РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
ФИЗИКО-ТЕХНИЧЕСКИЕ ПРОБЛЕМЫ РАЗРАБОТКИ ПОЛЕЗНЫХ ИСКОПАЕМЫХ

2016 № 6

ГЕОИНФОРМАТИКА

УДК 004.21, 004.9+551+622

ГОРНАЯ ИНФОРМАТИКА И ПРОБЛЕМА "БОЛЬШИХ ДАННЫХ" В ПОСТРОЕНИИ КОМПЛЕКСНЫХ МОНИТОРИНГОВЫХ СИСТЕМ БЕЗОПАСНОСТИ НЕДРОПОЛЬЗОВАНИЯ

И. В. Бычков[1], Д. Я. Владимиров[2], В. Н. Опарин[3], В. П. Потапов[4], Ю. И. Шокин[5]

[1]*Институт динамики систем и теории управления СО РАН,* E-mail: idstu@icc.ru,
ул. Лермонтова, 134, 664033, г. Иркутск, Россия
[2]*ОАО "ВИСТ групп",* E-mail: vladimirov@vistgroup.ru,
Докучаев пер., д. 3, стр. 1, 107078, г. Москва, Россия
[3]*Институт горного дела им. Н. А. Чинакала СО РАН,* E-mail: oparin@misd.ru,
Красный проспект, 54, 630091, г. Новосибирск, Россия
[4]*Кемеровский филиал Института вычислительных технологий СО РАН,* E-mail: ict@ict.nsc.ru,
ул. Рукавишникова, 21, 650025, г. Кемерово, Россия
[5]*Институт вычислительных технологий СО РАН,* E-mail: ict@ict.nsc.ru,
просп. Академика Лаврентьева 6, 630090, г. Новосибирск, Россия

Обсуждается актуальная проблема и перспективные пути ее решения в горной информатике, связанные с "большими данными" — потоками разнородной информации, сопровождающей процесс горного производства. Описывается технология BIG DATA и общая схема ее реализации на мини-кластерах с использованием программных средств Hadoop и MapReduce, иллюстрируемая конкретными примерами.

"*Большие данные*", *интеллектуальный анализ, вычислительные и мини-кластеры, неструктурированные массивы информации, потоковая обработка геомеханических и геодинамических данных, облачные технологии, распределенные вычисления, безопасное недропользование*

Горное дело и обеспечивающие стратегию его развития горные науки являются одним из двигателей научно-технического прогресса стран-лидеров мировой экономики, аккумулируя в себе передовые достижения науки и техники по широчайшему спектру (физика, химия, математика, информатика, машино- и приборостроение, экономика, геология, геомеханика, экология и др.) [1, 2].

Последние десятилетия в мировой горной науке и горном деле представляют период повышенного интереса к проблемам разработки месторождений полезных ископаемых в сложных горно-геологических, геомеханико-геодинамических и природно-климатических условиях. Развитие горных работ по освоению новых по глубине горизонтов залегания полезных ископаемых нередко сопровождается ростом уровня температур и горного давления, сопоставимого с пределом прочности пород на одноосное сжатие.

В таких условиях обеспечение безопасности ведения горных работ и автоматизация процессов горного производства, в том числе освоение гибких роботизированных систем добычи руды и угля при буровзрывных работах, погрузо-разгрузочных операциях, транспортировке горной массы, обогащении полезных ископаемых, а также сложных экономических расчетов и экологических последствий становятся стратегически важными направлениями исследований и инновационных разработок, где роль информационных технологий трудно переоценить.

В работах [1, 2] приведен анализ и дано обобщение достижений ведущих зарубежных и российских компаний и исследовательских центров в области автоматизации и роботизации подземных горных работ, что очень важно для развития "безлюдных технологий" ведения горных работ, особенно в условиях повышенной опасности осуществления горного производства. Велика роль и адекватных сложному производственному процессу специализированных информационных систем, работающих в режиме реального времени с огромными потоками разнородной информации.

В [3] представлен экспертно-аналитический обзор важнейших достижений в области нелинейной геомеханики и геофизики, геомониторинговых систем, а также современных информационных технологий для развития научных основ "Технологии предупреждения и ликвидации чрезвычайных ситуаций природного и техногенного характера", относимой к числу "критических технологий" для Российской Федерации. Ключевыми здесь являются направления исследований по физике и геомеханике формирования и развития очаговых зон разрушения горных пород в природных и горно-технических системах, а также научные и технико-технологические разработки по созданию многослойной геоинформационно-мониторинговой системы геомеханико-геодинамической и геоэкологической безопасности России и в мире.

Среди основных перспективных направлений исследований, научных задач и прикладных разработок выделяются: формирование методологических основ, разработка современных методов, контрольно-измерительных систем и алгоритмов автоматизированной обработки данных для комплексирования их в единую "*интеллектуальную оболочку с управляющими функциями принятия решений*", которая должна быть организована на принципах "*обучаемости*" и "*обратной связи*" с контролируемыми объектами по набору важнейших параметров недропользования. Для реализации современной геоинформационной среды в решении задач по комплексированию результатов геомеханико-геодинамических, геоэкологических и иных исследований по проблеме эффективного недропользования в [4] предложен новый подход, базирующийся на облачных информационных технологиях. Его практическая значимость показана в работах [5 – 18], посвященных актуальным проблемам недропользования в Сибири и особенно в Кузбассе.

Важно отметить "методологическую роль" энергетического подхода к комплексному анализу больших потоков разнородной информации по объектам недропользования. Необходимость именно такого подхода к анализу разнородных геоинформационных потоков с "неочевидной" их внутренней взаимосвязью обосновывается большим опытом комплексных и междисциплинарных исследований [19 – 27]. Принципиальной значимости — и следствие данных работ: о "модулирующем начале" геоэкологических процессов геомеханико-геодинамическими [3].

Цель настоящей работы — анализ современных достижений информационных технологий BIG DATA ("больших данных"), получивших существенное развитие и широкое распространение в мире [28], но уже применительно к решению сложных задач горного дела — на примере комплексного анализа геомеханико-геодинамического состояния массивов горных пород в регионах активного недропользования Сибири.

ГОРНАЯ ИНФОРМАТИКА И “БОЛЬШИЕ ДАННЫЕ”

С появлением новых мониторинговых систем и средств измерения физико-механических и иных характеристик горного массива и горного производства наблюдается качественное и количественное увеличение геоинформационных потоков. Если сравнительно недавно объемы геоинформации в несколько десятков мегабайт считались огромными, то в настоящее время информационные потоки с объемами в десятки терабайт становятся обычными. Новейшие системы лазерного сканирования, средства радарной интерферометрии, дистанционного зондирования Земли вместе с разнообразными специализированными комплектами датчиков, поставляющих информацию в режиме реального времени, производят очень большие потоки геоинформации, обработка и анализ которой традиционными методами становятся уже невозможными, даже при использовании систем облачных вычислений, которые в приложении к ряду задач горного дела рассмотрены в [4].

Существует несколько определений “*больших данных*” [28, 29], которые описывают их свойства, но при этом не сводятся только к их большим объемам. Понятие сути “большие данные” в общем случае опирается на ставшее уже классическим определение “четырех V”: “объем” (Volume), “скорость обработки” (Velocity), “достоверность” (Veracity) и “разнообразие” (Variety). В практических приложениях используют обычно лишь два-три из них, учитывая сложность одновременной обработки по всем параметрам сразу. В наиболее общем виде, в информационном контексте, под “**большими данными**” понимается совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных больших объемов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного роста количества информации и распределенных по многочисленным узлам вычислительной сети [28]. При этом в технологию “больших данных” включают средства массово-параллельной обработки плохо структурированных и неструктурированных данных (в том числе пространственных) за счет использования специальных алгоритмов, программных каркасов и библиотек проектов.

