

22 июня 2017



ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ



ПРАКТИЧЕСКИЕ
СЕМИНАРЫ
И КОНФЕРЕНЦИИ

Организатор



**ОТКРЫТЫЕ
СИСТЕМЫ**
Open Systems Publications



ДМИТРИЙ ВОЛКОВ,

ведущий цикла практических семинаров, главный редактор, **«Открытые системы.СУБД»**

Как монетизировать большие данные, преобразовав обнаруженные в них скрытые закономерности в конкретные бизнес-идеи? Как активизировать и персонифицировать работу с существующей и потенциальной клиентской базой? Как снизить риски при работе с клиентами, предотвратить мошенничество, создать персонализированные продукты и выбрать наиболее подходящий канал для взаимодействия?

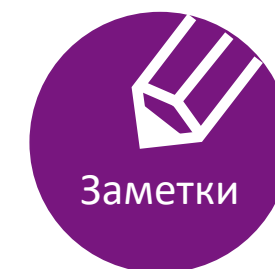
Конференция «Технологии больших данных 2017», продолжаящая цикл технологических мероприятий «Технологии блокчейна», «Технологии баз данных», «Технологии больших данных 2016» и других, организуемых издательством «Открытые системы», поможет аналитикам, специалистам по данным и разработчикам получить ответы на эти и многие другие вопросы. На одной площадке будет организовано обсуждение новых инструментов управления большими данными и опыта конкретных проектов их монетизации. Особое внимание будет уделено решению задач обработки больших массивов данных в реальном времени, использованию методов машинного обучения и искусственного интеллекта.



АНДРЕЙ НИКОЛАЕНКО,
IBS

ИНСТРУМЕНТЫ БОЛЬШИХ ДАННЫХ: ОТ КОНКУРЕНЦИИ К ИНТЕГРАЦИИ

Hadoop ворвался в мир корпоративных хранилищ — традиционную вотчину массово-параллельных реляционных СУБД, а на роль централизованных платформ данных сегодня претендуют резидентные гриды данных. С другой стороны, Hadoop и передовые NoSQL-системы «заговорили» на SQL, а реляционные СУБД научились выполнять задания MapReduce и овладели эффективным хранением JSON. Но такая конкуренция за ниши не только стала источником бурного развития самих инструментов, но и инициировала процессы интеграции инструментов. Уже сейчас можно при помощи одной методики проектировать хранилища и на HDFS, и на MPP-СУБД, а одним SQL-запросом можно «обойти» базы данных всех видов — от резидентных гридов и NoSQL-систем до РСУБД и Hadoop. В докладе приводится обзор современных технологий больших данных и анализируются тенденции их взаимодополнения и сосуществования.



Заметки



Следующая
секция



МАКСИМ САВЧЕНКО,
«Сбербанк-Технологии»

МОДЕЛИ НОВОГО ПОКОЛЕНИЯ: ПРИНЯТИЕ РЕШЕНИЙ НА ОСНОВЕ ПОВЕДЕНЧЕСКОЙ ИНФОРМАЦИИ РАЗЛИЧНОГО ТИПА

Половина населения планеты подключена сегодня к Интернету, а число цифровых устройств, генерирующих данные о поведении и предпочтениях клиентов, измеряется миллиардами, однако в большинстве отраслей и компаний не заметно адекватного роста эффективности за счет использования накапливаемых сведений, которые либо вообще не собираются с прицелом на дальнейший анализ, либо игнорируются и лежат мертвым грузом. До сих пор компании чаще оценивают то, что клиент сам сообщает о себе, чем ориентируются на его поведение – от анализа стиля его вождения автомобилем и потребительских предпочтений до анализа платежной дисциплины за пределами традиционного изучения кредитной истории. В результате промышленные системы, использующие источники данных нового типа для повышения эффективности бизнеса, по-прежнему редкость.

Доклад посвящен изложению опыта использования Сбербанком данных, полученных из различных источников, и иллюстрации того, как возникают потребности в новых подходах к их анализу и обработке.

