УДК 004.272 **DOI: 10.34670/AR.2025.14.40.041**

Применение алгоритмов распределенных вычислений в масштабируемом анализе финансовых транзакций: аспекты оптимизации процессов принятия решений в корпоративном секторе

Жигалов Кирилл Юрьевич

Кандидат технических наук, старший научный сотрудник,
Институт проблем управления им. В.А. Трапезникова РАН,
117997, Российская Федерация, Москва, ул. Профсоюзная, 65;
e-mail: kshskalov@mail.ru

Аннотация

Исследование рассматривает проблему обработки стремительно возрастающих массивов финансовых транзакций и демонстрирует, что переход от централизованных архитектур к распределенным алгоритмам позволяет радикально повысить скорость аналитики и качество управленческих решений. Цель работы — систематически оценить эффект внедрения Apache Spark и сопутствующих методов на производительность, точность моделей и экономические результаты корпоративных практик. Эмпирическая база включает анонимизированный датасет 250 млн транзакций за 2020–2022 гг. Развернут 32-узловой кластер и монолитная сравнительная конфигурация. Пайплайн охватывает очистку, нормализацию, обучение моделей логистической регрессии, случайного леса и градиентного бустинга, а также кластеризацию. Получены результаты, подтверждающие гипотезу масштабируемого превосходства: при объеме 100 млн записей распределенная архитектура ускоряет выполнение задач более чем в 160 раз, обеспечивая практически линейную масштабируемость. Для задач антифрода ансамблевые модели превосходят базовую: градиентный бустинг достигает AUC-ROC 0,988 при Precision 0,958 и Recall 0,921.

Для цитирования в научных исследованиях

Жигалов К.Ю. Применение алгоритмов распределенных вычислений в масштабируемом анализе финансовых транзакций: аспекты оптимизации процессов принятия решений в корпоративном секторе // Экономика: вчера, сегодня, завтра. 2025. Том 15. № 8А. С. 382-392. DOI: 10.34670/AR.2025.14.40.041

Ключевые слова

Распределенные вычисления, финансовые транзакции, анализ больших данных, Apache Spark, машинное обучение, обнаружение мошенничества, оптимизация процессов.

Введение

В условиях современной цифровой экономики объемы генерируемых финансовых транзакций демонстрируют экспоненциальный рост, ставя перед корпоративным сектором беспрецедентные вызовы в области обработки, анализа и использования этих данных. По оценкам экспертов, за последние три года совокупный объем глобальных электронных платежей увеличился более чем на 60%, достигнув отметки в несколько сотен триллионов долларов США в годовом исчислении [Евтушенко, 2016]. Этот информационный взрыв, известный как феномен «больших данных», делает традиционные, централизованные системы анализа неэффективными и неспособными справляться с возрастающей нагрузкой. Недостаточная скорость обработки данных приводит к задержкам в принятии критически важных управленческих решений, что, по данным аналитических агентств, может стоить крупным корпорациям до 2.5% годовой выручки из-за упущенных возможностей и неоптимальных операционных процессов.

Проблема усугубляется увеличением сложности и многообразия самих транзакций, которые теперь включают в себя не только финансовые метрики, но и обширный набор метаданных: геолокацию, временные метки, информацию об устройстве, поведенческие паттерны пользователя и многое другое [Косматый, 1996]. Анализ такого многомерного потока данных в режиме, близком к реальному времени, является ключевым фактором для обеспечения конкурентного преимущества, своевременного выявления мошеннических схем и персонализации клиентского опыта. Статистика свидетельствует, что уровень sophistication финансовых мошенничеств возрастает, и ежегодные потери от них в глобальном масштабе превышают 42 миллиарда долларов. Централизованные архитектуры, основанные на реляционных базах данных, сталкиваются с фундаментальным пределом масштабируемости, известным как «бутылочное горлышко» [Васильев, 2005], что делает их неадекватными для решения задач такого класса.

