Проблемы современной информатизации: Большие Данные и суперкомпьютинг

Contemporary Informatization Problems: Big Data and Supercomputing

Кияев В.И., СПбГЭУ, профессор, kiyaev@mail.ru
Газуль С.М., СПбГЭУ, заведующий сектором информационного обеспечения приёма студентов УОПСПР СПбГЭУ, sgazul@gmail.com

Ключевые слова: *большие данные, вычислительные проблемы, облачные технологии, гибридные вычислительные системы*

Большие Данные как явление появились в 70-е годы XX века, но работали с ними, как правило, в узкоспециальных научных областях при обработке структурированных данных на мейнфреймах – в основном, в сфере прогнозного моделирования. Проблемой Большие Данные стали в начале XXI века – появление, развитие и использование Интернет-пространства в экономике, бизнесе и социальной жизни общества (феномен социальных сетей) привело к экспоненциальному росту объема производимых данных и появлению нового и актуального направления в информационных технологиях – Big Data. Отметим, что этот термин до сих пор не имеет однозначного толкования, так как в настоящее время он не означает лишь объем накопленной информации, но включает в себя технологии сбора, хранения, обработки, а также инфраструктурные и сервисные услуги. В представленной ниже таблице показаны основные различия традиционной базы данных и базы Больших Данных.

Характеристика данных	*Традиционные базы данных*	*База Больших Данных*
Объем данных	От гигабайт (10^9) до терабайт (10^{12})	От петабайт (10^{15}) до эксабайт (10^{18})
Способ хранения	Централизованный	Децентрализованный
Степень структурированности	Высокая	Средняя и низкая
Модель хранения и обработки	Вертикальная – серверы с вертикальным масштабированием (большие SMP-системы)	Горизонтальная – кластерные и многоядерные суперкомпьютерные вычислительные системы с массовым параллелизмом (MPP)
Взаимосвязь данных	Сильная	Слабая
Вычислительная инфраструктура и технологии	Специализированные серверы и Центры обработки данных (ЦОД)	Суперкомпьютерные комплексы, параллельные вычисления, фабрики данных, облачные технологии

В настоящее время Большие Данные как сущность характеризуются пятью V:

- *Volume* – объемы данных, которые трудоемко и затратно собирать, обрабатывать и хранить традиционными способами и для которых требуются новый подход и усовершенствованные инструменты.
- *Variety* – многообразие, т.е. возможность одновременной обработки, структурированной и неструктурированной разноформатной информации. Неструктурированная информация включает в себя видео, аудио файлы, свободный текст, информацию, поступающую из профессиональных и социальных сетей. На сегодняшний день 80-85% информации входит в группу неструктурированной. Такая

информация нуждается в комплексном анализе и структуризации, чтобы сделать ее полезной для дальнейшей обработки.

- *Velocity* – скорость, которая требуется для обработки данных в реальном времени и доставки их пользователю.
- *Veracity* – достоверность данных. Эта характеристика стала одной из важнейших в проблеме обеспечения безопасной работы с данными.
- *Value* – ценность накопленной информации. Большие данные должны постоянно работать и быть максимально полезными, чтобы оправдать затраты, направленные на их сбор, обработку и анализ. Конечный результат использования больших данных – это всегда реальная ценность, направленная на решение государственных, экономических и социальных проблем.

Большие Данные широко используются во многих отраслях науки, экономики и бизнеса. Это – био и нанотехнологии, геофизика и геодинамика, горнодобывающая и нефтяная промышленность, метеорология, медицина, телекоммуникационные структуры, управление логистикой, финансовые услуги, государственное управление и многое другое. В связи с этим появляются и быстро развиваются новые технологии работы с Большими Данными. Если до середины 10-х годов XXI века при работе с данными основными были SQL-технологии, то в настоящее время получили развитие комплексные системы обработки Больших Данных, включающие в себя:

NoSQL (Not Only SQL – ряд подходов, направленных на реализацию базы данных, имеющих отличия от моделей, используемых в традиционных, реляционных СУБД. Их удобно использовать при постоянно меняющейся структуре данных. Например, для сбора, хранения и обработки неструктурированной информации, полученной из социальных сетей.