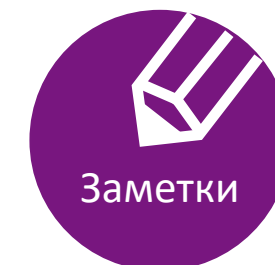


ДЕНИС АНИКИН,
Mail.ru

ХРАНИЛИЩЕ ДАННЫХ ИНТЕРНЕТА ВЕЩЕЙ

Без СУБД не обходится ни один интернет-сервис: базы данных являются основой любой современной ИТ-инфраструктуры, причем сегодня они должны автоматически распределяться по различным узлам, обладать средствами восстановления в случае отказа оборудования и масштабироваться в зависимости от текущей нагрузки. Все это особенно важно для Интернета вещей, для поддержки которых возможностей традиционных СУБД недостаточно. В докладе на примере открытой системы Tarantool, преимущества которой уже оценили такие компании, как Yota, «ВымпелКом», Badoo, Avito, QIWI и Wallarm, будет показано, какой должна быть СУБД для поддержки решений Интернета вещей, позволяющая обрабатывать информацию в реальном времени.

Предыдущая
секция



Следующая
секция



ОЛЕГ ФИНОШИН,
Vizex

ВИЗУАЛЬНЫЙ АНАЛИЗ: ВЫЯВЛЕНИЕ СКРЫТЫХ ПРОБЛЕМ И ИХ РЕШЕНИЙ В БОЛЬШИХ МАССИВАХ ДАННЫХ – МЕДИЦИНА, ПРОИЗВОДСТВО, ТРАНСПОРТ И ДР.

Визуальный анализ данных призван вовлечь человека в процесс обнаружения знаний в больших объемах сырых данных путем их представления в форме, обеспечивающей обнаружение скрытых закономерностей и аномалий, невозможное в случае применения какой-либо алгоритмической идентификации. В докладе показано, как с помощью новых технологий интерактивной визуальной работы с большими массивами оперативной и стратегической информации погрузить человека в данные и через визуальные образы дать ему инструменты восприятия существующих проблем и возможных путей их решения. Особое внимание будет уделено способам интегрированного представления множества влияющих на проблему аспектов с целью облегчения понимания информационных, управленческих и финансово-экономических процессов, а также обнаружения их взаимосвязей.

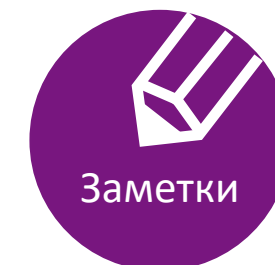


ЕВГЕНИЙ ЛИННИК,
«Техносерв»

БОЛЬШИЕ ДАННЫЕ В СИСТЕМАХ ПРОТИВОДЕЙСТВИЯ МОШЕННИЧЕСТВУ

В последние три года технологии больших данных все шире используются бизнесом для решения повседневных задач, и одна из проблем, с которой можно эффективно справиться с помощью, например, алгоритмов machine learning, – это борьба с мошенничеством в финансовом секторе. В докладе анализируется опыт решения задачи выявления мошенничества с помощью инструментов из экосистемы Hadoop и Spark MLlib.

Предыдущая
секция



Следующая
секция



ВАЛЕРИЙ АРТЕМЬЕВ,
Банк России

ПРИЧИНЫ НЕУДАЧ ПРОЕКТОВ БОЛЬШИХ ДАННЫХ

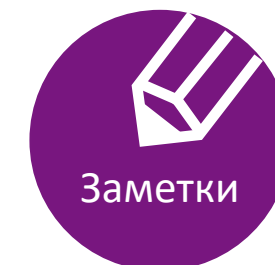
Сегодня произошла консолидация технологий, связанных с обработкой больших данных: уровень зрелости и освоения этих технологий стал важной составляющей успеха проектов в условиях вольницы Open Source и наличия корпоративных устоев. Однако оказалось, что одних лишь технологий для успеха проектов больших данных недостаточно.

