

Открытые системы

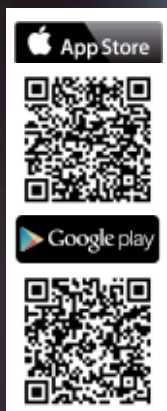
СУБД

№01
2016

ISSN 1028-7493

ИТ для бизнеса —
архитекторам
информационных систем

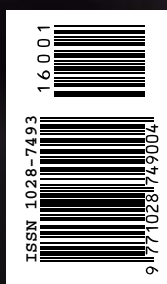
www.osmag.ru



МАШИННОЕ ОБУЧЕНИЕ

Распознавание мнений и отзывов в Сети

- Частное облако для широких масс
- Как увеличить быстродействие СУБД?
- Перспективы Открытой науки



Главный редактор
Дмитрий Волков, с.н.с., ИПМ РАН

Научный редактор
Наталья Дубова

Редакционный совет:

Валерий Аджиев, к.т.н., с.н.с.,
Национальный центр компьютерной анимации,
Университет Борнмута (Великобритания);

Фуад Алескерев, д.т.н., профессор, НИУ ВШЭ;

Михаил Горбунов-Посадов, д.физ.-мат.н.,
зав. отделом ИПМ РАН, доцент, МГУ;

Юрий Зеленков, д.т.н., зав. кафедрой прикладной
информатики, Финансовый университет
при Правительстве РФ;

Сергей Кузнецов, д.физ.-мат.н., профессор, МГУ;

Михаил Кузьминский, к.хим.н., с.н.с., ИОХ РАН;

Александр Легалов, д.т.н., профессор, СФУ;

Владимир Сухомлин, д.т.н., профессор, МГУ;

Павел Храмцов, к.т.н., доцент, МИФИ;

Игорь Федоров, к.т.н., профессор, МЭСИ;

Виктор Шнитман, д.т.н., профессор, МФТИ;

Леонид Эйсымонт, к.физ.-мат.н., научный
консультант, НИИ «Квант»

Корректор Ирина Карпушина

Верстка и графика Мария Рыжкова

Дизайн обложки Денис Киров

Адрес для корреспонденции:

127254, г. Москва, а/я 42

Телефоны:

+7 495 725-4780/84, +7 499 703-1854
+7 495 725-4785 (распространение, подписка)

Факс: +7 495 725-4783

E-mail: osmag@osp.ru

Подписной индекс:

99482 — «Каталог российской прессы» (МАП)

72773 — Объединенный каталог «Пресса России» АПР

59869 — «Каталог. Издания органов научно-технической
информации»

Реклама

000 «Рекламное агентство «Чемпион» +7 499 750-0467



**ОТКРЫТЫЕ
СИСТЕМЫ**
Open Systems Publications

© 2016 Издательство «Открытые системы»

Журнал зарегистрирован
в Министерстве РФ по делам печати,
телерадиовещания и средств массовых коммуникаций
03.07.2015
Свидетельство ПИ № ФС 77-62328

Журнал выходит 4 раза в год

Цена свободная

Выпуск издания осуществлен
при финансовой поддержке
Федерального агентства по печати
и массовым коммуникациям

Учредитель и издатель:

000 «Издательство «Открытые системы»
Россия, 127254, Москва,
проезд Добролюбова, дом 3, строение 3

Президент Михаил Борисов

Генеральный директор Галина Герасина

Директор ИТ-направления Павел Христов

Коммерческий директор Татьяна Филина

Все права защищены.
При использовании материалов
необходимо разрешение редакции и авторов.

В номере использованы иллюстрации
и фотографии: 000 «Издательство «Открытые
системы» и IEEE Computer Society.

Отпечатано в 000 «Богородский
полиграфический комбинат»
142400, Московская область,
г. Ногинск,

ул. Индустриальная, д. 40Б
(495) 783-9366, (49651) 73179

Тираж 4 000 экз.

Содержание №1 (211) 2016

НОВОСТИ. ФАКТЫ. ТЕНДЕНЦИИ.