Большие объемы данных обычно разделяют на две части, зависящие как от их природы, так и от методов работы с ними. Для данных, представляемых таблицами, могут быть использованы традиционные реляционные системы управления базами данных и соответствующие аналитические приемы. В то же время получаемые неструктурированные массивы информации, например при использовании систем лазерного сканирования или спутниковых снимков, не могут быть проанализированы с помощью традиционных статистических моделей. Здесь необходимо переходить к методам интеллектуального анализа, предсказательного моделирования, агентных систем и др. Говоря о большой скорости получения данных, имеются в виду большие скорости их обработки и возможность извлекать необходимые знания из формируемых потоков. Используются как традиционные методы фильтрации, так и технологии агрегирования данных по соответствующим (чаще всего эвристическим) алгоритмам, а также методы оперативной обработки транзакций.

В геомеханико-геодинамических исследованиях возникают качественно новые задачи по выяснению механизмов пространственно-временной упорядоченности и наличия “когерентности” структур в нелинейных средах [3], поэтому в поиске способов их обнаружения в природных диссипативных системах технология “больших данных” может сыграть решающую роль. Это, в частности, такой класс задач, как исследование закономерностей проявления природной и индуцированной сейсмичности, где наряду с традиционными геодинамическими данными, получаемыми от сейсмостанций и GPS полигонов [21 – 24], появляются новые типы “неструктурированной информации”, обработка которой также связана с *технологией “больших данных”*.

Отметим особо современные методы и геомеханические модели для диагностики и контроля изменений напряженно-деформированного состояния массивов горных пород в регионах с высокими техногенными нагрузками (Кузбасс, Норильск, Урал, Кольский полуостров и др.). Исходные результаты проводимых здесь на протяжении многих десятков лет экспериментальных исследований и теоретические расчеты по ним в настоящее время сложно найти и проанализировать, так как нередко они хранились "где угодно и как угодно". Сформировать современные системы на основе хранилищ данных (ETL системы-extract, transform, load) даже с учетом облачных сервисов вряд ли удастся. Многие организации хранили свою расчетную и экспериментальную информацию в различных, часто собственных форматах, в результате чего она либо утрачена, либо для ее использования требуется предпринять определенные усилия по упорядочению "хаоса данных". То же самое относится и к "дополнительной" технологической информации (по [3] — "технологический информационный слой"), связанной с конкретными горными предприятиями, где проводились натурные эксперименты.

Реализация технологии "больших данных" способна снять эти вопросы, а на основе единого подхода — создать общероссийскую систему для сбора, хранения и обработки разнородной и неструктурированной информации по отдельным горнопромышленным регионам России, опираясь на системы облачного сервиса [4]. В большинстве геомеханико-геодинамические данные представляют собой результаты либо экспериментальных замеров, либо соответствующих вычислений в рамках тех или иных механических моделей, использующих различные понятия о поведении и свойствах горного массива.

Так как в настоящее время объем и скорость поступления геоданных быстро нарастают (на несколько порядков!), то требуются качественно иные модели для хранения "произвольных" по форме представления научных данных, в том числе структурированных и неструктурированных, а также новые методы их обработки, основанные на DataScience [30], где преобладают синтезирующие теории, а статистические методы применяются к очень большим объемам и потокам информации.

Применительно к геомеханико-геодинамическим исследованиям выделяются следующие группы "больших данных":

— *экспериментальные данные*, получаемые в режиме реального времени и представляющие собой временные ряды, связанные с изменением характеристик состояния массива горных пород вследствие техногенных нагрузок;

— *облака точек*, получаемые с помощью приборов лазерного сканирования;

— *растровые изображения*, отражающие динамику изменения состояния массива и получаемые с помощью средств дистанционного зондирования Земли;

— *результаты расчетов*, в том числе сценарного типа, с различной пространственной размерностью — на основе геомеханических моделей, которые в большинстве случаев представляют собой различные наборы сеточных данных (grid модели);

— *аудио- и видео- мультимедийные данные*, полученные либо с помощью новейших приборов, либо на основе результатов расчетов;

— *архивные данные различной природы и форматов*, полученные ранее.

Такое многообразие информационных данных ставит не только ряд новых вопросов, связанных с их обработкой и хранением, но и свидетельствует о необходимости создания совершенно новых систем, ориентированных именно на большие потоки разнородной геоинформации. Развитие вычислительных кластеров, в том числе и локальных, позволяет развивать именно такие сложные системы. Естественно возникают вопросы, связанные с доступностью этих *Систем*, а также со специализированной операционной средой, для которой необходимо разра-

батывать соответствующие методы распараллеливания уже имеющихся задач и геомеханических моделей. Для большинства организаций-пользователей это является сложной проблемой, поэтому ныне трудно найти удачные примеры реализации геомеханических вычислительных комплексов, ориентированных на конкретных пользователей.

ГЕОИНФОРМАЦИОННЫЕ АСПЕКТЫ ТЕХНОЛОГИИ “БОЛЬШИХ ДАННЫХ”

Для решения отмеченных задач рассмотрим всю технологию получения разнородных потоков пространственной информации, генерируемых при изменении геомеханико-геодинамического состояния породных массивов, без обсуждения конкретных методов их инструментального измерения.

На рис. 1 приведена общая схема обработки и анализа пространственных данных, основанная на современном подходе, при котором данные, поступившие на вход информационной системы, превращаются в определенные (специализированные) решения. Данная схема позволяет определить процесс обработки как некоторую иерархическую систему, на верхнем уровне которой принимаются необходимые решения по предприятию-пользователю.

Рис. 1. Общая схема процессов обработки информационных данных

Для работы по этой схеме необходимо провести классификацию геомеханических данных, используя подходы BIG DATA. На рис. 2 показан один из возможных вариантов такой классификации, который позволяет определить возможные источники, виды данных, типы их хранения, последующей обработки и вычислений. Здесь достаточно четко выявляются элементы BIG DATA, которые могут использоваться при решении различных геомеханических задач.