Новые возможности для работы с большими данными вызвали всплеск интереса к продвинутой аналитике (машинному обучению, искусственному интеллекту и др.), data mining и исследованию данных. Открылись новые горизонты для предписывающей аналитики, анализа текстов, изменилось само отношение к качеству данных. Однако возник дефицит специалистов, обладающих знаниями и навыками в области анализа данных, программирования, управления данными и лингвистики. Вместе с тем включение data scientists в проектную команду — необходимое, но недостаточное для успеха проекта условие.

Любой проект больших данных — это конкретная предметная область, однако недостаток знаний его участников в конкретной прикладной сфере, отсутствие постановок задач анализа и критериев эффективности их решения, а также неосведомленность бизнес-аналитиков о потенциале технологий больших данных, методов и средств их анализа приводят к краху проекта.

Доклад посвящен изложению причин неудач проектов больших данных и рекомендациям по устранению возникающих на их пути барьеров.

Предыдущая
секция



Следующая
секция



ТЕХНОЛОГИИ
БОЛЬШИХ ДАННЫХ



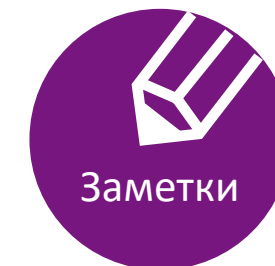
ПРАКТИЧЕСКИЕ
СЕМИНАРЫ
И КОНФЕРЕНЦИИ

ЗАМЕТКИ



ДЛЯ ЗАМЕТОК

Предыдущая
секция



Следующая
секция



АРТЕМ ДАНИЛОВ,
Avito.ru

РАЗВИТИЕ ХРАНИЛИЩА AVITO.RU – ОТ БОЛЬШИХ К ОЧЕНЬ БОЛЬШИМ ДАННЫМ

На первый взгляд проектирование хранилища данных – типовая задача: собрать требования, построить модель данных и реализовать ETL, однако лавинообразный рост объема данных и их сложности легко может нарушить работоспособность хранилища. В докладе показано, как, вслед за расширением бизнеса, хранилище «Авито» за несколько лет выросло до 76 Тбайт, откуда вообще у компании большие данные; в чем состоят основные принципы аналитики; как эволюционировало хранилище Avito.ru. Особое внимание уделено ответу на вопрос: почему следование ограничениям шестой нормальной формы (Anchor Modeling) помогает оперативно расширять хранилище, эффективно масштабировать нагрузку в среде MPP СУБД и развиваться в условиях микросервисов.

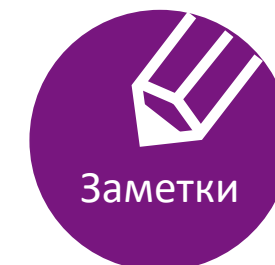


СЕРГЕЙ ЗОЛОТАРЕВ,
ArenaData

ЗАЧЕМ НУЖЕН «СТАНДАРТНЫЙ» HADOOP

Сегодня с Hadoop работают почти все современные платформы бизнес-аналитики, многие инструменты визуальной аналитики, разнообразное связующее программное обеспечение и даже некоторые реляционные СУБД. Однако существует множество дистрибутивов Hadoop, имеется масса различных версий и библиотек его экосистемы и часто собранная конструкция ведет себя непредсказуемо. Чтобы отметка «совместимо с Hadoop» обрела однозначный смысл, под эгидой Linux Foundation в 2015 году был создан консорциум разработчиков платформ Open Data Platform initiative (ODPi.org), объединяющий крупнейшие ИТ-компании и призванный обеспечить стандартизацию и взаимную совместимость дистрибутивов Hadoop путем утверждения их единой спецификации и тестирования на полное соответствие. Среди прошедших сертификацию в ODPi дистрибутивов Hadoop имеется российский — ArenaData Hadoop. Доклад посвящен анализу причин необходимости наличия стандартного дистрибутива, изложению особенностей сборки ArenaData и возможностям ее применения для снижения затрат и сроков разработки прикладного ПО.

