business Analytics

**DRAFT**

**of the paper**

**“The Development of Information System for Analyzing the Internet Resource openstreetmap.org”**

Student: N.S. Babchenko

Group: 475

Argument Consultant: Y.A. Perikov

Style and Language Consultant: K.A. Podobed

2013

**Abstract**

At the present, crowd-source-based resources have become increasingly popular in a world wide web. Every day these projects accumulate enormous amount of data during their activities. Thus, all this information has to be analyzed in order to get new valuable knowledge. This study consider one of such internet resource called OpenStreetMap, one of the most popular crowd-source project with totally over than one million of active followers who are involved in the development of it. As the main instrument of analysis will be used data warehouse technologies. As an example of knowledge, possible to be extracted from this source, in this study will be investigated the development degree of Russian regions by the number of social facilities.

Table of content

[Introduction 4](#_Toc348580829)

[Chapter 1. Theoretical base of the study 5](#_Toc348580830)

[1.1 Data warehouse base theory 5](#_Toc348580831)

[1.2 The description of ETL processes 5](#_Toc348580832)

[1.4 Introduction into OpenStreetMap 6](#_Toc348580833)

[1.4.1 Base knowledge about OpenStreetMap 6](#_Toc348580834)

[1.4.2 OpenStreetMap data structure 7](#_Toc348580835)

[Chapter 2. Practical part 7](#_Toc348580836)

[2.1 ETL process planning 8](#_Toc348580837)

[2.2 Extracting data 8](#_Toc348580838)

[2.3 Transformation of the data 9](#_Toc348580839)

[2.4 Loading data into data warehouse 9](#_Toc348580840)

[Chapter 3. Analysis of the developed data warehouse 9](#_Toc348580841)

[Conclusion 10](#_Toc348580842)

[Bibliography 10](#_Toc348580843)

# Introduction

This paper is devoted to the topic of data warehouse systems and the structure of my graduation work which is entitled “The development of Information System for Analysing Internet Resource Openstreatmap.org”.

OpenStreetMap (OSM) is non-commercial project that provides instruments to create free editable map of the world by community members. Nowadays, this project has collected a hundreds gigabytes of data, which publicly available for downloading. This feature provides enormous opportunities for data analysing and extracting valuable information.

The main purpose of this paper is to present the goal and the objectives of my graduation work, give the justification of the topic show the structure and describe the main stages of my investigation.

OpenStreetMap provides a large number of different possibilities to analyse almost all places in our Planet. That is why I decide to use these opportunities to investigate the degree of development of social construction in Russia. The aim of my study is to analyse Russian regions, comparing them by the amount of social constructions, such as hospitals, educational institutions, sports facilities and entertainment venues.

To achieve this goal I have to perform a number of tasks.

* First of all I should decide which instruments I need to use working with large amounts of data. I should analyze the most popular solutions in this field and choose the most appropriate one.
* Secondly, I need to analyze the structure of the data providing by OpemStreetMap and then, according to this information develop the plan of extracting this data and loading it into my storage system.
* The third task is to design data schema for data warehouse and to perform extraction, transformation and loading operations of the data.
* The fifth task is to analyze data in developed data warehouse. It means to design the reports which are able to provide the information about the development of social facilities in Russia.

Thus, consequently, the object of the study is an extraction of new knowledge by an analysis of OpenStreetMap internet source, by the development of data warehouse.

# Chapter 1. Theoretical base of the study

## Data warehouse base theory

Nowadays a large number of organizations around the world produce enormous amount of data during its activities. This data can be analyzed and used to get new knowledge. Data storage systems, designed to analyze large amounts of data called Data Warehouses. They are database-based systems configured for data analysis and that is why it has several particular qualities which distinguish it from transactional systems used in companies in their everyday operations.

* First of all, data in data warehouse is updated periodically instead of real time updating in transactional systems. It means that data is not transmitted to the system after adding one new record. It appears periodically for example every week. When organization accumulates certain amount of data, firstly it prepared for transaction and then uploaded into data warehouse.
* Secondly, the data is structured in such a way to ensure speed and ease of analysis.
* Thirdly, all the data related with a time. So, it enables to get the state of an object in particular time during whole the history of tracking.
* Finally, all information stored in data warehouse is unchangeable. It is impossible to update or delete data inside such information system, because the results of query made in different periods should be the same.