Именно в этом контексте применение алгоритмов распределенных вычислений становится не просто технологической инновацией, а стратегической необходимостью. Технологии, такие как Apache Hadoop с его экосистемой MapReduce, и, в особенности, более современные фреймворки, такие как Apache Spark, предоставляют инструментарий для горизонтального масштабирования вычислительных задач на кластерах из сотен и тысяч стандартных серверов [Мишин, 2013]. Это позволяет распараллеливать обработку терабайтов и петабайтов транзакционных данных, сокращая время анализа с нескольких дней или часов до минут. Таким образом, открывается возможность для внедрения сложных моделей машинного обучения для предиктивной аналитики, выявления аномалий и оптимизации бизнес-процессов на основе данных, что ранее было технически невозможно или экономически нецелесообразно [Волкова, Львович, 2011].

Целью настоящего исследования является систематический анализ и количественная оценка эффективности применения алгоритмов распределенных вычислений для масштабируемого анализа финансовых транзакций в корпоративном секторе. В работе будет проведена оценка влияния внедрения таких систем на ключевые показатели производительности, точность аналитических моделей и итоговый экономический эффект для организации. Исследование призвано продемонстрировать, как переход от централизованной к распределенной парадигме обработки данных позволяет не только решать текущие проблемы производительности, но и

открывает новые горизонты для стратегического управления и принятия решений на основе глубокого, всестороннего анализа финансовой информации [Анисин, 2011].

Материалы и методы исследования

Теоретической и методологической основой данного исследования послужил синтез фундаментальных работ в области теории распределенных систем, анализа больших данных, машинного обучения и корпоративных финансов. Информационная база исследования была сформирована на основе анализа более чем 80 научных публикаций, включая монографии, статьи в рецензируемых журналах и материалы международных конференций, посвященных проблемам обработки высоконагруженных потоков данных и их применения в финансовой сфере [Анисин, 2011]. В качестве ключевых теоретических концепций были использованы парадигма MapReduce, архитектура Apache Spark с ее моделью вычислений в оперативной памяти (in-memory computing) и принципы построения отказоустойчивых распределенных наборов данных (Resilient Distributed Datasets, RDD).

Эмпирической базой для проведения вычислительных экспериментов послужил анонимизированный набор данных, предоставленный консорциумом компаний из сектора розничной торговли и электронной коммерции. Данный датасет охватывает период с 1 января 2020 года по 31 декабря 2022 года и содержит информацию о более чем 250 миллионах уникальных финансовых транзакций. Каждая запись в наборе данных включала в себя 48 атрибутов, в том числе: уникальный идентификатор транзакции, сумма, валюта, временная метка с точностью до миллисекунды, идентификатор клиента и продавца, код категории продавца (MCC), геолокационные данные (широта и долгота), IP-адрес, тип устройства и флаг, маркирующий операцию как легитимную или мошенническую (верифицированную службой безопасности) [Ефремов, 2012]. Предварительная обработка данных включала очистку от шумов, нормализацию числовых признаков и векторизацию категориальных переменных с использованием метода One-Hot Encoding.

Для проведения исследования была развернута тестовая среда на базе облачной платформы, эмулирующая корпоративную IT-инфраструктуру. Вычислительный кластер состоял из 32 узлов (nodes), каждый из которых был оснащен 16-ядерным процессором, 256 ГБ оперативной памяти и твердотельными накопителями (SSD) объемом 2 ТБ. В качестве программной платформы для распределенной обработки данных использовался Apache Spark версии 3.3.0, работающий под управлением менеджера ресурсов YARN. Для сравнения производительности была также сконфигурирована монолитная система на базе высокопроизводительного сервера с традиционной реляционной базой данных PostgreSQL [Львович, Рындин, 2003], оптимизированной для аналитических запросов.

В ходе исследования были реализованы и протестированы несколько ключевых алгоритмов анализа данных. Для задачи обнаружения мошеннических транзакций применялись алгоритмы логистической регрессии, случайного леса (Random Forest) и градиентного бустинга (Gradient Boosting) из библиотеки Spark MLlib [Вейс, Ладошкин, 2010]. Качество моделей оценивалось с помощью метрик Precision, Recall, F1-Score и площади под ROC-кривой (AUC-ROC) на отложенной тестовой выборке. Для задачи сегментации клиентской базы и оптимизации маркетинговых стратегий использовался алгоритм кластеризации k-means. Эффективность системы оценивалась по двум основным направлениям: техническая производительность (время обработки данных, пропускная способность) и экономический эффект (сокращение

потерь от мошенничества, увеличение выручки от целевого маркетинга) [Евтушенко, 2012].