Предыдущая
секция



Следующая
секция



АЛЕКСЕЙ МЕРГАСОВ,
NoXA Data Lab

БЕРЕЖЛИВАЯ (LEAN) ИНФРАСТРУКТУРА ДАННЫХ

Доклад посвящен описанию уникальной архитектуры системы хранения данных емкостью свыше 50 Пбайт, обеспечивающей сквозную 100 процентную консистентность данных, линейную масштабируемость по емкости и производительности при крайне низкой стоимости владения. Особое внимание будет уделено изложению первого опыта применения таких инфраструктур в телекоммуникационных компаниях, ретейле, банках, на промышленных предприятиях и в государственном секторе.

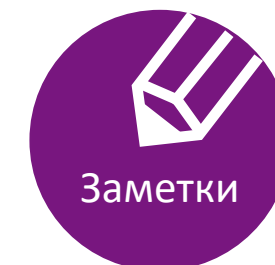


ЮРИЙ ПЕТРОВ,
BigBigData Group

ИНФРАСТРУКТУРА И ОРГАНИЗАЦИЯ РАБОТЫ НА РЕАЛЬНЫХ ПРОЕКТАХ БОЛЬШИХ ДАННЫХ

Стратегии развития многих компаний сегодня предусматривают запуск проектов больших данных, однако, как правило, их старт задерживается или вообще откладывается на неопределенное время из-за незнания оптимальных требований к аппаратно-программной инфраструктуре и особенностей организации работы. В докладе даны практические рекомендации и пояснения по выполнению проектов больших данных и развертыванию инфраструктуры, адекватной поставленным бизнесом задачам сбора, хранения и анализа больших массивов разнообразных данных, получаемых из различных источников.

Предыдущая
секция



Следующая
секция



ПАВЕЛ КЛЕМЕНКОВ,
Rambler&Co

МАШИННОЕ ОБУЧЕНИЕ: АРХИТЕКТУРА И КОНВЕЙЕР БОЛЬШИХ ДАННЫХ

Задачи бизнеса, определяемого данными, невозможно сегодня решить без специалистов по анализу данных (data scientists), однако найти и тем более привлечь к работе хороших специалистов в этой области очень трудно, но еще труднее обеспечить им условия для эффективного выполнения своих функций. В парадигме больших данных, распределенных систем и конкуренции за вычислительные ресурсы это сделать еще сложнее. В докладе рассмотрена архитектура информационной системы, центральное звено которой – математик-программист, методами машинного обучения решающий бизнес-задачи на больших данных, от неформальной постановки, подтверждения концепции proof-of-concept, вплоть до запуска кода в промышленную эксплуатацию. На примере бизнеса компании Rambler&Co показана эволюция такой информационной системы, позволившей создать комфортную среду для решения любой задачи обработки больших данных. Особое внимание уделено разбору реальных примеров использования методов машинного обучения для решения конкретных бизнес-задач.

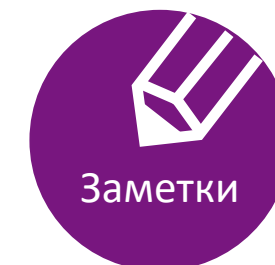


НИКИТА ЖАВОРОНКОВ,
НИЯУ МИФИ

БОЛЬШИЕ РАСПРЕДЕЛЕННЫЕ РЕЕСТРЫ. АНАЛИТИКА БЛОКЧЕЙНА

Сегодня о распределенных реестрах можно услышать даже от далеких от ИТ политиков: блокчейн, криптовалюты, умные контракты — все это обещает кардинально изменить уклад многих сфер жизни общества. Одно из преимуществ блокчейна – его прозрачность: у каждого участника сети есть доступ ко всей информации цепи блоков, которую можно анализировать, однако децентрализованные платежи, ведение распределенных реестров, сопровождение умных контрактов, и т. п., будут жизнеспособны лишь при наличии средств поиска и эффективного анализа огромных массивов неструктурированных данных из открытых реестров. В докладе разбираются подходы к извлечению полезной информации из таких реестров и технологии анализа больших данных, образуемых в огромных масштабируемых цепочках блоков.

