УДК 004.67

БОЛЬШИЕ ДАННЫЕ В НАБЛЮДЕНИИ ЗА ОКЕАНОМ: ВОЗМОЖНОСТИ И ПРОБЛЕМЫ

С.Х. Хабибов
Студент БГУИР

С.Н. Нестеренков
Кандидат технических наук, доцент, декан факультета компьютерных систем и сетей

А.Н. Марков
Старший преподаватель, магистр технических наук, заместитель начальника Центра информатизации и инновационных разработок БГУИР

Центр информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники, Республика Беларусь
Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: habibov200@gmail.com, s.nesterenkov@bsuir.by, a.n.markov@bsuir.by

С.Х. Хабибов
Студент 4 курса специальности "Программное обеспечение информационных технологий" БГУИР.

С.Н. Нестеренков
Кандидат технических наук, декан факультета компьютерных систем и сетей Белорусского государственного университета информатики и радиоэлектроники, доцент кафедры Программного обеспечения информационных технологий. Автор публикаций на тему машинного обучения, алгоритмов принятия решений, искусственных нейронных сетей и автоматизации

А.Н. Марков
Магистр технических наук, старший преподаватель кафедры ПИКС, заместитель начальника Центра информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники.

Аннотация. Наблюдение за океаном играет важную роль в исследовании океана. Наука об океане вступает в эпоху больших данных с экспоненциальным ростом информационных технологий и достижениями в области океанических обсерваторий. Океанические обсерватории представляют собой совокупность платформ, способных нести датчики для отбора проб океана в соответствующих пространственно-временных масштабах. Данные, собранные этими платформами, помогают ответить на целый ряд вопросов фундаментальных и прикладных исследований. Учитывая огромный объем, разнообразие типов, постоянное измерение и потенциальное использование океана данные наблюдений – это типичный вид больших данных, а именно морских больших данных. Традиционная инфраструктура, ориентированная на данные, недостаточна для решения новых задач, возникающих в науке об океане. В этом документе обсуждаются некоторые возможные новые стратегии для решения проблем морских больших данных на этапах хранения, обработки и анализа данных. Геологический пример иллюстрирует значительное использование морских больших данных.

Ключевые слова: Большие данные, наблюдение за океаном, морские большие данные, инфраструктура.

Введение.

Океан покрывает более 2/3 поверхности Земли. Фитопланктон в поверхностном океане производит половину кислорода в результате фотосинтеза на Земле. Девяносто процентов тепла от глобального потепления было поглощено океаном. Независимо от того, где мы живем, океан влияет на нашу жизнь. Однако 95% океана остается неисследованным и недооцененным человеком [1]. Это требует понимания всех аспектов океана, а также его сложных связей с атмосферой Земли, сушей, льдом, морским дном и жизнью, включая человечество. Это важно не только для расширения знаний о нашей планете, но и для обеспечения долгосрочного благосостояния общества и для того, чтобы помочь человеку управлять окружающей средой.

Океанография эволюционирует от экспедиционной науки на корабле к распределенному подходу, основанному на обсерватории, облегчающему сбор данных долгосрочных временных рядов и предоставляющему интерактивную возможность проводить эксперименты с использованием потоковой передачи данных в режиме реального времени [2]. Данные наблюдений за океаном из нескольких источников собираются и хранятся в беспрецедентном масштабе и с беспрецедентной скоростью [3]. Основываясь на определении больших данных Gartner [4], данные наблюдений за океаном действительно имеют характеристики трех V (volume-объем, velocity-скорость и variety-разнообразие). Таким образом, данные наблюдений за океаном можно рассматривать как типичный вид больших данных, т.е. морские большие данные.

Эти данные должны храниться в формате raw, анализироваться, калиброваться и обрабатываться для контроля качества, затем анализироваться и далее выводиться в другие продукты, такие как визуализации [5]. Благодаря уникальным характеристикам морских больших данных они превосходят возможности обработки и анализа обычных систем. Эта ситуация вызвала новые проблемы для традиционных технологий, таких как реляционные базы данных и масштабируемые инфраструктуры [6]. Текущие исследования, связанные с большими данными, в первую очередь касаются того, как более эффективно обнаруживать и осмысливать такие большие объемы данных и эффективно [7]. Ключевые исследуемые вопросы включают инфраструктуру [8], хранение [9], анализ [10], безопасность [11] и т.д.

