К концепции информационно-аналитической системы поддержки научных исследований, основанных на интенсивном использовании цифровых данных

А. В. Юрченко
Институт вычислительных технологий СО РАН, Новосибирск, Россия
Контактный e-mail: yurchenko@ict.nsc.ru

Рассматривается вопрос создания инфраструктуры и инструментов для работы с большими объемами научных данных. Актуальность вопроса усиливается из-за экспоненциального роста объемов научных данных и тренда на открытую науку и открытые данные.

Цель работы — формирование идеологии, описание концептуальных аспектов и некоторых технологических вопросов создания информационно-аналитической системы для исследователей, работающих с научными данными.

Определены место такой системы в исследовательском процессе, основные функциональные требования к ней, сформулированы ключевые установки для создания системы, в том числе концепт "просветления" данных. Перечислены базовые программные и аппаратные блоки системы и описано текущее состояние развития инфраструктуры Института вычислительных технологий для работы с научными данными.

Ключевые слова: инфраструктура научных исследований; наука, основанная на данных; информационная система; научные данные; хранение и обработка данных.

Введение

Цифровые данные стали одним из важнейших источников получения новых знаний. Тем не менее российские исследователи продолжают хранить свои научные данные преимущественно на персональных компьютерах, съемных устройствах, в лучшем случае — в персональных хранилищах данных. Организация данных, их каталогизация, как правило, осуществляются в ручном режиме, аналогично решаются и вопросы поиска. В результате в накопленных массивах данных способны разобраться лишь создавшие их исследователи. Инструменты для работы с данными, их обработки и анализа зачастую не упорядочены, а связь данных и таких инструментов существует лишь в голове пользователя. Все это снижает эффективность работы с данными, сильно ограничивает возможности их повторного использования, построения и анализа расширенных наборов данных, полученных из различных источников. Создание и предоставление ученым интегрированных информационно-аналитических и вычислительных инструментов, облегчающих им работу со своими данными и данными, находящимися в открытом доступе, — насущная задача развития инфраструктуры научных исследований в России.

В европейском и американском научных сообществах проблема развития инфраструктуры для хранения научных данных рассматривается в плоскости идеологий открытых данных и открытой науки. Кроме философских дискуссий о необходимости открытия научных данных [1, 2] и организационных вопросов о регламентах их открытия (опубликования) с сохранением научного приоритета [3] стоят и организационно-технические задачи формирования и принятия стандартов [4], а также ключевой вопрос — кто будет платить за инфраструктуру и ее содержание [5, 6]. Но, несмотря на нерешенные проблемы как организационного, так и технического плана, и Европа и Соединенные Штаты Америки двигаются в сторону открытой науки и открытых данных и строят соответствующую инфраструктуру.

Интеграция разнородных данных, построение каталогов и информационных систем для работы с данными, а также разработка методов и алгоритмов их анализа — одно из ключевых направлений деятельности Института вычислительных технологий (ИВТ) СО РАН [7, 8]. В ИВТ успешно создаются сложные информационно-аналитические системы для работы с библиотечными данными [9], данными дистанционного зондирования Земли [10] и др., созданы большие системы хранения и обработки научных данных. Кроме того, ИВТ является координатором развития информационно-телекоммуникационной инфраструктуры поддержки научных исследований в Сибири, объединяющей корпоративной компьютерной сетью все институты СО РАН.

Сегодня крупные сибирские исследовательские центры все интенсивнее используют измерительную технику и приборы, генерирующие большие объемы данных (см., например, [11]). Их потребности в ресурсах и инструментах для хранения и обработки этих данных постоянно растут. В складывающейся ситуации целесообразно создать и поддерживать такие инструменты и ресурсы именно в ИВТ, как обладающем достаточной компетенцией и необходимой для этого базовой инфраструктурой. Эти инструменты необходимо интегрировать в единую систему поддержки научных исследований, которая должна стать доступной для российских ученых.

1. Цели и задачи создания системы

Отдельные компоненты инфраструктуры для работы с данными научные учреждения в состоянии создавать самостоятельно или арендовать у коммерческих провайдеров, так как это не требует специальных научных компетенций и такие задачи могут быть реализованы квалифицированными системными администраторами. Однако, когда речь идет об интеграции данных, о проблемах их долговременного и надежного хранения, гибкого управления доступом к ним, имплементации с возможностью оркестровки комплексов программ и алгоритмов их обработки, проблема выходит за рамки возможностей типового отдела ИТ-поддержки научной организации и становится самостоятельным предметом исследований. Масштаб перечисленных задач позволяет говорить о необходимости создания целой платформы, включающей как программные продукты и средства их разработки и интеграции, так и аппаратные ресурсы для хранения, обработки и анализа данных.

Таким образом, формируется цель, ориентированная на поддержку научных исследований, основанных на интенсивном использовании цифровых данных, — создание и обеспечение функционирования информационно-аналитической системы на базе цифровой платформы распределенного типа для сбора и хранения, обработки и анализа, обмена и публикации научных данных.

Для создания востребованной информационной системы необходимо решить ряд таких традиционных задач разработки программного продукта, как сбор и анализ потребностей потенциальных пользователей, определение на их основе набора функциональных требований к системе, формирование стратегии развития продукта, выделение базового набора требований для создания минимального жизнеспособного продукта (MVP — Minimal Viable Product), разработка модели MVP и модели его реализации в рамках этой стратегии, подготовка технического задания на MVP и его реализация. Дальнейшее развитие продукта должно проводиться путем итерационной реализации новых возможностей на основе анализа изменений в потребностях пользователей, внесения соответствующих поправок в набор функциональных требований с целью максимально повысить эффективность использования.

2. Потребности исследователей — пользователей системы

Прежде чем определить потребности потенциальных пользователей системы, найдем ее место в рамках процесса научных исследований. Схема, представленная на рис. 1, в обобщенном и упрощенном виде описывает типовой процесс изучения некоторого объекта или процесса (либо их множеств/комплексов) в нотации IDEF0 [12]. Обладая некоторыми научными знаниями или доступом к ним, владея методиками проведения научных исследований, действуя в рамках заданных правил (или выходя за них при необходимости), исследователь использует специализированные инструменты для проведения научных исследований, чтобы получить новые научные знания об исследуемом объекте или процессе.