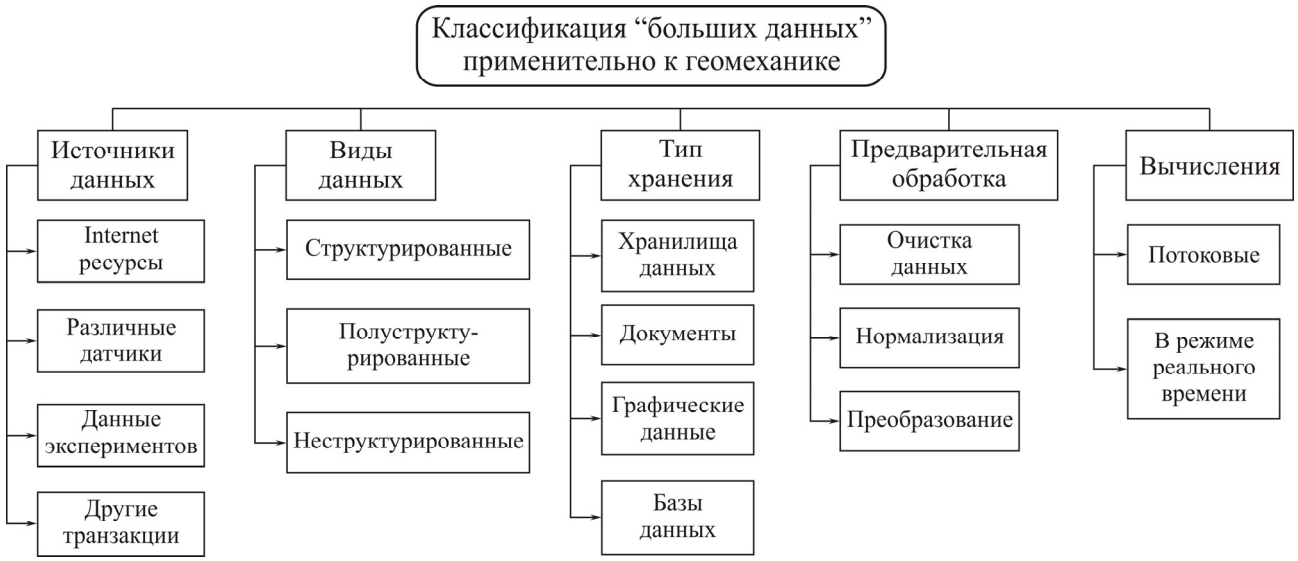

Рис. 2. Классификация геомеханических данных на основе подхода BIG DATA

Остановимся более подробно *на видах данных*. Информационные данные обычно классифицируются на структурированные, полуструктурированные и неструктурированные.

К структурированным данным относятся такие данные, которые имеют некоторую внутреннюю структуру, не обязательно иерархическую, но могут быть впоследствии представлены в форме таблиц или реляционных баз данных.

Полуструктурированные данные чаще всего используются в геоинформационных системах и могут представлять собой как векторные, так и растровые данные.

Неструктурированные данные — это новый вид данных, к которым можно отнести несегментированные изображения, облака пространственных точек (например, данные лазерного и радарного сканирования), различные типы мультимедийной информации.

Отдельно остановимся на обработке изображений, которые занимают все больше места в геомеханике и требуют подходов, значительно отличающихся от решения традиционных задач. В качестве примера можно указать схему анализа данных, полученных с помощью радарных систем дистанционного зондирования Земли, обработка которых позволяет получать такие характеристики, как смещения поверхности, происходящие в процессах разработки полезных ископаемых открытым или подземным способами. При обработке этих данных имеем несколько изображений — интерферограммы, карты смещений (как вертикальных, так и горизонтальных) и ряд других. Иными словами, имеются “потоки” изображений, которые нужно не только структурировать и дешифровать, но и организовать удобную систему их хранения, извлечения отдельных элементов информации, интересующих конкретного пользователя. При этом следует иметь в виду, что поскольку количество информации возрастает в разы и ее объемы даже для самых простых методов обработки превышают десятки гигабайт, то приходится выделять отдельные информационные элементы (на уровне площадных объектов) или использовать мощные вычислительные кластеры.

В связи с резким ростом информационных потоков решение сложных геомеханических задач по принципу “четырех V” весьма затруднительно, поэтому необходимо искать новые подходы, которые смогут решать подобного рода задачи с учетом “естественных ограничений”. Среди них и то, что потоки информационных данных формируются в разных регионах страны неодинаково и надо иметь специальную схему их интеграции, которая позволит реализовать информационные технологии, являющиеся стандартом во всем мире, где уже ставится вопрос о создании организаций, интегрирующих “большие данные” [31].

Рассматриваемый подход подразумевает достаточно простое решение этой задачи, основанное на преобразовании уже имеющихся в различных горнодобывающих предприятиях и организациях локальных вычислительных сетей и не требует дорогостоящих затрат, связанных с приобретением специализированных вычислительных кластеров или аренды соответствующих ресурсов.

ЗАДАЧА ИНТЕГРАЦИИ ИНФОРМАЦИОННЫХ ДАННЫХ

Предлагаемое решение основывается на создании вычислительных кластеров, использующих программно-технические комплексы на базе серверов стандартной архитектуры, с помощью технологии Hadoop [32] в рамках модели программирования MapReduce [33]. Это позволяет повысить не только производительность обработки различной, в том числе и неструктурированной, геоинформации за счет применения параллельных вычислительных систем, но и эффективность разработки новых приложений, связанных с обработкой и анализом “больших данных”. Преимуществами данного подхода являются [34]:

- ***автоматическое распараллеливание*** задачи на кластере из серверов стандартной архитектуры, создаваемых на уровне локальной вычислительной сети;

• ***распределение нагрузки*** между узлами кластера;

• ***защита от сбоев оборудования*** за счет перезапуска задачи на другом кластере;

• ***распределенная файловая система*** для хранения данных на внутренних серверах формируемого кластера.

В наиболее общем виде схема обработки информации на основе технологии Hadoop показана на рис. 3.

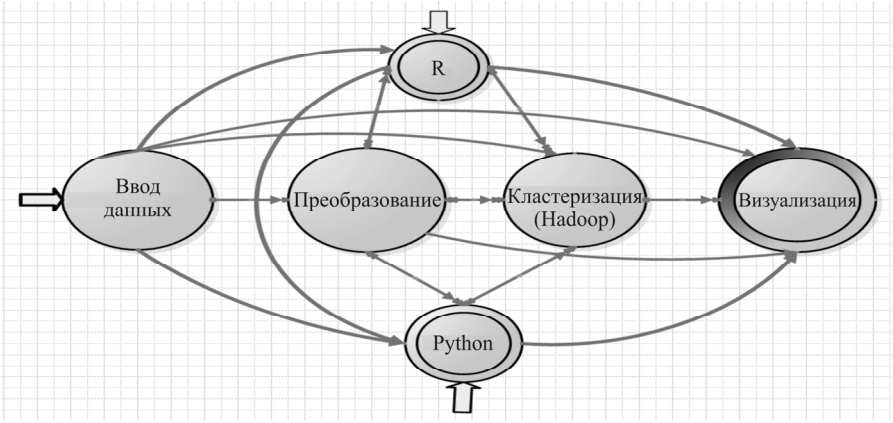

Рис. 3. Схема обработки информационных данных с использованием технологии Hadoop (R и Python — языки программирования приложений)

Отмеченные на этой схеме языки программирования приложений (Python и R) могут быть заменены на другие, например Mathlab, Mathcad, C# или аналогичные им.

В целом вся технология Hadoop представляет собой набор программных компонентов, образующих определенную “экосистему”, и освобождает пользователя от необходимости программирования “своих задач”, погружая расчетные модули в нее; автоматически маршрутизирует расчетные этапы, формируя необходимые информационные потоки. Рассматривая технологию более детально, можно выделить в ней следующие базовые компоненты.

HDFS — распределенная файловая система, позволяющая хранить файл или его компоненты на кластерах системы;

MapReduce — главный выполняемый модуль, программная модель для выполнения распределенных вычислений, состоящая из фаз разбиения и результирующей сборки результатов в среде HDFS;

Hbase — система управления базами данных, ориентированная на обработку столбцов с использованием NoSQL запросов для записи и чтения большого количества данных;

Zookeeper — система управления для координации различных модулей и выполнения соответствующих операций на вычислительных кластерах;

Oozie — система управления и масштабирования вычислительных потоков на кластерах, обеспечивающая комплексирование различных сервисов, включая MapReduce;

Pig — система программирования MapReduce, обеспечивающая как среду выполнения расчетов, так и соответствующий язык на уровне скриптов для анализа наборов данных;

Hive — система высокоуровневого языка последовательных запросов, позволяющая автоматизировать работу с MapReduce; подобно Pig, является абстрактным слоем, используемым для работы с базами данных.

