**Тезисы доклада**

Начало формы

1. **НАЗВАНИЕ ДОКЛАДА:**

(на русском языке) ‑ Особенности использования среды R в обучении IT-специалистов

(на английском языке) ‑ Using the R environment for education of IT professionals

1. **АВТОРЫ:**

Фамилия1 И. О., Фамилия2 И. О., Фамилия3 И. О.

(на русском языке) ‑ Румянцев А. С., Гусев В. А.

(на английском языке) - Rumyantsev A., Gusev V.

1. **ОРГАНИЗАЦИЯ (полное наименование, без аббревиатур):**

(на русском языке) ‑ Федеральное государственное бюджетное учреждение науки Институт прикладных математических исследований Карельского научного центра Российской академии наук

(на английском языке) ‑ Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences

1. **ГОРОД:**

(на русском языке) ‑ Петрозаводск

(на английском языке) ‑ Petrozavodsk

1. **ТЕЛЕФОН: 8 (8142) 76-63-12, доб. 33**
2. **ФАКС: 8 (8142) 76-63-13**
3. **E-MAIL: ar0@krc.karelia.ru**
4. **АННОТАЦИЯ**:

(на русском языке) ‑ Рассмотрен метод организации обучения IT-специалистов языку R для высокопроизводительного анализа данных на основе организации виртуальной рабочей среды Rstudio с использованием технологии контейнеризации Docker.

(на английском языке) ‑ Practical issues of teaching of high-performance data analysis in R language to the IT specialist are discussed. A method of arrangement of practical classes is suggested based on the Rstudio virtual environment and Docker containers technology.

1. **КЛЮЧЕВЫЕ СЛОВА**:

(на русском языке) ‑ высокопроизводительный анализ данных, организация обучения, Docker, R, Rstudio

(на английском языке) ‑ high performance data analysis, learning, Docker, R, Rstudio

1. **ТЕКСТ ТЕЗИСОВ ДОКЛАДА:**

Широкое распространение ИТ-инструментов и резкое удешевление стоимости хранения данных привели к взрывному росту объема создаваемой и хранимой информации. Одновременно, развитие методов и алгоритмов анализа данных обеспечило устойчивый интерес к аналитическим инструментам со стороны государственных структур и коммерческих компаний. В то же время, относительная легкость освоения и простота использования аналитических инструментов привели к появлению значительного числа начинающих аналитиков, которых исследовательское агентство Gartner объединило термином "citizen data scientist". По мнению директора по науке агентства Gartner Александра Линдена, скорость роста числа таких "городских аналитиков" к 2017 году в 5 раз превысит скорость роста специалистов по анализу данных. В этой связи обучение базовым навыкам анализа данных является насущной необходимостью в процессе обучения ИТ-специалиста.

Язык R является одним из наиболее распространенных и востребованных языков в сфере анализа данных. Развитые возможности статистического анализа, входящие в базовый функционал языка, эффективно дополняются обширной базой пакетов расширения (более 9000 на данный момент). Это позволяет применять аналитический пакет R исследователям в области биологии, химии, физики, медицины. Свободный (Open-Source) пакет R обладает богатыми возможностями визуализации данных, в том числе геопространственной информации, что делает R хорошей альтернативой геоинформационным системам (ГИС) в области базового функционала последних.

В области высокопроизводительного анализа больших данных пакет R эффективно дополняется такими открытыми платформами машинного обучения и анализа данных, как KNIME (knime.org) и Weka (http://www.cs.waikato.ac.nz/~ml/weka/). Параллельная обработка данных возможна благодаря развитым механизмам параллельной обработки, основанным как на технологии пересылки сообщений MPI, так и на более простых технологиях сокетов, реализованным в пакетах foreach, SNOW, doParallel.

Практическая часть обучения IT-специалиста, как правило, ведется на основе лабораторных работ. Выполнение лабораторных работ ведется в дисплейных классах, где актуальность и работоспособность среды программирования обеспечивается администратором. В то же время, при выполнении лабораторной работы с личного компьютера, а также при выполнении домашних заданий, настройку рабочей среды студент вынужден выполнять самостоятельно. Это порождает необходимость оказания преподавателем технической поддержи студентам по вопросам настройки рабочего окружения. Для работоспособности среды разработки в целом и успешного запуска подготовленных примеров в частности необходимо, чтобы на рабочем месте студента были установлены нужные версии компиляторов и библиотек, заданы правильные переменные окружения, пользователь должен обладать необходимыми правами и т.п.

