УДК 004.6:61

Использование OLAP-технологий для анализа данных о заболеваемости COVID-19

Шульдова Светлана Георгиевна, Змеева Юлия Викторовна
Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь
Минский инновационный университет, г. Минск, Беларусь
zmeeva2004@mail.ru

Аннотация

Статья посвящена решению прикладной задачи многомерного анализа данных о заболеваемости COVID-19 на основе OLAP-решения компании Microsoft. В качестве исходных данных использованы наборы данных, находящиеся в свободном доступе. Спроектирована витрина данных в среде Microsoft SQL Server Management Studio, разработан и реализован ETL-процесс средствами службы SQL Server Integration Services, построен и развернут куб в SQL Server Analysis Services, сформированы MDX-запросы. Визуализация результатов анализа выполнена с помощью MS Excel и служб SQL Server Reporting Services.

Ключевые слова: *OLAP-технология, многомерный анализ, куб, измерения, меры, COVID-19.*

Многие теоретические и прикладные задачи научных исследований сводятся к анализу многомерных данных, что обуславливается многомерностью описания процессов в различных предметных областях.

Технологии оперативной аналитической обработки многомерных данных (On-line Analytical Processing, OLAP) сформировались в середине 1990-х г. прошлого века, и за прошедшие годы в этой области были получены значительные фундаментальные и прикладные результаты [1].

В основе OLAP-технологии лежит понятие «гиперкуба» или «многомерного куба данных», в ячейках которого хранятся анализируемые данные, количественно характеризующие процесс, – меры (Measures). Осями многомерной системы координат служат основные атрибуты анализируемого процесса – элементы измерений (Dimensions). Вдоль каждой оси атрибуты могут быть организованы в виде иерархий, представляющих различные уровни их детализации. Благодаря такой модели данных пользователи могут формулировать сложные запросы, генерировать отчеты, получать подмножества данных. В качестве одного из измерений используется время.

Процесс OLAP-анализа представляется совокупностью операций с многомерными данными – детализации, консолидации (группировки), формирования среза и поворота. Операции консолидации определяют переход от детального представления данных к агрегированному, а в случае детализации осуществляется обратный переход. Формирование среза куба заключается в фиксации значения (значений) определенного измерения, при этом сокращается размерность куба. Срез представляет собой подкуб, в который входят все остальные измерения. Операция поворота заключается в изменении положения осей куба – измерений. В результате вращения меняется «точка зрения» на данные.

На основе OLAP-технологии выполнен анализ данных о заболеваемости COVID-19, доступных в открытых источниках [2]. Для реализации использовано решение компании Microsoft – SQL Server Analysis Services.

Первая задача, которую необходимо было решить, – создание витрины данных, позволяющей хранить данные и автоматизировать трудоемкий процесс сбора и обработки информации.

Исходные наборы данных содержат следующие сведения:

– о пациентах: пол, возраст, наличие хронических заболеваний и их перечень, проживание в Ухане;

– локации: страна, провинция, город;

– датах возникновения симптомов, госпитализации, подтверждения наличия коронавируса, окончания заболевания;

– болезни: симптомы, исход (выздоровление или смерть пациента).

В результате исследования наборов данных сформированы задачи анализа и определены измерения и меры. Задачи анализа:

– оценка географического распространения заболеваемости COVID-19;

– оценка влияния пола и возраста пациента, наличия хронических заболеваний на продолжительность и исход болезни;

– оценка влияния симптомов на продолжительность и исход болезни.

Схема витрины данных – «снежинка» – содержит измерения географии и времени, измерения с характеристиками пациентов и симптомами болезни, ссылочное измерение о хронических заболеваниях пациента и таблицу фактов, в которой зафиксирован каждый случай заболевания и его исход.

Для процедуры загрузки данных в хранилище используется процесс ETL (Extract, Transform and Load), реализация которого осуществлена средствами служб SQL Server Integration Services (SSIS).

Необходимо отметить, что данная процедура является наиболее трудоемкой и ответственной, а также наиболее затратной по времени.

