УДК 004.67 + 517.518.85 DOI: 10.30748/soi.2017.148.13

В.В. Федько

Харьковский национальный экономический университет им. С. Кузнеца, Харьков

ОБ ОДНОМ РАЦИОНАЛЬНОМ ПОДБОРЕ ДАННЫХ В БИЗНЕС-АНАЛИЗЕ

Рассмотрена проблема уменьшения нагрузки на сервер и сетевой трафик при решении задач визуализации данных в бизнес-анализе. Предлагается на стороне сервера уменьшить размер выборки данных из хранилища данных, а затем на стороне клиента аппроксимировать результаты запроса с помощью интерполяционных линейных сплайнов. Представлены результаты численных экспериментов визуализации при различных количествах данных, найдено значение константы аппроксимации и определены ее статистические характеристики.

Ключевые слова: *бизнес-анализ, визуализация данных, аппроксимация, численный эксперимент, статистические характеристики.*

Введение

Хорошо спроектированные приложения Business Intelligence (BI) помогут принимать более обоснованные решения, быстро понять различные информационные ресурсы в организации, и как они взаимодействуют друг с другом. Эти ресурсы могут включать в себя клиентские базы данных, информацию о цепочке поставок, данные персонала, производства, данные о продукции, продаж и маркетинговой деятельности, а также любой другой источник информации, критически важными для работы.

Важнейшим этапом в BI является визуализация данных. Она позволяет интегрально увидеть закономерности изучаемого процесса. Диаграмма дает только общее представление о процессе. Более точные данные представляют в числовом выражении в форме таблиц. Исходя из этой точки зрения, возникает вопрос о количестве данных, необходимых для построения диаграмм [1; 2].

Несмотря на все увеличивающуюся мощность современной вычислительной техники проблема достаточного количества данных для построения диаграмм в BI остается актуальной и в наши дни. Это обусловлено следующими причинами. Анализируемые данные, как правило, хранятся в хранилище данных, которое чаще всего расположено на сервере, к которому одновременно обращаются многие тысячи и даже миллионы клиентов. Поэтому задача их своевременного обслуживания остается важной даже при переводе хранилищ в облачные дата-центры [3–5]. Уменьшение количества выбираемых на сервере и передаваемых по сети данных позволяет уменьшить загрузку сервера и уменьшить трафик в сети, что в свою очередь приводит к улучшению масштабируемости дата-центра.

С другой стороны, для построения диаграммы аналитику не нужны все данные, которые хранятся в хранилище и связаны с изучаемым процессом. На практике там могут храниться миллионы и даже миллиарды записей. В настоящее время рост объема информации имеет экспоненциальную зависимость.

По данным компании IBS [6], до 2003 году мир было накоплено 5 эксабайтов данных (1 ЭБ = 1 млн терабайтов). Дальнейший рост объема информации представлен в следующей таблице:

Таблица 1

Рост объема информации

Год	2008	2011	2013	2015	2020 (прогноз)
Объем, ЗБ	0,18 ЗБ	1,76	4,4	6,5	40-44

Здесь 1 ЗБ = 1024 эксабайта

В работе исследуется подход, когда из большого количества данных выбирается незначительное их количество. Затем к выбранным данным применяется процедура аппроксимации, – данные интерполируются с помощью линейных сплайнов.

Основная часть

BI представляет собой набор стратегий, процессов, приложений, данных, продуктов, технологий и технических архитектур, которые используются для поддержки сбора, анализа, представления и распространения деловой информации [7].

Эффективность работы системы BI зависит от качества основных элементов ее архитектуры:

1. Системы-источники.
2. Хранилище данных.
3. BI-инструменты.
4. Пользователи.

Настоящая статья посвящена вопросам оптимизации второго и третьего элементов архитектуры. С точки зрения больших данных в работе исследуется подход к решению первых двух из трех проблем 3V (volume – объем данных, velocity – скорость и variety – разнообразие).

В качестве источника эмпирического ряда данных будем использовать функцию

$$y = \sin(x) \quad (1)$$

на интервале $[0, 2\pi]$.

Выбор обусловлен хорошей визуальной узнаваемостью функции и отсутствием совпадения с аппроксимантом (за исключением узлов интерполяции).

Построим эмпирические ряды для различного количества узлов n = 10, 20, 30, 40, 50, 100 с постоянным шагом

$$h = d / (n - 1), \quad (2)$$

где d – длина интервала. В нашем случае $d = 2\pi$.