The most common definition for data warehouse was given by Bill Inmon. According to him data warehouse is a “subject-oriented, integrated, non-volatile, time-variant collection of data in support of management’s decisions”.

## The description of ETL processes

ETL is an abbreviation for Extraction, Transformation and Loading data. It is a process of transferring data from one information system to another. This process consists of three man stages:

* Extraction of a data. At this stag
* Transformation of a data. It means that data format of a source and destination database can be different and it is necessary to configure original data into form compatible with data warehouse. This process also includes the assurance of data quality. To ensure the quality of data used data cleaning procedure. This process involves the adjustment of data format, data encoding, exclusion of unnecessary attributes, detection of the objects which have same name, but different meanings. All the data into data warehouse must meet the fallowing requirements of data quality.
  + Data must be accurate
  + Data must be complete
  + Data must be consistent
  + Data must be unique
  + Data must be current
* Loading data into data warehouse. The main goal of this step is to provide maximum downloading speed, because of large amounts of data. According to this aim, the process of loading has several features
  + First of all, it is impossible to use standard SQL commands, because they are too much slow. This operation usually performed by the integrated tools of database management system.
  + Secondly, table indexes are slow for downloading, that is why more appropriate solution is to remove them and re-create them in data warehouse.
  + Finally, it is highly recommended to use parallelism, while loading in the same time bought dimensions and facts.

## 1.4 Introduction into OpenStreetMap

### 1.4.1 Base knowledge about OpenStreetMap

OpenStreetMap(OSM) is a crowd-source-based project, which provides open geographical data. It is a free editable map of the world, which is available for use in many applications. OSM collects data with a help of over the million of its contributors and also organisations which provides aerial photography. Community members use GPS devices, which track routes, when users drive a car, ride a bike or just walking. Then they can transmit this data to OSM server, which will publish it for open access. Another way to participate in the development of open maps is to draw them, by using aerial photography provided by Bing. There are two main editors, which provide this functionality.

* First one called “Potlatch” and it is web-based application, which the easiest to get started with.
* Another one is “JOSM”. It is java-based desktop application, which is more powerful than Potlatch and is preferred by many experienced contributors but needs to be downloaded separately and takes a bit more effort to get started with.

The data form OpenStreetMap is used in over a hundred different applications developed by and for the people from different countries. All these applications can be divided into more than sixteen categories, which include gelocation services for meeting, services for routing, history, art, archaeology, biking, sport, public transport etc.

### 1.4.2 OpenStreetMap data structure

The data in OSM is presented by four main types of objects. They are Nodes, Ways, Closed Ways and Relations. All these elements are described by Tags which have key-value format and can provide any information about an object.

* A Node is a single geospatial point which has latitude, longitude and if necessary altitude. Nodes can be used to define standalone point features or be a part of other objects such as Ways and Relations.
* A Way is an ordered list of Nodes which normally has at list one tag or can be included in a Relation. A way can be open or closed. Closed one is whose first node is also last on the way. A closed way can be interpreted either as closed polyline or as area.
  + Open polyline is a way which describes linear object, and whose first node is not connected with last one. It usually used to describe roads, streams and railway lines.
  + Closed polyline is a way where the first node is shared with last node with suitable tagging. As an example it can be used to describe ringways.
  + An Area is a closed way which has special tag and it includes also the space inside it. It can be used to describe objects like buildings or places as parks.
* A Relation is an element which consists of one or more tags or also it can be an ordered list of nodes or ways as members which is used to define logical or geographical relationships between other elements. There are several types of relations which describes by attribute “role”. An example of using relations is administrative boundaries between countries, regions and cities.

Such structure of a data provides abilities to convert it into XML format for distribution in any platform.

# Chapter 2. Practical part

My practical part is divided into three following steps.

* Extraction, transformation and loading process planning
* Extracting data from OpenStreetMap source
* Transformation of a data in a convenient form for further work
* Loading data into data warehouse

## 2.1 ETL process planning

Data transferring process is one of the most significant stages in development of data warehouse. For this reason it has to be planned well.

It consists of fallowing stages:

* First of all, it develops general plan, which includes a list of data sources systems and data destination system.
* Next step is Data Flow Diagram design which is a graphical representation transferring data through information systems.
* Third step is to determine which transformations with data needed to be done:
  + Transformation of data type such as encoding or data format.
  + Normalization or de-normalization of database scheme.
  + Transformation associated with quality assurance.
* Finally it is determined the sequence of actions during all the process.

