УДК 37:004

Л. Ф. Панченко, ДЗ „Луганський національний університет імені Тараса Шевченка”

НОВІ ТРЕНДИ АНАЛІЗУ ДАНИХ

Панченко Л. Ф.
Нові тренди аналізу даних
Стаття присвячена питанням використання хмарних технологій аналізу даних. Виявлено можливості програми підтримки академічної спільноти SAS OnDemand for Academics, які дозволяють організувати роботу викладача й студента в курсах аналізу даних із використанням хмарних технологій: клієнтська частина працює на комп’ютерах викладачів та студентів, а сам аналіз здійснюється за допомогою хмарних серверів SAS. Аналізуються переваги такого підходу для університетських курсів: безкоштовність, використання передових технологій аналізу даних, формування умінь і навичок роботи з хмарними технологіями, можливість розташовувати у хмарах масиви даних для спільного аналізу. Розглядається зміст та організація навчання в курсі з аналізу даних „Passion Driven Statistics” проекту coursera.org.
Ключові слова: аналіз даних, хмарні технології, освіта, SAS.

Панченко Л. Ф.
Новые тренды в анализе данных
Статья посвящена вопросам использования облачных технологий анализа данных. Определены возможности программы поддержки академического сообщества SAS OnDemand for Academics, которые позволяют организовать работу преподавателя и студента в курсах анализа данных с использованием облачных технологий: клиентская часть работает на компьютерах преподавателей и студентов, а сам анализ осуществляется с помощью облачных серверов SAS. Анализируются преимущества такого подхода для университетских курсов: бесплатность, использование передовых технологий анализа данных, формирование умений и навыков работы с облачными технологиями, возможность располагать в облаках массивы данных для совместного анализа. Рассматривается содержание и организация обучения в курсе анализа данных „Passion Driven Statistics” проекта coursera.org .
Ключевые слова: анализ данных, облачные технологии, образование, SAS.

Серед найбільш вагомих трендів 2013 року за дослідженням агентства Gartner [6] – опрацювання великих масивів даних („big data”) та хмарні

обчислення. На рис.1 наведено динаміку пошукових запитів „big data”, отриману за допомогою сервісу Google Ngram. Запити „аналіз даних”, «великі дані» охарактеризовано як зверх популярні. Отже, важливим завданням вищої школи є підготовка конкурентоспроможного фахівця, обізнаного у відповідних технологіях.