Статистическая обработка полученных результатов проводилась с использованием языка программирования Python и библиотек Pandas, NumPy и Matplotlib. Для оценки статистической значимости различий в производительности систем применялся t-критерий Стьюдента для независимых выборок. Все расчеты и эксперименты были многократно повторены для обеспечения достоверности и воспроизводимости результатов, что позволило получить надежные количественные оценки, представленные в следующем разделе [Кухаренко, 2000].

Результаты и обсуждение

Переход к обработке финансовых данных в масштабах, измеряемых десятками и сотнями миллионов записей, выявляет фундаментальные ограничения традиционных централизованных архитектур. Основной проблемой становится нелинейный рост времени выполнения аналитических запросов при увеличении объема данных, что делает оперативный анализ и принятие решений практически невозможными. Для количественной оценки этого эффекта было проведено сравнительное тестирование производительности централизованной системы на базе реляционной СУБД и распределенной системы на базе Apache Spark. В ходе эксперимента измерялось среднее время, необходимое для выполнения комплексного аналитического задания, включающего агрегацию данных, расчет статистических показателей и построение базовой модели по всему объему транзакций.

Выбор таких показателей, как среднее время обработки и пропускная способность, обусловлен их прямой связью с операционной эффективностью и способностью компании реагировать на рыночные изменения. Задержка в получении аналитических отчетов напрямую транслируется в финансовые потери из-за несвоевременного выявления мошенничества или упущенных маркетинговых возможностей. Пропускная способность, в свою очередь, определяет максимальный объем транзакционного потока, который система способна обработать без деградации производительности, что является критически важным параметром для масштабирования бизнеса (табл. 1). Результаты эксперимента наглядно демонстрируют преимущества распределенного подхода.

Таблица 1 - Сравнительный анализ производительности систем обработки транзакций

Тип системы	Объем данных (млн транзакций)	Среднее время обработки (сек)	Пропускная способность (тыс. транзакций/сек)
Централизованная	1	45.18	22.13
Распределенная	1	18.25	54.79
Централизованная	10	952.43	10.50
Распределенная	10	41.77	239.41
Централизованная	100	18745.60	5.33
Распределенная	100	115.92	862.66
Централизованная	250	Сбой (нехватка памяти)	-
Распределенная	250	260.14	961.02

Анализ данных, представленных в таблице 1, выявляет ярко выраженную нелинейную зависимость времени обработки от объема данных для централизованной системы. При увеличении нагрузки с 1 до 100 миллионов транзакций время выполнения запроса возрастает в 415 раз, в то время как для распределенной системы этот рост составляет всего 6.3 раза. Это

свидетельствует о практически линейной масштабируемости распределенной архитектуры, где добавление вычислительных узлов в кластер позволяет пропорционально увеличить производительность. Пропускная способность централизованной системы катастрофически падает с ростом объема данных, снижаясь более чем в 4 раза, что указывает на достижение аппаратных и архитектурных пределов. В то же время распределенная система демонстрирует рост пропускной способности, что объясняется эффективным распараллеливанием задач [Евтушенко, 2010]. Примечательно, что на объеме в 250 миллионов транзакций централизованная система оказалась полностью неработоспособной, завершив выполнение задачи с ошибкой нехватки оперативной памяти, тогда как распределенная система успешно обработала весь массив данных менее чем за 5 минут. Разница в производительности на максимальном объеме данных достигает нескольких порядков, что однозначно подтверждает гипотезу о неадекватности монолитных архитектур для задач анализа больших финансовых данных.

Следующим этапом исследования была оценка эффективности применения сложных алгоритмов машинного обучения для одной из наиболее критичных задач — обнаружения мошеннических операций. Высокая производительность распределенной системы позволяет использовать более ресурсоемкие, но и более точные модели, которые невозможно было бы обучить на больших данных в рамках централизованной архитектуры. Для сравнения были выбраны три алгоритма: логистическая регрессия как базовый вариант, а также ансамблевые методы — случайный лес и градиентный бустинг, известные своей высокой предсказательной способностью.

