Правительство Российской Федерации

Государственное образовательное бюджетное учреждение

высшего профессионального образования

«Национальный исследовательский университет
Высшая школа экономики»

Факультет Бизнес-информатики

Кафедра Бизнес-аналитики

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

На тему: Разработка информационной системы для анализа интернет ресурса

Студент группы № 475

Бабченко Н.С.

Научный руководитель

Периков Ю.А.

Москва 2013

Оглавление

[Введение 3](#_Toc357993452)

[Глава 1. Применение концепции хранилищ данных при анализе интернет ресурса. 7](#_Toc357993453)

[1.1. Постановка задачи 7](#_Toc357993454)

[1.2. Анализ геоинформационной системы «OpenStreetMap» 8](#_Toc357993455)

[Глава 2. Реализация хранилища данных 14](#_Toc357993456)

[2.1. Проектирование хранилища данных 14](#_Toc357993457)

[2.2. Разработка механизма загрузки данных в хранилище 20](#_Toc357993458)

[2.3. Построение OLAP – куба 26](#_Toc357993459)

[Глава 3. Анализ данных 29](#_Toc357993460)

[3.1. Анализ данных OLAP-средствами 29](#_Toc357993461)

[Заключение 34](#_Toc357993462)

[Список использованной литературы 36](#_Toc357993463)

Введение

В настоящее время, в различных сферах деятельности, широкое распространение получили геоинформационные системы (ГИС), которые предоставляют собой широкий класс технологий, предназначенных для сбора, хранения и анализа географических данных и связанной с ними информацией. Такого рода системы помогают дать ответы на целый ряд вопросов различной направленности. В первую очередь, ГИС отвечает на вопросы о том, какие объекты находятся в заданной области или наоборот, помогает определить местоположение тех или иных географических объектов. Во-вторых, системы, содержащие временные измерения, помогают определить, то какие изменения происходили на заданной площади в течение некоторого промежутка времени. Одним из самых ценных вкладов, систем подобного рода, является возможность решения целого ряда задач по моделированию ситуаций и анализу последствий добавления новых объектов в существующие инфраструктуры населенных пунктов. Кроме того, геоинформационные системы могут использоваться как некого рода справочники, являющиеся неотъемлемой частью аналитических систем, используемых для анализа данных привязанных к географической информации, как например результаты Выборов Президента Российской Федерации, представленных в разрезе регионов, городов и отдельных избирательных участков. Иными словами, существует немало способов применения систем данного класса.

Однако, в тех случаях, когда речь не идет о системах, используемых на государственном уровне или уровне крупных корпораций, которые в состоянии самостоятельно собирать и обеспечивать свои системы данными, серьезной проблемой является именно добыча доступной и актуальной географической информации.

Произведенный анализ наиболее популярных ГИС, показал, что данную проблему можно решить, если использовать ресурсы сервисов, предоставляющих открытый доступ к своим данным. В общем случае, существует две разновидности доступа к информации: это использование API и выгрузка данных в текстовые файлы, с последующей загрузкой их в целевую систему. В первом случае, возможно получение готовой информации по какому-либо конкретному объекту, без необходимости загрузки и анализа всей базы. Второй случай удобен, когда необходимо проанализировать большой объем информации, поскольку скорость обработки данных на собственной системе будет выше, чем многократное обращение к удаленной базе данных, обслуживающей запросы миллионов пользователей. В качестве примера, систем, предоставляющих открытый доступ к своим данным, посредством, перечисленных выше способов, можно привести три популярных интернет сервиса.

Первый из них называется «OpenStreetMap» и представляет собой некоммерческий проект, который предоставляет инструменты для создания географической карты мира. Данный проект, базируется на модели краудсорсинга, поэтому каждый может принимать участие в создании карт, и каждый может получить доступ к имеющимся данным. На сегодняшний день, данный проект предоставляет одни из самых качественных и подробных карт. Общий объем данных превышает 200 ГБ и все эти данные доступны для скачивания в свободном доступе. Отличительной чертой является то, что все данные представлены в виде простых объектов: точек, линий и, так называемых, «отношений». Таким образом вся информация о географических объектах представлена в виде, вышеуказанных простых объектов и их атрибутов. Данных подход, облегчает задачу геопозиционирования, однако добавляет трудностей при использовании данных в аналитических целях.