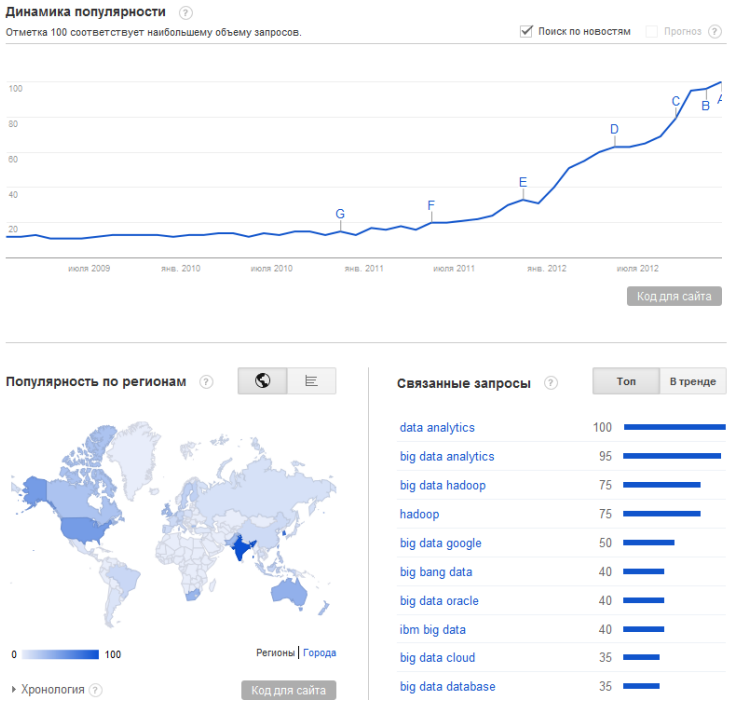

Рис.1. Динаміка популярності пошукового запиту *big data*

Освітяни, слідом за представниками бізнес структур, зацікавилися використанням хмарних сервісів. Хмарний сервіс Office 365 компанії Microsoft став платформою для інформаційної системи „Відкрита освіта”, спрямованої на автоматизацію освітнього процесу в російських школах. З її допомогою навчальні заклади зможуть організувати повноцінне віддалене навчання, оперативну комунікацію з батьками, електронний адміністративний облік [7]. Використанню хмарних технологій у освіті було присвячено нещодавно проведений Всеукраїнський науково-методичний семінар [4]. Матеріали семінару висвітлюють питання розвитку хмарних технологій, застосування їх у відкритій освіті, ВНЗ, післядипломній освіті, наводяться приклади застосування хмарних засобів навчання фізики, математики, інформатики. Аналіз публікацій

доповідей семінару [4], а також особистий досвід участі в ньому [2], дозволяє зробити висновок про недостатню розробленість питань використання хмарних технологій стосовно предметної галузі «аналіз даних».

Мета статті: розглянути можливості хмарних технологій стосовно галузі аналізу даних.

Аналіз застосування статистичного програмного забезпечення у магістерських програмах західних університетів, зокрема в галузі соціальних та гуманітарних наук [5], свідчить про наступну динаміку використання найбільш популярних пакетів (у % до загальної кількості курсів):

Таблиця 1

Динаміка використання статистичного програмного забезпечення у магістерських програмах університетів [5]

Програмне забезпечення	1996-2000 рр.	2001-2005 рр.	2006-2010 рр.
SPSS	33	31	31
Excel	30	29	17
SAS	33	32	24
STATA	4	8	27

У той же час, роботодавці вимагають від своїх співробітників вміння користуватися: тільки SPSS – 20%, SPSS або SAS – 22%, тільки SAS – 19%, SAS або STATA – 9% [5]. Таким чином, SAS і SPSS є найбільш затребуваними статистичними середовищами. Розроблений автором разом з О. Адаменко курс з комп'ютерного аналізу даних [1; 3] для підготовки фахівців різноманітних спеціальностей об'єднує виконання завдань з аналізу даних в середовищах Excel та SPSS. У пошуках розширення використання програмного забезпечення щодо статистичного аналізу даних ми звернулися до програми підтримки академічної спільноти від SAS. Ця програма (SAS OnDemand for Academics) включає безкоштовну можливість для студентів, що вивчають курси із використанням SAS та викладачів, які викладають курси із підтримкою в SAS,

завантажити на свій комп'ютер клієнт SAS та здійснювати аналіз даних за допомогою хмарних технологій із використанням серверів SAS [10].

Послідовність кроків для використання цієї можливості для викладача включає реєстрацію і створення індивідуального профілю SAS, обрання необхідного програмного забезпечення клієнта, завантаження його на свій комп'ютер, реєстрація курсу, розсилка інформації про майбутній курс студентам. Пройшовши усі ці кроки, ми отримали безкоштовний сертифікат на електронне навчання можливостям SAS строком на 1 рік. Електронне навчання включає курси: „Запити та звіти", „Основи програмування", „Програмування: техніка управління даними", „Дисперсійний аналіз. Регресія. Логістична регресія" з підтримкою в середовищі SAS Enterprise Guide; та „Прикладну аналітику" і „Передбачувальні моделі для бізнес-аналітиків" в SAS Enterprise Miner.