В узлах вычислим значение функции $y = \sin(x)$, а между ними воспользуемся линейной интерполяцией, т.е. на каждом отрезке между двумя соседними узлами аппроксимант определяется по формуле

$$y_a = y_i + (y_{i+1} - y_i)(x - x_i)/h, \quad (3)$$

где i = 1, 2,…, (n-1).

На рис. 1 – 6 представлены результаты построения графиков по выбранным данным.

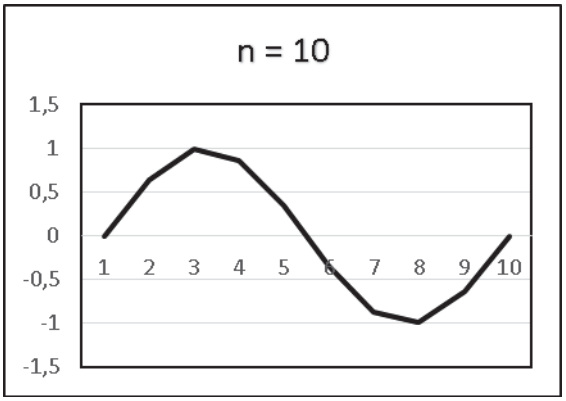

Рис. 1. График аппроксиманта при n = 10

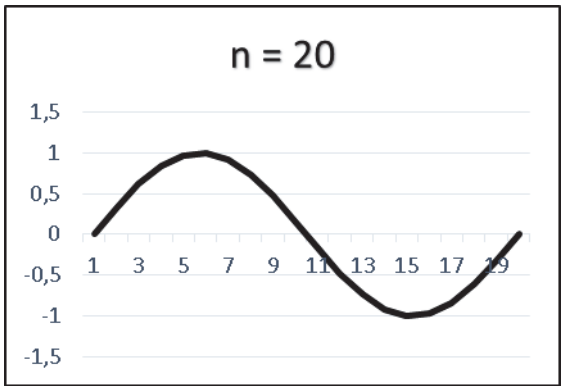

Рис. 2. График аппроксиманта при n = 20

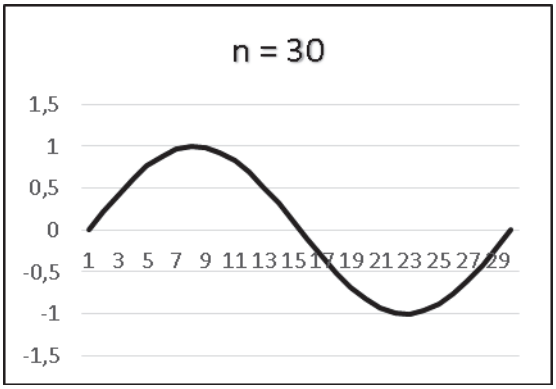

Рис. 3. График аппроксиманта при n = 30

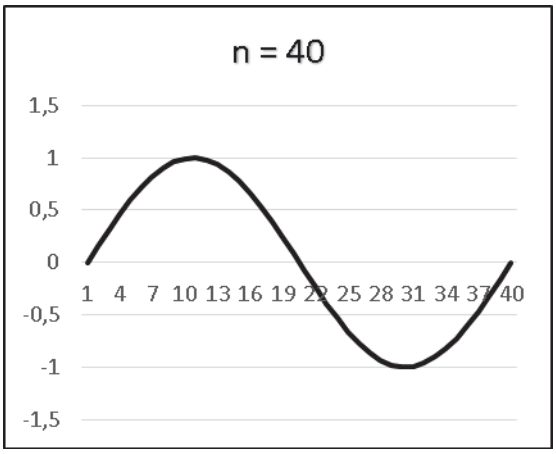

Рис. 4. График аппроксиманта при n = 40

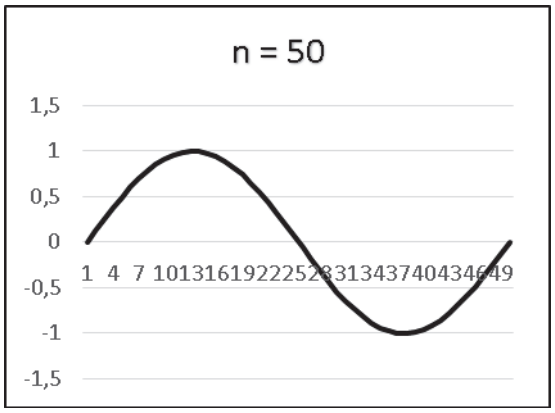

Рис. 5. График аппроксиманта при n = 50

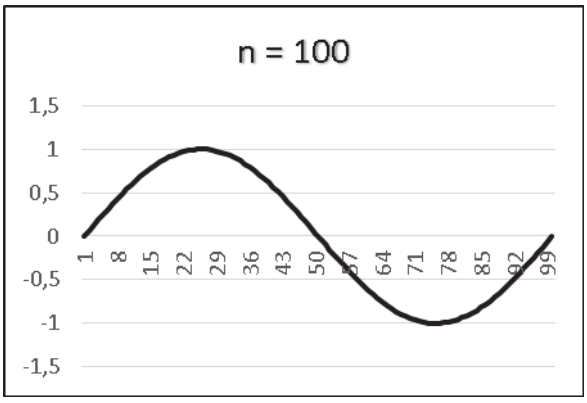

Рис. 6. График аппроксиманта при n = 100

Из приведенных выше графиков можно сделать следующие выводы:

1. Для задач анализа данных с целью интегрального представления о закономерностях изучаемого процесса достаточно остановиться на n = 20, а если с запасом – то n = 30.

2. Дальнейшее незначительное увеличение количества данных (n = 40, 50) слабо влияет на качество графика.

3. При значительном увеличении количества узлов (рис. 6) на графике появляется «рябь», связанная с погрешностями компьютерной графики.

Произведем оценку точности результатов численных экспериментов.

Как известно [8], кусочно-линейная аппроксимация имеет погрешность, которая определяется по следующей формуле:

$$\Delta = c / n^2 \qquad (4)$$

где c – константа аппроксимации,

n – количество узлов.

Чтобы определить константу аппроксимации, вычислим значения приближаемой функции и ее аппроксиманта в центре каждого отрезка между двумя соседними узлами. Для вычисления значений аппроксиманта используем формулу (3). После этого сравним их. В табл. 2 приведены максимальные значения отклонений аппроксиманта от приближаемой функции в зависимости от количества узлов, а также соответствующие значения относительных погрешностей.

