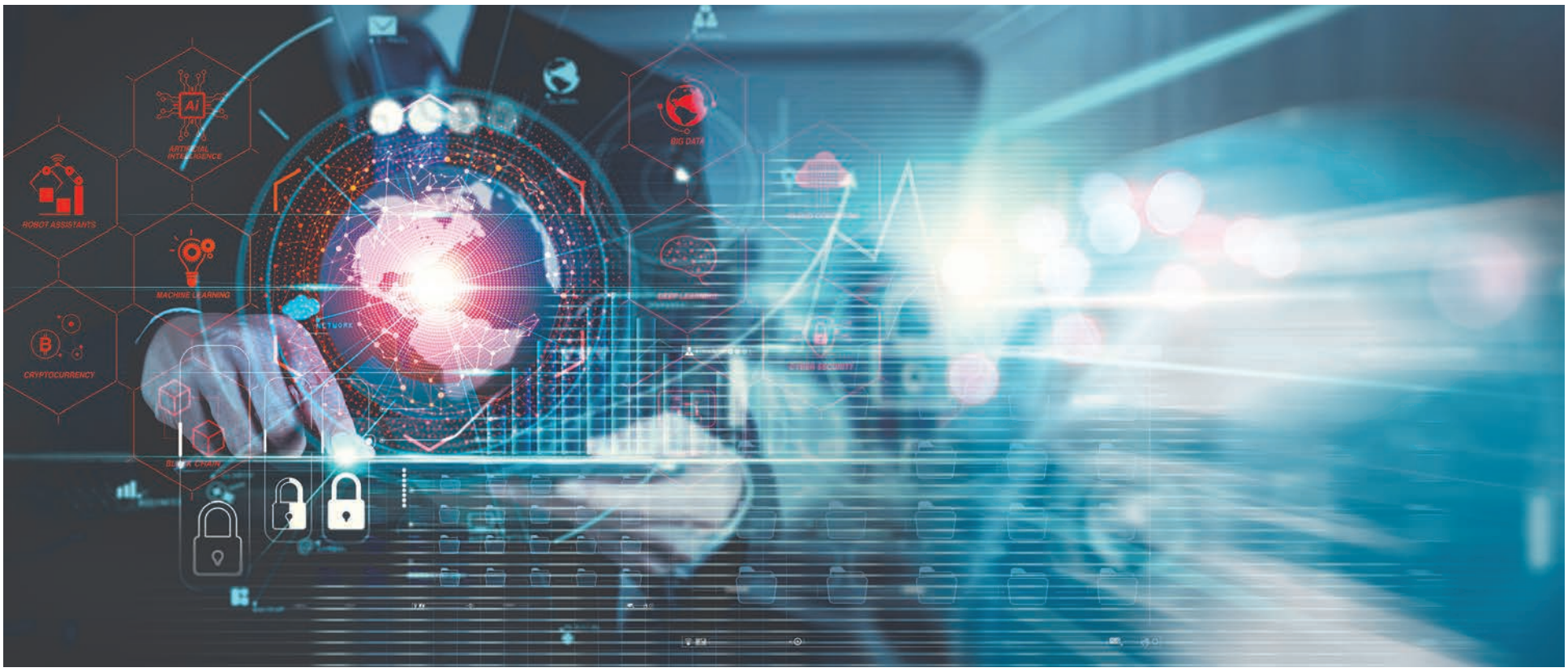

В Президиуме РАН

Время Больших

РАН созрела для создания Дата-центра

Андрей СУББОТИН

Про большие данные (Big Data) сегодня не говорит только ленивый, правда, не все до конца понимают, что же это такое. Под Big Data подразумеваются структурированные и неструктурированные данные огромных объемов и значительного многообразия, эффективно обрабатываемые горизонтально масштабируемыми программными инструментами. На прошедшем заседании Президиума РАН с докладом «Системный анализ больших данных для наук о Земле» выступил академик Алексей ГВИШИАНИ, который несколько прояснил ситуацию для тех, кто не в курсе.

Алексей Джерменович сразу отметил, что термин «большие данные» часто используется «волюнтаристски, интуитивно», поэтому, при написании доклада, который готовился вместе с академиком Владиславом Панченко, было решено подробнее остановиться на терминологии. Очевидно, потому, что иностранная лексика сегодня официально не в моде, А.Гвишиани нашел подходящий синоним Big Data - БоД.