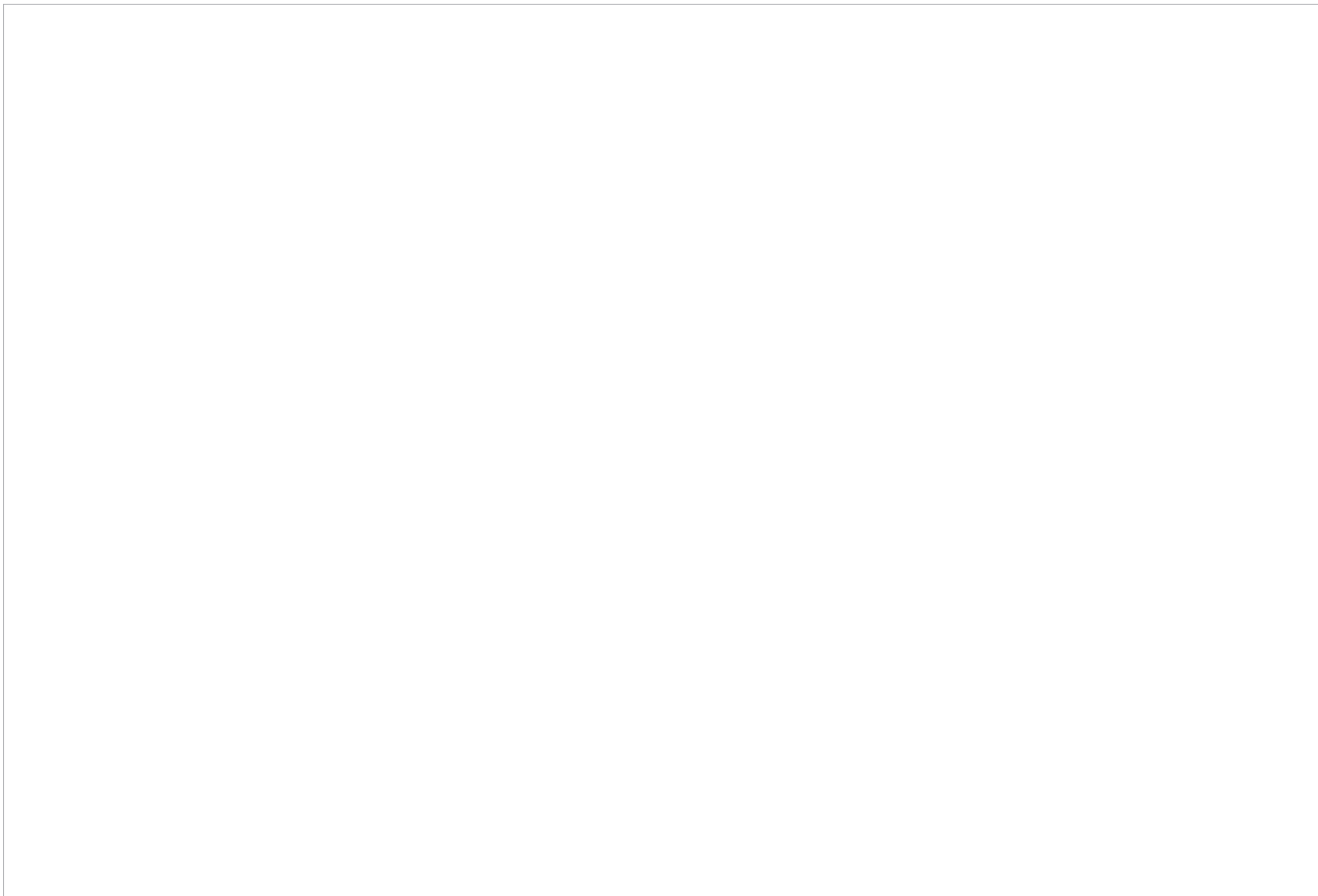
Предыдущая
секция



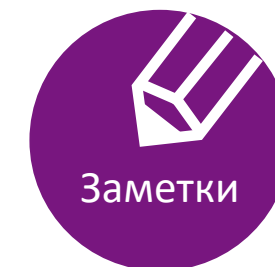
Следующая
секция



ДЛЯ ЗАМЕТОК



Предыдущая
секция



Следующая
секция



АЛЕКСЕЙ НАТЕКИН,
DM Labs, OpenDataScience

ВСЯ ПРАВДА О СПЕЦИАЛИСТАХ ПО ДАННЫМ

Промышленные системы работы с большими данными по-прежнему редкость: машинное обучение, Hadoop, искусственный интеллект сами по себе автоматически не трансформируются в дополнительную прибыль, а иногда инвестиции в технологии могут стать для бизнеса лишь бесполезной дорогостоящей игрушкой. Любой компании, решившей всерьез работать с большими данными, обязательно потребуется команда специалистов по данным (data scientists), способная реализовать планы и идеи руководителей компаний и организаций. Что это за специалисты, какими навыками и знаниями они должны обладать? Где и как их искать? Как поставить им задачу и сформулировать проблему, чтобы они не сбежали еще до конца испытательного срока? Доклад посвящен анализу проблемы поиска и подготовки кадров для эпохи больших данных – социально зрелых специалистов, обладающих не только математическими и инженерными знаниями, но и системным мышлением. Особое внимание будет уделено вопросам оптимизации труда data scientists в условиях, когда стоимость хранения и обработки данных неуклонно снижается, а стоимость труда специалистов по данным неуклонно растет.

