DOI: 10.24412/2413-2527-2023-234-27-33

Интегрированное и распределенное хранение и обработка данных с учетом кластеризации

А. А. Брызгалов, Е. С. Кутыева, Е. А. Петрова, д.т.н. А. Д. Хомоненко
Петербургский государственный университет путей сообщения Императора Александра I
Санкт-Петербург, Россия
shyrik777888@gmail.com, res19.01@gmail.com, katya26021984@rambler.ru, khomon@mail.ru

Аннотация. **Повсеместная интеграция информационных технологий ставит ряд сложных задач, связанных с хранением, обработкой и передачей данных. Одним из направлений развития здесь является разработка и усовершенствование систем интегрированной и распределенной обработки данных. Рассматриваются особенности и перспективы развития систем интегрированной и распределенной обработки данных. Приводится эталонная архитектура безопасности, которую предлагается рассматривать как базовую архитектуру масштабируемой системы хранения и обработки данных для возможного совершенствования характеристик ее производительности и/или надежности на основе кластеризации ее сетевой структуры. Обоснован выбор модели и метода, используемых для выявления влиятельных узлов сетевой структуры.**

Ключевые слова: **кластеризация, масштабируемая архитектура, обработка данных, интегрированная обработка, распределенная обработка.**

Введение

В условиях современного научно-технологического прогресса особую роль и актуальность имеет сегмент информационных технологий. Интенсивно повышающийся объем используемых данных в цифровом виде актуализирует разработку инновационных и модернизацию существующих технологий хранения и обработки информации. В результате развития этого сегмента улучшается качество и повышается эффективность производственных процессов, промышленных предприятий и технологических комплексов [1].

В современном мире непрерывно происходят усовершенствования в профессиональных и бытовых сферах жизнедеятельности человека. Движущей силой этих процессов являются инновационные разработки и открытия в рамках технологического прогресса на основе использования больших данных, которые, в свою очередь, собираются в процессе интегрированной и распределенной обработки данных сетевых структур. При этом особую актуальность получает вопрос, связанный с повышением информационного наполнения сетевых и телекоммуникационных структур. Наряду с этим отметим, что эффективность работы в этом случае напрямую связана с качественным наполнением больших данных [2].

Особенности интегрированной и распределенной обработки данных

Принцип современных технологий обработки данных состоит в том, что перед интеграцией данных на серверах происходит их накопление на вычислительных устройствах. Подобными устройствами могут являться рабочие станции или персональные компьютеры. Далее происходит распределение данных по вычислительным устройствам на основе запросов, поступающих на сервер. На основе указанных процессов происходит реализация интегрированной и распределенной обработки информации, характерной для сетевых структур.

При этом в сетевых структурах устройства формируются в ***кластеры***, которые, в свою очередь, группируются в пространства локальных, корпоративных и иных вычислительных сетей. Именно эти процессы способствует накоплению больших данных в сетевых структурах.

Одной из основных особенностей интегрированной и распределенной обработки данных является необходимость уметь делать нетривиальные выводы при решении практических задач. Эта особенность реализуется на основе интеллектуального анализа данных и когнитивного анализа данных.

Интеллектуальный анализ данных (ИАД) представляет собой процесс обнаружения пригодных к использованию сведений в больших наборах данных. В ИАД применяется математический анализ для выявления закономерностей и тенденций, существующих в данных. Обычно такие закономерности нельзя обнаружить при традиционном просмотре данных, поскольку связи слишком сложны из-за чрезмерного объема данных.

Когнитивный анализ данных (КАД) означает решение конкретной практической задачи пользователем, сопровождающееся познавательным процессом, в котором присутствует анализ данных.

Названные методы активно используются и взаимно дополняют друг друга в современных сетевых структурах, основанных на интегрированной и распределенной обработке данных. Это привело к тому, что термины ИАД и КАД используются в качестве синонимичных определений в ряде современных научных исследований.