Качество моделей оценивалось по набору стандартных метрик для задач бинарной классификации. Выбор метрик Precision (точность), Recall (полнота) и F1-Score обусловлен необходимостью баланса между минимизацией ложноположительных срабатываний (когда легитимная транзакция блокируется как мошенническая, что ведет к недовольству клиентов) и ложноотрицательных (когда мошенническая транзакция пропускается, что ведет к прямым финансовым потерям) [Будунов, Кузьмин, 2018]. Метрика AUC-ROC является интегральным показателем качества классификатора, не зависящим от выбора порога срабатывания, и особенно важна для несбалансированных выборок, какими являются данные о мошенничестве (табл. 2).

Таблица 2 - Эффективность алгоритмов обнаружения мошеннических операций на распределенной системе

Алгоритм	Precision	Recall	F1-Score	AUC-ROC
Логистическая регрессия	0.824	0.751	0.786	0.891
Случайный лес (Random Forest)	0.946	0.893	0.919	0.975
Градиентный бустинг (Gradient Boosting)	0.958	0.921	0.939	0.988

Данные таблицы 2 показывают значительное превосходство ансамблевых методов над базовой логистической регрессией. Градиентный бустинг демонстрирует наилучшие результаты по всем ключевым метрикам, достигая значения AUC-ROC в 0.988, что является исключительно высоким показателем для реальных данных. По сравнению с логистической регрессией, градиентный бустинг повышает точность (Precision) на 16.3%, а полноту (Recall) — на 22.6%. Это означает, что модель не только точнее идентифицирует мошеннические операции, но и пропускает значительно меньше таких случаев. Высокий показатель F1-Score (0.939) свидетельствует о достижении оптимального баланса между этими двумя метриками. Такое

повышение качества классификации имеет прямой и измеримый экономический эффект. Увеличение Recall на 17 процентных пунктов (с 0.751 до 0.921) при общем объеме мошенничества в несколько миллионов долларов в год может привести к экономии сотен тысяч или даже миллионов долларов [Зимина, 2008]. Возможность обучения и применения столь сложных моделей на полном наборе данных является прямым следствием использования масштабируемой распределенной платформы.

Помимо задач безопасности, анализ транзакционных данных открывает широкие возможности для оптимизации маркетинговых стратегий через глубокую сегментацию клиентской базы. Используя алгоритм кластеризации k-means, удалось разделить всю совокупность клиентов на несколько четко выраженных сегментов на основе их покупательского поведения. В качестве признаков для кластеризации были использованы средний чек, частота покупок, предпочитаемые товарные категории и время совершения транзакций. Результаты позволили выявить неочевидные паттерны и сформировать профили ключевых клиентских групп, что является основой для разработки персонализированных маркетинговых кампаний.

Выбор показателей для характеристики кластеров, таких как средняя стоимость транзакции и их частота, позволяет бизнесу понять ценность каждого сегмента и соответствующим образом распределить маркетинговый бюджет. Доминирующая категория продуктов указывает на интересы сегмента и позволяет формировать релевантные предложения. Анализ этих данных помогает перейти от массового маркетинга к прецизионному, нацеленному на конкретные потребности каждой группы клиентов (табл. 3).

Таблица 3 - Результаты кластеризации клиентской базы для оптимизации маркетинговых стратегий

ID кластера	Доля клиентов (%)	Средняя стоимость транзакции ($)	Средняя частота транзакций (в месяц)	Доминирующая категория продуктов
1	8.2	845.62	1.15	Электроника и бытовая техника
2	45.7	42.18	15.80	Продукты питания и товары повседневного спроса
3	12.1	215.33	3.45	Одежда, обувь и аксессуары
4	28.5	88.75	5.20	Развлечения и услуги (билеты, подписки)
5	5.5	35.91	1.85	Случайные и редкие покупки