Второй сервис называется «MapQuest Open Platform». Он базируется на выше упомянутом сервисе, однако в целом является самостоятельным и представляет собой открытую платформу, предоставляющую API к данным «OpenStreetMap». С помощью данной системы можно получать информацию, о том какой маршрут из пункта А в пункт Б будет наиболее быстрым или наиболее коротким. Функции данного сервиса можно использовать в системы-гида, которая в качестве результата выдает информацию о том где и когда нужно повернуть чтобы попасть из точки А в точку Б. Кроме того, сервис позволяет получать статистические снимки с карты, отображающие заданную область. Одной из самых полезных возможностей является поиск объектов по географическим координатам, а так же по названию и адресу.

Последний сервис, отличается тем, что позволяет выгружать большие объемы данных в виде текстовых файлов, в которых информация хорошо подготовлена для анализа. Сервис «GeoNames», предоставляет базу данных географических объектов, дополненных полезной информацией о них. Данные собираются из более, чем 60 различных источников. Однако наиболее полные данные доступны только при покупке подписки.

В отдельных случаях, вышеперечисленных сервисов, может быть вполне достаточно для решения широкого круга задач. *Проблема* состоит в том, что, если речь идет об аналитических задачах, то данных, открыто предоставляемых «GeoNames», может оказаться недостаточно, API открытых сервисов будут слишком медленными для обработки миллионов записей, а сервисы, базирующиеся на «OpenStreetMap» требуют дополнительной обработки данных, чтобы получить возможность использовать их по назначению. Поэтому потребность в качественных и доступных данных по-прежнему остается актуальной.

*Объектом исследования* являются геоинформационные сервисы, предоставляющие открытый доступ к своим данным. Соответственно, в качестве *предмета исследования* рассматриваются, непосредственно, географические данные.

Таким образом, *целью* данной работы, является получение геоинформационной системы, позволяющей объединить преимущества всех, вышеперечисленных систем, и получить, в конечном итоге, систему способную предоставлять максимально полные и подробные данные, подготовленные для различных целей анализа.

В теоретической части данной работы описываются особенности анализируемых интернет сервисов, рассматривается концепция хранилищ данных и возможность использования хранилища в качестве ядра разрабатываемой системы. Помимо этого, приводится общая структура системы, освящаются OLAP-технологии анализа. В практической части описываются особенности реализации системы, приводятся результаты ее работы на примере наиболее актуальных задач.

# Глава 1. Применение концепции хранилищ данных при анализе интернет ресурса.

## Постановка задачи

Система должна удовлетворять следующему ряду требований. Во-первых, данные должны выгружаться из открытых источников. В третьих, данные должны быть всегда актуальными, а значит система должна выполнять регулярные обновления из источника. В четверых, в системе должен быть заложен механизм хранения истории изменений данных связанный со временем. И последнее – данные должны быть структурированы для аналитической обработки и должны описывать конкретные географические объекты.

Для достижения заявленной цели, необходимо последовательно решить следующие задачи:

1. Анализ предметной области;
2. Проектирование хранилища данных;
3. Построение хранилища данных;
4. Описание источника данных, разработка механизма переноса данных;
5. Подготовка данных для анализа (создание OLAP-куба);
6. Просмотр срезов OLAP-куба;
7. Разработка необходимых отчетов;
8. Анализ данных с использованием системы.

В качестве инструментальных методов исследования были выбраны программные продукты корпорации Microsoft. В качестве целевой СУБД для хранилища данных выбрана Microsoft SQL Server 2005; построение OLAP-куба и разработка механизма переноса данных выполнены в среде Microsoft SQL Server Business Intelligence Development Studio. Для проектирования хранилища использован AllFusion ERwin Data Modeler 7 (ERwin), для построения отчетов – Crystal Reports

## 1.2. Анализ геоинформационной системы «OpenStreetMap»

OpenStreetMap(OSM) – это интернет проект, базирующийся на модели крауд-сорсинга, который предоставляет географические данные в свободном доступе. Основным элементом данного сервиса является географическая карта мира, в создании которой может принимать участие любой желающий.

Большая часть данных собирается силами участников сообщества, однако свой вклад вносят и некоторые крупные организации, которые предоставляют, например, аэрофотоснимки. Участники сообщества используют GPS устройства, которые отслеживают и запоминают, проделанные ими, маршруты, которые проделал пользователь, управляя автомобилем, катаясь на велосипеде или просто прогуливаясь по улице. Далее, сохраненные данные выгружаются на сервер OSM, где после некоторой обработки и модерации они публикуются на карте и становятся общедоступными.