Задача настройки рабочей среды студента является второстепенной относительно цели обучения, отнимает время у преподавателей и повышает порог вхождения для студентов, вынужденных решать одновременно задачи администрирования и программирования. Одним из возможных вариантов решения данной проблемы является выделение среды разработки в веб-приложение, которое находится под полным централизованным контролем преподавателя и/или администратора учебного заведения, и, в то же время, всегда доступно студенту с любого личного устройства.

Программирование на языке R, как правило, осуществляется с помощью интегрированной среды разработки Rstudio (rstudio.com). Данная среда имеет исполнение в виде веб-приложения, обладающего интерфейсом классического (автономного) приложения. Одним из преимуществ веб-приложения является запуск расчетов на сервере (на базе открытой операционной системы Linux), что позволяет существенно снизить требования к пользовательскому устройству (ноутбуку/персональному компьютеру). Особенностью веб-среды Rstudio является необходимость создания для пользователя системной учетной записи Linux для аутентификации, а также домашнего каталога для хранения пакетов расширения R (которые могут быть установлены из веб-приложения) и исходного кода.

Установленная в неизменном виде на физический сервер или на виртуальную машину веб-версия Rstudio в большей мере предназначена для личного использования и не обладает достаточной масштабируемостью в случае коллективного использования группой студентов. А именно, возникают следующие проблемы:

* 1. разграничение прав доступа и изоляция пользователей (в том числе от ошибочных действий других пользователей, приводящих к необходимости перезапуска сервера),
  2. резервное копирование,
  3. автоматизация подготовки рабочей среды для различных потоков студентов.

Решение перечисленных проблем возможно с помощью технологии контейнеризации приложений Docker. В соответствии с технологией Docker необходимо создать образ приложения, содержащий все необходимое для его работы в изолированной среде. На основе исходного (неизменяемого) образа приложения тиражируются рабочие копии - контейнеры. Образы и контейнеры имеют многослойную файловую систему, работающую по принципу copy-on-write (копирование при записи), вследствие чего создание множества одинаковых контейнеров не требует значительных ресурсов системы хранения (все файловые системы имеют единый исходный образ).

При создании образов приложений Docker принято придерживаться следующих принципов:

* 1. контейнер должен содержать лишь один работающий процесс (своеобразная реинкарнация принципа UNIX-way),
  2. удаление и повторное создание контейнера не должно приводить к потере результатов работы (результаты работы должны храниться вне контейнера).

Таким образом, организовать работу группы студентов предлагается следующим образом:

* 1. создается Docker-образ Rstudio с веб-сервером, набором необходимых пакетов и библиотек,
  2. для каждого студента тиражируется отдельный контейнер, располагающийся на отдельном поддомене (этот процесс может быть автоматизирован с помощью приложения docker-gen),
  3. домашний каталог студента монтируется в качестве домашнего каталога внутреннего (одинакового для всех) пользователя контейнера с помощью Docker-сервера, в нем хранится исходный код работ студента, в него же устанавливаются дополнительные пакеты.

Предлагаемый подход не зависит от операционной системы, и легко тиражируется. Работа в изолированных контейнерах исключает зависимость от ошибочных действий других пользователей. В случае необходимости, перезапуск контейнера (а также тиражирование из исходного образа) не влияет на работу других пользователей. В этой связи резервному копированию подлежат только домашние каталоги пользователей (в резервировании контейнера нет необходимости).

В качестве дальнейшего развития данной технологии можно предложить организацию аутентификации пользователей через используемую в организации систему (LDAP, OpenID и т.п.), а также создание Docker образа сервера Shiny (shiny.rstudio.com) в связке с контейнером Rstudio, что позволит вести разработку веб-приложений на языке R.

Работа поддержана РФФИ, гранты 15-07-02341, 15-07-02354, 15-29-07974, 16-07-00622, 16-47-100168, и Программой стратегического развития Петрозаводского государственного университета.

Разместите здесь текст на русском или английском языке объемом до 3-х страниц в формате MS Word (размер 12 пт, одинарный межстрочный интервал, верхнее, нижнее, правое поля 2 см., левое 3 см.) без переносов слов и повторяющихся пробелов.