Сбор данных.

На этапе сбора данных океанские обсерватории, оснащенные различными датчиками, используются для сбора необработанных данных из океана.

Океанические обсерватории представляют собой совокупность платформ, способных нести датчики для сбора данных в определенных пространственно-временных масштабах. Эти платформы включают корабли, спутники и ряд эйлеровых и лагранжевых систем.

– Корабли были основным инструментом океанографов на протяжении веков и останутся центральным элементом инфраструктуры в обозримом будущем. Возможности кораблей значительно улучшились в системах удержания станции и динамического позиционирования, многолучевых гидролокаторов и гидролокаторов бокового обзора.

– Спутники представляют собой наиболее важную инновацию в области океанографических технологий в наше время. Это новые инструменты для понимания различных океанических процессов и взаимодействия суши, воздуха и моря в десятилетних временных масштабах. Спутниковые данные выявили новые явления, которые ранее были недоступны при использовании только данных наблюдений на месте.

– Электрооптические кабели на морском дне с высокой пропускной способностью и устойчивой мощностью предлагают потенциальные средства для обеспечения непрерывного наблюдения в океане. Кабели морского дна успешно использовались для

изучения широкого спектра, таких как сейсмичность морского дна, цунами, динамика морского дна, продуктивность прибрежных экосистем и т.д.

– Дрифтеры и поплавки – это пассивные лагранжевы платформы с батарейным питанием, используемые для создания поверхностных и подповерхностных карт океанских течений и свойств океана соответственно.

– Причалы обеспечивают средства для размещения датчиков на фиксированных глубинах между морским дном и поверхностью моря. Они предоставляют высокочастотные данные о недрах с фиксированным местоположением в дополнение к пространственным данным, собираемым судами, автономными подводными аппаратами и дистанционным зондированием со спутников.

– Планеры – это тип автономного подводного аппарата, использующего движитель на основе плавучести для преобразования вертикального движения в горизонтальное. Благодаря очень низкому энергопотреблению планеры предоставляют данные в больших пространственно-временных масштабах.

- Автономные подводные аппараты (AUVS) обеспечивают столь необходимую гибкость при наблюдениях за океаном, поскольку они позволяют перемещать датчики по воде в трехмерном пространстве. Они могут систематически и синоптически обследовать определенные линии, области и/или объемы.

Многие страны и организации внесли свой вклад в создание глобальных, региональных или местных систем наблюдения за океаном, используя различные платформы с несколькими датчиками на борту. Далее представляется несколько национальных или международных проектов по долгосрочному наблюдению за океаном.

– Argo – это глобальный набор из более чем 3000 свободно дрейфующих профилирующих поплавков, которые собирают высококачественные профили температуры и солености с верхних 2000 м свободного ото льда мирового океана и течений со средних глубин. Это позволяет осуществлять непрерывный мониторинг температуры, солености и скорости в верхних слоях океана. Развертывание началось в 2000 году, и национальным программам необходимо предоставлять около 800 поплавков в год для обслуживания Argo array. Широкомасштабный глобальный массив уже превратился в основной компонент системы наблюдений за океаном. Он основывается на других сетях наблюдений за океаном в верхних слоях океана. Это единственный источник глобальных наборов данных о недрах, используемых во всех моделях и анализах ассимиляции океанических данных.

– Ocean Networks Canada (INC), инициатива Университета Виктории, управляет ведущими в мире обсерваториями НЕПТУНА и ВЕНЕРЫ в северо-восточной части Тихого океана у западного побережья Канады. Его цели заключаются в предоставлении научных знаний и информации для эффективного управления океаном и ответственного использования океана на благо канадцев. ONC подключенные к кабелю обсерватории собирают данные, которые помогают ученым и руководителям принимать обоснованные решения о прибрежных землетрясениях и цунами, изменении климата, управлении прибрежными районами, сохранении и безопасности на море.