Начало конца шифрования

Как поссорились разработчики баз данных in-memory

Исполнилось 25 лет первой в мире веб-странице

Операторы связи поддержат Open Compute Project

В DataCamp появится групповое обучение

основам обработки данных

За четыре года на неиспользуемое ПО

растрачено 30 миллиардов долларов

«Умные» счетчики

Данные «холодного» хранения

Для заблокированных по решению властей

веб-страниц ввели особый код

Большая группа по Большим Данным

Eclipse Che выходит в бета-версии

Алгоритм упростит общение с роботами

Watson в роли доктора и продавца

Oracle прекратит выпуск плагина Java

ПЛАТФОРМЫ

10 Рейтинг Top500 и прогресс

высокопроизводительных

вычислений

Эрих Штраймер, Ханс Мойер,

Джек Донгарра, Хорст Саймон

На протяжении двух десятилетий рейтинг

Top500 пользовался успехом в качестве мерила

производительности суперкомпьютеров и был

первоисточником, позволяющим проследить

технологические тенденции. Но насколько этот

рейтинг применим в эпоху экзамастических

суперкомпьютеров?

В ФОКУСЕ: МАШИННОЕ ОБУЧЕНИЕ

14 Практические аспекты машинного

обучения

Виктор Китов

Сегодня с машинным обучением связывают

много надежд, однако успешность его приме-

нения определяется не только выбором адек-

ватного задачи алгоритма, но и правильными

шагами на этапах планирования, разработки и

внедрения модели.

18 Машинное обучение для

понимания естественного языка

Павел Велихов

Технологии обработки естественного языка

сегодня шагнули далеко вперед, и немалая

заслуга в этом принадлежит машинному обу-

чению, применяемому, в частности, для пони-

хания текстов.

22 Машинное обучение

для планирования запросов

Олег Иванов

Рост объемов данных требует от СУБД уве-

личения производительности выполнения

запросов. Оптимизация плана выполнения

запроса с использованием средств машинного

обучения позволяет в разы уменьшить время

его обработки.

26 Машинное обучение

для Больших Данных

Андрей Орлов

Теоретические основы машинного обучения по-

явились практически одновременно с первыми

компьютерами, однако при его практическом

применении всегда приходится учитывать спе-

цифику конкретных систем. Работа с Большими

Данными средствами Nadoop требует инстру-

ментов адаптации алгоритмов машинного

обучения к этой платформе.

28 Оценка компетентностей студентов

на основе анализа социальных сетей

Михаил Захаров, Анатолий Карпенко,

Елена Смирнова

В третьем поколении Федеральных государ-

ственных образовательных стандартов выс-

шего профессионального образования задачи

обучения формулируются в терминах компе-

тентностей учащихся, что делает актуальной

задачу их оценки. На помощь приходят средст-

ва машинного обучения, позволяющие оценить

компетентности студентов на основе анализа

их поведения в социальных сетях.

ИНТЕРНЕТ ВЕЩЕЙ

31 Референтные архитектуры

для Интернета вещей

Михаэль Вейрих, Кристофер Эберт

Интернет вещей призван предоставить новые

возможности развития и повысить произво-

дительность труда благодаря свободе комму-

никации различных устройств друг с другом.

Однако воплощению этих идей мешает не-

хватка архитектурных шаблонов, описываю-

щих протоколы связи с учетом особенностей

конкретных отраслей.

34 Интернет вещей для энтузиастов:

метеостанция с визуализацией

Эдуард Лебедев

Интернет вещей развивается сегодня благо-

даря не только компаниям, предлагающим

решения, специально предназначенные для

работы в Интернете вещей, но и энтузиастам,

которым остро необходимы новые инстру-

менты работы с взаимодействующими «ум-

ными» вещами.

СУБД

36 Стоимостные оптимизаторы

для СУБД: вчера и сегодня

Леонид Борчук

Оптимизаторы запросов для реляционных

СУБД прошли долгий путь развития, превра-

тившись сегодня в сложную программу оценки

стоимости вариантов, однако в СУБД нового

поколения стоимость выполнения запросов

во внимание почти не принимается. Что это —

шаг назад или два вперед? И нужна ли вообще

преемственность для стоимостных оптимиза-

торов мира Больших Данных?