Кроме этого, есть и другие способы принимать участие в разработке OSM карт. Участники сообщества, могут рисовать карты, используя аэрофотоснимки, представляемые поисковой системой “Bing”.

На данный момент, существует два основных редактора карт, которые предоставляют эту функцию. Первый из них называется "Potach". Это веб-приложение, доступ к которому осуществляется через веб-браузер. Плюсом данной системы является, то что оно позволяет быстро освоить процесс разработки карт. Другой, не менее популярный редактор называется "JOSM". Это приложение для настольных компьютеров, написанное на языке “Java”, которое является более мощным, чем Potlach, поэтому его предпочитают многие опытные участники сообщества. Однако данных продукт требует скачивания и установки. Кроме того, он требует значительно больше усилий, для того чтобы начать с ним работать.

Данные, предоставляемые OpenStreetMap, используется в сотне различных приложениях, разработанных для людей из разных стран. Все эти приложения могут быть разделены на шестнадцать категорий, которые включают в себя услуги геолокации для встреч, услуги для маршрутов, истории, искусства, археологии, велосипедный спорт, спортивные площадки, общественный транспорт и т.д.

В данном исследовании, наибольший интерес представляет структура, которую используют разработчики OSM для хранения данных в системе. Таким образом, вся информация в системе представляется в виде трех основных видов объектов: «Node» или «Узлы», «Way» или «Путь» и «Relation» или «Отношение». Каждый из этих объектов может иметь неограниченное число атрибутов, которые помогают описывать конкретные географические объекты. Далее будут, более подробно, описаны все представленные элементы.

«Узел» представляет собой единственную геопространственную точку, которая описывается широтой, долготой и при необходимости, высотой. «Узлы» могут быть использованы как для описания одиночных объектов, так и быть частью других объектов: «Путей» или «Отношений». Наиболее простым примером демонстрирующим использование «Узлов» является описание небольших объектов, как например печатных киосков или автобусных остановок. Зачастую «Узлами» отмечаются центральные точки объектов, имеющих большую площадь, при этом контур объектов описывается при помощи «Путей» или даже «Отношений». Таким образом, один географический объект в системе может быть описан несколькими способами. Это относится к крупным зданиям, или административно-территориальным делениям, когда для упрощения поиска используются координаты центральной точки, а для описания контура объекта используются полилинии.

«Путем» называется сортированный список «Узлов», который обычно имеет список тегов. Другими словами, «Путь» можно описать как полилинию, состоящую из множества точек. «Путь» может входить в состав «Отношений». Выделяют две разновидности объектов подобного вида, то есть они могут быть открытыми или закрытыми. Закрытыми называют те, в которых первый узел, является так же и последним. Таким образом, закрытые пути описывают некоторые закрытые участки местности. Причем, в одном случае, такая полилиния может описывать только внешние границы объекта, а в другом может включать так же и внутреннюю область. Очень часто «Пути» используются для описания различного рода дорог: автомобильных, железнодорожных, пешеходных путей. Кольцевые автодороги или круговые перекрестки могут быть описаны при помощи закрытых «Путей», при этом внутреннее пространство не учитывается. В тех случаях, когда конец дороги не замыкается со своим началом, для описания могут использоваться «Пути» открытого вида. Помимо описания транспортных объектов, «Пути» могут использоваться для описания контуров зданий и сооружений или различных участков местности как, например, парки, леса, площади, населенные пункты и т.д.

Под «Отношениями» подразумевается элемент, который имеет один или более, описывающий его тэг, а так же содержит список элементов входящих в его состав из числа «Узлов», «Путей» или других «Отношений». «Отношения» необходимы, для того чтобы определять логические или географические взаимоотношения между элементами, входящими в их состав. Для определения типа взаимоотношений используется атрибут «роль». Выделяют четыре вида ролей:

* Маршрут, используется для описания различных маршрутов для передвижения пешеходов, велосипедистов и различного рода транспорта.
* Мультиполигон, используется для описания обширных участков местности имеющих сложную структуру, в частности в тех случаях когда внутри некоторой области имеются участки не принадлежащие данной области. Использование просто закрытого «Пути» не поможет описать такой объект, так как не вся внутренняя область будет в него входить.
* Граница, используется для описания административных или исторических границ.
* Ограничения, используются в основном для описания некоторых маршрутов. Они описывают правила движения на заданном участке дороги. В частности указывают такие места, где нельзя повернуть направо или налево или выполнить разворот.