Частичная детализация схемы применительно к процессам проведения измерений и обработки их результатов с использованием компьютерной системы представлена на рис. 2. Здесь опущены блоки, связанные с выделением из методики проведения исследований измерительных методик и стандартов, регламентов и инструкций, абстрактный “исследователь” заменен на процесс-специфичных экспериментатора, предметника и аналитика данных. На этой IDEF0-схеме место создаваемой системы обозначено

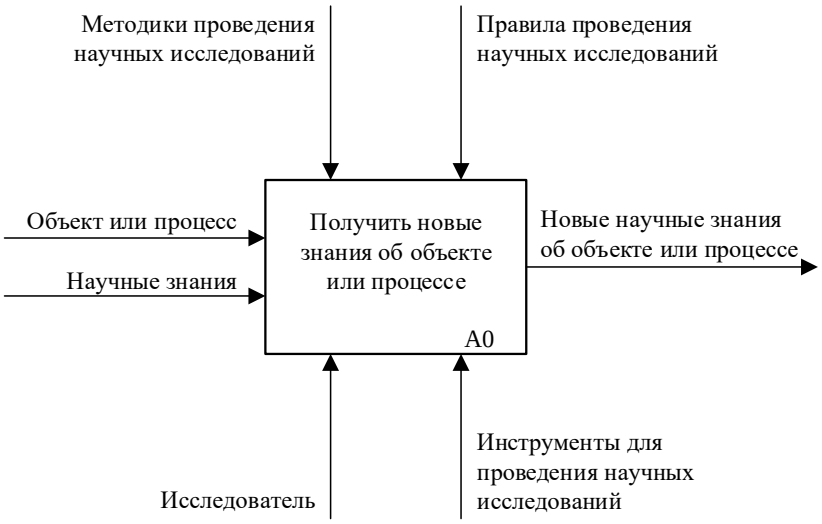

Рис. 1. Схема верхнего уровня исследовательского процесса

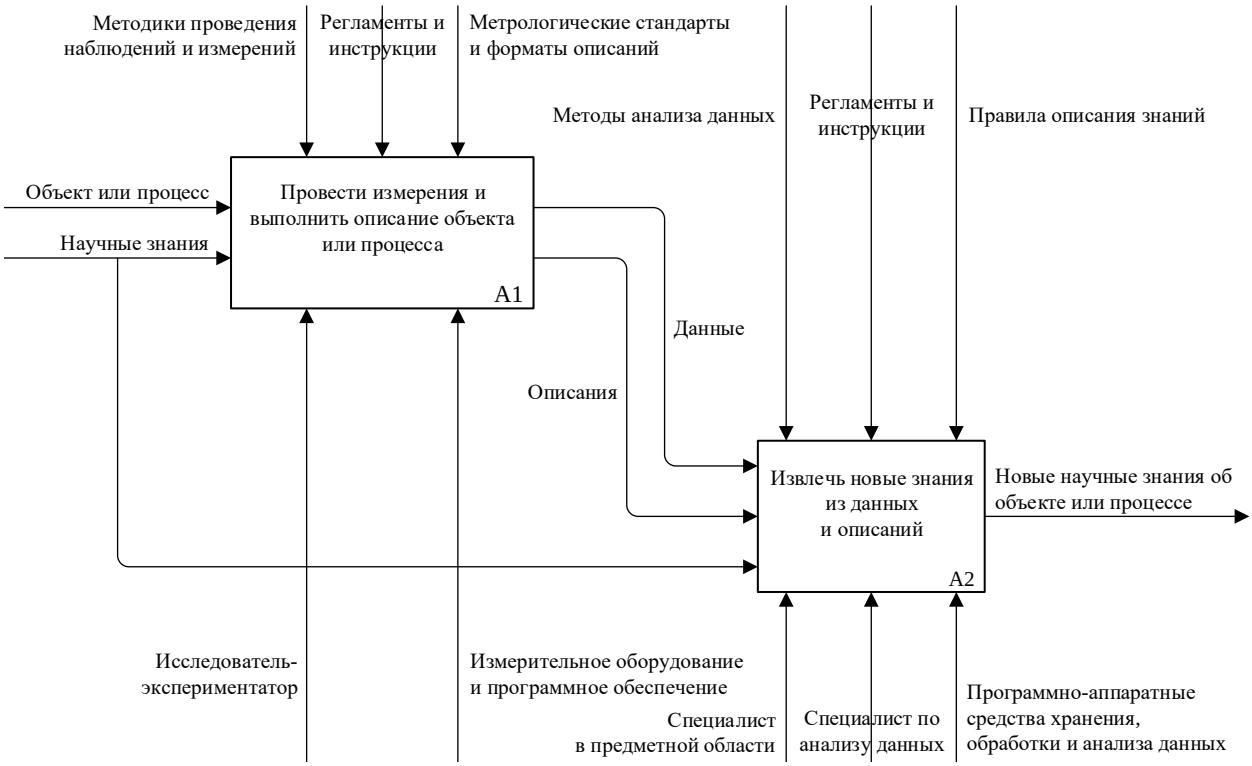

Рис. 2. Схема, детализирующая процесс получения и обработки данных

как один из механизмов извлечения новых знаний из данных и описаний (“Программно-аппаратные средства хранения, обработки и анализа данных”), которое и определяет круг основных пользователей системы и позволяет перейти к формированию списка функциональных требований к ней.

Список потребностей в порядке “цепочки использования” научных данных может выглядеть следующим образом: сбор и предварительная обработка — хранение — организация/систематизация — поиск и отбор — обработка и компьютерный анализ — визуализация — обмен и публикация.