ОПЫТ

40 Частное облако для масс

Владимир Сигунов

На базе OpenStack сегодня можно построить

частное облако. Однако опыт реальных проектов

показывает, что малым и средним предприя-

тиям требуются типовые масштабируемые аппа-

ратно-программные решения, поставляемые

вместе с технической поддержкой.

МИР

42 ДИСПАК — операционная система

атомного проекта

Ирина Крайнева

Атомный проект сыграл определяющую роль в

политическом и социально-экономическом раз-

витии СССР в послевоенный период, в частно-

сти, став стимулом к развитию вычислительной

техники, вычислительной математики и про-

граммирования. Одной из наиболее известных

программистских разработок, зародившихся в

недрах атомного проекта, стала операционная

система ДИСПАК для ЭВМ БЭСМ-6.

ИТ-УНИВЕРСИТЕТЫ

44 На пути к Открытой Науке

Сергей Парин

Новые информационные системы для органи-

зации научных коммуникаций позволяют по-

высить ответственность ученого за качество

выпускаемых работ и обеспечить более широ-

кое и открытое использование их результатов.

Одна из таких систем — «Соционет».

БИБЛИОТЕКА

46 ИТ на рубеже эпох

Александр Тьренко

Темы ноябрьского, декабрьского, январского

и февральского номеров журнала Computer

(IEEE Computer Society, Vol. 48, No. 11, 12 2015

и Vol. 49, No. 1, 2 2016) — суперкомпьютеры

экзафлопсной производительности, перспек-

тивы развития компьютеров после окончания

действия закона Мура, слияние человека с

кибернетическими системами и новые тен-

денции в мире облаков.

ОТКРЫТЫЕ СИСТЕМЫ СЕГОДНЯ

Начало конца шифрования

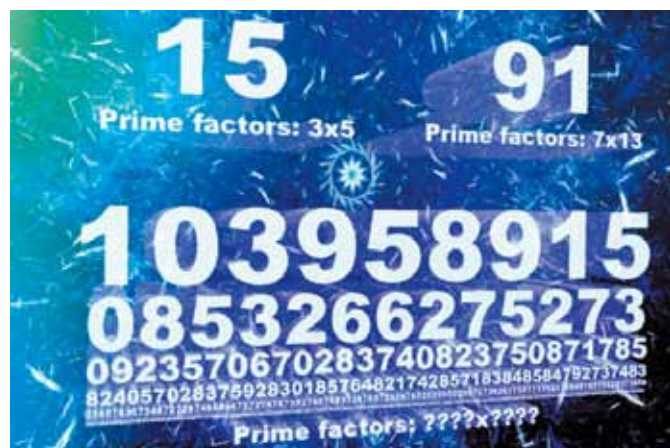
Строительство большой квантовой системы обойдется в колоссальные деньги, но речь уже идет «не о теоретической физике, а о практической инженерной задаче», утверждают исследователи из США и Австрии, построившие масштабируемый квантовый компьютер для разложения чисел на множители.

В 1994 году профессор МТИ Питер Шор разработал квантовый алгоритм разложения чисел на множители, способный работать практически с такой же скоростью, с которой действуют алгоритмы шифрования с открытым ключом. Надежность таких алгоритмов зависит от неспособности современных компьютерных систем раскладывать большие числа на множители за разумное время. За прошедшие годы исследователи проверили действенность алгоритма Шора на экспериментальных системах с малым числом кубитов, раскладывая на множители маленькие числа. Но для возможности взлома современных схем шифрования нужен квантовый компьютер с несколькими сотнями кубитов.