На рисунке 1 изображена часть логической модели данных «OpenStreetMap», показывающая отношения между элементами, которые используются для описания географических объектов. На представленной диаграмме наименования сущностей расшифровываются следующим образом:

* «Relation» – элемент «Отношение»
* «OldRelation» - используется для описания предыдущих версий элементов «Отношение»
* «Relation member» - используется для описания элементов, входящих в состав отношений
* «Way» - элемент «Путь»
* «OldWay» - используется для описания предыдущих версий элементов «Путь»
* «Node» - элемент «Узел»
* «OldNode» - используется для описания предыдущих версий элементов «Узел»
* «Way\_Node» - используется для описания элементов «Узел», входящих в состав элемента «Путь»



Рисунок 1. Логическая модель данных "OpenStreetMap"

 Как видно из представленной схемы, все связи являются неидетифицирующими. Так, например, связи между элементами и их атрибутами не требуют однозначной идентификации по номеру элемента, поскольку один и тот же элемент может описывать множеством тегов, и эти теги не обязательно должны различаться. Сущности, описывающие предыдущие версии элементов, так же не нуждаются в однозначной идентификации по номеру элемента, так как элемент может иметь множество «старых» версий.

 В целом подобная модель организации данных удобна для решения задач геопозиционирования, которые и являются основной цельной создания системы «OpenStreetMap». Кроме того, подобная структура позволяет легко экспортировать данные в формате XML файлов и таким образом использовать накопленные данные в любой другой системе. Однако, следует заметить, что такая структура не является удобной при решении аналитических задач, требующих возможность оперирования реальными географическими объектами. Одной из ключевых проблем, с которой можно столкнуться, при попытке анализа данных является то, что в таком формате данных, крайне сложно, без дополнительных вычислений определить иерархию объектов. В таком случае, становится проблематично, определить в каком городе, в каком регионе и в какой стране находится конкретное здание. Именно поэтому, одной из задач данного исследования будет решение данного вопроса.

# Глава 2. Реализация хранилища данных

## 2.1. Проектирование хранилища данных

Известно, что наиболее распространенными моделями при проектировании хранилища данных являются схема «звезда» и схема «снежинка», где модели измерений данных созданы не по третьей нормальной формой, а с учетом их тематики, характеризующую отдельную сущность предметной области. Основная цель пространственной модели заключается в минимизации времени выполнения запроса, и для этой цели схема «звезда» наиболее предпочтительна, чем схема «снежинка», поскольку «звезда» характеризуется большей денормализацией, когда таблицы измерений напрямую связаны с таблицей фактов и никаких консольных таблиц не используется.

 Схема «снежинка» (в схеме используются консольные таблицы, которые связаны с таблицей фактов через другие таблицы измерений) позволяет сокращать избыточность в таблицах измерений. Одним из преимуществ данной схемы является более быстрое выполнение запросов следующего типа: «выбрать все строки из таблицы размерностей на определенном уровне», которые часто используются при анализе данных и могут задерживать ход анализа. Плюсом также является возможность иметь таблицы фактов с разным уровнем детализации. И, конечно, данная схема позволяет сэкономить дисковое пространство.

Описываемая предметная область не отличается высокой сложностью структуры. Присутствуют лишь одна иерархия в виде дерева адресов, а так же одна консольная, описывающая класс, которому принадлежит тип географического объекта. Ввиду данных особенностей предметной области, для реализации многомерного представления данных с помощью реляционных таблиц была выбрана схема «Снежинка», которая обеспечит максимальную скорость выполнения запросов.

В рассматриваемой предметной области, в качестве факта было выбрано описание состояния географического объекта. Вся информация, описывающая характеристики географического объекта содержится в таблицах измерений, связанных с таблицей фактов радиальными связями. Всего было выделено восемь таблиц размерности и одна консольная таблица. На рисунке 1 представлена модель хранилища данных, описанного в нотации «Dimensional».

Стоит отметить, что таблица фактов содержит первичные ключи таблиц измерений. Все первичные ключи в таблицах размерности являются искусственными и не несут никакой смысловой нагрузки, так как используются исключительно в роли уникального индикатора. Все измерения можно охарактеризовать как «медленно изменяющиеся», поскольку информация, описывающая состояние объектов может со временем изменяться. Для того чтобы эффективно хранить всю историю изменений в каждом измерении предусмотрены поля, определяющие время, на момент которого запись была актуальна. Далее рассмотрим более детально каждую из таблиц.