– Инициатива океанских обсерваторий (OOI) – это финансируемый национальным научным фондом комплексный инфраструктурный проект, состоящий из научно обоснованных платформ и сенсорных систем, которые измеряют физические, химические, геологические и биологические свойства и процессы от морского дна до границы воздух-море. OOI изменила исследования океанов, создав сеть интерактивных, распределенных по всему миру датчиков с доступом к данным почти в режиме реального времени, расширяя

наши возможности для решения таких важнейших проблем, как изменение климата, изменчивость экосистем, подкисление океана и круговорот углерода.

Хранение данных.

Собранные морские данные будут переданы в инфраструктуру хранения данных для дальнейшей обработки и анализа. Долгосрочный устойчивый сбор данных из нескольких источников приводит к быстрому расширению и усложнению данных. Это создает огромные проблемы при хранении и обработке этих данных. Наборы данных, хранящиеся в центре обработки данных, поступают от множества различных датчиков, размещенных на платформах дистанционного зондирования или на месте. Для оптимизации системы и с учетом возможностей хранения и скорости отклика, метаданных и некоторых типов данные хранятся в реляционных базах данных, а некоторые другие типы данных хранятся в файлах. Обычно типы данных с широким диапазоном параметров, но не слишком большим объемом данных, таких как питательные вещества, загрязнители и любые другие измерения образцов, хранятся в реляционных базах данных. Однако типы данных с небольшим количеством параметров, но огромным объемом данных, такие как CTD, ADCP и датчики изображений, хранятся в двоичных файлах, ASCII или файлах изображений.

Файловые системы, нижний уровень механизмов хранения, являются основой приложений на верхних уровнях. У многих компаний и исследователей есть свои решения для удовлетворения различных требований к хранению больших данных. Например, GFS от Google – это расширяемая распределенная файловая система для поддержки крупномасштабных распределенных приложений с интенсивным использованием данных. HDFS и Kosmosfs являются производными от открытых исходных кодов GFS. Корпорация Майкрософт разработала Cosmos для поддержки своего поискового и рекламного бизнеса. Facebook использует Haystack для хранения большого количества фотографий небольшого размера.

Традиционные реляционные базы данных не могут справиться с проблемами категорий и масштабов, связанными с большими морскими данными. Базы данных NoSQL становятся основной технологией для хранения больших данных. Базы данных NoSQL имеют гибкие режимы, поддержку простого и легкого копирования, простой API, возможную согласованность и поддержку больших объемов данных. В этой статье будут представлены три основные базы данных NoSQL, основанные на различных моделях данных: базы данных с ключевыми значениями, базы данных, ориентированные на столбцы, и базы данных, ориентированные на документы.

Базы данных с ключевыми значениями.

Базы данных ключевых значений составлены на основе простой модели данных, и данные хранятся в соответствии с ключевыми значениями. Каждый ключ уникален, и клиенты могут вводить запрашиваемые значения в соответствии с ключами. Такие базы данных имеют простую структуру, а современные базы данных с ключевыми значениями имеют более высокую расширяемость и более короткие время отклика на запрос больше, чем у реляционных баз данных. За последние несколько лет появилось много баз данных с ключевыми значениями, основанных на системе Amazon Dynamo.

Базы данных, ориентированные на столбцы.

Базы данных, ориентированные на столбцы, хранят и обрабатывают данные в соответствии со столбцами, отличными от строк. Как столбцы, так и строки сегментированы в нескольких узлах для обеспечения возможности расширения. Многие базы данных, ориентированные на столбцы, в основном вдохновлены BigTable от Google. Базовая структура данных BigTable представляет собой многомерное отображение

последовательностей с разреженным, распределенным и постоянным хранилищем. Индексы сопоставления – это ключ строки, ключ столбца и временные метки, и каждое значение в сопоставлении представляет собой неаналитический массив байтов.

Обработка и анализ данных.

Из-за множества источников, массивных, разнородных и динамических характеристик прикладных данных, задействованных в распределенной среде, одной из наиболее важных характеристик больших данных является выполнение вычислений на уровне петабайта (PB), даже на уровне эксабайта (EB) со сложным вычислительным процессом. Таким образом, использование инфраструктуры параллельных вычислений для эффективного анализа и извлечения распределенных данных является важнейшей целью обработки больших данных.