По словам академика, БоД - это формализованная система понятий, имеющая очерченную область применения и снабженная для этого оригинальным программным обеспечением и лежащей в его основе математикой (частью системного анализа). Алексей Джерменович отметил, что термин «БоД» появился недавно - в статье 2008 года Клиффорда Линча (журналиста-ученого) в журнале Nature по аналоги с терминами «большая нефть», «большие деньги» и т. д.

Два года спустя БоД впервые обрел основные свои характеристики, свойства, известные три V: Volume («объем»), Velocity («скорость») и Variety («разнообразие»). 3V должны при этом иметь значения, существенно превышающие стандартные пороги выборок. Velocity понимается здесь как скорость сбора, передачи, архивирования, научного анализа и роста разнообразия информации. По словам ученого, последнее особенно важно, потому что, если имеется очень много данных, но они одинаковы, не разнообразны, это не есть БоД.

Стало ясно, что в случае с БоД меняется парадигма научных исследований. Изучение информации с целью ответить на вопрос «почему?» сменяется поиском ответа на вопросы «что именно?» и «как действовать?». Для этого во главу угла ставится распознавание корреляций и трендов в гигантских массивах БоД. Так появилось программное обеспечение обработки БоД, альтернативное Системе управления базами данных (СУБД): без требований, жесткой иерархии и однородности данных. Это ПО может обрабатывать как структурированные, так и неструктурированные данные, например, тексты. БоД стали эффективным инструментом внедрения научных результатов в реальный сектор экономики, сокращающим путь от анализа a posteriori к прогнозированию.

Преобразование количества накопленной информации в качество решений называют «феноменом БоД». Так, с появлением в 2010-х годах шахматных БоД, программы, построенные на принципах 1960-х годов (например, «Каисса» Михаила Ботвинника), сегодня играют на уровне гроссмейстеров (Deep Blue, AlphaZero). Аналогично феномен БоД ярко проявляется в прогрессе продуктов искусственного интеллекта, таких как Google Translate или Chat GPT, когда система искусственного интеллекта может продолжать писать текст самостоятельно.

Такой прогресс достигнут не от того, что «поумнели» алгоритмы, пояснил ученый. «В основном это происходит потому, что данные стали просто гигантскими по объему, разнообразными и они очень быстро поступают и обрабатываются», - отметил он.

- Важным базовым понятием является «масштабируемость», - отметил А.Гвишиани. - Информационная система (система данных и их обработки) называется масштабируемой, если она обладает возможностью наращивания вычислительных и системных ресурсов без структурных изменений системы и способна увеличить свою производительность пропорционально дополнительным ресурсам.

Масштабирование бывает горизонтальным и вертикальным. Вертикально масштабируемая система увеличивает свою производительность за счет усиления каждого компонента системы, без изменения программ (простыми словами, усиливается каждый компьютер системы). Горизонтально масштабируемая система делает это за счет добавления к системе новых компонентов. При этом может требоваться модификация программ для полноценного использования добавленных ресурсов (то есть система позволяет добавлять новые компьютеры, и это может происходить физически или виртуально).

Создание БоД - это процесс, разворачивающийся во времени. БоД формируются из первоначального ядра обычных, «маленьких», данных развертыванием во времени одновременно каждого из трех V. По словам А.Гвишиани, «процесс этот имеет начало, но не имеет конца». Классическим и очевидным примером БоД является сеть Интернет со всеми ее составляющими информационными потоками. Если, например, за отправную точку взять 2010 год, когда объем данных в мире составлял несколько зетабайт (1 зетабайт - это 10 в 21-й степени байт), то к 2025 году прогнозируется его рост свыше 160 зетабайт.

Алгоритмическая модель работы с БоД - международная платформа Map Reduce (Google, 2004), созданная совместными усилиями множества ученых из разных стран. Это модель параллельной обработки БоД на компьютерных кластерах путем разделения задачи на независимые части: в процессе вычислений Map входные пары ключ/значение преобразуются в выходные пары ключ\значение. Эта задача реализуется с помощью программы HADOOP для структурированных и неструктурированных БоД. Она альтернативна СУБД (Oracle, SQL и проч.). Такой подход практически гарантирует неуничтожаемость данных и отказоустойчивость системы.

