УДК 550.8.056

Елкин Иван Сергеевич, доцент, к.т.н.
(КузГТУ, г. Кемерово)
Михайлов Иван Владиславович, студент гр. ИИб-251, II курс
(КузГТУ, г. Кемерово)
Почуева Ангелина Сергеевна, студент гр. ИИб-251, II курс
(КузГТУ, г. Кемерово)
Панин Егор Алексеевич, студент гр. ИИб-251, II курс
(КузГТУ, г. Кемерово)
Elkin Ivan Sergeevich, associate professor, Ph.D.
(KuSTU, Kemerovo)
Mikhailov Ivan Vladislavovich,student, 2nd year
(KuSTU, Kemerovo)
Angelina Sergeyevna Pochueva, student, 2nd year
(KuSTU, Kemerovo)
Panin Egor Alekseevich, student, 2nd year
(KuSTU, Kemerovo)

АНАЛИЗ ПРИМЕНИМОСТИ BIG DATA ТЕХНОЛОГИЙ ДЛЯ ОБРАБОТКИ ГЕОФИЗИЧЕСКИХ ДАННЫХ ПРИ РАЗВЕДКЕ УГОЛЬНЫХ МЕСТОРОЖДЕНИЙ

ANALYSIS OF THE APPLICABILITY OF BIG DATA TECHNOLOGIES FOR PROCESSING GEOPHYSICAL DATA IN COAL DEPOSIT EXPLORATION

Аннотация. Представлены исследования применимости инструментария Big Data в ходе обработки массивов геофизической информации, получаемой при поиске и разведке угольных залежей. Выполнен разбор природы и свойств геофизических данных, собираемых в процессе изучения угольных месторождений. Проведён аналитический обзор современных Big Data-технологий, среди которых отдельно рассмотрены экосистема Hadoop, платформа Apache Spark и инструментарий машинного обучения. Создан экспериментальный образец программного решения на базе языка Python и библиотеки PySpark, ориентированный на обработку синтетических каротажных записей по 500 скважинам.
Ключевые слова: Big Data; геофизические данные; угольные месторождения; Apache Spark; PySpark; распределённые вычисления; машинное обучение; обработка каротажа.

Abstract. This paper examines the applicability of Big Data tools for processing geophysical datasets acquired during the exploration and prospecting of coal deposits. The nature and properties of geophysical data collected during the study of coal deposits are analyzed. An analytical review of contemporary Big Data technologies is provided, with a specific focus on the Hadoop ecosystem, the Apache Spark platform, and machine learning toolkits. An experimental software solution was developed using Python and the PySpark library, designed for processing synthetic well-logging records from 500 boreholes.
Keywords: Big Data; geophysical data; coal deposits; Apache Spark; PySpark; distributed computing; machine learning; well-log processing.

На современном этапе геологоразведка угольных месторождений сопровождается стремительным ростом объёмов геофизической информации. В ходе полевых работ активно используются различные способы изучения недр: сейсморазведка (в том числе в двух- и трёхмерном исполнении); геофизические исследования скважин (известные как каротаж); гравиметрические и магнитометрические измерения. Применение новейших цифровых приборов с высокой частотой оцифровки и многоканальными системами сбора позволяет накапливать информационные массивы размером до терабайт. При этом процедуры обработки и последующей интерпретации собранных сведений становятся всё более трудоёмкими и требовательными к вычислительным ресурсам, занимают много времени [1-4].

Одна из центральных трудностей, с которой сталкивается современная геофизика, – это разрыв между скоростью накопления данных и возможностями их обработки классическими методами. Значительная часть геофизических материалов анализируется с помощью специализированных программных продуктов, функционирующих на однопроцессорных архитектурах, и стандартных вычислительных алгоритмов, которые зачастую не справляются с обработкой многогигабайтных и терабайтовых массивов [2].

В зарубежных научных кругах вопросы внедрения распределённых вычислений в геофизику освещены в публикациях Chambers и Zaharia (2018), Hall (2020) [5, 6]. Российские исследования представлены работами Никитина и Петрова (2018), Титовым и Бобровым (2021) [3, 4]. Вместе с тем системный анализ того, насколько Big Data-технологии применимы именно к задачам разведки угольных месторождений с учётом специфики угленосных толщ, недостаточно изучен.