И вот на днях исследователи из МТИ и Инсбрукского университета опубликовали в Science доклад о том, что им удалось построить масштабируемый квантовый компьютер на ионных ловушках, который с помощью алгоритма Шора разложил на множители число 15. Как утверждают ученые, увеличивать количество кубитов в их компьютере можно просто добавляя атомы и лазеры. Таким образом, говорят исследователи, они создали первую масштабируемую реализацию алгоритма Шора. Попутно ученые придумали, как уменьшить количество кубитов, необходимое для реализации алгоритма. К примеру, для разложения на множители числа 15 в обычном случае нужно 12 кубитов, а аппарату на ионных ловушках достаточно пяти, каждый из которых представлен одиночным атомом, удержи-

ваемым в состоянии суперпозиции двух энергетических уровней. Стабильность квантовой системы обеспечивается за счет удаления одного электрона из каждого атома и удержания атомов на месте с помощью электрического поля. По словам исследователей, квантовые вентили, используемые в алгоритме Шора, создаются путем воздействия на четыре из пяти атомов лазерными импульсами, а передача результатов выполняется через пятый атом.

В IBM, где некогда тоже проверили действенность алгоритма Шора на системе с малым числом кубитов, с похвалой отозвались о разработке ученых, отметив, что для создания компьютера достаточного масштаба понадобится на порядки больше кубитов, причем их нужно будет перебрасывать между ловушками сложного устройства с помощью тысяч одновременных управляющих лазерных импульсов.



Источник: Jose-Luis Olivares/MIT

Oracle покупает разработчика технологии переноса приложений в облако

Oracle приобретает стартап Ravello Systems, основанный около четырех лет тому назад создателями гипервизора KVM. Компания предлагает программное обеспечение, при помощи которого в организациях могут использовать популярные общедоступные облака для «расширения» своих центров обработки данных. Oracle вышла на облачный рынок с опозданием, уже после того, как на нем закрепились крупные игроки, такие как Amazon Web Services и Microsoft Azure. Аналитики не исключают, что Ravello как разработчик перспективной технологии была «превентивно» куплена Oracle, чтобы не позволить приобрести ее кому-то вроде IBM или HP. Как сообщают в Ravello, разработки компании будут объединены с IaaS-решениями Oracle, чтобы помогать заказчикам корпорации быстро переносить сложные приложения в публичные облака без затратной доработки кода. Сотрудники Ravello перейдут в подразделение Oracle по публичным облакам.

Как поссорились разработчики баз данных in-memory

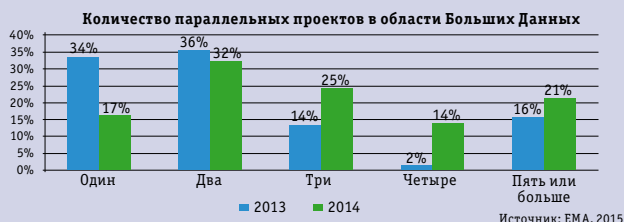
Разработчики баз данных с обработкой в памяти Hazelcast и GridGain обвинили друг друга в нечестности. СУБД GridGain представляет собой коммерческую реализацию системы с открытым кодом Apache Ignite. Компания Hazelcast первой опубликовала результаты тестирования производительности двух разработок, из которых следовало, что Hazelcast значительно опережает GridGain. Как утверждает директор Hazelcast, в последующем сравнении (показавшем противоположный результат) компания GridGain использовала другую конфигурацию тестирования и по существу фальсифицировала результаты. Такое поведение, подчеркивает он, недостойно участника проекта Apache и нарушает правила программы Apache Incubator. В GridGain заявляют, что повторили тесты, но результаты оказались точно такими же. Тесты выполнялись в среде Amazon AWS, и их исходные коды опубликованы. Между тем в Hazelcast тестирование проводили на своем оборудовании.

Все дело в автоматизации

Аналитики компании Enterprise Management Associates опубликовали собранную ими статистику по корпоративным проектам в области Больших Данных. Как отмечают авторы, сегодня данные все шире используются в организациях в процессах принятия решений. В частности, почти 63% участников проведенного ЕМА опроса сообщили, что стратегии, основанные на данных (data-driven strategies), занимают важное место в их организациях. Растет число организаций, где проекты в области Больших Данных находятся на стадии ближайших планов или непосредственного осуществления, а компаний, где такие проекты на начальной стадии, становится меньше. Среди участников опроса, проведенного ЕМА в 2014 году, 55% сообщили, что стратегия в сфере Больших Данных принята и играет в организации либо «ключевую», либо «важную» роль.