Вычислительная модель. Большие данные обычно хранятся на сотнях и даже тысячах коммерческих серверов. Таким образом, традиционные параллельные модели, такие как интерфейс передачи сообщений (MPI) и открытая мультипроцессорная обработка (OpenMP) может оказаться недостаточной для поддержки таких крупномасштабных параллельных программ. В последнее время некоторые предложенные модели параллельного программирования эффективно повышают производительность NoSQL и сокращают разрыв в производительности по сравнению с реляционными базами данных. Поэтому эти модели стали краеугольным камнем для анализа массивных данных.

– MapReduce – это простая, но мощная программная модель для крупномасштабных вычислений с использованием большого количества кластеров коммерческих ПК для достижения автоматической параллельной обработки и распределения. В MapReduce вычислительная модель имеет только две функции, то есть отображение и уменьшение. Функция Map обрабатывает входные пары ключ-значение и генерирует промежуточные пары ключ-значение. Затем MapReduce объединит все промежуточные значения, относящиеся к одному и тому же ключу, и передаст их в функцию Reduce. Пользователю нужно только запрограммировать две функции для разработки параллельного приложения.

Анализ данных.

Анализ данных является заключительным и наиболее важным этапом в цепочке создания ценных больших данных с целью извлечения потенциальных полезных значений и предоставления предложений или решений. Однако анализ данных – это обширная область, которая часто меняется и является чрезвычайно сложной. Многие традиционные методы анализа данных все еще могут использоваться для анализа больших данных, такие как кластерный анализ, факторный анализ, корреляционный анализ, регрессионный анализ, A / B-тестирование, статистический анализ, интеллектуальный анализ данных и т.д. Некоторые методы анализа больших данных могут быть использованы для ускорения извлечения ключевой информации из массивные данные. В настоящее время основными методами обработки больших данных являются: фильтр, хеширование, индекс, триэль, параллельные вычисления и т.д.

Для приложений анализа морских данных интеллектуальный анализ данных является важным методом извлечения скрытой, неизвестной, но потенциально полезной информации и знаний из массивных, неполных, зашумленных, нечетких и случайных данных. В 2006 году Международная конференция IEEE по серии интеллектуального анализа данных (ICDM) определила десять наиболее влиятельных алгоритмов интеллектуального анализа данных, включая C4.5, k-means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes и CART. Эти десять алгоритмов охватывают классификацию, кластеризацию, регрессию, статистическое обучение, анализ ассоциаций и

интеллектуальный анализ связей, все из которых являются наиболее важные темы в области исследований и разработок в области интеллектуального анализа данных. Чтобы адаптироваться к многоисточниковым, неопределенным, динамичным морским большим данным, существующие методы интеллектуального анализа данных должны быть расширены во многих отношениях.

Параллельная обработка была основным направлением разработки эффективных платформ обработки данных, чтобы данные могли обрабатываться распределенным и параллельным образом, повышая пропускную способность обработки данных. MapReduce - наиболее репрезентативная парадигма. Современные исследования в области анализа больших данных сосредоточены в основном на использовании парадигмы программирования MapReduce и экосистемы Hadoop, что привело к появлению ряда СУБД, которые могут быть развернуты в распределенной облачной среде, таких как Pig и Hive.

После распараллеливания алгоритмов традиционные программные средства анализа получат возможность обработки больших данных. Das и др. интегрировали R, инструмент статистического анализа с открытым исходным кодом, и Hadoop для улучшения слабой масштабируемости традиционного инструмента анализа и слабых аналитических возможностей Hadoop. Глубокая интеграция переводит обработку данных на параллельную обработку, что обеспечивает мощные возможности глубокого анализа для Hadoop. Стандартный Weka, инструмент машинного обучения и интеллектуального анализа данных с открытым исходным кодом, может запускать только на одном компьютере с ограничением в 1 ГБ памяти. Wegener и др. интегрировали Weka и MapReduce, чтобы преодолеть ограничения, используя преимущества параллельных вычислений для обработки данных объемом более 100 ГБ в кластерах MapReduce. В последние годы извлечение ценной информации и глубоких знаний из больших данных стало насущной необходимостью во многих дисциплинах. Из-за его высокого влияния во многих областях было разработано много систем и аналитических инструментов для анализа больших данных, таких как Apache Mahout, MOA, CAMOA и Vowpal Wabbit.