MapReduce – модель распределения вычислений при обработке массивов Больших Данных. Используется для параллельных вычислений с очень большими объемами данных (петабайты и эксабайты). Особенностью такой модели является то, что в программном интерфейсе не данные передаются на обработку программе, а наборы данных «отыскивают» необходимые совокупности программ, необходимых для их обработки. Принцип работы заключается в последовательной обработке данных двумя методами Map и Reduce. Map выбирает предварительные данные, Reduce агрегирует и и осуществляет поиск программных инструментов.

Hadoop – свободно распространяемый набор утилит, библиотек и фреймворков для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Объединенная кластерная система защищена от выхода из строя любого из узлов кластера, так как каждый блок имеет, как минимум, одну копию данных на другом узле.

SAP HANA – высокопроизводительная NewSQL платформа для хранения и обработки данных. Обеспечивает высокую скорость обработки запросов. Еще одним отличительным признаком является то, что SAP HANA упрощает системный ландшафт, уменьшая затраты на поддержку аналитических и прогнозных систем.

Web Crowler – сервис, позволяющий извлекать и анализировать неструктурированную информацию из сетей, извлекать и формировать новые знания из web-ресурсов, а также использовать их для обогащения своих данных и построения новых бизнес-моделей. Инструмент Web Crawler As A Service включает в себя NoSQL, Hadoop и специализированный язык программирования R для статистической обработки данных и работы с графикой, а также свободную программную среда вычислений с открытым исходным кодом в рамках проекта GNU.

Появление такого инструментария работы с Большими Данными и требования рынка систем обработки Больших Данных в реальном времени привело формированию новой платформы Big Data-As-A-Service, которая предоставляет возможности для работы с Большими Данными в облачных структурах.

Экспоненциальное возрастание объема данных достаточно быстро привело к вычислительным проблемам при их обработке. В начале XXI века ставку делали на увеличение вычислительной мощности, на кластерные и суперкомпьютерные многоядерные системы, распараллеленные методы вычислений (MPP и OpenMP), которые предназначены для программирования многопоточных приложений на многопроцессорных системах с общей памятью. Однако лавинообразное нарастание объемов данных и требование к их обработке в режиме реального времени привело к реальным вычислительным противоречиям – многоядерный суперкомпьютер, используемый для обработки Больших Данных, во многих случаях становился неэффективным, так как критичным становилось время сбора и передачи данных в вычислительную систему и доставки результатов пользователю. Кроме того, технологии распараллеливания, основанные на классической схеме вычислений, с некоторых пор сами стали становиться тормозом для высокопроизводительных вычислений (High Performing Computing – HPC). В связи с этим появилась насущная необходимость говорить о таких новых важнейших аспектах вычислительных процессов, как: модели программирования, степень и уровни параллельности, неоднородность программных и аппаратных систем, сложность иерархии памяти и трудности одновременного доступа к ней в распределенных вычислениях, стек системного и прикладного ПО, надежность, энергопотребление, сверхпараллельный ввод/вывод и т. д. [1]. На наш взгляд это неизбежно приведёт к смене парадигмы высокопроизводительных вычислений. Среди новых характерных черт будущей вычислительной парадигмы все более отчетливо проступают следующие: стохастичность, гибридность, асинхронность, кластерность (отсутствие жесткой централизации и динамическая кластеризация на классы связанных моделей). Ниже дано наше видение составляющих этой парадигмы.

Стохастичность. С одной стороны, хорошо известно, что компьютеры становятся все миниатюрнее и миниатюрнее, размер элементарного вычислительного элемента (вентиля) приближается к размеру молекулы или даже атома. На таком уровне законы классической физики перестают работать и начинают действовать квантовые законы, которые в силу принципа неопределенности Гейзенберга принципиально не дают точных ответов о состоянии. С другой стороны, стохастичность – это известное свойство сложных динамических систем, состоящих из огромного числа компонент. Вычислительная система, использующая этот принцип, превращается в динамическую систему, которая способна сама настаиваться на задачу и подключать необходимые вычислительные мощности и инструменты.

Гибридность будущих процессов вычислений можно понимать необходимость рассмотрения комбинации непрерывных и дискретных процессов, т. е. учет непрерывной эволюции протекания физических процессов при работе той или иной модели и скачкообразное переключение с одной модели на другую.

Увеличение быстродействия вычислительных устройств и уменьшение их размеров с неизбежностью приводит к необходимости операций с «переходными» процессами, серьёзным ограничением классической модели вычислений является разбиение памяти на изолированные биты, потому как, во-первых, сокращение длины такта и расстояний между битами с определенного уровня делает невозможным рассматривать их изолированно в силу законов квантовой механики. Вместо примитивных операций с классическими битами в будущем было бы естественно перейти операциям, задаваемыми теми или иными динамическими моделями

микромира, оперирующими с наборами взаимосвязанных «битов» [3]. При этом простейшими «моделями» могут остаться классические операции с битами.