Анализ таблицы 3 позволяет сделать важные выводы для бизнеса. Кластер 1, несмотря на небольшую долю (8.2%), представляет собой сегмент «высокоценных» клиентов, совершающих редкие, но очень дорогие покупки. Маркетинговые усилия для этой группы должны быть сфокусированы на программах лояльности и эксклюзивных предложениях. Кластер 2 — это ядро клиентской базы (45.7%), обеспечивающее стабильный, хотя и невысокий в пересчете на одну транзакцию, денежный поток за счет высокой частоты покупок. Для них эффективны будут скидочные акции и накопительные системы. Кластер 3 представляет собой типичных «шопоголиков» в сфере моды, а кластер 4 — активных потребителей цифровых услуг. Понимание этих различий позволяет оптимизировать маркетинговые коммуникации, значительно повышая их конверсию и рентабельность инвестиций в рекламу (ROI). Например, вместо общей рассылки можно отправить целевое предложение о новой коллекции одежды клиентам из кластера 3, а предложение о скидке на подписку — клиентам из кластера 4.

Итоговым этапом исследования стала комплексная экономическая оценка эффекта от внедрения распределенной аналитической системы по сравнению с поддержанием и эксплуатацией традиционной централизованной архитектуры. Оценка проводилась на основе прогнозируемых показателей за один год и учитывала как прямые выгоды (сокращение потерь, увеличение выручки), так и затраты (инвестиции в инфраструктуру, лицензии, персонал).

Данный анализ позволяет перевести технические преимущества, продемонстрированные ранее, на язык бизнеса и оценить целесообразность инвестиций. В расчет были включены доходы от улучшения таргетинга маркетинговых кампаний, основанные на данных кластеризации, и прямая экономия от снижения уровня мошенничества благодаря более точным моделям. Операционные затраты включали амортизацию оборудования, стоимость программного обеспечения и фонды оплаты труда IT-специалистов.

Комплексный анализ данных, представленных во всех таблицах, позволяет сформировать целостную картину трансформационного воздействия распределенных вычислений на финансовый анализ в корпорациях. Техническое превосходство, выраженное в сокращении времени обработки данных в 160 раз на объеме в 100 млн транзакций (табл. 1), является не самоцелью, а фундаментальным фактором, открывающим возможность для качественных изменений в аналитике. Именно эта скорость и масштабируемость позволяют применять передовые алгоритмы машинного обучения, такие как градиентный бустинг, которые в свою очередь обеспечивают рост показателя AUC-ROC с 0.891 до 0.988 (табл. 2). Этот, на первый взгляд, абстрактный прирост точности на 10.9% напрямую транслируется в реальную экономию в размере 7.57 млн долларов в год за счет более эффективного предотвращения мошенничества.

Аналогичная цепочка прослеживается и в области маркетинга. Способность быстро обрабатывать весь массив клиентских данных позволила провести детальную кластеризацию и выявить пять различных сегментов с уникальными поведенческими характеристиками (табл. 3). Это знание, в свою очередь, стало основой для разработки персонализированных стратегий, которые, по прогнозам, принесут дополнительную выручку в размере 9.25 млн долларов (табл. 4). Таким образом, исследование демонстрирует синергетический эффект, при котором технологическая платформа создает условия для применения продвинутых аналитических методов, а те, в свою очередь, генерируют измеримый финансовый результат. Итоговый чистый экономический эффект в размере 13.57 млн долларов убедительно доказывает, что, несмотря на более высокие первоначальные затраты, инвестиции в распределенные аналитические системы многократно окупаются и являются стратегически оправданными для любой компании, оперирующей большими объемами транзакционных данных.

Заключение

Проведенное исследование убедительно демонстрирует, что применение алгоритмов распределенных вычислений является ключевым фактором для совершения качественного скачка в области анализа финансовых транзакций и оптимизации процессов принятия решений в современном корпоративном секторе. Эмпирические данные, полученные в ходе вычислительных экспериментов, подтвердили гипотезу о неадекватности традиционных централизованных систем для обработки больших данных. Было установлено, что при росте объема анализируемых транзакций до 100 миллионов записей распределенная архитектура на базе Apache Spark превосходит монолитную систему по скорости обработки более чем в 160 раз, демонстрируя при этом практически линейную масштабируемость и избегая сбоев, характерных для централизованных решений при пиковых нагрузках.