Последние годы отмечены активным развитием технологий Big Data в различных научных и промышленных областях. Они дают возможность обрабатывать значительные информационные объёмы за счёт распределённых вычислений, масштабируемых систем хранения и методов машинного обучения. Их применение предполагает новые перспективы

для анализа геофизических данных, повышения достоверности геологических прогнозов и ускорения обработки результатов геологоразведочных изысканий [8].

Особую значимость рассматриваемая тема приобретает для угледобывающей отрасли Кузбасса, поскольку качественная обработка геофизической информации имеет первостепенное значение при изучении строения угольных пластов, выявлении тектонических нарушений и оценке целесообразности разработки месторождений [9].

В ходе геологоразведочных работ на угольных месторождениях востребован целый ряд геофизических методов. Сейсморазведка (2D и 3D) даёт возможность изучать строение горных пород и выявлять тектонические разрывы на базе анализа распространения упругих колебаний, отражающихся от границ геологических слоёв, различающихся по плотности и упругим свойствам. Геофизические исследования скважин (каротаж) включают электрический каротаж, с помощью которого измеряют удельное сопротивление пород; радиоактивный каротаж, позволяющий оценивать плотность и литологический состав; а также акустический каротаж для изучения упругих характеристик и степени трещиноватости горных пород [1].

Геофизическим данным, получаемым в рамках этих исследований, присущ ряд черт: значительный информационный объём (трехмерная сейсмическая съёмка на площади 50 кв. км с шагом дискретизации 2 мс даёт до 500 Гбайт исходных данных); разнородность форматов (SEG-Y, LAS, DLIS, CSV); сложное внутреннее устройство, включающее временные ряды и пространственные привязки; присутствие шумов и погрешностей измерений (уровень помех может достигать 10–15% от полезной составляющей); неравномерность пространственного распределения; многопараметровый характер [2].

Основные характеристики геофизических данных при разведке угольных месторождений систематизированы в табл. 1.

Таблица 1.

Характеристики геофизических данных при разведке угольных месторождений

Параметр	Значение/Характеристика	Примечание
Объём информации	До 500 ГБ на одну 3D-съёмку	При многоволновой сейсморазведке - кратно больше
Форматы данных	SEG-Y, LAS, DLIS, CSV, TXT	Требуется предварительная конвертация и нормализация
Структура	Временные ряды,	Сложность

	пространственные координаты	согласования разных типов данных
Уровень шума	10–15% от полезного сигнала	Необходима фильтрация
Пространственное распределение	Неравномерное (сейсмика - сетка, скважины - точки)	Требуется интерполяция
Количество параметров	5-20 и более на одну точку измерения	Необходимы многомерные методы анализа

Для обработки больших объёмов геофизической информации использовалась платформа Apache Spark в локальном режиме с 8 потоками. Spark обеспечивает вычисления в оперативной памяти (in-memory computing) и горизонтальное масштабирование, что даёт существенный выигрыш в скорости по сравнению с классической моделью MapReduce [8]. В качестве эталона для сравнения выбран традиционный подход на базе библиотеки pandas, работающей в однопоточном режиме.

Для проверки предложенного подхода разработан прототип на языке Python с использованием PySpark [7].

Характеристики тестового набора данных приведены в табл. 2.

Таблица 2

Характеристики тестового набора данных

Параметр	Значение
Количество скважин	500
Интервал глубин	0–500 м
Шаг дискретизации	0,2 м
Точек на скважину	2500
Всего точек данных	1 250 000
Используемые параметры	ГК, ПС, сопротивление, плотность, пористость
Уровень искусственного шума	До 10%
Внесено аномальных зон	~12% скважин

Прототип реализует последовательный алгоритм обработки данных из шести этапов (рис.1). Ниже приведено описание каждого этапа с указанием математических выражений.