Применение морских больших данных.

Большие данные о море имеют основополагающее значение для различных областей исследований в области биологии, наук о земле, наук об океане и атмосфере.

Цунами – это массивная, быстро движущаяся волна, созданная подводным землетрясением или оползнем. Большой объем воды, вытесненный внезапным движением морского дна, создает импульс в океане, который исходит из своего источника со скоростью до 500 миль в час и простирается на тысячи футов ниже поверхности. Хотя и редкие, цунами, подобные тем, что произошли в марте 2011 года в Японии и в декабре 2004 года вокруг Индийского океана был трагическим напоминанием о разрушительной силе океана. В результате правительства стран, прилегающих к Тихому и Индийскому океанам, с помощью ученые со всего мира постоянно следят за океанским дном на предмет возможной сейсмической активности, вызывающей цунами, и быстро меняющихся признаков цунами в открытом океане. Даже предупреждение за несколько минут может означать разницу между широкомасштабной катастрофой и спасением сотен или тысяч жизней.

По данным геологической службы США, 1 апреля в 4:46:45 по тихоокеанскому дневному времени (23:46:45 UTC) у тихоокеанского побережья Чили произошло землетрясение магнитудой 8,2 балла. Приборы Ocean Networks Canada зафиксировали как сотрясение грунта, так и очень небольшое цунами, когда они пересекали северо-восточную часть Тихого океана (показано на рис. 1).

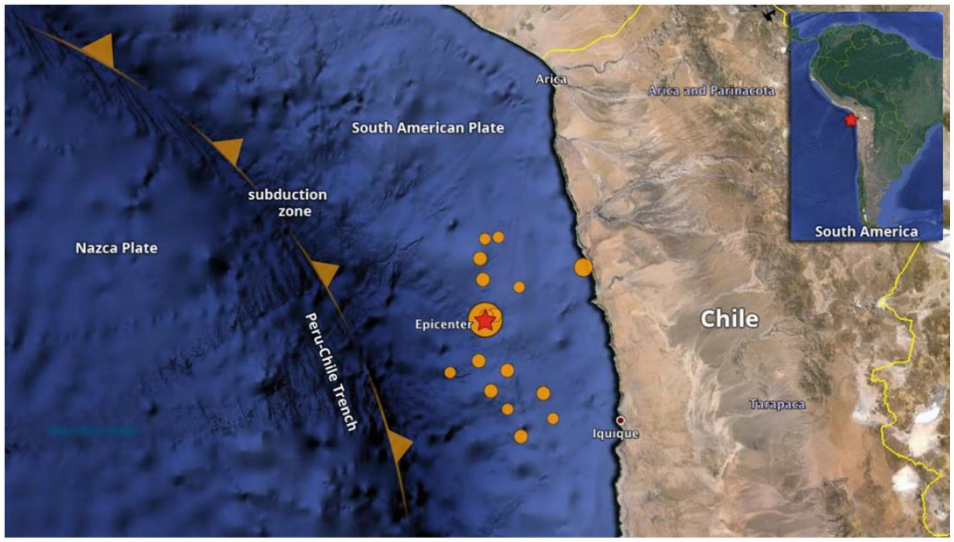

Рисунок 1 – Карта эпицентра и 16 подземных толчков вдоль зоны субдукции между плитами Наска и Южной Америки, 1 апреля 2014 г.

Заключение.

Наука об океане вступает в эпоху больших данных с экспоненциальным ростом информационных технологий и достижениями в области океанических обсерваторий. Однако морские большие данные все еще находятся в зачаточном состоянии. Многие ключевые технические вопросы, такие как хранение больших данных, вычислительная модель, метод анализа и прикладная система, поддерживающая принятие решений, должны быть полностью исследованы.