> **БоД стали эффективным инструментом внедрения научных результатов в реальный сектор экономики, сокращающим путь от анализа a posteriori к прогнозированию.**

- Map Reduce - открытая система, ее не затронули санкции, - отметил А.Гвишиани. - Так что эти идеи можно использовать. HADOOP - тоже открытая система, но она находится под санкциями, и это - вызов отечественным ученым, инженерам и коммерческим компаниям. Работы в этой области ведутся.

Докладчик «помечтал» о том, как можно создать БоД Российской академии наук.

- Для этого нужны три вещи, - сказал он. - Первая - источники данных. Замечу, что такие источники у нас есть: кроме огромного запаса данных, которые производятся институтами РАН и Минобрнауки, в самой РАН есть колоссальное количество БоД. Это тексты документов, которые экспертирует Академия наук, - сказал Алексей Джерменович. - Я бы предложил оставить эти тексты в РАН (пока они уничтожаются спустя несколько лет), а из них делать БоД.

По словам ученого, при наличии БоД нужно также создать саму оболочку дата-центра. То есть необходимо помещение с большими электрическими ресурсами, мощными резервными энергогенераторами, непрерывным, неуничтожаемым кондиционированием, системой гарантированного пожаротушения и круглосуточным сервисом дежурных инженеров.

И, конечно, необходима команда специалистов.

Также академик рассказал об использовании больших данных в науках о Земле. Сегодня к ним можно отнести метеорологические данные, которые составляют петабайты (10^{15} Б) информации, данные дистанционного зондирования Земли, которые к 2025 году должны достичь объема 300 экзабайт (10^{18} Б). Сотни терабайт (10^{12} Б) информации накоплены в международной сети экологических наблюдений станций SMEAR, данные глобального сейсмического мониторинга, включая временные ряды сейсморегистрации, измеряются в петабайтах (10^{15} Б), информация геофизической разведки и поиска полезных ископаемых копится со скоростью 100 петабайт в сутки, данные горнодобывающих и перерабатывающих комплексов - 130 терабайт в год. Региональные мультидисциплинарные данные по Арктической зоне РФ составляют экзабайты. Многие другие данные также могут стать большими, подчеркнул ученый.

А.Гвишиани напомнил, что РАН является членом Международного института прикладного системного анализа (IIASA). Он был создан в 1972 году в Австрии и расположен на правах аренды за 1 евро в год в здании Лаксенбургского дворца (пригород Вены). Учредителями выступили СССР, США, Австрия. Сегодня членами IIASA являются 22 страны. В институте постоянно работают 300 ученых из разных стран, включая порядка трех десятков специалистов из России. IIASA вносит весомый вклад в решение глобальных проблем, выполняя междисциплинарный системный анализ информации по экологическим, социальным, технологическим, экономическим и сельскохозяйственным тематикам. Сотрудничество с IIASA координирует Комитет по системному анализу РАН (КСА), председателем которого является вице-президент РАН В.Панченко.

В заключение академик рассказал о БоД в реальном секторе экономики, отметив, в частности, что более половины крупных российских компаний различной направленности (маркетинг, торговля, банковская сфера и страхование, машиностроение и проч.) уже более трех лет используют БоД в своей деятельности. Заинтересованность в работе с БоД при взаимодействии с РАН и ее институтами обозначили ПАО «Уралкалий», АО «НИИАС» (РЖД), компания «Металлинвест», АК «АЛРОСА (ПАО), ПАО «Газпром», Национальная компьютерная корпорация и др.

Трудности при работе с БоД-проектами тоже есть. Самые значительные из них - нехватка кадров для ведения проектов, сложность выбора подходящей архитектуры/инструментов и необходимость капитальных вложений.

В.Панченко поблагодарил своего содокладчика «за академическую подачу проблемы».

- Мы в Академии наук переходим на сложную систему экспертных оценок практически всех российских научных проектов, в том числе крупных, - сказал Владислав Яковлевич. - Поэтому мне представляется крайне важным развитие подхода для создания дата-центра РАН, в котором смогут «впитать» всю ту информацию, которой мы располагаем, для проведения корректной и быстрой экспертизы. IIASA доступно очень широкое информационное поле, и организация не исключала Россию из своих рядов. Это открытый канал в Европу и весь мир, позволяющий российским ученым до сих пор получать уникальную информацию о глобальных проектах. ■