Дополнительно в системе Hadoop имеются средства разработки разнообразных приложений для интеграции с многочисленными программными комплексами:

Sqoop — программное средство для трансляции данных между реляционными базами, хранилищем и Hadoop экосистемой. Оно облегчает систематизацию генерируемых потоков информации при импортировании и экспортировании различных данных при их распараллеливании для MapReduce;

Flume — подсистема для организации распределенного сервиса по предварительной обработке потоков данных, включая их очистку, агрегирование и перемещение большого количества от индивидуальных компьютеров в среду HDFS;

Mahout — библиотека алгоритмов для реализации методов извлечения знаний в распределенной вычислительной среде.

Одна из возможных архитектур экосистемы Hadoop с ориентацией на обработку геомеханических данных приводится на рис. 4.

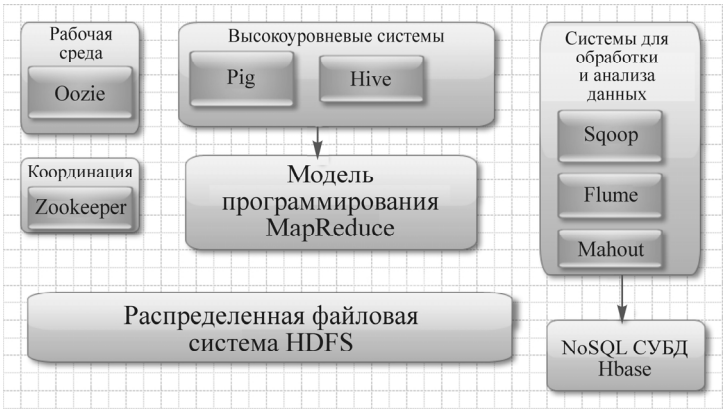

Рис. 4. Архитектура набора технологий Hadoop

Стек технологий состоит из нескольких уровней и включает в себя как обязательные элементы на уровне программирования MapReduce, так и отдельные встраиваемые подсистемы, связи между которыми показаны стрелками. Такая архитектура обеспечивает определенную гибкость и легко настраивается на различные классы задач самим пользователем.

При построении прикладных информационных систем одним из основных является вопрос об источнике информационных данных. Источники таких данных могут быть как локальными, так и распределенными в глобальной вычислительной сети — практически в любом географически удаленном пункте. Технология Hadoop не предусматривает работу с распределенными источниками, за исключением тех, которые находятся в HDFS среде. Для решения вопросов сбора данных наиболее целесообразно использовать облачные технологии [4]. При дальнейшем анализе работы с кластерной технологией будем считать, что вопрос сбора данных из распределенных источников уже решенным.

Особенность предлагаемого решения — потоковая обработка данных, которая дает возможность не только анализировать их большие объемы, но и организовывать работу в реальном масштабе времени. Это позволяет создавать различные комплексные мониторинговые системы, в том числе для анализа геомеханического состояния массива горных пород, подверженного техногенному воздействию, обеспечивая процессы сбора и передачи информационных данных системами облачного сервиса. Предлагаемый подход может вести обработку и анализ больших объемов разнородных данных, отвлекаясь от специфики конкретной задачи.

Для пояснения рассмотрим схему потоковой обработки спутниковых радарных снимков для расчета смещений поверхности Земли в районах с высокими техногенными нагрузками. Эта схема достаточно проста для реализации средствами Hadooop и применяется в наиболее распростра-

ненных пакетах обработки радарных данных методами радарной интерферометрии, например Sarscape, Erdas [35]. Этапы обработки радарных снимков показаны на рис. 5. Однако такие расчеты требуют значительных вычислительных ресурсов, что препятствует обработке уже нескольких десятков снимков, формируемых при потоковой обработке.

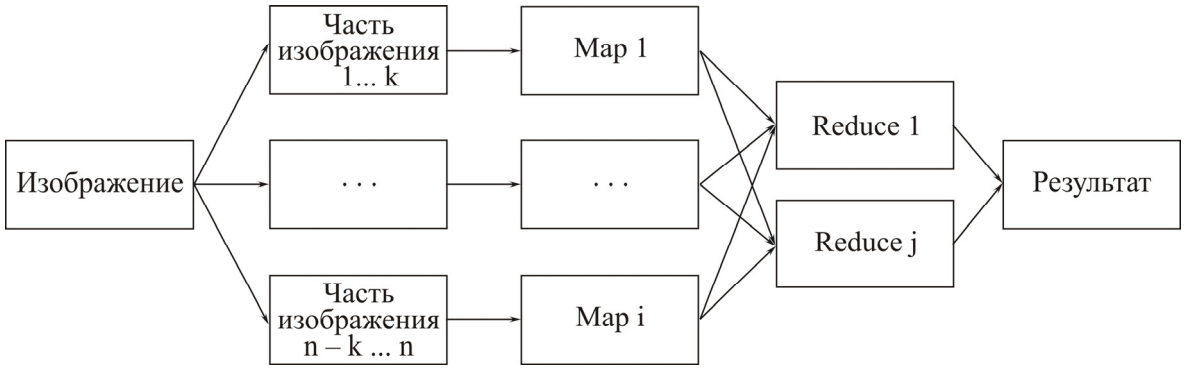

Рис. 5. Схема обработки одного спутникового изображения

Групповая обработка и анализ такой информации носит нетривиальный характер и часто невозможна без применения средств высокопроизводительных вычислений. Для решения подобных задач применяются средства массово-параллельной обработки неопределенно структурированных данных алгоритмами MapReduce [33], программными каркасами и библиотеками проекта Hadoop [32]. Это эффективно в условиях непрерывного прироста данных за счет их распределения по многочисленным узлам вычислительной сети.

В радарной интерферометрии можно выделить общую концепцию по принципу разделения препроцессинга на ресурсоемкие задачи (параллельные) и задачи, не требующие значительного процессорного времени (последовательные). К параллельным задачам можно отнести расчет интерферограммы и свойства когерентности. Такие задачи актуальны, когда речь идет о современных аппаратах типа дистанционного зондирования Земли Cosmo-SkyMed, Sentinel-1A и др. В этих случаях объем предоставляемых информационных ресурсов может измеряться десятками гигабайт, а “локальный” подход к решению проблемы будет использовать большие объемы оперативной памяти и дискового пространства вычислительной системы.

В общем случае может рассматриваться задача обработки изображений дистанционного зондирования Земли, особенностью которой является работа с большими объемами данных (несколько терабайт), а также длительное время выполнения вычислительных операций на отдельном вычислительном устройстве, где предполагается два сценария практического применения:

(1) изображения дистанционного зондирования Земли собираются для каждого контрольного участка за несколько лет, а обработка изображений происходит по запросу пользователя Системы с учетом заданных требований и параметров. Результаты уже проведенных расчетов сохраняются при этом для быстрого доступа;

(2) изображения дистанционного зондирования Земли поступают в виде “потока” (примерно одно-два изображения на каждую территорию). Для поступивших изображений проводится предварительная обработка, в том числе выявление изображений, непригодных для дальнейшего анализа. Пользователь Системы сможет получать доступ к уже имеющимся результатам и выполнять обработку изображений с заданными параметрами.

В обоих случаях необходимо большое количество дискового пространства (от нескольких терабайт) для хранения исходных и полученных данных, а выполнение отдельной операции по обработке изображения может занимать до нескольких часов на отдельном компьютере.

Для решения таких задач предлагается использовать систему для распределенных вычислений Hadoop. Данный подход позволяет отказаться от применения дорогостоящих систем хранения данных в пользу распределенной файловой системы HDFS, которая является частью проекта Hadoop. HDFS предназначена для хранения больших объемов данных (несколько терабайт или петабайт), распределяя информационные данные между большим количеством вычислительных узлов. HDFS предоставляет надежность хранения данных путем резервирования, а также быстрый доступ к ним и легкую масштабируемость путем введения дополнительных узлов в кластер [36].