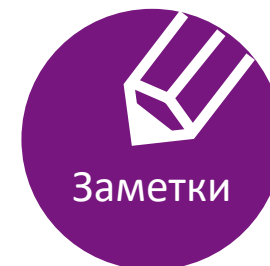


ДМИТРИЙ БАБАЕВ,
MTC, OpenDataScience

ОРГАНИЗАЦИЯ БОЛЬШИХ ДАННЫХ ДЛЯ ИХ ЭФФЕКТИВНОГО ИСПОЛЬЗОВАНИЯ В МАШИННОМ ОБУЧЕНИИ

В докладе пойдет речь о типах данных, образующих нынешние большие данные на примере проблем, решаемых сегодня телеком-операторами и интернет-компаниями. Будут рассмотрены способы представления данных применительно к задачам машинного обучения. Особое внимание в докладе будет уделено алгоритмам машинного обучения, позволяющим получить компактные представления больших данных, экономящим компьютерные ресурсы без потери качества решения реальных задач.

Предыдущая
секция



Следующая
секция

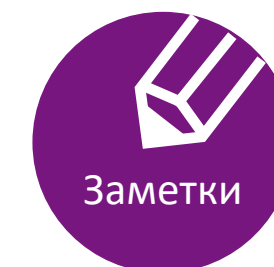


АЛЕКСАНДР СИДОРОВ,
HeadHunter

МАШИННОЕ ОБУЧЕНИЕ: УМНЫЙ ПОИСК РАБОТНИКОВ И РАБОТОДАТЕЛЕЙ

Много времени у специалистов по подбору кадров обычно уходит на разбор резюме и определение кого из претендентов на вакансию следует приглашать на очное или телефонное собеседование. Аналогичная ситуация возникает и при выборе работодателя – на личной странице соискателя необходимо выводить перечень вакансий, наиболее точно отражающий его запросы. В докладе показано, как технологии машинного обучения позволили компании HeadHunter упростить поиск работы и наем сотрудников, на десятки процентов повысив точность рекомендаций и в направления будет развитие проекта обработки больших данных.