Другой особенностью интегрированной и распределенной обработки данных является необходимость решения задачи, связанной с учетом ***гетерогенности*** (разнородности) данных. Решение этой задачи зависит от каждого конкретного случая и приложения, в основном осуществляется с помощью разнообразных метрик Дейка, Хэмминга, Евклида и множества других [3–5].

В частности, наиболее распространенной мерой для определения расстояния между двумя точками на плоскости, образованной координатными осями x и y, является метрика Евклида — расстояние между двумя точками евклидова пространства. Квадрат евклидова расстояния:

$$d(X,Y) = \sum_{i=1}^{n} (X_i - Y_i)^2.$$

При возведении в квадрат лучше учитываются большие разности.

Каждая из указанных технологий обработки данных имеет свои достоинства и недостатки в зависимости от области использования и индивидуальных особенностей объекта. Для более детального определения различия и степени родства интегрированного и распределенного способа обработки данных представим каждый из по отдельности.

Ключевые варианты построения распределенных систем обработки данных

Каждая из ЭВМ, находящихся в распределенных системах обработки, является специальной, предназначенной для решения определенных задач. Таким образом, распределенный метод обработки данных базируется на градации нескольких функций обработки между определенным числом электронно-вычислительных машин или комплексом процессоров, которые объединены в единую сеть.

Возможные ***варианты реализации распределенной обработки*** данных:

1. Посредством установки одной или нескольких вычислительных машин (ЭВМ) ***на каждом уровне*** (узле) системы. В этом случае обработка данных осуществляется одной или несколькими вычислительными машинами в зависимости от реальных потребностей системы и ее возможностей, а система баз данных в общем смысле состоит из узлов, каждый из которых является системой управления базой данных (СУБД), а узлы взаимодействуют между собой таким образом, что база данных любого узла доступна пользователю так, как если бы она являлась по отношению к нему локальной.

2. Путем организации работы множества процессоров ***внутри одной системы***. Такой вариант, как правило, применяется в системах обработки банковской и финансовой информации для создания сети обработки данных (отделения, филиалы и т. д.).

Основными ***преимуществами распределенного способа обработки*** данных являются: высокий уровень надежности, наличие возможности обработки любого объема данных в заранее назначенные сроки, сокращение времени и ресурсов, направляемых для манипуляции с данными. Кроме того, отметим высокую гибкость и улучшение эксплуатационных характеристик систем при использовании этого метода обработки данных.

Относительно организации данных в распределенном алгоритме классифицируют два основных вида параллельного поведения распределенной системы: параллелизм в пространстве и параллелизм во времени. Одной из основных характеристик таких систем является ускорение R решения задачи параллельной системы относительно последовательной однопроцессорной системы (базовый закон Амдала):

$$R = \frac{T_1}{T_n}, \qquad (1)$$

где T_1 — время решения задачи на однопроцессорной системе;

T_n — время решения той же задачи на n-процессорной системе.

Выразим указанные значения через основные параметры алгоритмов. Пусть W — общее число операций в алгоритме решения задачи. Множество всех операций можно разделить на два подмножества: последовательно и параллельно выполняющихся операций. Пусть $W_{\text{посл}}$ — число операций в первом подмножестве, $W_{\text{пар}}$ — число операций во втором подмножестве. Очевидно, что $W = W_{\text{посл}} + W_{\text{пар}}$.

Закон Амдала не учитывает специфику и типы операций, с помощью которых описывается алгоритм, но использует среднее время t выполнения одной операции как последовательной, так и параллельной. Благодаря этому соотношение (1) можно переписать в виде

$$R = \frac{W_t}{\left(W_{\text{посл}} + \frac{W_{\text{пар}}}{n}\right) \times t} = \frac{1}{a + \frac{1-a}{n}},$$

где $a = W_{\text{посл}}/W$ — доля последовательных операций в общем числе операций алгоритма.

При числе процессоров n, стремящемся к бесконечности, ускорение R стремится к величине $1/a$.