Применение технологии MapReduce дает возможность выполнять отдельные операции обработки данных параллельно, что приводит к общему снижению длительности вычислительных операций. Эта технология была представлена компанией Google вначале как модель параллельного программирования с использованием вычислительных кластеров на базе персональных компьютеров с низкой стоимостью. Программа, построенная на MapReduce, состоит из двух основных этапов: Map и Reduce. На первом этапе происходит предварительная обработка данных, разделенных на части, каждая из которых обрабатывается на отдельном компьютере. На втором этапе данные, полученные на отдельных компьютерах, собираются вместе для получения конечного результата. Это позволяет снизить общее время длительности вычислительных операций, используя один из указанных выше подходов, либо их комбинацию:

— изображение дистанционного зондирования Земли (ДЗЗ) разделяется на несколько частей и проводится операция по его обработке параллельно на нескольких вычислительных устройствах. Данные, полученные на каждом из узлов, объединяются для формирования конечного результата (см. рис. 5);

— осуществляется одновременно несколько операций по обработке изображений с различными параметрами на нескольких компьютерах. Данный подход позволяет перейти к анализу результатов обработки изображений без ожидания последовательного выполнения каждой операции (рис. 6).

Таким образом, решение задачи обработки ДЗЗ средствами Hadoop позволяет провести предварительную оценку возможностей интеграции технологий обработки спутниковых изображений и распределенных вычислений, а также разработать систему по обработке пространственных данных в распределенной среде. По аналогии можно обрабатывать облака точек, получаемых при лазерном или радарном сканировании элементов породного массива, подвергаемого техногенному воздействию.

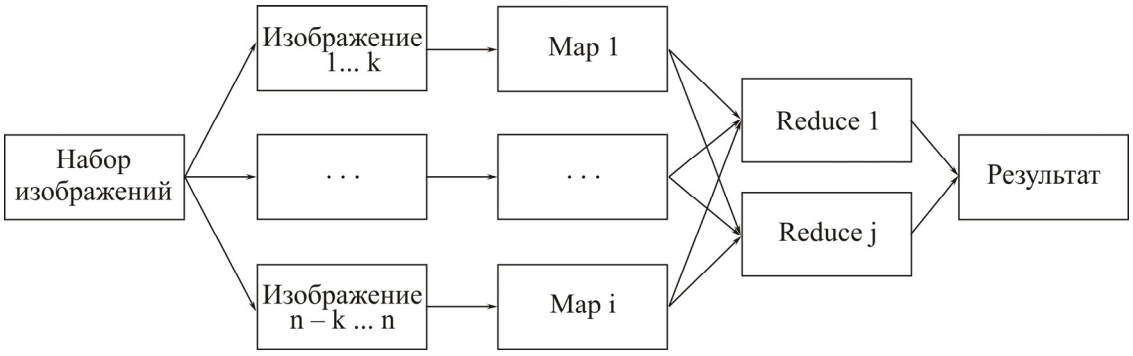

Рис. 6. Схема одновременной обработки нескольких спутниковых изображений

Отмеченное выше подтверждает также возможности обработки больших потоков информации, формируемых в различных разделах горной науки и горного дела. Наиболее эффективно применение подхода для обработки различных данных дистанционного зондирования Земли, объемы которых стремительно растут и становятся более доступными. Отличительная осо-

бенность таких данных — возможность описывать и сравнивать различные характеристики, связанные как с состоянием горного массива, так и окружающей среды на сравнительно больших площадях, что особенно важно для оценок степени техногенного воздействия [3, 4].

Существуют готовые системы, позволяющие выполнять это в виде сервисов, но не решаются вопросы разделения и сборки пространственных данных, а также управления распределенным вычислением для применения этих систем без программирования. Для решения проблемы предлагается технология обработки растровых изображений в рамках модели распределенных вычислений MapReduce, которая может использовать инструменты пространственной обработки в распределенной вычислительной среде без их модификации. Операции над пространственными данными, которые применяются в Map и Reduce, повторяются для различных инструментов геообработки ввиду общности обрабатываемых данных. В настоящей работе предлагается метод, включающий в себя обработчики для операций Map и Reduce и спецификации, на основе которых будет происходить процесс распределения и сбора данных среди вычислительных узлов.

Map и Reduce обработчики представляют собой библиотеки для выполнения WPS-сервисов. Для каждого WPS-сервиса существует спецификация. В зависимости от настроек распределения входных данных, определенных в спецификации, проводится их разделение с последующим вызовом копий сервисов на удаленных узлах. Модуль выполнения сценариев последовательно опрашивает выполняемые копии сервисов, и как только последняя копия сервиса завершает свою работу, все результаты работы копий скачиваются модулем и происходит процесс сборки результата в соответствии с правилами, определенными в спецификации. На рис. 7. показана схема работы метода обработки пространственных данных.

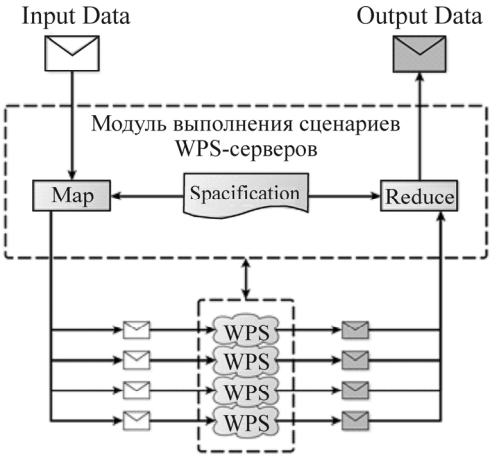

Рис. 7. Схема технологии обработки растровых изображений в рамках модели MapReduce

Обработчик операции Map включает реализованные функции чтения спецификаций, на их основе формируются параметры для распределения растровых данных между вычислительными узлами. Для разделения данных формируются запросы/параметры запуска утилиты GDAL TRANSLATE, предназначенной для конвертации растров, с возможностью получения части растра. Обработчик операции Reduce включает реализованные функции чтения спецификаций и объектно-ориентированные обработчики сбора данных. Обработчики данных выполняют стандартные функции обработки конфликтных ситуаций, возникающих в процессе сбора данных, например при сборе частей мозаики растра в одно целое.

К таким ситуациям можно отнести поступление повторяющихся или неоднозначных данных. В этом случае обработчик применяет к ним операцию, указанную в спецификации. В текущей версии доступны следующие операции: max — установить максимальное значение из

двух перекрывающихся пикселей, min — установить минимальное значение из двух перекрывающихся пикселей, avg — вычислить среднее значение из двух перекрывающихся пикселей. Спецификации написаны в формате Java Script Object Notation (JSON). Настройки спецификаций указывают минимальные и максимальные размеры ячейки для обработки, позволяя операции Map самостоятельно определять размер ячеек для оптимальной загрузки вычислительных узлов вызываемыми сервисами (при этом в обработчик также сообщается число вычислительных узлов). Расчет ячейки проводится на основе стратегии равномерной загрузки вычислительных узлов, т. е. обработчик стремится занять как можно большее число узлов, максимизируя размер ячейки и минимизируя число вызовов сервиса на каждом узле в целях минимизации расходов на соединение и передачу данных.

Спецификации для операции Map содержат следующую информацию: ширина и высота ячейки данных, ширина полосы перекрывающихся пикселей для соседних ячеек. Спецификации для операции Reduce содержат название метода, применяемого на шаге сбора полученных результатов, для обработки перекрывающихся пикселей.