Предыдущая
секция



Следующая
секция



РУСЛАН ХОХЛОВ,
ВТБ24

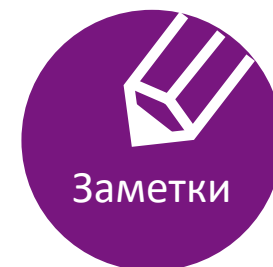


ДАНИИЛ ЗАЙЦЕВ,
«Глоубайт Консалтинг»

СИМБИОЗ РЕЛЯЦИОННЫХ СУБД И HADOOP В КРУПНОМ БАНКЕ

До недавнего времени в банковской сфере главенствовали традиционные подходы к построению хранилищ – сбор данных из систем-источников по четко сформулированным требованиям для решения заранее определенных задач. Однако, при сохранении всех традиционных функций, современному банку требуется осваивать более гибкие подходы к обработке данных, предусматривающие применение широкого спектра инструментов анализа и моделирования. Сегодня появляются все новые источники данных с динамической структурой, а скорость поддержки изменений в уже существующих перестает соответствовать ожиданиям бизнес-подразделений. Все это, в совокупности с быстро растущими объемами данных, подтолкнуло банк ВТБ24 к концепции Data Lake, которая гармонично вписалась в развивающуюся экосистему Hadoop. Доклад посвящен изложению опыта развития хранилища данных, как удовлетворяющего требованиям решаемых задач, так и учитывающего все особенности технологического стека Hadoop.

Предыдущая
секция



Следующая
секция



АРТЕМ ВЕДЕРНИКОВ,
Data-Centric Alliance

ПЕРСОНАЛИЗАЦИЯ РЕКЛАМНЫХ КАМПАНИЙ В РЕАЛЬНОМ ВРЕМЕНИ

Основа любой системы, позволяющей рекламодателям управлять целевыми аудиториями рекламных кампаний, в условиях омниканальности показывая только адресную рекламу заинтересованной целевой аудитории, – аналитическая платформа обработки аудиторных данных. Аналитическая платформа от Data-Centric Alliance, в реальном времени позволяющая обрабатывать данные о более 600 млн анонимных пользователей Интернета, основана на Apache Kafka и уже более года находится в промышленной эксплуатации, демонстрируя высокую производительность и отказоустойчивость. В докладе обсуждается архитектура системы сегментации пользователей, опыт использования распределенной службы высокой готовности Flume для сбора, накопления и перемещения в хранилище Kafka больших объемов потоковых данных, получаемых из множества источников.

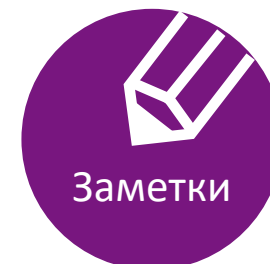


ДЕНИС ТУРДАКОВ,
ИСП РАН

АНАЛИЗ СОЦИАЛЬНЫХ СЕТЕЙ: ОХОТА НА БОТОВ И ТРОЛЛЕЙ

Социальные сети – огромный неконтролируемый мир общения и распространения сведений, где любой может размещать произвольные данные о себе, притворяясь, например вымышленной личностью. Возможность оперировать не одним аккаунтом, а сотнями виртуальных персонажей, распространять практически любую информацию, включая фейковую стала почвой для развития целой отрасли на стыке рекламы и PR. В докладе будут представлены технологии «Талисман» и «Текстерра», предназначенные для анализа больших массивов данных из социальных сетей, восстановления неполных профилей пользователей, выявления умышленных искажений, обнаружения виртуальных личностей (ботов) и анализа информационных кампаний с целью выявления троллей.