Рисунок 2 Модель хранилища данных в нотации Dimensional

 **«DIM\_ADR» -** измерение, описывающее *адрес географического объекта.* Данное измерение описывает одну из основных характеристик любого географического объекта, его местоположение. Измерение описывает полный адрес от страны до улицы, на которой находится объект. В том случае, когда объект не имеет какой-либо составляющей адреса, поля заполняются пустыми значениями. В состав таблицы входят следующие поля:

* *«AdressKey»(int)* - уникальный идентификатор адреса
* *«Street»(nvarchar(255))* – название улицы, к которой относится объект.
* *«City»(nvarchar(255))* – название города, в котором располагается объект.
* *«County»(nvarchar(255))* – для Российской Федерации, это название области в которой находится объект
* *«State»(nvarchar(255))* – для Российской Федерации, это название округа в котором находится объект.
* *«Country»(nvarchar(255))* – название страны, в которой располагается объект.
* *«Country code»(nvarchar(2))* – двузначный код страны.
* *«PostCode»(int)* – почтовый индекс.
* *«ValidFromDTTM»(datetime)* – дата начала действия записи.
* *«ValidToDTTM»(datetime* )– дата окончания действия записи.

**«DIM\_TYPE» -** справочник типов географического объекта. В базе данных «OpenStreetMap» выделяется порядка 180 классов и более 900 типов объектов. Класс представляет собой некоторое смысловое объединение типов. Например, класс «Медицина» включает в себя такие типы объектов, как больницы, стоматологии, аптеки и ветеринарные клиники. Класс «Образование» включает в себя детские сады, школы, колледжи и университеты. Измерение содержит следующие поля:

* *«Id»(int)* - уникальный идентификатор типа объекта.
* *«ClassId»(int)* – внешний ключ на таблицу, описывающую класс, которому принадлежит данный тип.
* *«Name»(nvarchar(255))* - название типа
* *«Description»(nvarchar(255))* - описание типа

**«DIM\_CLASS»** - справочник классов географических объектов. Является родительской таблицей для справочника «DIM\_TYPE». Таблица включает в себя следующие поля:

* *«Id»(int)* - уникальный идентификатор класса объекта.
* *«Name»(nvarchar(255))* - название класса
* *«Description»(nvarchar(255))* – описание класса.

 **«DIM\_GEOGRAPHY» -** измерение, характеризующее объект с точки зрения географических координат. В данной таблице хранятся объекты специального класса SQL Server для описания географических объектов: «SQLGeography». Данный класс представляет данные в геодезических координатах, что облегчает процедуру хранения множества географических координат объекта, а так же дальнейшие манипуляции с данными. Таблица включает в себя следующие поля:

* *«GeographyKey»(int)* - уникальный идентификатор
* *«Point»(geography)* – численность населения
* *«Line»(geography)* – численность населения
* *«Polygon»(geography)* – численность населения
* *«ParentKey»(int)* – ссылка на родительский объект, в который входит данный объект
* *«ValidFromDTTM»(datetime)* – дата начала действия записи.
* *«ValidToDTTM»(datetime* )– дата окончания действия записи.

**«DIM\_NAME»** - измерение, характеризующее название объекта. Таблица включает в себя следующие поля:

* *«NameKey»(int)* - уникальный идентификатор названия объекта
* *«Name»(nvarchar(255))* – краткое название объекта
* *«FullName»(nvarchar(255))* – полное название объекта
* *«ValidFromDTTM»(datetime)* – дата начала действия записи.
* *«ValidToDTTM»(datetime* )– дата окончания действия записи.

**«FCT\_GEO\_OBJECT»** - основная таблица фактов, характеризующая состояние некоторого географического объекта. Все поля таблицы представляют собой внешние ключи на таблицы измерений:

* *«NameKey»(int)* – внешний ключ на измерение «DIM\_NAME»
* *«TypeKey»(int)* – внешний ключ на измерение «DIM\_Type»
* *«AdressKey»(int)* – внешний ключ на измерение «DIM\_Adress»
* *«GeographyKey»(int)* – внешний ключ на измерение «DIM\_Geography»
* *«Importance»(float)* – показатель, характеризующий важность объекта.
* *«Place\_Rank»(int)* – показатель, характеризующий ранг объекта
* *«Area»(float)* – показатель, характеризующее площадь объекта. В качестве единиц измерения площади принимается квадратный метр.
* *«Levels»(int)* – показатель, характеризующий этажность здания
* *«Population»(float)* – показатель, характеризующий численность населения в заданной области.