Интегрированные системы обработки данных для информационных структур

Этот способ обработки данных состоит в организации информационной модели управляемого объекта, другими словами, создании распределенной базы данных. Основной особенностью его является то, что интегрированные технологии в распределенных системах подразумевают коллективное пользование и централизованное управление базами данных. При этом наличие большого объема данных и разнообразие задач требуют разграничить базы данных [6].

Несмотря на это, интегрированная обработка данных выполняется посредством единого информационного массива. На основе этого в разы повышается качество, а также достоверность и скорость обработки информации. Также одной из особенностей интегрированной обработки данных является отделение процедуры обработки данных от процедур их сбора. На основе этого способа обеспечивается максимальный уровень удобства работы пользователя.

Особую актуальность использования этот метод обработки данных получил при создании интегрированных автоматизированных систем управления (ИАСУ). Важным этапом при разработке таких систем является определение эффективности и качества ИАСУ. Расчет экономической эффективности предполагает, что каждая компонента ИАСУ имеет свою эффективность, а эффективность системы обусловлена степенью интеграции компонентов. С учетом данных особенностей эффективность ИАСУ можно определить следующим образом:

$$Э = К_С \times \sum Э_i ,$$

где $Э_i$ — экономическая эффективность компоненты ИАСУ;

$К_С = f(k_j, a_j)$ — системный коэффициент;

k_j — частные значения показателей интеграции по одному из показателей интеграции всех компонентов ИАСУ;

a_j — коэффициент предпочтительности показателя k_j.

Рассматривая перспективы развития интегрированной обработки данных, отметим, что одним из основных направлений развития и интеграции этой технологии явля-

ются ERP-системы (англ. *Enterprise Resource Planning* — планирование ресурсов предприятия). К примеру, одним из перспективных направлений является поставка в комплекте с этим продуктом инструментариев и ускорителей, предназначенных для реализации аналитических средств под нужды каждого отдельного заказчика.

Другой перспективой из этого же направления является создание ***готового бизнес-контента и дополнительных приложений*** на основе интегрированной аналитики. Примерами их являются карты сбалансированных показателей, CRM (англ. *Customer Relationship Management* — управление взаимоотношениями с клиентами), инструменты визуализации данных, планирования и многих других, значительно повышающих эффективность ведения бизнеса.

Автоматизированная система обработки данных в ОАО «РЖД»

В качестве примера интегрированной обработки данных рассмотрим автоматизированную систему, используемую в ОАО «РЖД». Единая автоматизированная система актово-претензионной работы (ЕАСАПР М) предназначена для автоматизации актово-розыскной работы и сопутствующей работы, выполняемой линейными работниками станций [7].

Объектами внедрения ЕАСАПР М являются: грузовые станции (рабочие места приемосдатчиков груза и багажа, сотрудников актово-розыскных групп станций, приемосдатчиков пунктов коммерческого осмотра поездов и вагонов, приемосдатчиков коммерческих постов безопасности, операторов станционных технологических центров (СТЦ), дежурных по железнодорожной станции (ДСП), приемосдатчиков контейнерных площадок, работающих в структурных подразделениях Центра фирменного транспортного обслуживания (ЦФТО) и в «ТрансКонтейнере»); отделы коммерческой работы АФТО; службы грузовой и коммерческой работы ТЦФТО; отделы таможенной и брокерской деятельности; Управление коммерческой работы ЦФТО; Управление по таможенно-брокерской деятельности ОАО «РЖД». Система ЕАСАПР М состоит из комплекса подсистем.

Система ЕАСАПР М реализована как одноуровневая система, серверный комплекс которой функционирует на сетевых серверах приложений и сетевых серверах баз данных, находящихся в Московском ИВЦ — структурном подразделении ГВЦ — филиала ОАО «РЖД».