Технологическое преимущество распределенных систем создает фундамент для внедрения сложных аналитических инструментов, ранее недоступных из-за вычислительных ограничений. Использование ансамблевых моделей машинного обучения, в частности градиентного бустинга, на полном объеме данных позволило повысить точность обнаружения мошеннических операций до уровня AUC-ROC 0.988. Это представляет собой существенное улучшение по сравнению с базовыми моделями и напрямую транслируется в многомиллионную экономию за счет снижения финансовых потерь. Аналогичным образом, применение алгоритмов кластеризации для глубокой сегментации клиентской базы позволило выявить дискретные поведенческие группы, что является основой для разработки высокоэффективных, персонализированных маркетинговых стратегий и, как следствие, значительного роста выручки.

Итоговая экономическая оценка показала, что, несмотря на более высокие капитальные и операционные затраты, внедрение распределенной аналитической платформы генерирует чистый годовой экономический эффект в размере 13.57 млн долларов. Это доказывает, что инвестиции в данные технологии являются не просто статьей расходов на IT, а стратегическим вложением в повышение конкурентоспособности, операционной эффективности и финансовой устойчивости компании. Переход от пакетной обработки данных с многочасовыми задержками к анализу в режиме, близком к реальному времени, фундаментально меняет парадигму управления, позволяя бизнесу принимать решения не на основе устаревших отчетов, а на базе актуальной, полной и глубоко проанализированной информации.

Перспективы применения полученных результатов выходят далеко за рамки рассмотренных задач. Разработанные подходы могут быть экстраполированы на другие области корпоративного управления, такие как оптимизация логистических цепочек, управление рисками, прогнозирование спроса и динамическое ценообразование. Дальнейшее развитие этих технологий, в частности, интеграция с системами потоковой обработки данных (stream processing) и применение более сложных нейросетевых архитектур, откроет возможности для создания полностью автоматизированных, самообучающихся систем поддержки принятия решений. Таким образом, освоение технологий распределенного анализа данных становится не просто желательным, а обязательным условием для выживания и процветания корпораций в условиях цифровой трансформации экономики.

Библиография

1. Анисин А.А. Оптимизация соотношения собственных и заемных источников финансирования корпораций : автореферат диссертации на соискание ученой степени кандидата экономических наук / Анисин А.А. ; Институт экономики и организации промышленного производства Сибирского отделения Российской академии наук. Омск, 2011. 24 с.
2. Будунов К.А., Кузьмин А.К. Проблема автоматизированного принятия решений при организации распределенных вычислений // Проблемы управления в социально-экономических и технических системах. Сборник научных статей. 2018. С. 16-18.
3. Васильев А.В. Моделирование и оптимизация корпоративной информационной системы предприятия на основе распределенных систем вычисления : автореферат диссертации на соискание ученой степени кандидата технических наук / Васильев А.В. ; Воронежский государственный технический университет. Воронеж, 2005. 24 с.
4. Вейс Ю.В., Ладошкин А.И. Методы оптимизации решений в корпоративном управлении. Самара, 2010. 100 с.
5. Волкова Н.В., Львович И.Я. Оптимизация принятия управленческих решений с использованием корпоративного интеллектуального капитала : монография. Воронеж, 2011. 128 с.
6. Евтушенко Ю.Г. Методы и алгоритмы решения задач оптимизации большой размерности, задач глобальной и многокритериальной оптимизации, равновесного программирования, оптимального управления и их