В соответствии с разработанной структурой, было сгенерировано хранилище данных на MSSQLServer сервере. Код процедуры, выполняющей создание хранилища, представлен в приложении.

## Разработка механизма загрузки данных в хранилище

Геоинформационный сервис «OpenStreetMap» предоставляет свои данные в виде выгрузок в файлы формата XML. Существует несколько способов выгрузки данных. Первый способ заключается в использовании непосредственно сайта openstreetmap.org, на котором расположен основной сервис системы – географическая карта. На данном сайте расположен раздел «Экспорт». Этот раздел представляет собой интерфейс, при помощи которого пользователь может выделить необходимую ему область карты и скачать ее в виде файла формата XML. Другой способ заключается в то, чтобы использовать готовые выгрузки карт, которые формируются ежедневно. Непосредственно «OpenStreetMap», предоставляет выгрузки только всей карты мира, виде фала с названием «OSMPlanet», который занимает порядка 250 ГБ пространства на жестком диске. Однако, конечным пользователям нет необходимости, скачивать выгрузку всей карты. Существует множество сервисов, которые занимаются тем, что вырезают из общего файла отдельные блоки, соответствующие конкретным странам, регионам и городам мира. В данном исследовании будет использован именно такой подход. В качестве источника выгрузки, будет использоваться сервис «GIS-LAB.info», который предоставляет «вырезки» из файла «OSMPlanet» по Российской Федерации, ее Федеральным округам, областям и городам.

Для того, чтобы заполнить, разработанное хранилище данных необходимо разработать механизм загрузки, который позволит автоматизировать данный процесс. В качестве инструмента, используемого для реализации подобного механизма, была выбрана служба SQL Server Integration Services среды Microsoft SQL Server Business Intelligence Development Studio. Стоит отметить, что в настоящее время на рынке присутствует множество ETL средств, отличающихся, прежде всего, набором инструментов. В частности наиболее популярными программными продуктами являются «SAS Data Integration Studio», а так же пакет программных средств, для интеграции данных от компании «Informatica». Выбор инструментов от компании Microsoft был обоснован прежде всего богатым набором инструментов для работы с данными, выгружаемыми из файлов, в частности крайне простой и удобный механизм работы с XML. Одним из решающих факторов в выборе инструментов была возможность использования специального класса «SqlGeography», который используется SQL Server. Данный класс позволяет оперировать географическими координатами как объектами типа «Точка», «Линия» или «Полигон», что существенно облегчает вычисления и процесс написания SQL процедур.

Весь процесс загрузки данных разделен на два этапа:

* Во время первого этапа, все данные, в полном объеме загружаются в базу данных без каких либо изменений. Это необходимо для того, чтобы впоследствии облегчить и ускорить более трудоемкие операции, связанные с вычислениями.
* Второй этап подразумевает, непосредственную загрузку данных в хранилище данных. Большая часть этого процесса будет проходить с использованием разработанных SQL процедур, которые производят необходимые вычисления и заполняют справочники и измерения

.

Типичный процесс загрузки данных состоит из следующих этапов:

* Исполнение SQL процедуры, создающей необходимые таблицы в базе данных, в том случае, если они не были созданы до этого. Кроме того возможна очистка таблицы, если это требуется, например при обновлении таблиц-справочников.
* Подключение к источнику: файлу с данными или базе данных.
* Сопоставление полей источника и таблицы назначения.
* Произведение необходимых преобразований с данными.
* Подключение к объекту назначения, то есть к таблице в базе данных, в которую будет производиться выгрузка.
* Выполнение процесса выгрузки данных и перенаправление ошибок выполнения в отдельный файл, для последующего анализа.

На рисунках 3 и 4 представлен процесс загрузки данных из XML файла в базу данных. Пакет использует три вида объектов: Источник – XML файл, инструмент преобразования данных и назначение. Сначала данные, при помощи XML схемы считываются из файла. В файле, все данные представлены в следующем виде:

<?xml version='1.0' encoding='UTF-8'?>

<osm version="0.6" generator="Osmosis 0.41-44-g988e1a9">

<node id="1556926787" version="1" timestamp="2011-12-21T10:25:45Z" lat="45.0514773" lon="38.9548741">

 <tag k="highway" v="crossing"/>

 </node>

<way id="4528185" version="4" timestamp="2011-10-06T21:40:16Z"