В качестве сервера баз данных используется Microsoft SQL Server 2008. Система ЕАСАПР М взаимодействует с различными системами ОАО «РЖД». Блок-схема программно-технического комплекса (ПТК) системы и информационных потоков приведена на рисунке 1.

Что касается распределенной обработки информации, то специалисты некоторых фирм ведут работы над переосмыслением термина «база данных». В этом процессе специалисты сталкиваются со множеством ключевых проблем, решение которых и определяет будущее баз данных и систем распределенной обработки [8].

Исходя из этого, одним из ключевых направлений развития распределенной обработки данных является усовершенствование распределенных баз данных. При обеспечении достаточной эффективности и качества работы таких систем ожидается их интеграция практически во всех профессиональных областях жизнедеятельности человека, использующих электронные формы представления и обработки информации. Наиболее важной и перспективной задачей из данного направления является обеспечение автоматизированной и высоконадежной работы с хорошо сбалансированной распределенной инфраструктурой [9].

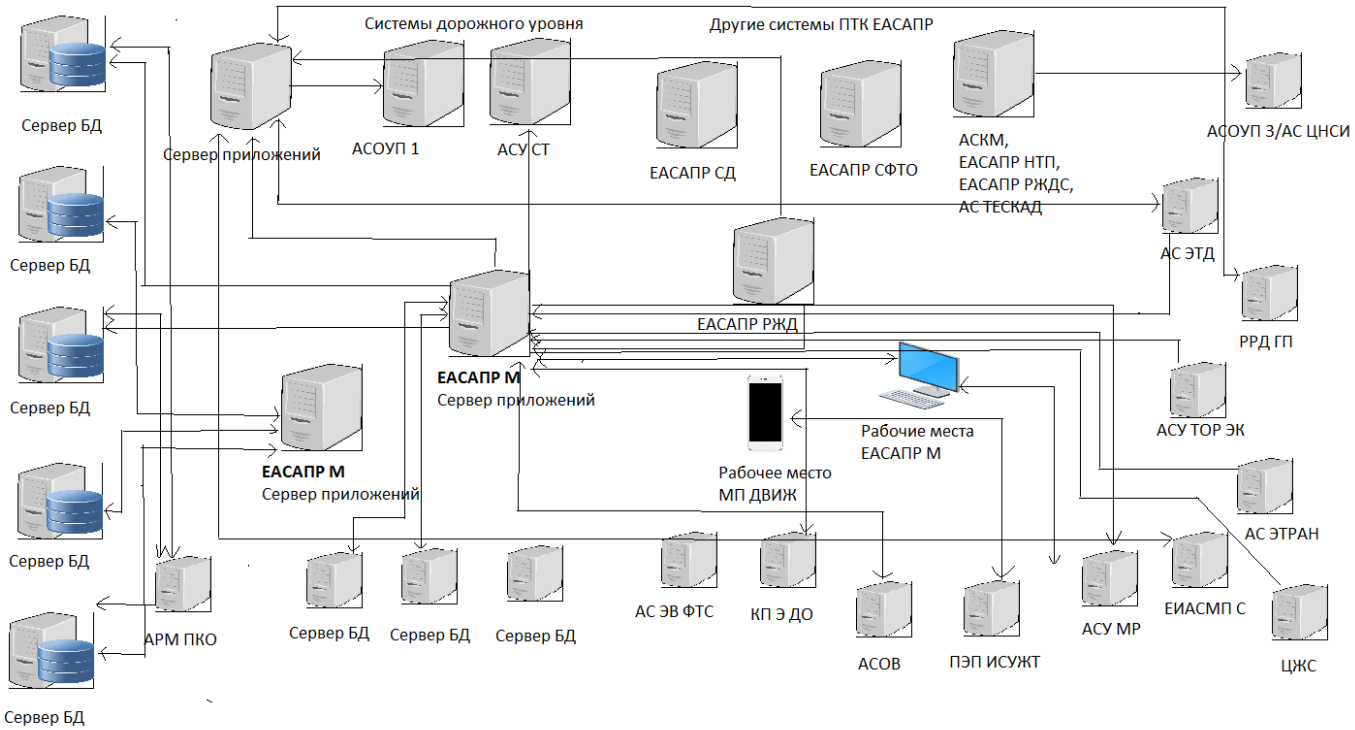

Рис. 1. Блок-схема ПТК с информационными потоками

Управление данными в информационных системах с масштабируемой архитектурой

Для хранения данных в современных информационных системах широкое распространение получили хранилища данных. Современные тренды указывают на необходимость переосмысления способов управления данными и их интеграции. В частности, требуется перейти от объединения всех данных в одном хранилище к подходу, позволяющему легко и безопасно распределять, собирать и использовать данные.

Решением проблем эффективного хранения и обработки разрозненных данных в настоящее время становится ***масштабируемая*** архитектура: типовая и предметно-ориентированная архитектура с набором схем, проектов, принципов, моделей и наилучших практик, она упрощает и интегрирует управление данными в организации на основе распределения. В частности, такая архитектура во многом улучшает управление данными, предлагая единый механизм управления безопасностью, основными данными, метаданными и моделирования данных; она позволяет работать с несколькими облачными провайдерами и локальными платформами и дает необходимые контроль и гибкость.