В качестве базовой потребности при работе с данными будем рассматривать их хранение. Ее ключевыми атрибутами являются:

- быстрая загрузка данных в хранилище;
- быстрое получение данных из хранилища;
- гибкое управление доступом;
- надежность и безопасность хранения.

К порожденным и дополнительным атрибутам этой потребности относятся возможности: систематизации и поиска, хранения данных “как можно ближе к потребителю”, распространения и публикации данных, быстрого “прямого” доступа к данным, автоматического или автоматизированного сбора данных, обеспечение юридической чистоты при хранении и публикации данных.

Второй базовой потребностью является обработка и анализ данных с такими ключевыми атрибутами, как:

- быстрая обработка больших объемов и выборок данных;

- наличие специализированных методов и алгоритмов для этих целей и возможность использования (интеграции) собственных методов и алгоритмов;
- возможность работы с разнородными данными.

К порожденным и дополнительным атрибутам относятся возможности: использования данных других пользователей, построения выборок данных и работы с ними, построения цепочек преобразований данных, использования вычислительных ресурсов различной архитектуры, распределенной обработки данных с оркестровкой вычислительных ресурсов, в том числе ресурсов третьих лиц (сторон).

Третья базовая потребность организация данных основывается на их описании и работе с описаниями. К ее атрибутам относятся возможности:

- структурированного и произвольного описания;
- расширения и уточнения (изменения) описаний;
- автоматизированного и автоматического извлечения метаданных и структуризации описаний;
- автоматического описания типовых данных, например собираемых в автоматическом режиме.

На основе этого списка с учетом разнообразия научных данных, их пользователей и источников можно сформулировать ряд ключевых установок, которым должна удовлетворять система:

- приоритет доступности и удобства использования (*Usability@Top*);
- хранение всех видов научных данных в любых форматах (*StoreEverything*);
- гибкое управление доступом к данным с обеспечением различных уровней их приватности и доступности (*FromPrivate2Public*);
- интеграция данных из всех возможных источников для решения конкретных задач (*IntegrateData*);
- объединение всех доступных ресурсов для обработки данных (*CombineResources*);
- автоматизированное формирование пула инструментов для работы с данными на основе любой доступной информации о них (*UseEverythingKnown*);
- автоматическое и автоматизированное расширение представления о данных из всех доступных источников (*EnlightFromAnywhere*);
- сохранение всех действий с данными и создание шаблонов действий и последовательностей действий (*RememberEverything*);
- построение системы взаимосвязей данных друг с другом, алгоритмами и методами, описаниями, источниками, объектами и субъектами исследований и т. д. (*OntologizeAll*).

3. Ключевые установки системы

Вопросы доступности и удобства использования при создании научным сообществом для собственных нужд программных продуктов и информационных систем традиционно остаются на втором плане. Однако именно эти аспекты являются ключевыми при оценке пригодности продукта для большинства пользователей, не исключая исследователей. Закладывая принцип *Usability@Top* в качестве ключевого, мы руководствуемся тем, что удобством пользователя нельзя пренебрегать, а также признаем, что востребованность системы определяется не только ее функциональностью, но и тем, насколько удобно ей пользоваться. Реализация принципа может осуществляться путем анализа частоты использования того или иного функционала и построения иерархии функци-

ональных возможностей системы с учетом этого фактора так, чтобы наиболее востребованные функции были наиболее доступны в пользовательском интерфейсе. В основу пользовательского интерфейса для реализации установки *Usability@Top* необходимо положить принцип разумного минимализма для того, чтобы исключить его перегруженность маловостребованными функциями, что так характерно для продуктов с открытым кодом.

Установка *StoreEverything* — шаг в сторону от специализированных систем хранения научных данных, таких как базы геномов или каталоги спутниковых снимков и результатов их обработки. Полагая высокую степень универсальности в отношении загружаемых данных, с одной стороны, мы уменьшаем возможности и усложняем имплементацию узкоспециализированных методов, алгоритмов и программ для работы с конкретными типами данных, с другой — закладываем потенциал для совместного комплексного анализа данных разных типов из различных источников, связанных, например, через объекты или субъекты исследований либо другим образом. Реализация этой установки требует, чтобы единицей хранения была абстракция высокого уровня, такая как каталог файлов. Чтобы применять к этим данным различные встроенные в систему и внешние средства обработки и анализа данных, потребуется дополнительное уточнение типов данных и другой информации, позволяющей определить допустимые виды операций с этими данными. Тем не менее *StoreEverything* в сочетании с остальными ключевыми установками может открыть исследователям новые горизонты по работе с научными данными для поиска закономерностей в традиционно неинтегрируемых наборах данных.

Наличие подсистемы управления доступом к данным — естественное требование к любой системе, где могут храниться данные многих пользователей. При разработке и реализации системы управления доступом к научным данным возникает противопоставление потребностей научного сообщества в открытости данных и желания их владельцев сохранить научный приоритет при использовании таких данных путем ограничения к ним доступа, а также требований законодательства о защите персональных данных. Наиболее простым путем решения этой проблемы остается, с одной стороны, делегирование управления доступом к данным и возложение ответственности за их распространение непосредственно на пользователя-владельца, с другой — разработка и соблюдение регламентов принудительного открытия или ограничения доступа к данным. Пользователь-владелец должен иметь возможность как хранить данные в закрытом виде, так и предоставлять их научному сообществу в различных формах, от обмена с избранными коллегами до публикации в открытом виде, соблюдая при этом установленные регламенты и действующие законодательные нормы. В этом суть установки *FromPrivate2Public*.

Вопрос интеграции данных имеет два аспекта. Первый — уже упомянутая интеграция разнородных данных, связанных, например, через объекты или субъекты исследований, для совместного анализа. Второй — интеграция данных из внешних источников. Для последнего необходимо предусмотреть реализацию стандартных механизмов обмена, протоколов и внешних программных интерфейсов, позволяющих подключать данные из других доступных пользователю источников к системе и манипулировать ими. Первый аспект становится важнейшим инструментом развития науки, основанной на данных. Совместный комплексный анализ разнородных данных и данных, полученных из различных источников, возможен при наличии соответствующих аналитических программных средств и инструментов агрегации и интеграции таких данных в виртуально

единые наборы данных. Необходимость реализации того и другого закладывается в систему установкой *IntegrateData*.