ETL-пакеты создаются в среде MS Visual Studio в виде проекта. Поток управления, представляющий собой последовательность задач, выполняемых пакетом, определяет общий ход функционирования всего пакета. Наиболее используемая задача в потоке управления – задача потока данных, предназначенная для извлечения данных из источников, а затем преобразования и загрузки данных в целевые таблицы витрины данных. Спроектированный управляющий поток содержит три последовательные задачи потока данных – две задачи загрузки данных в измерения, а затем задачу загрузки в таблицу фактов. Следует отметить, что данные в исходных наборах потребовали использования большого количества преобразований с целью их «очистки» и унификации. Также в процессе загрузки создан атрибут для категории возраста по классификации Всемирной организации здравоохранения.

Разработка проекта SQL Server Analysis Services в среде Visual Studio осуществляется в несколько этапов:

1) подключение к источнику данных (витрина с загруженными данными);
2) формирование представления источника данных;
3) проектирование измерений и создание иерархий;
4) построение куба, включая вычисляемые элементы;
5) развертывание куба;
6) проектирование и выполнение MDX-запросов.

Были созданы естественные иерархии времени и географии, а также иерархия для пациента пол – возрастная группа – наличие хронических заболеваний. На рисунке 1 показана структура куба. Развернутый куб представлен на рисунке 2 с результатами выполнения запроса по географическому распространению вируса.

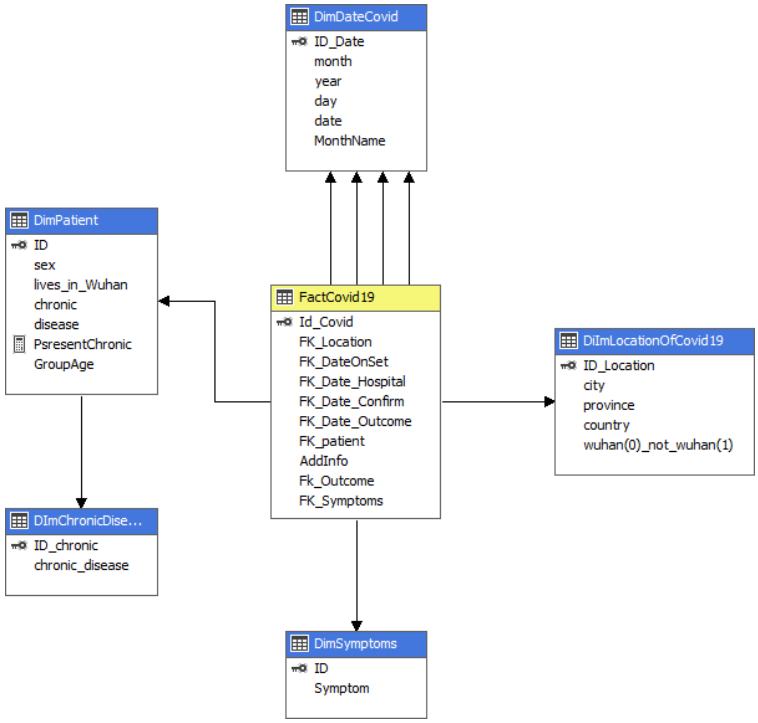
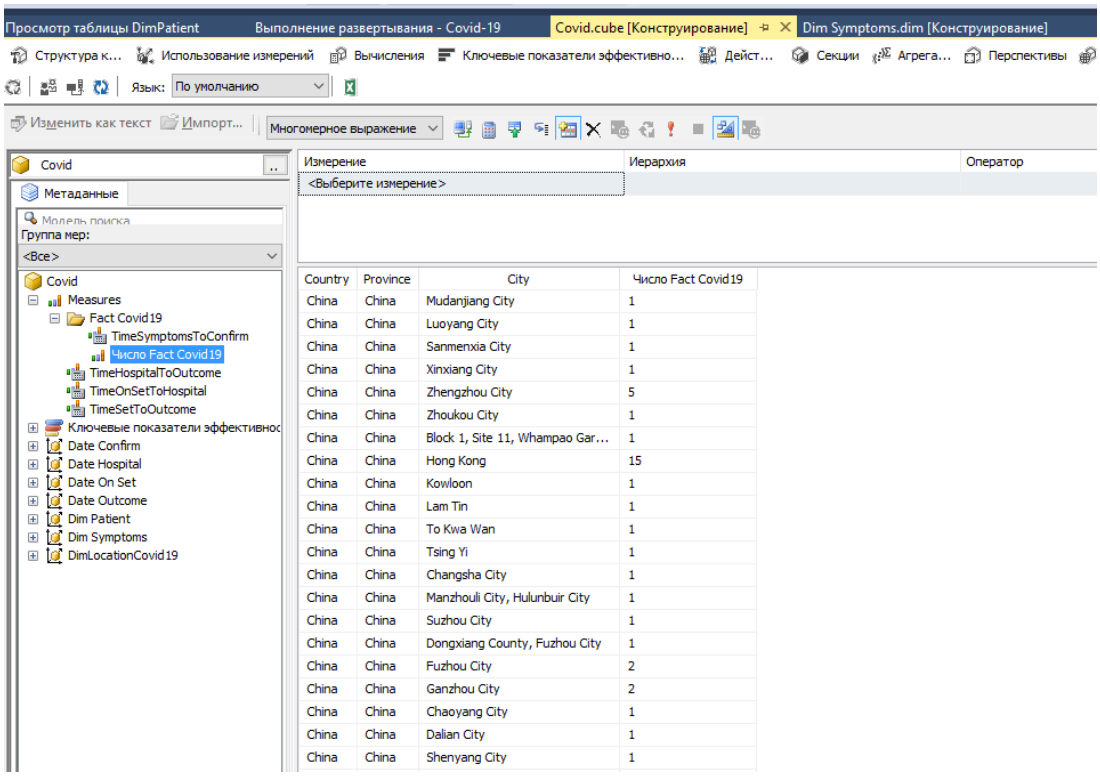