Масштабируемая архитектура предоставляет не зависящие от предметной области и повторно используемые архитектурные блоки, одновременно обеспечивает гибкость, сочетая различные способы доставки данных с разнообразными технологиями. На рисунке 2 приведена эталонная архитектура безопасности масштабируемой архитектуры системы хранения и обработки данных [10].

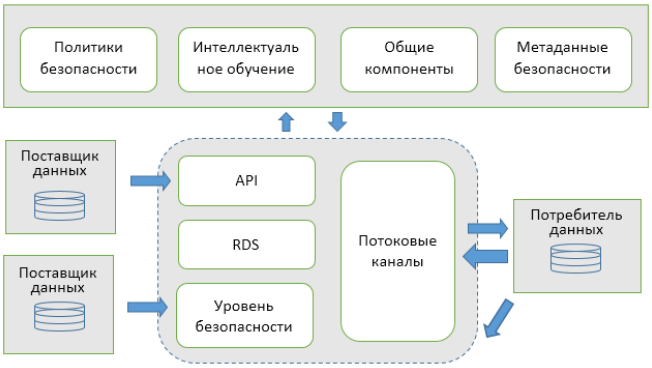

Рис. 2. Эталонная архитектура безопасности для масштабируемых систем

В архитектуре данных ***только для чтения*** (Read-only DataStores, RDS) автоматически создается ориентированное на потребителя «представление», которое он может использовать. Определенные данные могут быть скрыты, замаскированы или зашифрованы. Для такой функциональности RDS могут работать вместе с различными инструментами защиты данных, например: Apache Ranger, Apache Sentry и др.

Архитектура API применяется для подключения сервисов и распределения небольших объемов данных для использования в реальном времени и с малой задержкой. В отличие от архитектуры RDS, она упрощает операции записи, обновления и удаления.

Потоковая архитектура ориентирована на потоковую передачу больших объемов событий и сообщений в реальном времени. Потоковая передача асинхронна, отличается от API высокой пропускной способностью, может использоваться для копирования состояния приложения.

Отметим, что приведенная на рисунке 2 эталонная архитектура безопасности может рассматриваться как базовая архитектура масштабируемой системы хранения и обработки данных для дальнейшего совершенствования характеристик ее производительности и/или надежности на основе кластеризации ее сетевой структуры.

О кластеризации сетевых структур и выявлении влиятельных узлов

При обработке данных в сетевых структурах крайне важным параметром является выделение конкретных сетевых узлов и их кластеров, которые несут большую часть нагрузки при работе с большими данными. Имея информацию о сетевых узлах и кластерах, представляется возможной организация профилактических мероприятий с целью исключения потенциальной возможности отказа вычислительных устройств узлов, а также повышения защищенности данных.