Комбинирование вычислительных систем предполагает возможность для пользователя подключения к решению исследовательских задач всех доступных ему вычислительных ресурсов. Под вычислительными ресурсами нужно понимать и низкоуровневые сервисы, связанные с выделением вычислительных серверов или кластеров, и высокоуровневые, предоставляющие готовые методы и алгоритмы обработки и анализа данных. Ключевым аспектом реализации установки *CombineResources* должны стать возможности создания сложных вычислительных схем в виде рабочих процессов (workflows) и автоматической или автоматизированной оркестровки вычислительных сервисов под каждую конкретную задачу.

Для реализации установок *IntegrateData* и *CombineResources* в сочетании со *StoreEverything* и для того, чтобы мотивировать использование системы не только с целью долговременного хранения данных и обмена ими, но и применения встроенных и внешних аналитических инструментов различного уровня сложности, добавляется блок установок, связанных с описанием данных, их организацией и формированием связей, в том числе с аналитическими инструментами.

Так, установка *UseEverythingKnown* ориентирует систему на накопление "знаний" о данных, которые позволяют подбирать комплексы инструментов обработки и анализа, специфичные для конкретных данных или их наборов. Различные формы фильтрации инструментов в зависимости от объекта применения практикуются во всех современных системах работы с документами и данными. Не до конца решенный вопрос, который предстоит исследовать в рамках создания системы, — это связка интеграции разнородных данных и инструментов для работы с ними. Выборки данных, в том числе разного типа, но связанных через объект исследований, географически или по времени и т. д., представляют особый интерес для комплексного анализа научной проблемы и поиска сложных взаимосвязей между событиями и явлениями. Поэтому для такого анализа задача подбора аналитических инструментов чрезвычайно важна и требует оперативного решения. Осуществлять такой подбор можно, имея максимально глубокие представления о данных и их наборах, получить которые за один раз и из одного источника практически невозможно. Кроме того, каждое использование данных дает о них новую информацию, новые знания, которые также целесообразно использовать при их дальнейшем анализе и подборе для него специализированных аналитических инструментов.

Установка *EnlightFromAnywhere* ориентирует систему на сбор всей возможной информации о данных и их наборах из всех возможных источников в автоматическом и автоматизированном режимах. Процесс расширения представлений о данных и их наборах в ходе жизненного цикла использования назовем словом "просветление". Разберем подробнее этот термин.

Система изначально будет принимать данные как "черный ящик" (установка *StoreEverything*). Далее она будет позволять работать с данными, применяя различные инструменты в зависимости от имеющихся знаний об этих данных (установка *UseEverythingKnown*). Состояние "белый ящик" — когда данные описаны полностью, установлено, какого они типа, известны методы работы с этим типом данных, доступны соответствующие инструменты, также известно, кем и когда, с какой целью получены данные, к какому объекту исследований относятся и т. д., что достижимо далеко не всегда. Кроме того, особую ценность данные приобретают, обрастая "соседями", т. е.

данными и информацией, связанными с основными данными, но, возможно, не явно и не напрямую. Состояние “белый ящик” — идеальное и труднодостижимое, к которому система и пользователь стремятся привести данные в процессе “просветления”.

Просветление — это процесс “превращения” “черного ящика” в “белый ящик” и наполнения его информацией и системой связей с другими накопленными данными и информацией. Типичное состояние данных или наборов данных в ходе этого процесса — “серый ящик”, который “светлеет” со временем при обращении к этим данным, их использовании, добавлении новой информации о них, формировании и расширении описаний. Так можно объяснить термин “просветление”. Здесь можно проследить аналогию с накоплением данных дистанционного зондирования об участке Земли, часто закрываемом облаками. Для такого участка “чистый” снимок получить за один раз крайне сложно или невозможно, но можно сформировать его путем многократных наблюдений и дополнения новыми “чистыми” фрагментами в ходе таких наблюдений. Основные механизмы “просветления” при работе с произвольными данными — это дополнение, расширение, уточнение самих данных и их описаний, в том числе в результате анализа, а также построение связей с другими данными и накопленной в системе информацией.

Для реализации установки *EnlightFromAnywhere* необходимо предоставить пользователям как средства для дооописания данных и наборов данных (при этом на начальном этапе описание может быть даже пустым), так и инструменты для построения, указания связей данных с другими объектами в системе, в том числе с другими данными, описаниями данных, документами и др. Также необходимо реализовать методы и алгоритмы, автоматизирующие этот процесс. В результате работы этих механизмов данные начнут обрастать связями, которые со временем могут приобретать самостоятельную ценность, формируя своеобразные онтологии для данных и их наборов.

Создание комплексов связей — новый инструмент для работы с данными. Его целью является построение аналога социальной сети, но только для данных, документов, субъектов и объектов исследований, иных единиц хранения в системе. Каждая такая единица хранения или их набор могут быть связаны с другими единицами хранения через какой-либо признак или систему признаков. Типичные признаки, которые могут связывать единицы хранения в системе, — это объект исследований (измерений и т. п.), географическое место, время, субъект исследований (человек или группа, прибор или комплекс приборов), тип измерений и др. Система связей может быть представлена в виде сложного графа с неэквивалентными ребрами, которые могут быть направленными или ненаправленными, его узлы могут иметь, например, транзитивные свойства либо нет и т. д. Построение связей может не иметь прямого влияния на данные и их описание, однако оно расширяет возможности исследователя, так как “притягивает” к набору данных другие источники информации и знаний (данные, описания, документы и пр.) для их уже совместного анализа.

Построение таким образом своеобразных онтологий вокруг набора данных — одна из ключевых установок системы, которую назовем *OntologizeAll*. Ее реализация должна приблизить систему к выполнению ее миссии: способствовать накоплению критической массы информации и данных ключевых типов, чтобы у исследователя могла произойти “революция в понимании” результатов исследований и он смог перейти от описания отдельных случаев к синтетическому восприятию исследуемых объектов и процессов, выявлению и описанию закономерностей на основании анализа накопленных данных и дальнейшему обобщению результатов в виде новых законов.