 <nd ref="28077048"/>

</way>

<relation id="3009" version="2" timestamp="2008-03-12T17:05:57Z"

 <member type="way" ref="4530679" role="inner"/>

 <tag k="type" v="multipolygon"/>

 </relation>

 </osm>

То есть сначала, в теле документа перечисляются в «Узлы», затем все «Пути» и в конце список всех «Отношений». Каждый из этих тегов может иметь дочерние элементы. Для узла это только атрибуты, заключенные в тег *<tag/>*. «Путь» может включать как атрибуты, так и ссылки на «Узлы». «Отношение» может включать атрибуты и члены, входящие в состав отношения, которые могут быть любыми элементами: «Узел», «Путь» или другое «Отношение».



Рисунок 3 Процесс загрузки данных из XML файла в БД Часть 1



Рисунок 4 Процесс загрузки данных из XML файла в БД Часть 2

******

Рисунок 5 Сопоставление полей источника и целевой таблицы

## Построение OLAP – куба

 При построении OLAP – куба необходимо выполнение следующих шагов:

 1) **определение источника данных (Data Source)**

 Так, в качестве источника данных используется созданное нами ранее хранилище данных на SQL сервере.

 2) **определение среза реляционной базы данных (Data Source View)**

 В качестве среза реляционной базы данных понимаются таблицы, которые необходимы для заполнения куба и с которыми мы будем непосредственно работать в дальнейшем. В данной работе используются все таблицы, представленные в хранилище данных.

 3) **выделение размерностей данных**

 С точки зрения своих возможностей размерности могут быть:

* регулярными (Regular)
* из таблицы фактов (Fact Dimension)
* ссылочными (Reference)
* многие-ко-многим (Many-to-Many)

Для создания куба, прежде всего, необходимо выделить регулярные размерности, которые определяют регулярные отношения между членами размерности и показателем. Они представляют собой внешние ключи, которые ссылаются на таблицу фактов.

Так, таблица фактов «FCT\_Geo\_Object» содержит внешние ключи на восемь таблиц измерений, которые могут быть представлены в качестве регулярных размерностей (таблицы: Dim\_Geography, Dim\_Levels, Dim\_Adr, Dim\_Name, Dim\_Population, Dim\_Area, Dim\_Length, Dim\_Class, Dim\_Type).

Вся информация касательно содержимого таблиц была представлена выше при описании полей таблиц. Необходимо только отметить, что для анализа в размерности DIM\_ADR была создана иерархия «Адреса», которая состоит из следующих элементов: Country (Страна), State (Округ), County (Область), City (Город), Street (Улица).



Рисунок 6 Иерархия "адреса"

После того как все измерения и сам куб были развернуты на сервере, становится возможным протестировать его функционирование. Для того чтобы сделать это, необходимо извлечь тестовый срез куба. Используем измерение «DIM\_ADR» (адреса) и иерархию по городам. Кроме того для среза используем показатели таблицы фактов: площадь, численность населения, важность объекта, ранг объекта. Результат представлен на рисунке 7.

#

Рисунок 7 Срез куба

# Глава 3. Анализ данных

## 3.1. Анализ данных OLAP-средствами

Прежде всего, стоит отметить, что в данной работе не ставилась цель проведения детального анализа какой-либо конкретной области. В данной главе, будут сделаны попытки, демонстрации возможностей построенной системы, на примере нескольких регионов России. В качестве инструмента проведения анализа, будет использоваться среда инструмента SQL Server Business Intelligence Development Studio, с помощью которого производилось разворачивание целевого куба. В качестве задач анализа, будут поставлены следующие:

1. *Проанализировать количество объектов в заданных регионах России по шести областям:*
	* Образование
	* Развлечения
	* Финансы
	* Здравоохранение
	* Общественное питание
	* Транспорт
2. Анализ сетей быстрого питания в Москве и Московской области из числа наиболее популярных отечественных и зарубежных компаний.

Прежде чем приступить к анализу, необходимо определиться с тем, что включают в себя исследуемы области:

* *Образование* включает в себя все детские сады, школы, заведения среднего-профессионального образования, высшие учебные заведения и т.д.
* К развлечениям относятся театры, кинотеатры, ночные клубы, центры творчества и т.д.
* Под финансами подразумеваются банки, банкоматы и обменные пункты валют
* Здравоохранение – это больницы, поликлиники, аптеки, ветеринарные клиники и т.д.
* Общественное питание – бары, рестораны, кафе, точки быстрого питания.
* Транспорт – парковки для автомобилей и велосипедов, автозаправки, автомойки, автобусные станции и т.д.