Множество методов кластеризации классифицируют на плоские и иерархические (строящие систему вложенных разбиений на непересекающиеся кластеры), а также на четкие и нечеткие [11].

Четкие методы кластеризации разбивают исходное множество объектов X на несколько непересекающихся подмножеств. При этом любой объект из X принадлежит только одному кластеру. Нечеткие методы кластеризации позволяют одному и тому же объекту принадлежать одновременно нескольким (или даже всем) кластерам, но с различной степенью уверенности экспертов.

Четкие методы являются наиболее изученными. Одним из классических методов кластеризации является метод COBWEB концептуальной кластеризации объектов, предложенный в работах [12, 13]. По сути он означает инкрементную систему для иерархической концептуальной кластеризации объектов. В дальнейшем предложено множество различных вариантов методов кластеризации данных в информационных системах.

В частности, в статье [14] предложен метод оптимальной энтропийной кластеризации высокоразмерных данных в информационных системах, основанный на энтропийном подходе к выбору состояния элементов сообщений — на классическом принципе «малая величина энтропии соответствует большому количеству информационного содержания» и позволяющий добиться сокращения признакового пространства.

В качестве примера нечеткой кластеризации отметим предложенный в статье [15] метод, который позволяет работать с объектами, характеризуемыми нечеткими параметрами и строить модель концептуальной кластеризации для объектов нечеткой природы.

В работах [16–18] предложены математические модели, алгоритмы по выявлению влиятельных по информативности узлов информационно-вычислительных сетей. Используется метод идентификации узлов KDEC, требующий знания положения узлов, расстояния между ними, метод идентификации влияния узла на основе анализа иерархий, требующий учета мультиатрибутности данных сложных сетей. Для этого предлагается метод измерений локальной центральности, основанный на топологической

структуре и характеристиках взаимодействия узлов и окружения. При этом не учитываются особенности данных по активности сетевых устройств во времени.

В статье [19] для устранения перечисленных выше ограничений предложена математическая модель, учитывающая особенности данных по вычислительной активности устройств сетевого узла при расчете его информативной нагрузки за расчетное время и формирование наиболее информативных кластеров узлов в вычислительных сетях.

С учетом изложенного, для выявления влиятельных узлов информационно-вычислительных сетей целесообразно применение метода и модели, предложенных в работе [19]. В качестве базовой модели информационно-вычислительных сетей нами в рассматриваемой ситуации предлагается использовать эталонную архитектуру безопасности для масштабируемых систем [10].

Заключение

Рассмотренные методы обработки информации представляются актуальными. Несмотря на достаточно высокий уровень развития названных технологий, постоянно возникающие практические задачи требуют непрерывного усовершенствования и повышения эффективности каждой из технологий обработки информации. Приведенную эталонную архитектуру безопасности информационных систем, на наш взгляд, целесообразно рассматривать как базовую архитектуру масштабируемой системы хранения и обработки данных для дальнейшего совершенствования характеристик на основе кластеризации ее сетевой структуры. В частности, при решении задачи определения влиятельных и важных сетевых узлов по обработке данных.

Литература

1. Казьмина, И. В. Специфика совершенствования производственных процессов при проведении реинжиниринга на основе современных информационных технологий // Экономинфо. 2016. № 25. С. 52–56.

2. Портнов, М. С. Потенциал применения современных информационных технологий в бизнес-аналитике / М. С. Портнов, А. В. Речнов, В. П. Филиппов // Вестник Российского университета кооперации. 2020. № 2 (40). С. 87–92.

3. Foss, A. H. Distance Metrics and Clustering Methods for Mixed-Type Data / A. H. Foss, M. Markatou, B. Ray // International Statistical Review. 2019. Vol. 87, Is. 1. Pp. 80–109. DOI: 10.1111/insr.12274.