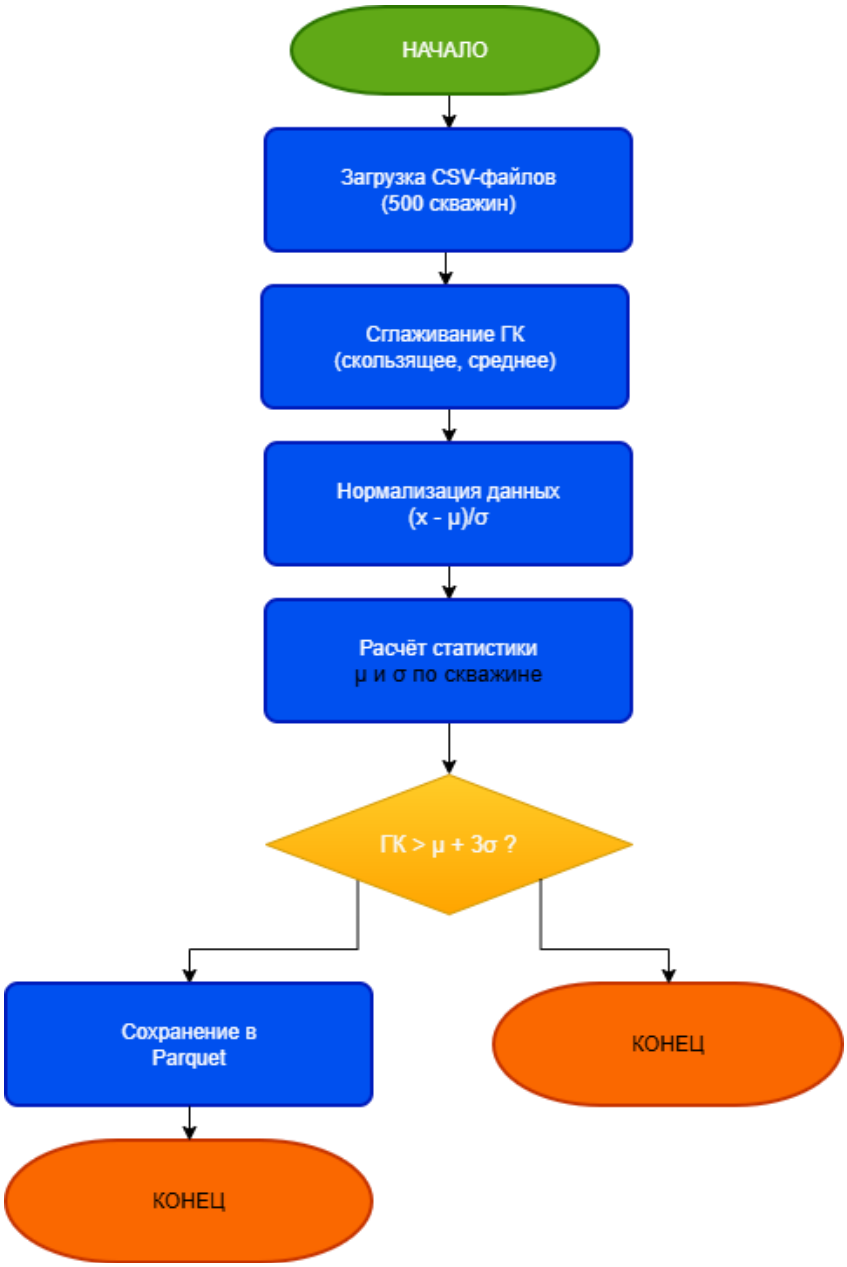

Рисунок 1 – Блок-схема алгоритма обработки

Этап 1. Загрузка данных. Программа считывает 500 CSV-файлов. Spark автоматически распределяет чтение между доступными ядрами.

Этап 2. Фильтрация шумов (сглаживание). Для снижения влияния помех применяется скользящее среднее с окном в 5 точек:

$$GR_{\text{сглаж(i)}} = \frac{GR(i-2)+GR(i-1)+GR(i)+GR(i+1)+GR(i+2)}{5},$$

где $GR_{\text{сглаж}(i)}$ – сглаженное значение гамма-каротажа на той же глубине; $GR_{(i)}$ – значение гамма-каротажа на глубине с индексом i (исходные данные); $i-2$, $i-1$, $i+1$, $i+2$ – индексы двух соседних точек выше и двух ниже по глубине.

Операция выполняется независимо для каждой скважины.

Этап 3. Нормализация данных. Выполняется Z-нормализация для приведения всех признаков к единому масштабу:

$$x_{norm} = \frac{x-\mu}{\sigma},$$

где x – исходное значение параметра (например, значение ГК, ПС или сопротивления в конкретной точке); x_{norm} – нормализованное значение того же параметра; μ – среднее арифметическое значение данного параметра по всей выборке; σ – среднеквадратическое отклонение данного параметра по всей выборке.

