ИССЛЕДОВАНИЕ ПРОЦЕССА МИГРАЦИИ БАЗ ДАННЫХ И СКРИПТОВ С ПОМОЩЬЮ ДИСПЕРСИОННОГО АНАЛИЗА

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Зубко В.В.

Отвагин А.В. − к. т. н., доцент

Наиболее распространенная проблема, возникающая при расширении компании и оптимизации бизнес-процессов – низкая эффективность баз данных и приложений. Для решения этой проблемы и повышения продуктивности информационных систем компании прибегают к миграции. Компания Ispirer Systems разрабатывает и поставляет на рынок SQLWays – высокоадаптивное решение для конвертации баз данных, поддерживающее широкий круг СУБД и намного уменьшающее время и стоимость миграции баз данных.

Интерес представляет исследование автоматизации процесса миграции (конвертации) баз данных и скриптов из исходной СУБД в целевую, а именно: выявление факторов, существенно влияющих на производительность данного процесса, с целью оценки времени и сложности миграционного проекта.

Для поиска зависимостей в экспериментальных данных был выбран метод дисперсионного анализа [1]. Данный метод заключается в выделении и оценке отдельных факторов, вызывающих изменчивость изучаемой случайной величины, путём исследования значимости различий в средних значениях.

В качестве входных выбраны три параметра: тип исходной СУБД, вид проекта и количество SQL-кода в проекте. Выходным параметром будет являться степень автоматизации миграционного программного модуля (процент правильно сконвертированного SQL-кода, по сути КПД программы).

Полный перебор возможных сочетаний параметров системы потребует чрезмерно большого количества опытов. Эксперимент, в котором пропущены некоторые сочетания уровней анализируемых факторов, называют дробным факторным экспериментом (ДФЭ). Число опытов можно значительно сократить, если воспользоваться ДФЭ по схеме латинского квадрата, введенного впервые Фишером [2]. Латинский квадрат $n \times n$ – это квадратная таблица, составленная из *n* элементов (чисел или букв) таким образом, что каждый элемент повторяется в каждой строке и каждом столбце только один раз.

План и результаты эксперимента данного исследования представлен на базе латинского квадрата 4×4 (таблицы 1 и 2 соответственно). Изучается влияние трех факторов на процесс конвертации, каждый из которых изменяется на четырех уровнях. Факторы:

1) тип СУБД (A) – по строкам: DB2 (a_1), MySQL (a_2), Oracle (a_3), Sybase (a_4);

2) вид проекта (B) – в ячейках: конвертация DDL базы данных (b_1), простого скрипта (b_2), наличие API вызовов (b_3) и dynamic SQL (b_4);

3) количество SQL-кода (C) – по столбцам: 100 (c_1), 500 (c_2), 1000 (c_3), 5000 (c_4) строк SQL-кода.

Табл. 1 – План эксперимента

Тип СУБД	Количество кода			
	b_1	b_2	b_3	b_4
a_1	c_1	c_2	c_3	c_4
a_2	c_4	c_1	c_2	c_3
a_3	c_3	c_4	c_1	c_2
a_4	c_2	c_3	c_4	c_1

Табл. 2 – Результаты эксперимента

Тип СУБД	Количество кода			
	b_1	b_2	b_3	b_4
a_1	0,91	0,88	0,86	0,82
a_2	0,91	0,87	0,95	0,92
a_3	0,82	0,79	0,87	0,85
a_4	0,91	0,88	0,84	0,90

После проведения дисперсионного анализа входных данных латинского квадрата по соответствующему алгоритму значимость линейных эффектов проверяются по критерию Фишера [3]. Если дисперсионное отношение удовлетворяет неравенствам:

$$\frac{S_A^2}{S_{ош}^2} < F_{1-p}(f_1, f_2), \qquad \frac{S_B^2}{S_{ош}^2} < F_{1-p}(f_1, f_2), \qquad \frac{S_C^2}{S_{ош}^2} < F_{1-p}(f_1, f_2)$$

где, p – уровень значимости; f_1, f_2 – число степеней свободы, равные $f_1 = n-1$; $f_2 = (n-1)(n-2)$, принимаются нулевые гипотезы: $\alpha_i = 0$, $\beta_i = 0$, $\gamma_i = 0$.

Если какое-нибудь дисперсионное отношение оказывается больше табличного, соответствующая нулевая гипотеза отвергается и влияние фактора считается значимым. Приняв гипотезу о значимости

влияния фактора, то есть гипотезу о значимости различия в средних. обычно выясняют, какие именно средние значимо различаются между собой при помощи критерия Стьюдента [5] или множественного рангового критерия Дункана [6]. Если же согласно условиям задачи один или два фактора являются источниками неоднородностей, влияние которых надо исключить при подсчете главного эффекта (это обеспечивается планированием по схеме латинского квадрата), то средние по источникам неоднородностей не подсчитывается и не проверяется значимость их различия по статистическим критериям.

Обозначим дисперсионные отношения для факторов A, B и C как F_A, F_B и F_C соответственно. По результатам расчетов значения дисперсионных отношений оказались следующими: $F_A = 6.398$, $F_B = 1.092$ и $F_C = 3.675$. Так как для 5% уровня значимости, число степеней свободы $\gamma_1 = 3$ и $\gamma_2 = 6$, а $F_{табл} = 4.76$, то ненулевая гипотеза отвергается только для первого фактора и данный фактор считается значимым, остальные гипотезы принимаются нулевыми.

Таким образом, было произведено планирование эксперимента по исследованию процесса миграции баз данных и скриптов. С помощью дисперсионного анализа и метода латинских квадратов было выяснено, что только первый из рассматриваемых факторов (тип исходной СУБД) оказался значимым, остальные (вид проекта и количество SQL-кода в проекте) оказались незначимыми факторами для данного процесса. Также была произведена проверка соответствующих гипотез на языке программирования R в среде разработки RStudio.

Список использованных источников:

1. Изаков, Ф. Я. Планирование эксперимента и обработка опытных данных / Ф. Я. Изаков // Уч. пособие для магистрантов и аспирантов. – Челябинск, 1997. – 128 с.
2. Fisher, R. A. Statistical methods for research workers – Edinburgh, 1925.
3. Батрак, А. П. Планирование и организация эксперимента / А. П. Батрак // Уч. пособие к изучению теоретического курса для студентов направления 220500. – Красноярск, 2010. – 60 с.
4. Шеффе, Г. Дисперсионный анализ / Пер. с англ. – 2-е изд. – М., 1980. – 512 с.
5. Student. The probable error of a mean // Biometrika. 1908. № 6 (1). P. 1-25.
6. Зедгинидзе, И. Г. Планирование эксперимента для исследования многокомпонентных систем / И. Г. Зедгинидзе. – М., 1976. – 390 с.