Таблица 2

Максимальные значения отклонений и относительных погрешностей

n	Макс. отклонение	Макс. отн. погр., %
10	0,0594	6,0
20	0,0136	1,4
30	0,0059	0,6
40	0,0032	0,3
50	0,0018	0,2
100	0,0036	0,05

Используя данные табл. 1 и формулу (4), вычисляем значения константу аппроксимации в зависимости от количества узлов. В табл. 3 представлены полученные результаты.

Определим следующие статистические характеристики полученных значений [9].

Среднее значение константы аппроксимации как ее значение, при вычислении которого в совокупности сохраняется неизменным и определяется формулой (5):

$$\overline{c} = \frac{\sum_{k=1}^{m} c_k}{m}, \qquad (5)$$

где m – количество экспериментов. В нашем случае m = 6.

Таблица 3

Экспериментальные значения константы аппроксимации

n	c
10	600
20	560
30	540
40	480
50	500
100	500

Среднее линейное отклонение, которое представляет собой среднее из абсолютных (по модулю) отклонений от средней арифметической в анализируемой совокупности данных. Математическая формула имеет вид:

$$a = \frac{\sum_{k=1}^{m} |c_k - \overline{c}|}{m}. \qquad (6)$$

Дисперсия как мера, характеризующая разброс данных вокруг математического ожидания. Вычисляется по формуле (7).

$$s^2 = \frac{\sum_{k=1}^{m} |c_k - \overline{c}|}{m}. \qquad (7)$$

Среднеквадратичное отклонение, вычисляемое как корень из дисперсии. Определяется по формуле (8).

$$s = \sqrt{\frac{\sum_{k=1}^{m} |c_k - \overline{c}|}{m}}. \qquad (8)$$

Коэффициент вариации, который используется получения относительной меры разброса данных. Рассчитывается путем деления среднеквадратичного отклонения на среднее арифметическое. Формула коэффициента вариации имеет вид

$$V = \frac{s}{\overline{c}}. \qquad (9)$$

Результаты вычислений по формулам (5–9) приведены в табл. 4.