4. О модели системы и ее проектировании

Процесс работы исследователя с данными направлен непосредственно на получение с их помощью новых знаний. Однако в реальности работа с данными имеет множество, в том числе отложенных, активностей, реализация которых и может привести к желаемому результату — получению новых знаний. Эти активности включают процессы получения (сбора), передачи и хранения научных данных, их обработку и анализ, обмен и публикацию (распространение). В реализацию активностей могут быть вовлечены как владельцы данных, так и сторонние исследователи — пользователи и эксперты, администраторы информационно-аналитических систем и привлекаемых ресурсов, разработчики ресурсов и их модулей-сервисов (рис. 3). А для обеспечения высокого уровня доступности системы в основе ее программной части должна лежать надежная аппаратно-системная инфраструктура, включающая хранилище, вычислительную часть и высокоскоростную телекоммуникационную среду.

Информационная система будет состоять из таких трех традиционных частей (уровней реализации), как:

- аппаратно-системный (infrastructure layer);
- программная платформа (backbone layer);
- приложения (frontend layer).

Схематически эти уровни представлены на рис. 4.

Аппаратно-системный уровень состоит из телекоммуникационной среды (TelecomResources), обеспечивающей высокоскоростной обмен данными между прочими компонентами инфраструктуры, инженерной инфраструктуры центров обработки данных (DatacenterInfrastructure), систем хранения данных (StorageResources), серверов и кластеров для обработки данных и вычислений (ComputingResources), а также операционных систем и программно-задаваемых систем управления потоками данных и их хранением (SystemLayerApps).

Программная платформа промежуточного уровня содержит файловые каталоги и базы данных (Storage&DB), подсистему управления доступом и защиты информации,

Рис. 3. Облако системы и ее основные внешние связи

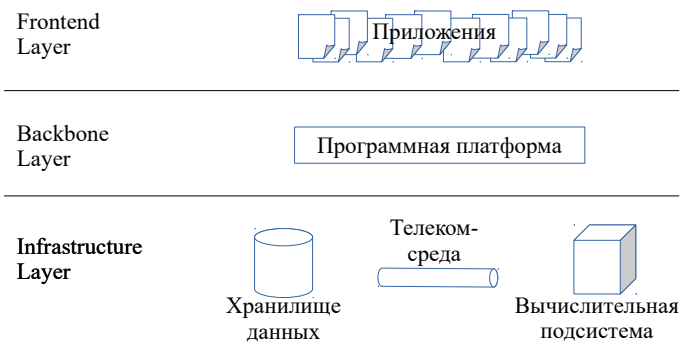

Рис. 4. Уровни реализации информационной системы

в том числе разграничения прав и шифрования (Authx2Ware), базовую подсистему доступа к данным (AccessWare), подсистему обеспечения обработки данных (ProcessWare), подсистему описания данных (DescriptWare), подсистему поиска и агрегации данных (Search&CollectEngine).

Приложения условно можно разделить на базовые (BasicApps), “простые” приложения для преобразования данных (SimpleDPApps — Simple Data Processing Applications), и “продвинутые” (AdvDPApps — Advanced Data Processing Applications). Базовые приложения должны предоставлять ключевые функции по работе с любыми данными, в том числе находящимися в состоянии “черного ящика”, — это аутентификация пользователя, загрузка и выгрузка данных, управление доступом, поиск и построение выборок, создание описаний и их изменение, подключение внешних источников данных и вычислительных ресурсов. “Простые” приложения для преобразования данных должны работать с базовыми типами данных и предоставлять возможности, условно независимые от области происхождения данных, такие как простые математические преобразования, неспецифичная кластеризация и устранение шумов и др. “Продвинутые” приложения существенно преобразуют данные, учитывая их специфику, порождают на их основе новые продукты, извлекают информацию и формируют новые “знания”. Это различные методы и механизмы анализа данных, такие как кластеризация и классификация, редукция/аппроксимация, синтаксический, статистический, интеллектуальный анализ и др. Список приложений для работы с данными должен включать алгоритмы классификации и кластеризации данных по типу, верификации типа данных, классификации, кластеризации и верификации самих данных, синтаксического и семантического анализа документов и описаний, детерминированного и статистического анализа, интеллектуального анализа, в том числе механизмы глубокого обучения.

Эти пожелания весьма широки, и их список и набор желательных атрибутов будут быстро дополняться при дальнейшем изучении запросов потенциальных пользователей. Однако потребность в системе поддержки исследований, основанных на использовании цифровых данных, уже назрела, и потенциальные пользователи вынуждены осваивать существующие продукты и услуги, чаще всего не дающие исследователю необходимых специальных возможностей, которые в итоге приходится искать и приобретать дополнительно. Поэтому, для того чтобы максимально оперативно ввести систему в эксплуатацию и предоставить пользователям некоторый базовый набор сервисов, создавать систему нужно в рамках технологий гибкой разработки. На первом этапе необходи-

мо создать и запустить с минимальными временными затратами MVP, далее полученный продукт итеративно приближать к искомому идеалу, с одной стороны, опираясь на имеющиеся методические, технологические и аппаратные возможности, с другой — стремясь к эффективному удовлетворению потребностей пользователей.

5. Инфраструктурная и системная основа минимального жизнеспособного продукта

Вопросы развития вычислительной инфраструктуры для выполнения исследований и разработки научных программных продуктов не являются для ИВТ новыми [13, 14]. В конце 2000-х годов на повестке уже стоял вопрос о возможностях телекоммуникационной инфраструктуры и потенциале ее использования для построения распределенной вычислительной инфраструктуры на основе грид-технологий [15], тогда же начало оформляться понимание того, что такое ресурсный центр в области информационно-вычислительного обеспечения научных исследований [16]. Развитие вычислительной части инфраструктуры суперкомпьютерных и высокопроизводительных вычислений в СО РАН застыло в состоянии 2011 г. [17] на долгие шесть лет. Однако благодаря поддержке со стороны ФАНО России 2017 г. обогатил ее новыми вычислительными комплексами в Сибирском (`http://www.sscc.icmmg.nsc.ru/`) и Иркутском (`http://hpc.icc.ru/`) суперкомпьютерных центрах. Развитие телекоммуникационной среды и ресурсов для хранения данных продолжалось при этом своим чередом.