Этап 4. Для каждой скважины вычисляются среднее и стандартное отклонение:

$$\mu_{well} = \frac{1}{n}\sum_{j=1}^{n} GR_j,\ \sigma_{well} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(GR_j - \mu_{well})^2},$$

где μ_{well} – среднее значение гамма-каротажа по конкретной скважине; n – количество точек измерения в данной скважине (в эксперименте n=2500); j – индекс точки измерения (от 1 до n); GR_j – значение гамма-каротажа в j-й точке; σ_{well} – стандартное отклонение гамма-каротажа по конкретной скважине.

Этап 5. Точка считается аномальной, если её значение превышает среднее более чем на три сигмы:

$$is_{anomaly(i)} = \begin{cases} 1, \text{ если } GR(i) > \mu_{well} + 3\sigma_{well} \\ 0, \text{ иначе} \end{cases},$$

где $is_{anomaly(i)}$ – бинарный признак: 1 означает «аномалия», 0 – «норма»; $GR(i)$ – значение гамма-каротажа в точке с индексом i; μ_{well} – среднее значение гамма-каротажа по данной скважине; σ_{well} – стандартное отклонение гамма-каротажа по данной скважине.

Доля аномальных точек в выборке:

$$P_{anomaly} = \frac{N_{anomaly}}{N_{total}} 100\%,$$

где $P_{anomaly}$ – процент аномальных точек в общем массиве данных; $N_{anomaly}$ – количество точек, отмеченных как аномальные; N_{total} – общее количество точек в выборке (в эксперименте 1 250 000).

По результатам эксперимента $P_{anomaly}$≈12%.

Этап 6. Обработанные данные сохраняются в формате Parquet, обеспечивающем сжатие и быстрый доступ.

В рамках экспериментального сопоставления получены следующие показатели. На вычислительной машине с процессором AMD Ryzen 5 3400G (4 ядра, 8 потоков) и 16 ГБ оперативной памяти классический подход на базе библиотеки pandas показал время обработки 14,3 секунды. Созданный прототип на Apache Spark в локальном режиме справился с аналогичной задачей за 7,3 секунды. Следовательно, даже в однокомпьютерной конфигурации удалось добиться ускорения в 1,96 раза.

Таблица 3

Сравнение производительности методов обработки данных

Метод обработки	Конфигурация	Время выполнения	Максимальный объём данных
Pandas (традиционный)	Ryzen 5 3400G, 16 ГБ RAM	14,3 сек	Ограничен ОЗУ (~5-8 млн строк)
PySpark (Big Data)	Ryzen 5 3400G, 16 ГБ RAM	7,3 сек	Масштабируется добавлением узлов

Анализ результатов дополнительно показал, что в 12% скважин присутствуют интервалы, отнесённые к категории аномальных, что вероятно свидетельствует о наличии зон с изменёнными коллекторскими свойствами либо тектоническими нарушениями.

Полученные экспериментальные данные подтверждают, что даже в рамках одного компьютера технологии распределённых вычислений способны ускорять обработку геофизических материалов. При переходе к полноценному четырёхузловому кластеру, согласно теоретическим оценкам, ускорение может возрасти до трёх и более раз. Добавление дополнительных узлов (10–20) позволяет снизить время обработки практически пропорционально числу вводимых вычислительных ресурсов. Это преимущество становится особенно весомым при работе с крупными месторождениями, где объёмы геофизической информации измеряются терабайтами [4].

Основное достоинство Big Data-технологий – горизонтальная масштабируемость. Традиционные методы, базирующиеся на использовании одного сервера, при росте объёмов данных неизбежно упираются в ограничения по доступной оперативной памяти и процессорной мощности. Единственным выходом в такой ситуации становится приобретение более мощного оборудования, что связано с большими финансовыми затратами. Распределённые системы предлагают иной путь: наращивание вычислительной силы за счёт добавления в кластер новых узлов, причём время обработки практически не увеличивается, а нагрузка автоматически перераспределяется между всеми доступными ресурсами [8].