Аналізу даних із використанням в SAS присвячений також курс „Passion Driven Statistics" який проходив на Coursera в березні 2013 р. [9]. У якості базового статистичного програмного забезпечення в ньому використовувався також SAS OnDemand for Academics. Сілабус курсу складався з наступних тем:

- Інсталяція статистичного програмного забезпечення, множини даних та їх документація.
- Управління даними.
- Описова статистика та візуалізація даних.
- Порівняння середніх (ANOVA), тести незалежності (Chi Square) і кореляція.
- Модерування.
- Представлення статистичних результатів.

Діяльність студентів у цьому курсі включала:

- дослідження документації даних;
- виконування основних програм;
- управління даними;
- побудову графіків і діаграм;

- перевірку гіпотез;
- написання звіту за результатами дослідження.

Студентам надавалися декілька множин даних з різних галузей: 1) характеристики кратерів Марса; 2) здоров'я підлітків; 3) психічне здоров'я дорослих та його розлади; 4) соціальні, економічні та медичні показники країн світу.

Цікаво, що спочатку всі дані було представлено в форматі SAS. А потім за дискусією на форумі було розширено їх також для аналізу в SPSS, R, Stata, а також подано у вигляді csv-файлів. На підставі вибору студентом даних, кожен генерував статистичні гіпотези для перевірки, готував дані для аналізу, проводив описовий та аналітичний аналіз, а також оцінював, інтерпретував і представляв результати досліджень.

Оцінка студента в курсі складалася з тестів, завдань з аналізу даних і остаточного представлення результатів за допомогою індивідуальних блогів студентів. За тести студенту начислялося 40%, за завдання з аналізу даних – 20%; фінальний проект оцінювався в 40%.

Допоміжні матеріали було представлено у вигляді тексту, а також відео-лекцій і демонстрацій. Список найкращих блогів було представлено іншим студентам для ознайомлення [8].

Рис. 2. Хмарина слів, складених з відповідей студентів курсу [8]

Аналіз хмарини слів з рис. 2 свідчить, що студенти пов’язують поняття аналізу даних з „навчанням”, формуванням „важливих умінь”, „розумінням”, „покращанням”, „відновленням”.

У блозі викладача курсу Лізи Діркер (Lisa Dierker) ми знайшли широкий список причин, чому студенти приєдналися до курсу [8]. Ось деякі з них:

- Я намагався знайти себе!
- Я люблю статистику.
- Я вірю, що розуміння статистики забезпечує міцну основу для високої громадянськості в епоху „великих даних”.
- Мені подобається формат, тобто застосування знань для вирішення проблем.
- Мені подобається розповідати історії через статистику.
- Щоб побудувати Start Up.
- Отримати навички, необхідні, щоб стати вченим у галузі аналізу даних.
- Нудьга від програми вищої школи.
- Це безкоштовно!
- Для вивчення з перших рук досвіду, педагогіки, що використовуються в МООС викладачем.
- Я викладач прикладної статистики, і я хочу дізнатися більше про статистику та знайти цікаві ідеї для застосування їх у класі.
- Осмислення даних по-новому слугуватиме мені не тільки професійно, а й творчо.
- Я люблю дані!

Таким чином, хмарні технології аналізу даних об’єднують провідні тенденції розвитку інформаційних технологій: аналіз великих масивів даних, так звані „великі дані” („big data”) та хмарні обчислення. Ресурси програми підтримки академічної спільноти SAS OnDemand for Academics дозволяють

організувати роботу студентів і викладачів у курсах аналізу даних із використанням хмарних технологій. Клієнтська частина працює на комп'ютерах викладачів та студентів, а сам аналіз здійснюється за допомогою хмарних серверів SAS. Переваги такого підходу для університетських курсів: безкоштовність, використання передових технологій аналізу даних, формування умінь і навичок роботи з хмарними технологіями, можливість розташовувати у хмарах масиви даних для спільного аналізу.

Подальші напрямки дослідження: розширення програми та навчально-методичного забезпечення курсів з аналізу даних для студентів університету за рахунок використання можливостей SAS OnDemand for Academics та хмарних технологій.