Рисунок 1 – Структура куба

Для оценки продолжительности болезни и влияния на нее различных факторов были созданы вычисляемые меры: периоды времени между датами появления симптомов, подтверждения коронавируса, госпитализации и окончания заболевания.

В качестве OLAP-клиента используется MS Excel. В книгу MS Excel импортирован развернутый куб, построена сводная таблица, созданы срезы.

Исходные наборы данных содержат WKT-координаты точки фиксации каждого случая заражения. Для визуализации географического распространения вируса был создан проект SQL Server Reporting Services, в котором, в отличие от SQL Server Analysis Services, поддерживаются геопространственные данные. Созданный отчет содержит цветовую карту распространения вируса COVID-19.



Country	Province	City	Число Fact Covid19
China	China	Mudanjiang City	1
China	China	Luoyang City	1
China	China	Sanmenxia City	1
China	China	Xinxiang City	1
China	China	Zhengzhou City	5
China	China	Zhoukou City	1
China	China	Block 1, Site 11, Whampao Gar...	1
China	China	Hong Kong	15
China	China	Kowloon	1
China	China	Lam Tin	1
China	China	To Kwa Wan	1
China	China	Tsing Yi	1
China	China	Changsha City	1
China	China	Manzhouli City, Hulunbuir City	1
China	China	Suzhou City	1
China	China	Dongxiang County, Fuzhou City	1
China	China	Fuzhou City	2
China	China	Ganzhou City	2
China	China	Chaoyang City	1
China	China	Dalian City	1
China	China	Shenyang City	1

Рисунок 2 – Развернутый куб в браузере

Таким образом, было обработано более 13 000 записей о заболеваемости коронавирусом в январе-феврале 2020 г. С учетом планируемого интеллектуального анализа данных, возможности для выполнения которого поддерживаются SQL Server Analysis Services, можно сделать вывод, что использование OLAP-решения компании Microsoft обеспечивает полное видение анализируемой ситуации и непрерывный оперативный контроль за ее развитием.

Следует отметить, что используемые программные компоненты относятся к бесплатно распространяемому программному обеспечению.

Литература

1. Барсегян, А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян. – СПб.: БХВ-Петербург, 2009. – 336 с.
2. Epidemiological data from the COVID-19 outbreak, real-time case information [Электронный ресурс] – Режим доступа: https://www.nature.com/articles/s41597-020-0448-0. – Дата доступа: 10.05.2020.