## 2.2 Extracting data

OpenStreetMap provides an open access to the data they store. Anybody can download a single file, which contains the data about the whole Planet. This file called Planet.osm. It is in XML format, so it can be easy converted into any other format, or uploaded into database. It is a large file, which contains over 250 GB of uncompressed data and about 16 GB in a compressed form. So it is rather difficult to work with such amount of data. There are also files called Extracts which contain OpenstreetMap data for individual continents, countries and regions. There are several providers that update their Extracts every week or day or even hour. So in my study I will use one of such providers to get latest data of Russian regions.

The next task I need to perform is to upload obtained data from XML file into database. There are several ways to do it. The first way is to develop simple program which will analyse data from XML and put it into database. The other method is to use existing solution – the software designed to load data OpenStreetMap data into database. One problem is that, this program works with PostgreSQL database while in my study I use Microsoft SQLServer. So I have to add staging PostgreSQL database into my project, as temporary storage for a data. Then I will use ODBC interface to upload data from PostgreSQL into MsSQLServer database.

## 2.3 Transformation of the data

When data is uploaded into database it will have the structure same as in XML file. It will consist from nodes, ways and relations. This form is appropriate for storing geographical data, but not for analysis. So I need to transform all these nodes, ways and relation into more convenient objects as Country, Region, City and simple object like a house or a lake. But there is also one problem here. When data stores in XML format it does not support hierarchy. So to determine in what country is located any house it has to make some calculations with coordinates. More suitable solution is to organize database such way when any object has relation with a place when it locate, cities have relations with region and regions have relation with countries. After this step I suppose to get a database with a structure appropriate for further work.

## 2.4 Loading data into data warehouse

For further investigation, transformed on the previous stage data has to be downloaded into data warehouse database. At the same moment there is a possibility, that a part of data could not be uploaded, because of incompatible types or other limitations. Such data has to be stored in a separate file for further analysing and improving data quality.

# Chapter 3. Analysis of the developed data warehouse

The result of the practical part of my study will be the developed data warehouse with complete geographical information about all Russian regions. The next step of my study will be an analysis of this information.

The aim of analysis in my study is to get information about the degree of development of social construction of Russian regions.

There are 8 Federal Districts and 83 subjects of Russia today. Firstly, I am going to investigate every subject of Russia and every Federal District separately and then to compare all subjects and all Districts to each other. Of course, all Russian regions have different size and population. For this case, I plan to get this information by analysing the number of social institutions per a thousand of people.

My research will include the following categories for analysis:

* Education
* Healthcare
* Entertainment, Arts and Culture
* Sport

I am going to get the following results:

* The number of education, healthcare, entertainment and sport facilities for each subject of Russia and for each Federal District.
* The relative number of social facilities for each subject and district per a thousand of people.
* Total number and relative total number of social facilities for whole country.

# Conclusion

Based on what have been discussed, the study can draw the following conclusions.

1. OpenStreetMap as other crowd-source-based projects becomes more and more popular today. It is trance-continental resource uniting millions of users. It accumulates large amount of data every day and such information has to be analyzed to get useful knowledge.
2. Such amount of data can be analyzed just with a help of special tools. The most suitable one in this case is a creation of data warehouse. But such approach has a set of appropriate features related with database specific structure and organization of the process.

The research is not at its final stage yet. At the current moment completed preparation works, involving collection and analysis of theoretical material related to the topic. The next stage is a preparing to the practical part of a study, which includes planning of extraction, transformation and loading processes and the development of data warehouse schema. And the final stage of this investigation will be analytical part, consisting of the development of data warehouse queries, which will help to answer the questions about the degree of the social development of Russian regions.

# Bibliography

1. Проектирование хранилищ данных для систем бизнес-аналитики: учебное пособие / В.Е. Туманов. — М.: Интернет-Университет Информационных Технологий: БИНОМ.
2. Auth, G. & Maur, E. (2002). A Software Architecture for XML-Based Metadata Interchange in Data Warehouse Systems. Proceedings of EDBT 2002 Workshop on XML Data Management (XMLDM)
3. <http://wiki.openstreetmap.org/wiki/Potlatch>
4. <http://wiki.openstreetmap.org/wiki/JOSM>
5. http://wiki.openstreetmap.org/wiki/Map\_Features