ИВТ СО РАН более 20 лет координирует деятельность по развитию корпоративной академической компьютерной сети, объединяющей учреждения, ранее входившие в Сибирское отделение РАН. Это позволило создать мощную разветвленную телекоммуникационную инфраструктуру более чем в десяти городах Сибири, включая крупнейшие научные центры: Новосибирский, Иркутский, Красноярский, Томский. Развитием и поддержкой внутригородских сетей традиционно занимались организации-координаторы в регионе, а общие коммуникационные каналы и самый крупный новосибирский кластер сети — зона ответственности ИВТ. В конце 2000-х годов был заложен высокоскоростной десятигигабитный сегмент корпоративной сети, изначально предназначенный для интенсивного обмена данными между крупными вычислительными ресурсами. Созданная академическая сеть и сервисы на ее основе стали мощным инструментом поддержки междисциплинарных научных исследований, способствуя научному и экономическому развитию региона, за что коллективу авторов во главе с научным руководителем ИВТ академиком Ю.И. Шокиным присвоена премия Правительства РФ в области науки и техники за 2012 г.

В последние годы при активной поддержке со стороны ФАНО России обеспечиваются функционирование созданной сети и развитие специализированного высокоскоростного сегмента. В 2017 г. десятигигабитные каналы связи объединили крупные научные учреждения в Новосибирске, Иркутске, Красноярске, Томске, Кемерове, на скорости 1 Гбит в сеть включен центр обработки данных дистанционного зондирования Земли в г. Барнауле. Летом 2017 г. запущено тестирование десятигигабитного выделенного канала между центральным узлом связи в ИВТ СО РАН (г. Новосибирск) и Межведомственным суперкомпьютерным центром (г. Москва) через узел обмена трафиком MSK-IX, что позволило объединить вычислительные ресурсы Межведомственного суперкомпьютерного центра с ресурсами Сибирского суперкомпьютерного центра (г. Новосибирск).

В ИВТ поступательно развивается инфраструктура для хранения и обработки данных. К 2016 г. совокупная сырая емкость установленных в ИВТ систем хранения данных (СХД) составила около 1 ПБ. В 2017 г. в опытную эксплуатацию запущена первая очередь новой СХД (рис. 5), которая строится на основе платформы с открытым исходным кодом Ceph (`http://ceph.com/`). Система предназначена для размещения, обмена и долговременного хранения научных данных. Первая очередь системы состоит из 12 узлов хранения и трех управляющих узлов, объединенных интерконнектом на базе специализированного десятигигабитного оборудования для центров обработки данных. Каждый узел дает системе по 96 ТБ сырого дискового пространства, он подключен к сети двумя независимыми интерфейсами по 10 Гбит. В итоге объем сырого дискового пространства составляет 1.15 Пб, что с учетом резервирования дает пользователям до 0.57 ПБ для размещения научных данных. Тестирование показало высокую производительность построенной системы, скорость обмена данными с которой в настоящее время ограничивается скорее скоростью подключения пользователя. Система защищена от существенных сбоев и потери данных за счет различных уровней резервирования как самих данных, так и аппаратных, в том числе сетевых ресурсов.

Создание новой системы хранения научных данных стало возможным благодаря поддержке предложенного ИВТ интеграционного проекта в рамках программы реструктуризации сети научных учреждений, подведомственных ФАНО России, и реализации программы развития ИВТ. В соответствии с этой программой в ИВТ образован и развивается центр научных ИТ-сервисов, закупается дорогостоящее оборудование для оснащения центра обработки данных. В 2017 г. запланирована реализация второго этапа работ по созданию и развитию центра научных ИТ-сервисов, который включает расширение новой СХД в 2–2.5 раза (до 2.5–3 ПБ сырого дискового пространства), что позволит размещать в ней до 1.5 ПБ научных данных.

На базе новой СХД формируется иерархия ИТ-сервисов хранения, обмена и совместной работы с научными данными и документами, которые можно рассматривать как системно-аппаратную основу MVP создаваемой системы.

Базовым сервисом комплекса является выделение дискового пространства на отказоустойчивой СХД (BSS — Basic Storage Service). Выделенное дисковое простран-

Рис. 5. Дисковый массив новой системы хранения данных

ство может быть подключено к любой компьютерной системе в поддерживаемом ИВТ СО РАН сибирском сегменте академической корпоративной компьютерной сети организаций ФАНО России с использованием специальных блочных и файловых протоколов.

Сервисы второго уровня связаны с запуском виртуальных машин в отказоустойчивом распределенном кластере ИВТ (BVS — Basic Virtualization Service). Они могут применяться, в частности, для организации доступа и использования выделяемых на СХД дисковых пространств или для обработки научных данных.

Сервисами третьего уровня являются платформы для хранения, обмена и совместной работы с научными данными и документами. Первый из них — сервис автоматизации совместной деятельности рабочих групп (SGCS — Scientific Groups Collaboration Service). Он строится на базе платформы с открытым исходным кодом Zimbra (`https://www.zimbra.com/`). Сервис позволяет обмениваться электронными сообщениями (e-mail), управлять списками контактов, вести ежедневник (календарь), управлять задачами как для отдельных пользователей, так и для групп пользователей с возможностью открытия (sharing) доступа другим пользователям системы к документам и папкам, в том числе почтовым, событиям календаря и задачам. Сервис имеет современный веб-интерфейс, у него есть возможность работы с большинством почтовых клиентов стационарных и мобильных платформ через стандартные протоколы. Доступны два варианта использования сервиса: для организации-пользователя может быть выделена отдельная виртуальная машина с развернутой платформой либо предоставлена возможность регистрации пользователей в общей системе. В первом случае администрирование платформы осуществляют специалисты организации-пользователя, во втором — регистрация и управление пользователями выполняются службой ИВТ.