О МЕТОДАХ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА В ЗАДАЧАХ ГОРНОГО ДЕЛА, МИНИ-КЛАСТЕРНЫЕ СИСТЕМЫ

Предлагаемый подход к созданию мини-кластерных систем для обработки “больших данных” также может эффективно использоваться для методов интеллектуального анализа в различных областях знаний [37]. Эти методы пока не нашли применения для решения задач горного дела, однако они позволяют исследователю совершенно по-новому взглянуть на результаты своего научного анализа и обнаружить закономерности, о существовании которых он даже не предполагал. Широко используемые в горном деле модели, связанные с граничными задачами механики горных пород, по своей природе являются многопараметрическими, но нередко возникает вопрос о их адекватности или взаимном влиянии тех или иных параметров. Для такого анализа приходится применять сложные численные методы, в том числе и для оценки степени влияния параметров решения при одновременном изменении некоторых из них. Здесь методы интеллектуального анализа данных (ИАД, datamining), обладающие простотой и наглядностью, легко решают подобные задачи на персональных компьютерах.

Для того, чтобы определить, какие из многочисленных известных параметров наиболее существенно влияют на газоносность угольных пластов, необходимо проводить достаточно сложные комплексные вычисления или применять некоторую идеализированную математическую модель. Тем не менее это не гарантирует необходимой полноты физико-химических моделей, поскольку изменение горнотехнологических условий отработки месторождений оказывает большое влияние и необходимо заново проводить вычисления. Также сложно учесть многообразие и множественность геологических и геомеханических характеристик на развитие геомеханико-геодинамических и геоэкологических процессов даже в случае отработки простых по строению пластовых месторождений полезных ископаемых.

Так как в настоящей работе не представляется возможным достаточно полно рассмотреть все аспекты применения технологии “больших данных” к обработке и анализу горнотехнологических (включая геодинамические) данных, приведем лишь ее общую концептуальную схему (рис. 8).

Схема показывает существование множества методов ИАД, которые можно использовать в работе с “большими данными” при решении задач горного дела. К достоинствам предлагаемого подхода можно отнести и тот факт, что исследователь имеет возможность выбрать наиболее удобный для него и разработать некоторый вычислительный шаблон, по которому в дальнейшем будут

обрабатываться аналогичные данные, в том числе и потоковые. При этом необязательно использовать только один из методов (например, “деревья решений”), можно применять их комбинацию, которая позволит получить наиболее адекватные знания для конкретного набора данных.

Рис. 8. Общая концептуальная схема интеллектуального анализа информационных данных

Применение ИАД существенно упрощается в связи с тем, что существует большое количество программных комплексов, их реализующих [38]. Программные комплексы обработки “больших данных” на основе ИАД позволяют не только проектировать определенные шаги по их обработке, но и запоминать общий алгоритм при потоковой обработке информации. Используемые данные могут быть распределены по нескольким источникам (например, в узлах сети Интернет) и обрабатываться в одном месте с последующей передачей результатов пользователям. На этой основе достаточно просто реализовывать такие режимы облачного сервиса, как SAAS, DAAS, PAAS [4].

Постоянно нарастающие объемы пространственных экспериментальных данных различной природы и форматов обеспечивают более полное описание геомеханико-геодинамических и геоэкологических процессов, однако для этого требуются новые подходы к их хранению, обработке и анализу. Предлагаемый нами подход при его детализации для конкретных видов природных и техногенных процессов потребует, очевидно, значительных объемов дополнительных исследований, связанных как с выбором соответствующих методов интеллектуального анализа, так и их настройкой на конкретный тип потоков данных.

НЕКОТОРЫЕ ПРИМЕРЫ РЕАЛИЗАЦИИ ТЕХНОЛОГИИ BIG DATA В ГОРНОМ ДЕЛЕ

Рассмотрим результаты обработки большого массива данных по газоносности угольных пластов Кузбасса, который получен из различных пространственно-распределенных баз данных и сведен в таблицу, анализ которой проведен средствами Data Mining. Для анализа данных использованы как свободно распространяемые программные комплексы WEKA [39], ORANGE [40], так и коммерческий IBM SPSS Modeler [41]. Из всего многообразия методов (см. рис. 8), выбраны прогностические и описательные. Отдельные результаты выполненных расчетов приведены в качестве иллюстраций на рис. 9 – 12.

В расчетах использована информация по угольным пластам Кузбасса, содержащая 23 000 записей, включающих в себя: название серии пластов (текстовый параметр SERIA), название подсерии (текстовый параметр PODSER), имя свиты (текстовый параметр SVIT), название пласта (текстовый параметр PLAST), название шахтоучастка (текстовый параметр SH_UCH), глубину от поверхности (H, м), влажность пласта (VA, %), зольность пласта (AA, %), среднее значение выхода летучих на глубине 100 м (AP, $м^3$/т), коэффициент метаноемкости (VP, 1/МПа), петрографический коэффициент (PETRGRF, %), экспериментально замеренные значения метаноемко-

сти горючей массы угля (см^3/г) при температуре (°C) и давлениях 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0 МПа, (M10-M40), плотность пород пласта (DI, DK, г/см^3). Использовались экспериментальные и расчетные данные, полученные из различных источников. Они содержали как текстовую, так и числовую информацию. Это существенно осложнило бы построение моделей при использовании других методов, бóльшая часть из которых не способна работать с таким разнообразием.

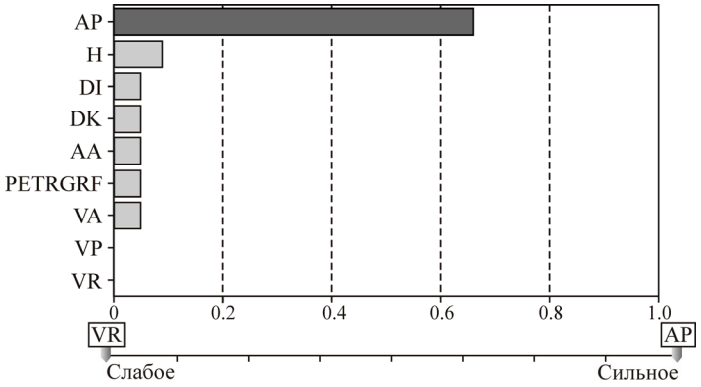

Рис. 9. Расчет влияния различных физико-механических и иных показателей на метаноемкость угольных пластов Кузбасса

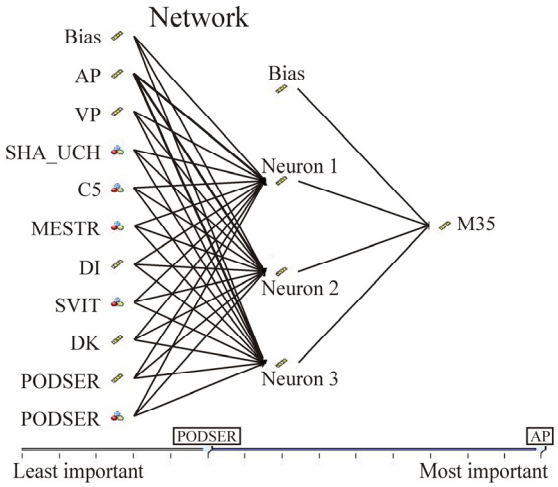

Рис. 10. Модель нейронной сети для расчета метаноемкости угольных пластов Кузбасса (при давлении 3.5 МПа)