Таблица 4

Статистические характеристики константы аппроксимации

$\overline{c}$	a	s^2	s	V
530	36,7	2040	45,2	8 %

Полученные результаты свидетельствуют о достаточной надежности примененной методики.

Выводы

В работе рассмотрены вопросы повышения эффективности выполнения анализа данных в приложениях бизнес-анализа данных. Основное внимание уделено решению проблем объем данных (volume) и скорости (velocity) на этапе визуализации данных в системе BI.

Предложен подход, когда из большого количества данных выбирается незначительное их количество. Затем к выбранным данным применяется процедура аппроксимации, – данные интерполируются с помощью линейных сплайнов.

Представлены результаты численных экспериментов визуализации при различных количествах данных. Они показывают, что для задач анализа данных с целью интегрального представления о закономерностях изучаемого процесса достаточно остановиться на количестве узлов n = 20, а если с запасом – то n = 30. Дальнейшее незначительное увеличение количества данных (n = 40, 50) слабо влияет на качество графика. При значительном увеличении количества узлов (n = 100) на графике появляется «рябь», связанная с погрешностями компьютерной графики.

Найдено значение константы аппроксимации и определены ее статистические характеристики. В частности, коэффициента вариации не превышает 8 %, что вполне допустимо для задач обработки данных.

Уменьшение количества выбираемых на сервере и передаваемых по сети данных дает синергетический эффект. Оно позволяет уменьшить загрузку сервера и уменьшить трафик в сети, что в свою очередь приводит к улучшению масштабируемости дата-центра.

Список литературы

1. Data Visualization. [Electronic resource]. – Access mode: http://www.sas.com/en_us/insights/big-data/data-visualization.html.

2. Федько В.В. Основи інформаційних технологій. Електронні таблиці MS Excel 2010 : навчальний посібник / В.В. Федько, В.І. Плоткін. – Х. : Вид. ХНЕУ, 2012. – 288 с.

3. Amazon QuickSight. [Electronic resource]. – Access mode: https://quicksight.aws/.

4. Boost your organization's IQ with our BI tools. [Electronic resource]. – Access mode: http://www.sap.com /solution/platform-technology/analytics/business-intelligence-bi.html.

5. Visualize Business Insights. [Electronic resource]. – Access mode: https://www.oracle.com/solutions/business-analytics/business-intelligence/data-visualization.html.

6. Мир big data в 8 терминах. [Электронный ресурс]. – Режим доступа: http://rb.ru/howto/big-data-in-8-terms/.

7. Business intelligence. [Electronic resource]. – Access mode: https://en.wikipedia.org/wiki/Business_intelligence.

8. Марчук Г.И. Методы вычислительной математики / Г.И. Марчук. – Санкт-Петербург: Изд. Лань, 2009. – 608 с.

9. Малярец Л.М. Теория вероятностей и математическая статистика : учебно-практическое пособие для иностранных студентов / Л.М. Малярец, А.А. Егоршин. – Х. : ХНЭУ, 2013. – 304 с.

Поступила в редколлегию 10.03.2017

Рецензент: д-р техн. наук, проф. М.И. Сидоренко, Институт радиофизики и электроники НАН Украины, Харьков.

ПРО ОДИН РАЦІОНАЛЬНИЙ ПІДБІР ДАНИХ В БІЗНЕС-АНАЛІЗІ

В.В. Федько

Розглянуто проблему зменшення навантаження на сервер і мережевий трафік при розв'язанні задач візуалізації даних в бізнес-аналізі. Пропонується на боці сервера зменшити розмір вибірки даних зі сховища даних, а потім на боці клієнта апроксимувати результати запиту за допомогою інтерполяційних лінійних сплайнів. Подано результати чисельних експериментів візуалізації при різних кількостях даних, знайдено значення константи апроксимації та визначено її статистичні характеристики.

Ключові слова: *бізнес-аналіз, візуалізація даних, апроксимація, чисельний експеримент, статистичні характеристики.*

ON A RATIONAL SELECTION DATA IN BUSINESS ANALYSIS

V.V. Fedko

Is discussed the problem of reducing the server load and network traffic at the decision of tasks in the data visualization business analysis. It is proposed to reduce the data on the server-side sample size from the data store, after that on the client-side approximate the request results using linear interpolation splines. The results of numerical experiments for visualization of different data quantities, found the value of constant approximation and determined its statistical characteristics.

Keywords*: business analysis, data visualization, approximation, numerical experiment, the statistical characteristics.*