Предыдущая
секция



Следующая
секция

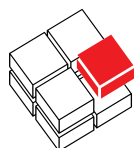


ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ



ПРАКТИЧЕСКИЕ СЕМИНАРЫ И КОНФЕРЕНЦИИ

ОРГАНИЗАТОРЫ



**ОТКРЫТЫЕ
СИСТЕМЫ**
Open Systems Publications

Издательство «Открытые системы», созданное в 1993 году для комплексной информационной поддержки профессионалов, отвечающих за построение масштабных компьютерных систем, своей основной целью видит предоставление полной и качественной информации для различных категорий читателей. Сегодня «Открытые системы» – ведущее издательство в России, производящее высокопрофессиональные издания для специалистов и руководителей из таких отраслей, как информационные технологии, компьютерные системы, нефтегазовая отрасль, телекоммуникации, сети передачи данных, полиграфия и медицина.

Подробнее: www.osp.ru

computerworld^{россия}

Computerworld Россия – ведущий международный еженедельник, посвященный информационным технологиям. Важнейшие события ИТ-индустрии в России и в мире, новые технологии, продукты и услуги, примеры успешных внедрений информационных систем на отечественных предприятиях.

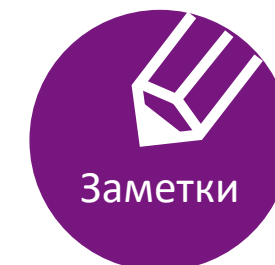
Подробнее: <http://www.osp.ru/cw/>

**Открытые
системы**
ЖУРНАЛ ДЛЯ АРХИТЕКТОРОВ ИНФОРМАЦИОННЫХ СИСТЕМ

Журнал «Открытые системы. СУБД» издается с 1993 года и сегодня является одним из самых авторитетных в России периодических изданий, посвященных построению гетерогенных информационных систем корпоративного уровня; архитектурам современных компьютерных систем и микропроцессоров; операционным системам; СУБД и хранилищам информации, а также веб-технологиям. Большое внимание на страницах журнала уделяется освещению практического опыта разработки и внедрения конкретных решений на промышленных предприятиях, в научных и государственных учреждениях, на транспорте, в медицине и в телекоммуникационных компаниях. В 2007 году журнал вошел в список изданий, рекомендованных Высшей аттестационной комиссией (ВАК).

Подробнее: <http://www.osmag.ru/>

Предыдущая
секция



Заметки



ТЕХНОЛОГИИ
БОЛЬШИХ ДАННЫХ



ПРАКТИЧЕСКИЕ
СЕМИНАРЫ
И КОНФЕРЕНЦИИ

ИНФОРМАЦИОННЫЕ ПАРТНЕРЫ

Информационные партнеры

Открытые
системы
ЖУРНАЛ ДЛЯ АРХИТЕКТОРОВ ИНФОРМАЦИОННЫХ СИСТЕМ

computerworld^{россия}

интерфакс
INTERFAX

Системный
администратор
ежемесячный журнал www.vest.ru

бизнес & информационные технологии
БИТ

ML Boot Camp

APPS  ALL



SQL
COM
СТИЛЬ ЖИЗНИ SQL

яИТЫ
КЛУБ ИТ-ДИРЕКТОРОВ

Pro
Bitcoin.
Главные новости из
мира криптовалют
и технологии блокчейн
VK.COM/PRO.BITCOIN



ПРАКТИЧЕСКИЕ
СЕМИНАРЫ
И КОНФЕРЕНЦИИ



ТЕХНОЛОГИИ
БОЛЬШИХ ДАННЫХ



ТЕХНОЛОГИИ
БАЗ ДАННЫХ



ТЕХНОЛОГИИ БЛОКЧЕЙНА

Russian
ITMF
IT Management Forum

**BIG
DATA**

**SMART
COMPANY**

RUS.NET
Российский сетевой форум
Ether.NET Open.NET Secure.NET IoT



форум
Бизне\$видео

По вопросам участия: Ольга Пуркина



+ 7 495 725 47 80



kon@osp.ru



www.osp.ru