Если в 2013 году у большинства таких организаций были один-два активных проекта в области Больших Данных, то в 2014-м — уже два-три; выросла доля респондентов с четырьмя и более проектами. Одновременно растут и сложности. Больше всего участников, 22%, назвали главным препятствием для осуществления стратегии неэффективное управление данными (в ЕМА управление данными определяют как «разработку и внедрение архитектур, политик, практик и процедур эффективного управления жизненным циклом информации на предприятии»). На втором месте (12,7%) — трудности внедрения платформ мультиструктурных данных, обусловленные прежде всего наличием большого числа источников.



Как убедились аналитики ЕМА, во многих случаях автоматизация помогает справляться с трудностями, обеспечить необходимую координацию и снизить затраты благодаря повышению результативности работы ИТ-персонала. Но в ИТ-службах нередко не справляются с проектами автоматизации. Как показало исследование, одной из главных помех является нехватка соответствующих инструментов. В частности, в 44% организаций либо вообще не имеют средств автоматизации, либо пользуются несовершенными инструментами собственной разработки.

С ростом применения Больших Данных будет расширяться автоматизация инфраструктуры, используемой для сбора огромных объемов информации. В ЕМА полагают, что предприятиям стоит уделять больше внимания интеграции такой инфраструктуры с корпоративными приложениями и подключению сторонних инструментов, самосовершенствующихся путем обучения на обширной базе пользователей. В ЕМА уверены, что использование оптимальных инструментов автоматизации позволит ИТ-службе остаться «в водительском кресле» управления стратегией, основанной на данных.

Исполнилось 25 лет первой в мире веб-странице

Исполнилось 25 лет сайту Европейской организации по ядерным исследованиям (ЦЕРН). Первая в истории Интернета веб-страница, разработанная Тимом Бернерсом-Ли, была запущена 20 декабря 1990 года; она не была тогда публичной, но стала плацдармом для дальнейшего развития интернет-коммуникаций и Глобальной паутины в целом. Сайт ЦЕРН рассказывал о самой Сети, о том, как искать информацию и как создавать свой сервер. Бернерс-Ли, которого теперь принято считать «отцом» Всемирной сети, предложил спецификации языка HTML, протокола HTTP и идентификаторы URL, которыми до сих пор пользуются для ссылки на конкретный интернет-адрес. Он же разработал и первый браузер под названием WorldWideWeb, а также программное обеспечение для сервера.

World Wide Web

The WorldWideWeb (W3) is a wide-area, [hypertext](#) information retrieval system using a [distributed](#) access to a large number of documents. Everything there is either about W3 or linked directly or indirectly to this document, including an [extensive](#) summary of the project, [Mailing lists](#), [Policy](#), [What's new here?](#), [Features](#) to the world's online information, [subjects](#), [W3 services](#), etc.

[How](#) to use the system you are using

[Software Products](#) A list of W3 project components and their current state (e.g. [Last Modified](#), [XSL](#), [Yield](#), [NoXSL](#), [Server](#), [Tools](#), [Mail](#), [editor](#), [Libraries](#))

[Technical](#) Details of protocols, formats, program semantics etc.

[Bibliography](#) Paper documentation on W3 and references etc.

[People](#) A list of some people involved in the project

[History](#) A summary of the history of the project

[How can I help?](#) If you would like to support the web.

[Getting only](#) Getting the code by anonymous FTP, etc.

Европейская организация по ядерным исследованиям поделилась новостью в своем Twitter, предлагая всем желающим посмотреть, как начинался современный Интернет.

Первая веб-страница располагалась по адресу info.cern.ch; любопытно, что эта ссылка до сих пор функционирует.