Література

1. **Адаменко О. В.** Теоретико-методичні засади навчання студентів аналізу даних з використанням комп'ютера / О. В. Адаменко // Вісн. Луган. нац. ун-ту імені Тараса Шевченка. – 2010. – № 17. – С. 31 – 35.

2. **Адаменко О. В.** Хмарні технології аналізу даних / О. В. Адаменко, Л. Ф. Панченко // Хмарні технології в освіті : Матер. Всеукр. наук.-метод. сем. – Кривій Ріг : Видавничий відділ КМІ, 2012. – С. 143 – 144.

3. **Панченко Л. Ф**. Компьютерный анализ данных : учеб. пособие для студентов высш. учеб. заведений / Л. Ф. Панченко, Е. В. Адаменко ; Гос. учрежд. „Луган. нац. ун-т имени Тараса Шевченко". – Луганск : Изд-во ГУ „ДЗ ЛНУ імені Тараса Шевченка", 2010. – 188 с.

4. **Хмарні** технології в освіті : Матер. Всеукр. наук.-метод. сем. – Кривій Ріг : Видавничий відділ КМІ, 2012. – 173 с.

5. **Adams W. C**. Statistical Software for Students: Academic Practices and Employer Expectation /William C. Adams, Donna Lind Infeld, Carli M. Wulff. – [Електронний ресурс]. – Режим доступу : http:// umdcipe.org/conferences/ Classroom/Presentations/APPAM%20Teaching%20Workhop_Adams%20Infeld_pres_11_1109.pdf

6. **Gartner:** Top 10 Strategic Technology Trends For 2013 [Електронний ресурс]. – Режим доступу : http:// www.forbes.com/ sites/ericsavitz/2012/10/23/ gartner-top-10-strategic-technology-trends-for-2013/

7. **Microsoft** Office 365 [Електронний ресурс]. – Режим доступу : //www.ferra.ru/ru/techlife/news/2013/05/27/Microsoft-Office-365-ru-school/

8. **Passion** Driven Statistics. – [Електронний ресурс]. – Режим доступу : http://passiondrivenstatistics.tumblr.com/

9. **Passion** Driven Statistics. – [Електронний ресурс]. – Режим доступу : https://www.coursera.org/course/pdstatistics

10. **SAS** OnDemand for Academics [Електронний ресурс]. – Режим доступу : https://support.sas.com

Panchenko L. F.

New Trends in Data Analysis

The article is devoted to the use of cloud technology of data analysis. The opportunities of programs for the academic community SAS OnDemand for Academics are discussed. This program allows to organize the work of a teacher and a student in the course of data analysis with cloud computing: the client part works on the computers of teachers and students, and the analysis is carried out with cloud servers SAS. The advantages of this approach for university courses are analyzed: free, using of advance data mining techniques, the skills formation to work with cloud technologies, the ability to have a cloud data sets for joint analysis.

The content and organization of learning in the course of data analysis „Passion Driven Statistics" of project coursera.org are considered. Course syllabus consists: statistical software set up; data sets and data documentation; data management; descriptive statistics and data visualization; comparing means; tests of categorical independence and correlation. The different data sets (characteristics of Mars craters, adolescent health, adult psychiatric, social, economic and health indicators of countries worldwide) to analyze are proposed. Evaluations included quizzes (40%), applied data assignments (20%) and a final project via individual student blog sites (40%). Learning materials has been presented as text, video lectures and demonstrations.

Analysis of the clouds of words with student's answers shows that students relate concepts of data analysis with „learning", „new skills formation", „understanding", „improvement", „refreshing".

Key words: data analysis, cloud computing, education, SAS.

Відомості про автора

Панченко Любов Феліксівна – доктор педагогічних наук, професор кафедри теоретичної і прикладної інформатики ЛНУ імені Тараса Шевченка. Наукові інтереси: інформаційно-комунікаційні технології в освіті, інформаційно-освітнє середовище університету, статистичний аналіз даних.

Стаття надійшла до редакції 08.04.2013 р.
Прийнято до друку 26.04.2013 р.