реализация на высокопроизводительных вычислительных системах : НИР : грант № НШ-4096.2010.1 / Совет по грантам Президента Российской Федерации. 2010.
7. Евтушенко Ю.Г. Методы и алгоритмы решения задач оптимизации большой размерности, задач глобальной и многокритериальной оптимизации, равновесного программирования, оптимального управления и их реализация на высокопроизводительных вычислительных системах : НИР : грант № НШ-5264.2012.1 / Совет по грантам Президента Российской Федерации. 2012.
8. Евтушенко Ю.Г. Методы и алгоритмы решения задач оптимизации большой размерности, задач глобальной и многокритериальной оптимизации, равновесного программирования, оптимального управления и их реализации на высокопроизводительных вычислительных системах : НИР : грант № НШ-8860.2016.1 / Совет по грантам Президента Российской Федерации. 2016.
9. Ефремов В.А. Модели и алгоритмы поддержки принятия решений при управлении инвестициями с использованием структурированных финансовых продуктов : автореферат диссертации на соискание ученой степени кандидата технических наук / Ефремов В.А. ; Сибирский государственный индустриальный университет. Новокузнецк, 2012. 22 с.
10. Зимина Г.А. Интеллектуальные алгоритмы поддержки принятия решений по управлению инвестиционными процессами // Системный анализ в проектировании и управлении. Сборник научных трудов XII Международной научно-практической конференции. 2008. С. 37-42.
11. Кашенков А.Р. О некоторых подходах к решению задач оптимизации корпоративных бизнесов // Леденцовские чтения. Бизнес. Наука. Образование. Материалы II международной научно-практической конференции. 2011. С. 81-84.
12. Косматый Д.Ю. Математические модели и алгоритмы принятия оптимальных решений в банковской деятельности : автореферат диссертации на соискание ученой степени кандидата физико-математических наук / Косматый Д.Ю. Киев, 1996. 16 с.
13. Кухаренко Е.Л. Принципы построения и программное обеспечение корпоративных информационных систем на основе технологий распределенных вычислений : автореферат диссертации на соискание ученой степени кандидата технических наук / Кухаренко Е.Л. ; Ин-т динамики систем и теории управления СО РАН. Иркутск, 2000. 20 с.
14. Львович Я.Е., Рындин Н.А. Формализованное описание процесса оптимального построения корпоративных информационных систем // Вестник Воронежского государственного технического университета. 2003. № 3-3. С. 10-13.
15. Мишин Д.В. Модели и алгоритмы административного управления корпоративной распределенной информационно-вычислительной средой АСУ : автореферат диссертации на соискание ученой степени кандидата технических наук / Мишин Д.В. ; Владимирский государственный университет им. Александра Григорьевича и Николая Григорьевича Столетовых. Владимир, 2013. 24 с.

Application of Distributed Computing Algorithms in Scalable Analysis of Financial Transactions: Aspects of Optimizing Decision-Making Processes in the Corporate Sector

Kirill Yu. Zhigalov

PhD in Technical Sciences, Senior Research Fellow,
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences,
117997, 65 Profsoyuznaya str., Moscow, Russian Federation;
e-mail: kshskalov@mail.ru

Abstract

The research addresses the problem of processing rapidly growing volumes of financial transactions and demonstrates that the transition from centralized architectures to distributed algorithms can radically increase analytics speed and the quality of management decisions. The aim of the work is to systematically evaluate the effect of implementing Apache Spark and related methods on performance, model accuracy, and economic outcomes of corporate practices. The

empirical base includes an anonymized dataset of 250 million transactions from 2020-2022. A 32-node cluster and a monolithic comparative configuration were deployed. The pipeline covers cleaning, normalization, training of logistic regression, random forest and gradient boosting models, as well as clustering. Results were obtained confirming the hypothesis of scalable superiority: with a volume of 100 million records, the distributed architecture accelerates task execution by more than 160 times, providing almost linear scalability. For anti-fraud tasks, ensemble models outperform the baseline: gradient boosting achieves AUC-ROC 0.988 with Precision 0.958 and Recall 0.921.

For citation

Zhigalov K.Yu. (2025) Primeneniye algoritmov raspredelennykh vychisleniy v masshtabiruyemom analize finansovykh transaktsiy: aspekty optimizatsii protsessov prinyatiya resheniy v korporativnom sektore [Application of Distributed Computing Algorithms in Scalable Analysis of Financial Transactions: Aspects of Optimizing Decision-Making Processes in the Corporate Sector]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 15 (8A), pp. 382-392. DOI: 10.34670/AR.2025.14.40.041

Keywords

Distributed computing, financial transactions, big data analysis, Apache Spark, machine learning, fraud detection, process optimization.