Другой сервис третьего уровня построен на платформе с открытым исходным кодом NextCloud (`https://nextcloud.com/`), он предназначен для совместной работы с файлами и документами, а также среднесрочного и долговременного хранения общих данных (CDSS — Cloud Data Store&share Service). Сервис является расширенным аналогом Dropbox, он позволяет загружать и хранить файлы и папки, предоставлять к ним доступ другим пользователям системы, совместно редактировать документы (с помощью интегрированной подсистемы на основе LibreOffice, `http://libreoffice.org/`) с поддержкой версионности, публиковать файлы, папки и документы, подключать хранилище к компьютерным системам в виде внешних дисков, использовать его для резервного копирования и автоматической синхронизации данных с помощью клиентских приложений для стационарных компьютеров и мобильных устройств. Через интерфейсы платформы возможны подключение и использование различных внешних хранилищ данных. Как и для SGCS, использование сервиса реализуется в двух формах: развернутой на отдельной виртуальной машине и управляемой специалистами организации-пользователя платформы либо на общей центральной платформе, управляемой службой ИВТ.

И SGCS, и CDSS позволяют настраивать идентификацию и авторизацию пользователей на основе различных служб каталогов: Active Directory, LDAP и др. Сервис CDSS поддерживает создание федераций, позволяя организовать обмен данными и совместную работу пользователей других инсталляций платформы NextCloud.

ИВТ СО РАН продолжает работы по расширению списка ИТ-сервисов поддержки научных исследований для организаций, подведомственных ФАНО России. Миссия ИВТ — дать ученым и исследователям удобные инструменты для работы с их цифровыми данными, организовать среду для совместной работы с такими данными, предо-

ставить возможности для их публикации в рамках концепции OpenScience. Ключевая задача, которую запланировано решить в рамках этой миссии, — создать и открыть доступ к информационно-аналитической системе поддержки научных исследований, основанных на интенсивном использовании цифровых данных, позволяющей реализовать долговременное хранение, обработку и анализ, обмен и публикацию различных научных данных.

Заключение

Формирование комплекса инструментов для работы с научными данными является одной из наиболее актуальных проблем развития инфраструктуры научных исследований. Проблема создания такого комплекса лежит далеко за рамками компетенций традиционных поставщиков "облачных" услуг хранения данных и вычислительных ресурсов. Интеграция различных инструментов для анализа данных, построение онтологий научных данных с помощью системы взаимосвязей между всеми объектами системы, реализация идеологии открытых данных и открытой науки могут существенно расширить возможности исследователей, дать им новые средства для выявления и описания закономерностей в данных и, соответственно, в объектах и процессах для дальнейшего обобщения результатов наблюдений в виде новых законов природы и развития социума.

Сформулированные в работе требования к информационной системе, интегрирующей такой комплекс инструментов, и изложенные ключевые установки ее создания позволяют приступить к созданию модели информационно-аналитической системы поддержки научных исследований, основанных на интенсивном использовании цифровых данных. Построенная, поддерживаемая и развиваемая в ИВТ СО РАН аппаратно-системная инфраструктура и формируемый на ее основе комплекс научных ИТ-сервисов могут стать хорошей базой для реализации такой системы, тем более что набор базовых сервисов для хранения, обмена и совместной работы с научными данными и документами уже доступен исследователям в сибирском регионе.

Благодарности. Работа выполнена при финансовой поддержке Президентской программы "Ведущие научные школы РФ" (грант № НШ-7214.2016.9).

Список литературы / References

[1] **Nathan, L.Yo., Stephen, F.S., Pardis, C.S.** Data sharing: Make outbreak research open access // Nature. 2015. Vol. 518, iss. 7540. P. 477–479.

[2] **Cutcher-Gershenfeld, J., Baker, K.S., Berente, N., Flint, C. et al.** Five ways consortia can catalyse open science // Nature. 2017. Vol. 543, iss. 7647. P 615–617.

[3] **Litton, J.-E.** We must urgently clarify data-sharing rules // Nature. 2017. Vol. 541, iss. 7638. P 437.

[4] **Gibney, E.** European labs set sights on continent-wide computing cloud // Nature. 2017. Vol. 523, iss. 7559. P. 136–137.

[5] Nature editorial. Empty rhetoric over data sharing slows science // Nature. 2017. Vol. 546, iss. 7658. P. 327.

[6] Nature editorial. Don't let Europe's open-science dream drift // Nature. 2017. Vol. 546, iss. 7659. P. 451.

[7] **Жижимов О.Л., Федотов А.М., Шокин Ю.И.** Основные принципы, архитектура и реализация информационных систем ИВТ СО РАН // Изв. Кыргызского гос. техн. ун-та им. И. Раззакова. 2016. № 3(39). Ч. 1. С. 348–352.
Zhizhimov, O.L., Fedotov, A.M., Shokin, Yu.I. Basic principles, architecture and realization of information systems ICT SB RAS // Izv. Kyrgyzskogo Gos. Tekhn. Un-ta im. I. Razzakova. 2016. No. 3(39). Pt 1. P 348–352. (In Russ.)

[8] **Шокин Ю.И., Федотов А.М., Жижимов О.Л.** Технологии создания распределенных информационных систем для поддержки научных исследований // Вычисл. технологии. 2015. Т. 20, № 5. С. 251–274.
Shokin, Yu.I., Fedotov, A.M., Zhizhimov, O.L. Technologies for designing of distributed information systems to support research // Comput. Technologies. 2015. Vol. 20, No. 5. P. 251–274. (In Russ.)