Кроме этого гибридность может означать сочетание в единой системе суперкомпьютер с облачными структурами, очень хорошо приспособленными для операций с разнообразными низко структурированными наборами Больших Данных.

В наших исследованиях мы синтезировали определение гибридной информационной и вычислительной системы, а также предложили методику перехода к таким системам. Гибридная система – это система, которая в своей структуре имеет ресурсы, расположенные в вычислительных облаках различного типа (публичные, частные), а также в локальной сети предприятия или в кластерной сети, рис 1. Для такой системы необходим механизм, обеспечивающий возможность переноса информационных ресурсов в вычислительные облака и локальную инфраструктуру [2].

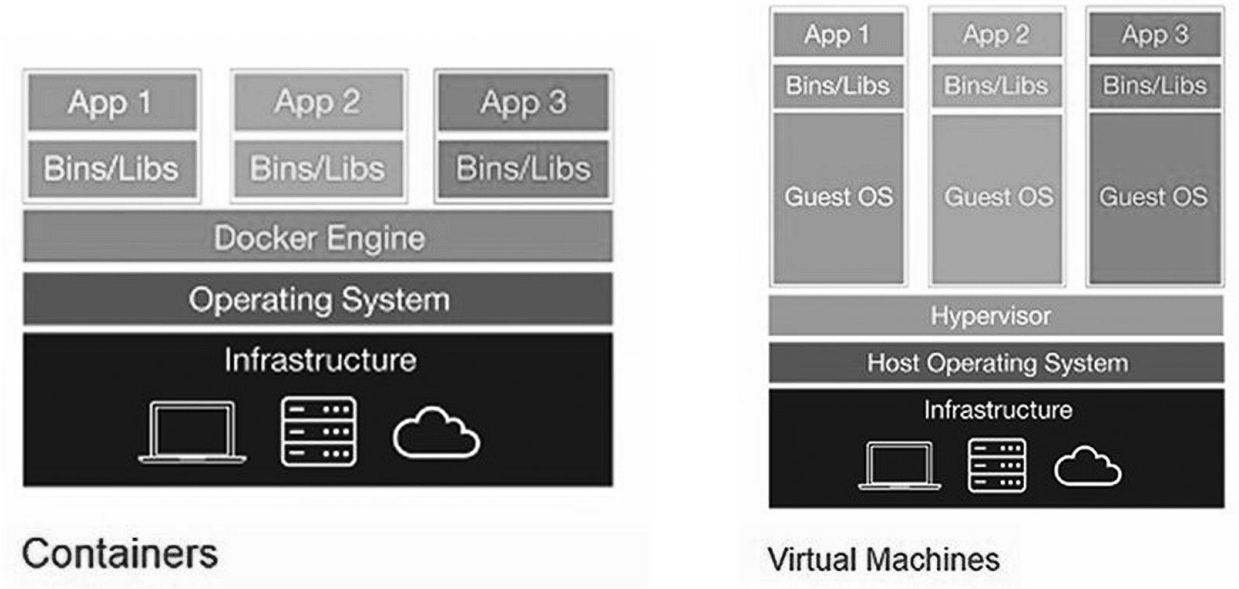

Рис. 1. Реализация принципа виртуализации при построении гибридных систем (с применением контейнеров Docker или виртуальных машин)

Асинхронность. Отказ от унифицированных простых вычислительных элементов неизбежно приводит к отказу от синхронизации работы различных компонент, имеющих существенно отличающиеся физические характеристики и свои длительности «тактов». В рамках классической теории множеств противоречивый смысл понятия единого «такта» выражается в рамках неразрешимости проблемы континуума в рамках аксиоматики Френкеля-Цермело.

Кластерность. Одним из неожиданных результатов многочисленных попыток в разработках (создании, адекватном описании поведения и управлении) сложных стохастических систем оказалась перспективность модели мультиагентных систем, в которой топология связей агентов между собой меняется со временем. При этом понятию агента может соответствовать как некоторая динамическая модель (компонент системы), так и определенный набор моделей. При отсутствии жесткой централизации такие системы способны эффективно решать достаточно сложные задачи, разбивая их на части и автономно перераспределяя ресурсы на “нижнем” уровне, эффективность часто повышается за счет самоорганизации агентов и динамической кластеризация на классы связанных моделей.