Инфраструктура: Различные океанические обсерватории непрерывно собирают и передают данные. Объем данных достигает беспрецедентных масштабов, которые превысят возможности хранения и обработки существующих инфраструктур. Традиционная инфраструктура, ориентированная на данные, в которой центральная система управления данными принимает данные и предоставляет их пользователям на основе запросов, недостаточна для выполнения целого ряда научных задач, например, сбора данных в реальном времени, анализа данных и моделирования океана в различных масштабах и обеспечение возможности адаптивных экспериментов в океане. Все более растущие данные и их требования в режиме реального времени вызывают проблемы с хранением и управлением такими огромными разнородными наборами данных при умеренных требованиях к аппаратной и программной инфраструктуре.

Список литературы

[1] Hole Oceanographic Institution. http://www.whoi.edu/Woods.

[2] Schofield, O., et al.: Automated sensor network to advance ocean science. EOS Trans. Am. Geophys. Union 91(39), 345–346 (2010).

[3] Chave, A.D., et al.: Cyberinfrastructure for the US Ocean Observatories Initiative: enabling interactive observation in the ocean. In: OCEANS 2009 – EUROPE, Bremen. IEEE (2009).

[4] Beyer, M.A., Laney, D.: The Importance of ‘Big Data’: A Definition. Gartner Inc., Stamford (2012).

[5] Farcas, C., et al.: Ocean Observatories Initiative scientific data model. In: OCEANS 2011, Waikoloa, HI. IEEE (2011).

[6] Park, K., Nguyen, M.C., Won, H.: Web-based collaborative big data analytics on big data as a service platform. In: 2015 17th International Conference on Advanced Communication Technology (ICACT), Seoul. IEEE (2015).

[7] Bellatreche, L., Furtado, P., Mohania, M.K.: Guest editorial: a special issue in physical design for big data warehousing and mining. Distrib. Parallel Databases 34(3), 289–292 (2015).

[8] Demchenko, Y., Laat, C., Membrey, P.: Defining architecture components of the Big Data Ecosystem. In: 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN. IEEE (2014).

[9] Du, Y., et al.: Study of migration model based on the massive marine data hybrid cloud storage. In: 2012 First International Conference on Agro-Geoinformatics (AgroGeoinformatics), Shanghai. IEEE (2012).

[10] Huang, D., et al.: Modeling and analysis in marine big data: advances and challenges. Math. Probl. Eng. 2015, 1–13 (2015).

[11] Yang, K., et al.: Enabling efficient access control with dynamic policy updating for big data in the cloud. In: 2014 Proceedings IEEE INFOCOM, Toronto, ON. IEEE (2014).

BIG DATA IN OCEAN OBSERVATION: OPPORTUNITIES AND CHALLENGES

S.KH. HABIBOV
Student of Belarusian State University of Informatics and Radioelectronics

S.N. NESTERENKOV
PhD, Associate Professor
Dean of Faculty of Computer Systems and Networks

A.N. MARKOV
Senior lecturer of the department, Deputy head of the Center for Informatization and Innovative Developments

Center for Informatization and Development of the Belarusian University of State Informatics and Radioelectronics, Republic of Belarus.
Belarusian State University of Informatics and Radioelectronics, Republic of Belarus.
E-mail: habibov200@gmail.com, s.nesterenkov@bsuir.by, a.n.markov@bsuir.by

Abstract. Ocean observation plays an essential role in ocean exploration. Ocean science is entering into big data era with the exponentially growth of information technology and advances in ocean observatories. Ocean observatories are collections of platforms capable of carrying sensors to sample the ocean over appropriate spatio-temporal scales. Data collected by these platforms help answer a range of fundamental and applied research questions. Given the huge volume, diverse types, sustained measurement and potential uses of ocean observing data, it is a typical kind of big data, namely marine big data. The traditional data-centric infrastructure is insufficient to deal with new challenges arising in ocean science. This paper discusses some possible new strategies to solve marine big data challenges in the phases of data storage, data computing and analysis. A geological example illustrates the significant use of marine big data.

Keywords: Big data, ocean observation, marine big data, infrastructure.