Большим потенциалом обладает и автоматизация процессов интерпретации геофизических данных. В разработанном прототипе для выявления аномалий применялось простое правило трёх сигм. Однако дальнейшее использование методов машинного обучения позволяет значительно расширить круг автоматизируемых операций. В их числе - автоматическое распознавание литотипов пород (песчаник, алевролит, уголь) по комплексу каротажных кривых, корреляция разрезов скважин, прогнозирование физико-механических свойств горных пород [9].

Нельзя обойти вниманием и экономическую составляющую. Традиционная вычислительная инфраструктура для геофизической обработки, подразумевающая применение суперкомпьютеров или специализированных рабочих станций, отличается крайне высокой стоимостью. Кластерные решения на базе стандартных серверов, напротив, позволяют снизить капитальные вложения при сохранении или даже росте производительности [6].

В качестве следующего перспективного развития видится разработка специализированных методов обработки геофизических данных с опорой на Big Data-технологии и машинное обучение, в том числе построение нейросетевых моделей для автоматической интерпретации сейсмических разрезов, прогноза качества углей и обнаружения зон геологических нарушений. Актуальным остаётся и создание отказоустойчивых распределённых систем для мониторинга геодинамических процессов в реальном времени на действующих горных предприятиях.

Таким образом, при выполнении исследований было сделано:

– Проанализировано современное состояние технологий работы с большими данными. Показано, что такие платформы, как Hadoop, Apache Spark, NoSQL-базы данных и инструменты машинного обучения, обеспечивают распределённое хранение, параллельные вычисления и интеллектуальный анализ информации, недоступные классическим методам.

– Создан демонстрационный прототип на Python с использованием PySpark, реализующий загрузку, предварительную обработку, фильтрацию, статистический анализ и выделение аномальных зон в каротажных данных. Прототип подтвердил практическую пригодность предложенных решений.

– Проведённая оценка эффективности показала сокращение времени обработки в 1,96 раза на однокомпьютерной конфигурации, повышение масштабируемости, возможность автоматизации анализа и экономическую оправданность внедрения Big Data-технологий в практику геологоразведочных работ.

– Определены главные характеристики геофизических массивов, обуславливающие необходимость перехода к распределённым вычислительным схемам. Экспериментальный образец показал сокращение длительности вычислений в 1,96 раза на однокомпьютерной системе с процессором AMD Ryzen 5 3400G и 16 ГБ оперативной памяти. По результатам теоретической оценки установлено, что масштабирование до четырёхузлового кластера способно дать ускорение в три и более раза.

Представленные выводы базируются на синтетических данных, имитирующих каротажные измерения. Для подтверждения эффективности в реальных условиях необходима апробация на производственных материалах с участием геофизиков.

Технологии, относящиеся к классу Big Data, несут в себе высокий потенциал для применения в геофизической практике при изучении угольных месторождений. Их внедрение даёт возможность результативно работать с крупными массивами информации, ускорять процесс геологоразведочных изысканий и улучшать качество интерпретации геологических сведений. Разработанный прототип служит доказательством работоспособности предложенного подхода.

Список литературы

1. Бондаренко В.М., Демура Г.В., Савенко Е.И. Общий курс разведочной геофизики. – Москва: Недра, 2019. – 350 с.
2. Бондаренко В.М., Лумпов Е.Е., Лыхин А.А. Интерпретация геофизических данных: учебное пособие. – Москва: МГРИ, 2020. – 280 с.
3. Уайт Т. Hadoop: подробное руководство. – Санкт-Петербург: Питер, 2019. – 672 с.
4. Лубенская Е.В. Большие данные: технологии обработки и хранения. – Москва: Юрайт, 2021. – 320 с.
5. Chambers B., Zaharia M. Spark: The Definitive Guide. – O'Reilly Media, 2018. – 600 p.
6. Hall M.A. A survey of machine learning for big data processing // Journal of Big Data. – 2020. – Vol. 7, no. 1. – P. 1–35.
7. Барсегян А.А., Куприянов М.С., Степаненко В.В. Технологии анализа данных: Data Mining, Text Mining, Visual Mining. – Санкт-Петербург: БХВ-Петербург, 2019. – 420 с.
8. Федоров А.В. Введение в Apache Spark: обработка больших данных на Python. – Санкт-Петербург: БХВ-Петербург, 2022. – 350 с.
9. Вьюгин В.В. Математические основы машинного обучения. – Москва: МЦНМО, 2019. – 380 с.