В DataCamp появится групповое обучение основам обработки данных

Сайт DataCamp с 2013 года предлагает онлайн-обучение основам обработки и анализа данных. Курсы состоят из серии коротких видеоуроков и практических заданий на закрепление пройденного. Все материалы студентам предоставляются через Интернет. DataCamp не выдает официальных дипломов, вместо них студенты получают значки и свидетельства об окончании курсов. Некоторые курсы на DataCamp бесплатны, но доступ ко всем имеющимся курсам стоит 25 долл. в месяц или 250 долл. в год. В феврале DataCamp объявила о получении финансирования еще на 1 млн долл. и о реализации новых возможностей для корпоративных пользователей. Если до сих пор студенты могли проходить обучение только индивидуально, то теперь на сайте можно создавать группы учащихся. Сведения об изучении материалов и выполнении заданий отображаются в единой таблице. Через обучение на сайте DataCamp прошло уже около 300 тыс. человек, и преподаватели чаще всего интересовались именно возможностями работы с группами учеников, поясняют в компании.

Операторы связи поддержат Open Compute Project

Крупнейшие телекоммуникационные компании (Verizon, AT&T, Deutsche Telekom, SK Telecom и Equinix) присоединились к возглавляемому Facebook проекту Open Compute Project, в рамках которого конечные пользователи сами разрабатывают открытое оборудование для своих нужд — серверы, системы хранения данных и сетевые устройства. До этого среди

участников ОСР преобладали провайдеры облачных сервисов и ведущие ИТ-компании, теперь же ожидается, что операторы связи предоставят для инициативы проектную документацию по мощным коммутаторам и другому оборудованию магистрального класса. Как отмечают в IDC, операторам связи сейчас приходится непросто — из-за увеличения объемов видео и другого трафика потребности в пропускной способности ежегодно возрастают на 30—40%, а доходы операторов растут не так быстро, поэтому им нужен доступ к недорогому оборудованию. Кроме того, предприятия переносят больше задач в общедоступные облака Microsoft, Amazon и Google, ставя под угрозу доход от управляемых сетевых услуг, которые предоставляются операторами. Чтобы соединиться с публичными облаками и продолжать предоставлять такие услуги, телекоммуникационным компаниям нужно осваивать новые сетевые технологии, используемые облачными провайдерами, в том числе программно-конфигурируемые сети и виртуализацию сетевых функций.

Традиционная и современная аналитика: найти пять отличий

По прогнозам Gartner, в этом году объем глобального рынка инструментов бизнес-аналитики вырастет по сравнению с 2015 годом на 5,2% и составит 16,9 млрд долл. К числу таких инструментов аналитики относят индивидуальные аналитические приложения и платформы бизнес-аналитики, а также инструментарий управления эффективностью деятельности компании и средства выполнения сложной аналитики. По оценкам, данный рынок сейчас находится на последнем этапе многолетнего перехода от аналитических отчетов, которые готовятся преимущественно силами отдела ИТ, к аналитике, осуществляемой по принципу самообслуживания самими бизнес-пользователями. В Gartner отмечают, что аналитика стала стратегически важной для всех предприятий и представителей всех бизнес-специальностей; например, директорам по маркетингу нужно быть не только экспертами по брендингу и размещению рекламы, но и специалистами в области анализа данных по заказчикам. Навыки в области аналитики столь же важны для руководителей, отвечающих за кадровый отдел, цепочку поставок, финансы и т. д. Современный бизнес требует быстрых темпов получения аналитических знаний, поэтому происходит демократизация аналитики — переход на простые в использовании, быстрые и адаптивные современные платформы, которые обеспечивают более глубокой анализ данных из разнообразных источников.

Аналитическая операция	Традиционная аналитика	Современная платформа бизнес-аналитики
Источник данных	Требуется предварительное многомерное моделирование (звездоподобные схемы и т. п.)	Предварительное моделирование не требуется (работа с двумерными таблицами)
Загрузка и подготовка данных	Осуществляется отделом ИТ	Осуществляется с помощью средств, предоставляемых отделом ИТ
Создание контента	В основном осуществляется ИТ-персоналом, частично — опытными пользователями	Осуществляется бизнес-пользователями
Анализ	Структурированные специальные отчеты на каждый случай, анализ по заранее подготовленной модели	Свободное исследование данных
Передача результатов	Плановая рассылка уведомлений и отчетов, предоставление через портал	Через сервисы совместной работы, в виде презентаций и с использованием открытых API

Источник: Gartner, 2016