[9] **Жижимов О.Л., Федотов А.М., Шокин Ю.И.** Платформа ZooSPACE — организация доступа к разнородным распределенным ресурсам // Электронные библиотеки. 2014. Т. 17, № 2. `http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2014/part2/ZFS`
Zhizhimov, O.L., Fedotov, A.M., Shokin, Yu.I. Platform ZooSPACE — providing access to heterogeneous distributed resources // Russ. Digital Libr. J. 2014. Vol. 17, No. 2. `http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2014/part2/ZFS` (In Russ.)

[10] **Шокин Ю.И., Добрецов Н.Н., Мамаш Е.А., Кихтенко В.А., Воронина П.В., Смирнов В.В., Чубаров Д.Л.** Информационная система приема, обработки и доступа к спутниковым данным и ее применение для решения задач мониторинга окружающей среды // Вычисл. технологии. 2015. Т. 20, № 5. С. 157–174.
Shokin, Yu.I., Dobretsov, N.N., Mamash, E.A., Kikhtenko, V.A., Voronina, P.V., Smirnov, V.V., Chubarov, D.L. An information system for acquisition, processing and access to satellite data and its applications in environmental monitoring // Comput. Technologies. 2015. Vol. 20, No. 5. P. 157–174. (In Russ.)

[11] **Белов С.Д., Зайцев А.С., Каплин В.И., Король А.А., Сковпень К.Ю., Сухарев А.М., Адакин А.С., Никульцев В.С., Чубаров Д.Л., Кучин Н.В., Ломакин С.В., Калюжный В.А.** Использование виртуализованной суперкомпьютерной инфраструктуры Новосибирского научного центра для обработки данных экспериментов физики высоких энергий // Вычисл. технологии. 2012. Т. 17, № 6. С. 36–46.
Belov, S.D., Zaytsev, A.S., Kaplin, V.I., Korol, A.A., Skovpen, K.Yu., Sykharev, A.M., Adakin, A.S., Nikultsev, V.S., Chubarov, D.L., Kuchin, N.V., Lomakin, S.V., Kalyuzhny, V.A. Using the virtualized HPC infrastructure of Novosibirsk Scientific Center for production analysis of HEP experiments data // Comput. Technologies. 2012. Vol. 17, No. 6. P. 36–46. (In Russ.)

[12] Systems engineering fundamentals. Fort Belvoir, Virginia: Defense Acquisition Univ. Press, 2001. 222 p.

[13] **Шокин Ю.И., Федорук М.П., Чубаров Д.Л., Юрченко А.В.** Высокопроизводительные вычисления в ИВТ СО РАН // Вычисл. технологии. 2006. Т. 11. Спецвыпуск 6. С. 17–26.
Shokin, Yu.I., Fedoruk, M.P., Chubarov, D.L., Yurchenko, A.V. High performance computations in ICT SB RAS // Comput. Technologies. 2006. Vol. 11. Special issue 6. P. 17–26. (In Russ.)

[14] **Shokin, Yu.I., Fedoruk, M.P., Chubarov, D.L., Yurchenko, A.V.** Computing facility of the Institute of Computational Technologies SB RAS // Notes on Numer. Fluid Mech. and Multidisciplinary Design. 2008. Vol. 101. P. 1–7.

[15] **Шокин Ю.И., Федорук М.П., Чубаров Д.Л., Юрченко А.В.** О перспективах Grid в Сибирском регионе // 6-е Совещание Российско-казахстанской рабочей группы по вычисл. и информ. технологиям. Алматы, Казахстан, 16.03–18.03.2009: Тр. совещания. Алматы, 2009. С. 324–338.
Shokin, Yu.I., Fedoruk, M.P., Chubarov, D.L., Yurchenko, A.V. On the future of Grid in Siberian region // 6-th Russian-Kazakhstan Workshop on Comput. and Inform. Technologies. Almaty, Kazakhstan, 16–18 march 2009: Proc. of the Work. Almaty, 2009. P. 324–338. (In Russ.)

[16] **Shokin, Yu.I., Fedoruk, M.P., Chubarov, D.L., Yurchenko, A.V.** Building a resource center for the grid infrastructure // Intern. Conf. "Mathematical and Informational Technologies MIT-2009". Kopaonik, Serbia, Budva, Montenegro, 27.08–05.09.2009: Zbornik Radova. Urednik: Dragon Acimovic, 2010. P. 377–380. (In Russ.)

[17] **Шокин Ю.И., Федорук М.П., Чубаров Д.Л., Юрченко А.В.** О развитии инфраструктуры суперкомпьютерных и распределенных вычислений в СО РАН // Информ. технологии и вычисл. системы. 2011. № 3. С. 9–19.
Shokin, Yu.I., Fedoruk, M.P., Chubarov, D.L., Yurchenko, A.V. Development of the supercomputing and distributed computing infrastructure in the Siberian Branch of the Russian Academy of Sciences // Inform. Technologies and Comput. Sys. 2011. No. 3. P. 9–19. (In Russ.)

Поступила в редакцию 14 июля 2017 г.

On the concept of information-analytical system for supporting data intensive science

Yurchenko, Andrey V.

Institute of Computational Technologies SB RAS, Novosibirsk, 630090, Russia
Corresponding author: Yurchenko, Andrey V., e-mail: yurchenko@ict.nsc.ru

We consider the problem of developing the infrastructure and special tools for manipulating big volumes of scientific data. Actuality of the problem is increasing due to exponential growth of data volume and the emerging open science and open data trends.

The purpose of this work is to form and describe an ideology, some conceptual aspects and technological issues of developing of the information-analytical system for researchers who deal with scientific data.

The place of this system in the research process and it's basic functional requirements are specified. The key settings, including the concept of data "enlightenment" are described. The basic hardware and software blocks of the system are listed and the current state of the IT infrastructure at the Institute of Computational Technologies SB RAS regarding the data intensive science is reported.

Keywords: science infrastructure, data intensive science, information system, scientific data, storing and processing data.

Acknowledgements. This research was partly made within the Grant of the President of Russian Federation for supporting Leading Scientific Schools (NSh-7214.2016.9).

Received 14 July 2017