Отходу от классических информационных и вычислительной моделей во многом способствуют широкое применение мультиагентных технологий и методов рандомизации [4,6].

Мультиагентные технологии решения задач возникли как ответ на потребности решения сложных задач в условиях, максимально приближенных к тем, в которых функционируют реальные системы. В этом случае на первый план выступает способность быстрого отклика на непредсказуемые изменения, то есть адаптивность метода решения задачи или системы, реализующей этот метод.

На практике очень часто оказывается, что классические методы решения задач либо неприменимы к реальной жизни, либо они требуют огромных объемов расчетов (для которых не хватит мощности современных суперкомпьютеров), либо они вовсе отсутствуют. Во многих таких случаях альтернативой оказываются мультиагентные технологии, суть которых заключается в принципиально новом методе решения задач. В отличие от классического способа, при котором проводится поиск некоторого четко определенного (детерминированного) алгоритма, позволяющего найти наилучшее решение проблемы, в мультиагентных технологиях решение получается автоматически в результате взаимодействия множества самостоятельных целенаправленных программных модулей – агентов.

Одной из важнейших характеристик мультиагентных технологий является отказ от традиционной для информационных технологий парадигмы разделения процессов получения информации и принятия управленческих решений. В случае сложных систем, состоящих из огромного числа взаимодействующих динамических объектов, возможность получения реальной «мгновенной картины мира» можно вообразить себе только теоретически, на практике во время сбора всей необходимой информации «картина мира» может существенно измениться, то есть время сбора информации становится критичным фактором. При использовании мультиагентных технологиий компоненты системы начинают взаимодействовать и реализовывать те или иные управляющие воздействия самостоятельно, не дожидаясь «команды из центра». В большинстве случаев такого центра может просто не быть, так как динамический ансамбль агентов на основе консенсуса «выбирает» агента, который исполняет роль базового управляющего модуля. По мере выполнения общей задачи роль «управленца» также на основе поиска консенсуса может передаваться другому агенту и т.д. [6].

Схожей концепцией в принятии решений «на лету» является концепция рандомизации, в которой также имеется механизм «учета» в процессе решения непредвиденных событий, появляющихся в работе систем. В этом случае можно говорить о том, что неопределенности не только не мешают, но и дают дополнительные возможности в решении проблем.

Примером повышения эффективности процессов обработки данных и управления при изменении парадигмы является рандомизация управляющих воздействий при решении задач оценивания неизвестных параметров системы при наблюдениях с произвольными внешними помехами и использование замкнутых стратегий управления в условиях неопределенностей. Такие стратегии оказываются практически незаменимыми в случае, если помехи не являются случайными (статистическими), например, это значения некоторой неизвестной функции [5].

Таким образом, использование гибридных методов вычислений и гибридных вычислительных систем позволяет решать ранее недоступные для решения задачи для эффективного использования Больших Данных.

Литература

1. Амелин К.С., Граничин О.Н., Кияев В.И., Корявко А.В. Взгляд на перспективы развития принципиально новых компьютерных устройств и систем // В сб. тр. Всерос. научн. конф. «Научный сервис в сети Интернет: эксафлопсное будущее» — г. Москва-Новороссийск, — 2011. — С. 28-31.

2. Газуль С.М., Ананченко И.В., Кияев В.И. Совершенствование образовательного процесса в ВУЗе: активные методы обучения и гибридные информационные системы на основе виртуализации // Современные проблемы науки и образования. – 2015. – № 2; URL: www.science-education.ru/122-20856 (дата обращения: 12.09.2016).
3. Граничин О.Н., Молодцов С.Л. Создание гибридных сверхбыстрых компьютеров и системное программирование. СПб, 2006, 108с.
4. Граничин О.Н. Характеристики перспективных принципиально новых компьютерных устройств и систем // Механика, управление и информатика. 2011. № 5, с. 147-161.
5. Граничин О.Н. Обратные связи, усреднение и рандомизация в управлении и извлечении знаний // Стохастическая оптимизация в информатике. 2012. Том 8. Вып. 2, с. 3-48.
6. Ерофеева В.А., Иванский Ю.В., Кияев В.И. Управление роем динамических объектов на базе мультиагентного подхода. // Компьютерные инструменты в образовании – 2015, №6 – с. 36-44.