Для проведения анализа необходимо использовать измерение «DIM\_Domain» (Направление), измерение «DIM\_ADR» (адреса) для выбора необходимых регионов, а так же агрегированный показатель количества объектов.

В результате, были получены данные, которые показали, что больше всего в заданных регионах располагается объектов образования и пунктов общественного питания. Меньше всего объектов по направлению культуры и отдыха. Наибольшее количество объектов располагается в Москве и Московской области. На рисунках 8 и 9 представлены результаты выполнения данных запросов.



Рисунок 8 Количество объектов различных направлений по четырем регионам России



Рисунок 9 Количество объектов различных направлений для каждого из выбранных регионов

 Следующая задача для анализа состоит в том, чтобы выяснить, какие из сетей общественного питания наиболее широко представлены в Москве и Московской области. В анализе принимали участие следующие компании: «KFC», «Subway», «Бургер Кинг», «Макдоналдс», «Теремок», а так же другие сети, слабо представленные данном регионе. Для проведения анализа, необходимо было воспользоваться измерениями «DIM\_ADR» для определения заданных регионов, а так же измерением «DIM\_Name» для определения названия объекта. Кроме того, использовался агрегированный показатель количества объектов. Результаты анализа показали, что наиболее широко представлены сети компаний «KFC» и «Макдоналдс». Кроме того, в заданном регионе сформирована сеть общественных столовых, которые так же занимают весомую долю среди предприятий общественного питания.



Рисунок 10 Количество заведений общепита в Москве и Московской области

 Следует так же отметить, что проводимый анализ основывается на данных, географической карты, подробность которой зависит во многом от активности участников сообщества в заданном регионе. Исходя из этого, нужно заметить, что результаты являются лишь косвенными отражениями реальной картины по причине неполноты данных.

# Заключение

В рамках данной работы было разработано хранилище данных для хранения информациии, связанной с геоинформационным сервисои «OpenStreetmap». Данные, составившие его основу, описывают ключевые характеристики географических объектов таких как страны, регионы и горорда, а так же целый ряд различных типов объектов из направллений образование, медицина, развлечение, транспорт, общественное питание. Каждый объект может быть описан, в зависимости от своего типа и класса, по таким характеристикам как площадь и численность населения для объектов, представляющих нанеселенные территории. Различного рода здания описываются широтой, долготой, свойе этажностю и прочими характеристиками. Тем не менее, данные в хранилище не являются исчерпывающими, поскольку детальность информации напрямую зависит от уровня детальности карты, на основе которой извлекались данные. Основной технической задачей было преобразование структуры данных, в которой информация хранилась на источнике, более подготовленная для решения задач геопозиционирования, в структуру, позволяющую решать аналитические задачи, оперируя реальными геграфическими объектами и их свойствами. Для автоматизации процесса загрузки, была разработана библиотека классов на яыке программирования C#. Это было необъодимо, по той причине, что процесс загрузки был разбит на несколько этапов, позволяющих получить более качественные данные. Однако, основыне операции выгрузки данных из источника и конечная загрузка данных в хранилище данных, осуществлялись средствами среды Microsoft SQL Server Business Inteligence Studio.

После того как было разработано хранилище данных, на основе его данных был сформирован OLAP-куб, демонстрирующий аналитические возможности разработанной системы.

Далее стоит выделить основные направления для дальнейшего развития разработанной системы. В первую очередь необходимо повышать качество и полноту данных за счет интерграции сисемы с другими геоинформационными сервисами. Для повышения аналитической ценности информации, возможна привязка географических данных к данным в различных предметных областях. Так например итеграция системы с данными о стоимости покупки/аренды жилых и не жилых помещений позволит реализовать систему по быстрому поиску наиболее выгодных предложений. Еще одним направлением соверешенствования системы, может послужить развитие ее в качестве инструмента моделирования последствий строительства новых соорежений в существующей инфраструктуре. Для это может потребоваться потребоваться интеграция со статистическими даннми о движении потоков транспорта и т.д. В целом, стоит отметить, что существует достаточно широкий круг направлений дальнейшего развития системы, однако первоначальная задача состоит в том, чтобы обеспечить максимальную полноту и качество данных.

# Список использованной литературы

1. Федоров А., Елманова Н. Введение в OLAP-технологии Microsoft. – М.: Диалог-МИФИ, 2002. – 272 с.