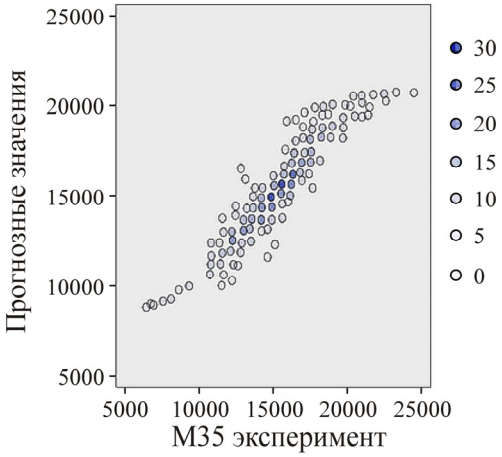

Рис. 11. Соответствие прогнозных и экспериментальных значений метаноемкости угольных пластов при давлении 3.5 МПа, полученное на основе модели нейронной сети

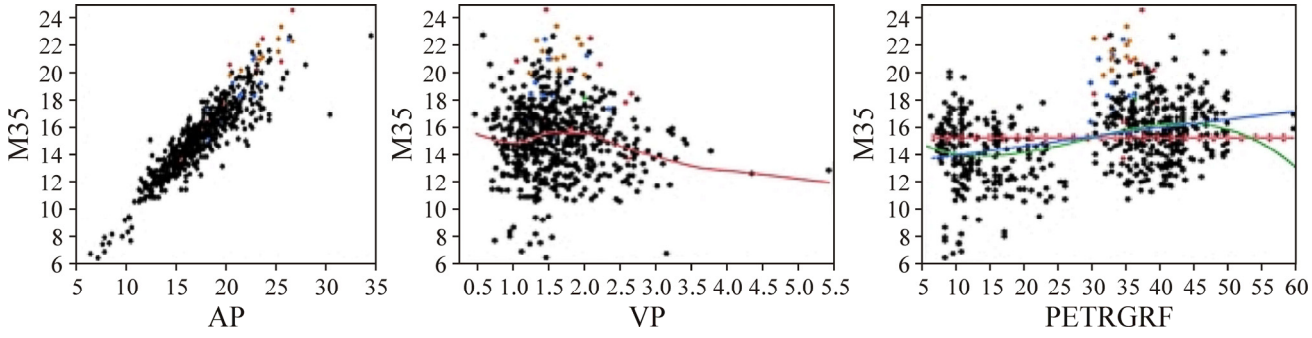

Рис. 12. Покомпонентные расчеты зависимости метаноемкости угольных пластов от среднего выхода летучих (AP), коэффициента метаноемкости (VP) и петрографического коэффициента (PETRGRF)

Из выполненных расчетов следует, что основное влияние на метаноемкость угольных пластов оказывают средний выход летучих (АР) и глубина залегания (Н), а зольность, плотность и петрографические характеристики оказывают менее значительное влияние на ее величину, что согласуется с соответствующими экспериментальными данными по геомеханическим и физико-химическим процессам в угольных пластах [42 – 46].

ВЫВОДЫ

Представлен подход к обработке и анализу информационных данных больших и сверхбольших объемов, генерируемых мониторинговыми системами в горном деле. Для таких систем уже на стадии проектирования требуются новые подходы, позволяющие вести обработку потоков информации в реальном масштабе времени. Соответствующая информация должна накапливаться и впоследствии обрабатываться на основе описываемой технологии BIG DATA. Рассматриваются особенности этой технологии и предлагается общая схема ее реализации на миникластерах с использованием программных средств Hadoop и MapReduce. Приводятся примеры работы информационной системы с неструктурированными данными, что позволяет не только вести потоковую обработку регистрируемых данных, но и значительно сокращать общее время их анализа.

Показана необходимость использования методов интеллектуального анализа геоинформационных данных для больших массивов. Это позволяет в существующем многообразии информационных потоков организовать их гибкую комплексную обработку с целью получения новых знаний. Приводится общая схема методов интеллектуального анализа данных, которые могут быть использованы в горных науках и горном деле.

СПИСОК ЛИТЕРАТУРЫ

1. **Мировой опыт** автоматизации горных работ на подземных рудниках / В. Н. Опарин, Е. П. Русин, А. П. Тапсиев, А. М. Фрейдин, Б. П. Бадтиев; отв. ред. Н. Н. Мельников. — Новосибирск: Изд-во СО РАН, 2007. — 99 с.
2. **Трубецкой К. Н., Кулешов А. А., Клебанов А. Ф., Владимиров Д. Я.** Современные системы управления горно-транспортными комплексами / под ред. акад. РАН К. Н. Трубецкого. — СПб.: Наука. — 2007. — 344 с.
3. **Адушкин В. В., Опарин В. Н.** От явления знакопеременной реакции горных пород на динамические воздействия — к волнам маятникового типа в напряженных геосредах // ФТПРПИ. — Ч. I, 2012. — № 2. — С. 3 – 27; Ч. II, 2013. — № 2. — С. 3 – 46; Ч. III, 2014. — № 4. — С. 10 – 38; Ч. IV, 2016. — № 1. — С. 3 – 49.
4. **Бычков И. В., Опарин В. Н., Потапов В. П.** Облачные технологии в решении задач горной информатики // ФТПРПИ. — 2014. — № 1. — С. 138 – 152.
5. **Потапов В. П.** Математическое и информационное моделирование геосистем угольных предприятий. — Новосибирск: Изд-во СО РАН, 1999. — 156 с.
6. **Опарин В. Н., Потапов В. П., Юшкин В. Ф. и др.** К вопросу формирования информационной геомеханической модели строения Кузнецкого угольного бассейна // ФТПРПИ. — 2006. — № 3. — С. 27 – 49.
7. **Опарин В. Н., Потапов В. П., Попов С. Е., Замараев Р. Ю., Харлампенков И. Е.** Разработка распределенных ГИС-средств мониторинга миграций сейсмических проявлений // ФТПРПИ. — 2010. — № 6. — С. 88 – 95.
8. **Потапов В. П., Опарин В. Н., Логов А. Б., Замараев Р. Ю., Попов С. Е.** Геоинформационная система регионального контроля геомеханических ситуаций на основе энтропийного анализа сейсмических событий (на примере Кузбасса) // ФТПРПИ. — 2013. — № 3. — С. 148 – 156.
9. **Логов А. Б., Опарин В. Н., Потапов В. П., Счастливцев Е. Л., Юкина Н. И.** Энтропийный метод анализа состава техногенных вод горнодобывающего региона // ФТПРПИ. — 2015. — № 1. — С. 168 – 179.