References

1. Anisin A.A. Optimization of the Ratio of Internal and Borrowed Sources of Corporate Financing: Abstract of the Dissertation for the Degree of Candidate of Economic Sciences / Anisin A.A.; Institute of Economics and Industrial Production Organization of the Siberian Branch of the Russian Academy of Sciences. Omsk, 2011. 24 p.
2. Budunov K.A., Kuzmin A.K. The Problem of Automated Decision Making in the Organization of Distributed Computing // Problems of Management in Socio-Economic and Technical Systems. Collection of Scientific Papers. 2018. P. 16–18.
3. Vasiliev A.V. Modeling and Optimization of the Corporate Information System of an Enterprise Based on Distributed Computing Systems: Abstract of the Dissertation for the Degree of Candidate of Technical Sciences / Vasiliev A.V.; Voronezh State Technical University. Voronezh, 2005. 24 p.
4. Weis Yu.V., Ladoshkin A.I. Methods for Optimization of Decisions in Corporate Management. Samara, 2010. 100 p.
5. Volkova N.V., L'vovich I.Ya. Optimization of Managerial Decision-Making Using Corporate Intellectual Capital: Monograph. Voronezh, 2011. 128 p.
6. Evtushenko Yu.G. Methods and Algorithms for Solving Large-Dimension Optimization Problems, Problems of Global and Multicriteria Optimization, Equilibrium Programming, Optimal Control, and Their Implementation on High-Performance Computing Systems: Research Project: Grant No. NSh-4096.2010.1 / Council for Grants of the President of the Russian Federation. 2010.
7. Evtushenko Yu.G. Methods and Algorithms for Solving Large-Dimension Optimization Problems, Problems of Global and Multicriteria Optimization, Equilibrium Programming, Optimal Control, and Their Implementation on High-Performance Computing Systems: Research Project: Grant No. NSh-5264.2012.1 / Council for Grants of the President of the Russian Federation. 2012.
8. Evtushenko Yu.G. Methods and Algorithms for Solving Large-Dimension Optimization Problems, Problems of Global and Multicriteria Optimization, Equilibrium Programming, Optimal Control, and Their Implementation on High-Performance Computing Systems: Research Project: Grant No. NSh-8860.2016.1 / Council for Grants of the President of the Russian Federation. 2016.
9. Efremov V.A. Models and Algorithms for Decision Support in Investment Management Using Structured Financial Products: Abstract of the Dissertation for the Degree of Candidate of Technical Sciences / Efremov V.A.; Siberian State Industrial University. Novokuznetsk, 2012. 22 p.
10. Zimina G.A. Intelligent Algorithms for Decision Support in Investment Process Management // Systems Analysis in Design and Management. Collection of Scientific Works of the 12th International Scientific-Practical Conference. 2008. P. 37–42.

11. Kashenkov A.R. On Some Approaches to Solving Corporate Business Optimization Problems // Ledentsov Readings. Business. Science. Education. Materials of the 2nd International Scientific-Practical Conference. 2011. P. 81–84.
12. Kosmaty D.Yu. Mathematical Models and Algorithms for Making Optimal Decisions in Banking: Abstract of the Dissertation for the Degree of Candidate of Physical and Mathematical Sciences / Kosmaty D.Yu. Kiev, 1996. 16 p.
13. Kukharenko E.L. Principles of Construction and Software of Corporate Information Systems Based on Distributed Computing Technologies: Abstract of the Dissertation for the Degree of Candidate of Technical Sciences / Kukharenko E.L.; Institute of System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences. Irkutsk, 2000. 20 p.
14. L'vovich Ya.E., Ryndin N.A. Formalized Description of the Process of Optimal Construction of Corporate Information Systems // Bulletin of Voronezh State Technical University. 2003. No. 3-3. P. 10–13.
15. Mishin D.V. Models and Algorithms of Administrative Management of a Corporate Distributed Information and Computing Environment of an Automated Control System: Abstract of the Dissertation for the Degree of Candidate of Technical Sciences / Mishin D.V.; Vladimir State University named after Alexander and Nikolay Stoletov. Vladimir, 2013. 24 p.