10. **Опарин В. Н., Потапов В. П., Гиниятуллина О. Л., Харлампенков И. Е.** Фрактальный анализ траекторий миграций геодинамических событий в Кузбассе // ФТПРПИ. — 2012. — № 3. — С. 75 – 81.
11. **Потапов В. П., Опарин В. Н., Гиниятуллина О. Л., Харлампенков И. Е.** Разработка сервиса облачных вычислений и обработки данных о сейсмособытиях в геомеханико-геодинамически активных угледобывающих районах Кузбасса // ФТПРПИ. — 2015. — № 3. — С. 162 – 168.
12. **Потапов В. П., Опарин В. Н., Гиниятуллина О. Л., Харлампенков И. Е.** Облачный сервис обработки сейсмособытий на основе диаграмм Вороного с использованием технологии GOOGLE APP ENGINE // ФТПРПИ. — 2015. — № 5. — С. 169 – 178.
13. **Опарин В. Н., Потапов В. П., Гиниятуллина О. Л., Счастливцев Е. Л.** Исследование процесса зарастания отвалов предприятий горного производства по данным дистанционного зондирования // ФТПРПИ. — 2013. — № 6. — С. 133 – 141.
14. **Опарин В. Н., Потапов В. П., Гиниятуллина О. Л., Андреева Н. В.** Мониторинг загрязнений бассейна районов активной угледобычи с использованием данных дистанционного зондирования // ФТПРПИ. — 2012. — № 5. — С. 181 – 188.
15. **Опарин В. Н., Потапов В. П., Гиниятуллина О. Л.** О комплексной оценке состояния окружающей среды по данным дистанционного зондирования Земли в регионах с высокой техногенной нагрузкой // ФТПРПИ. — 2014. — № 6. — С. 199 – 209.
16. **Опарин В. Н., Потапов В. П., Гиниятуллина О. Л., Андреева Н. В., Счастливцев Е. Л., Быков А. А.** Оценка пылевого загрязнения атмосферы угледобывающих районов Кузбасса в зимний период по данным дистанционного зондирования Земли // ФТПРПИ. — 2014. — № 3. — С. 126 – 137.
17. **Потапов В. П., Опарин В. Н., Счастливцев Е. Л., Гиниятуллина О. Л., Харлампенков И. Е., Сидоренко П. В.** Об одном подходе к построению многослойной геоинформационной системы экологической оценки горнопромышленных регионов на примере их биоразнообразия // ФТПРПИ. — 2016. — № 4. — С. 186 – 195.
18. **Опарин В. Н., Потапов В. П., Логов А. Б., Счастливцев Е. Л., Юкина Н. И.** Выделение кластеров загрязняющих ингредиентов в промышленных водных объектах Кузбасса // ФТПРПИ. — 2016. — № 5. — С. 181 – 190.
19. **Опарин В. Н., Тапсиев А. П., Востриков В. И., Усольцева О. М и др.** О возможных причинах увеличения сейсмической активности шахтных полей рудников “Октябрьский” и “Таймырский” Норильского месторождения в 2003 г. // ФТПРПИ. — Ч. I: Сейсмический режим, 2004. — № 4. — С. 3 – 22; Ч. II: Рудник “Октябрьский”, 2004. — № 5. — С. 3 – 25; Ч. III: Рудник “Таймырский”, 2004. — № 6. — С. 5 – 22; Ч. IV: Влияние площадей подработки налегающих породных массивов, 2005. — № 1. — С. 3 – 8.
20. **Опарин В. Н., Еманов А. Ф., Востриков В. И., Цибизов Л. В.** О кинетических особенностях развития сейсмоэмиссионных процессов при отработке угольных месторождений Кузбасса // ФТПРПИ. — 2013. — № 4. — С. 3 – 22.
21. **Современная геодинамика** массива горных пород верхней части литосферы: истоки, параметры, воздействие на объекты недропользования / В. Н. Опарин, А. Д. Сашурин, Г. И. Кулаков, А. В. Леонтьев, Л. А. Назаров и др. — Новосибирск: Изд-во СО РАН, 2008. — 449 с.
22. **Методы** и системы сейсмодеформационного мониторинга техногенных землетрясений и горных ударов / В. Н. Опарин, С. Н. Багаев, А. А. Маловичко и др. — Новосибирск: Изд-во СО РАН. — Т. 1. — 2009. — 304 с; Т. 2. — 2010. — 261 с.
23. **Методы** и измерительные приборы для моделирования и натурных исследований нелинейных деформационно-волновых процессов в блочных массивах горных пород / В. Н. Опарин, Б. Д. Аннин, Ю. В. Чугуй и др. — Новосибирск: Изд-во СО РАН, 2007. — 320 с.
24. **Деструкция** земной коры и процессы самоорганизации в областях сильного техногенного воздействия / В. Н. Опарин, А. Д. Сашурин, А. В. Леонтьев и др. — Новосибирск: Изд-во СО РАН, 2012. — 632 с.
25. **Опарин В. Н., Юшкин В. Ф., Акинин А. А., Балмашнова Е. Г.** О новой шкале структурно-иерархических представлений как паспортной характеристики объектов геосреды // ФТПРПИ. — 1998. — № 5. — С.

26. **Опарин В. Н., Танайно А. С.** Каноническая шкала иерархических представлений в горном породоведении. — Новосибирск: Наука, 2011. — 259 с.
27. **Опарин В. Н., Потапов В. П., Танайно А. С.** К проблеме информационного обеспечения мониторинга геодинамических процессов в условиях интенсивного недропользования в Кузнецком бассейне // ФТПРПИ. — 2006. — № 5. — С. 40 – 66.
28. **Hrushikesha Mohanty Prachet, Bhuyan Deepak Chenthati.** Editors, Big Data, A Primer, Springer, New Delhi, Heidelberg, New York, Dordrecht, London © Springer India, 2015. — 184 p.
29. **Warden P.** Big Data Glossary, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, 2011. — 42 p.
30. **Schutt R., O'Neil C.** Doing DataScience. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, 2014. — 375 p.
31. **Andersen C.** Creating a Data-Driven organization, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2015. — 285 p.
32. **White T.** Hadoop: the definition guide, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, 2009. — 497 p.
33. **Gunarthne Th.** Hadoop MapReduce v 2 Cookbook. Second edition, Packt Publishing, Birmingham UK, 2015. — 215 p.
34. **Warden P.** Big Data Glossary, Published by O'Reilly Media, Inc., 1005, Gravenstein Highway North, Sebastopol, 2011. — 42 p.
35. **Kampes B. M.** Radar Interferometry. Persistent Scaterrer Technique, Published by Springer, 2005. — 212 p.
36. **Lublinsky B., Smith K. T., and Jakubovich A.** Professional Hadoop Solutions, John Wiley&Sons Inc. Indianapolis, Indiana, 2013. — 477 p.
37. **Zaki M. J., Vagner M. Jr.** Data Mining and analysis. Fundamental Concepts and Algorithm, Cambridge University Press, New York, 2014. — 607 p.
38. **Text** mining and visualization: case studies using open-source tools, Edited M. Hoffman and A. Chisholm, CRC Press, Taylor&Francis Group. Boca Raton, London, New York, 2016. — 295 p.
39. **Witten I. H., Frank E.** Data mining: practical machine learning and techniques, second edition. Morgan Kaufman Publishers is an imprint of Elsiver, Amsterdam, Boston, Heidelberg, London, New York, Paris, San Diego, Sydney, Tokio, 2005. — 525 p.
40. **Orange software**. https://en.wikipedia.org/wiki/Orange_(software).
41. **McCormick K., Abbot D., Brown M. S., Khabaza T., and Mutchler S. R.** IBM SPSS Modeler Cookbook. Packt Publishing. Birmingham Mumbai, 2013. — 360 p.
42. **Опарин В. Н., Киряева Т. А., Гаврилов В. Ю., Шутилов Р. А., Ковчавцев А. П., Танайно А. С., Ефимов В. П., Астраханцев И. Е., Гренев И. В.** О некоторых особенностях взаимодействия между геомеханическими и физико-химическими процессами в угольных пластах Кузбасса // ФТПРПИ. — 2014. — № 2. — С. 3 – 30.
43. **Опарин В. Н., Киряева Т. А., Гаврилов В. Ю., Танашев Ю. Ю., Болотов В. А**. К проблеме возникновения очаговых зон подземных пожаров // ФТПРПИ. — 2016. — № 3. — С. 155 – 175.
44. **Бобин В. А.** Сорбционные процессы в природном угле и его структура. — М.: ИПКОН АН СССР, 1987. — 135 с.
45. **Эттингер И. Л., Шульман Н. В.** Распределение метана в порах ископаемых углей. — М.: Наука, 1975. — 111 с.
46. **Ходот В. В., Яновская М. Ф., Премыслер Ю. С. и др.** Физикохимия газодинамических явлений в шахтах. — М.: Наука, 1973. — 139 с.

Поступила в редакцию 29/IX 2016