4. Финогеев, А. А. Распределенная обработка данных в беспроводных сенсорных сетях на основе мультиагентного подхода и туманных вычислений / А. А. Финогеев, А. Г. Финогеев, И. С. Нефедова // Надежность и качество: Труды Международного симпозиума: в 2 т. (Пенза, Россия, 23–31 мая 2016 г.) / под общ. ред. Н. К. Юркова. — Пенза: Пензенский гос. ун-т, 2016. — Т. 1. — С. 258–260.

5. Дударев, В. А. Анализ методов интеграции для разработки информационно-аналитических систем по свойствам неорганических соединений / В. А. Дударев, И. О. Темкин, В. Ф. Корнюшко // Программные продукты и системы. 2020. Т. 33, № 2. С. 283–296.
DOI: 10.15827/0236-235X.130.283-296.

6. Болодурина, И. П. Подходы к идентификации сетевых потоков и организации маршрутов трафика в виртуальном центре обработки данных на базе нейронной сети / И. П. Болодурина, Д. И. Парфенов // Программные продукты и системы. 2018. Т. 31, № 3. С. 507–513.
DOI: 10.15827/0236-235X.123.507-513.

7. Единая автоматизированная система актово-претензионной работы хозяйства коммерческой работы в сфере грузовых перевозок нового поколения (ЕАСАПР М НП) // РЖД Software. URL: http://software.rzd.ru/catalog/122 (дата обращения 31.05.2023).

8. Разработка параллельного модуля генерации защищенной картографической базы данных / Р. Ф. Гибадуллин, А. А. Новиков, Н. В. Хевронин, М. Ю. Перухин // Вестник технологического университета. 2016. Т. 19, № 10. С. 102–105.

9. Поленов, М. Ю. Модифицированная распределенная архитектура обработки данных для геоинформационных систем / М. Ю. Поленов, Д. А. Иванов // Известия Южного федерального университета. Технические науки. 2020. № 6 (216). С. 99–108. DOI: 10.18522/2311-3103-2020-6-99-108.

10. Стренгхольт, П. Масштабируемые данные. Лучшие шаблоны высоконагруженных архитектур = Data Management at Scale: Best Practices for Enterprise Architecture / пер. с англ. С. Черникова. — Санкт-Петербург: Питер, 2022. — 368 с. — (Бестселлеры O'Reilly).

11. Ершов, К. С. Анализ и классификация алгоритмов кластеризации / К. С. Ершов, Т. Н. Романова // Новые информационные технологии в автоматизированных системах: Материалы Девятнадцатого научно-практического семинара (Москва, Россия, 21 апреля 2016 г.) / под ред. В. А. Галактионова. — Москва: ИПМ им. М. В. Келдыша, 2016. — С. 274–279.

12. Michalski, R. S. Knowledge Acquisition Through Conceptual Clustering: A Theoretical Framework and an Algorithm for Partitioning Data into Conjunctive Concepts // International Journal of Policy Analysis and Information Systems. 1980. Vol. 4, No. 3. Pp. 219–244.

13. Fisher, D. H. Knowledge Acquisition Via Incremental Conceptual Clustering // Machine Learning. 1987. Vol. 2, Is. 2. Pp. 139–172. DOI: 10.1023/A:1022852608280.

14. Аскерова, Б. Г. Оптимальная энтропийная кластеризация в информационных системах // Программные продукты и системы. 2017. Т. 30, № 4. С. 643–646.
DOI: 10.15827/0236-235X.030.4.643-646.

15. Назаров, А. О. Модель и метод концептуальной кластеризации объектов, характеризуемых нечеткими параметрами // Фундаментальные исследования. 2014. № 9-5. С. 993–997.

16. Identifying Influential Nodes in Complex Networks Based on Local Effective Distance / J. Zhang, B. Wang, J. Sheng, [et al.] // Information. 2019. Vol. 10, Is. 10. Art. No. 0311. 15 p. DOI: 10.3390/info10100311.

17. Bian, T. Identifying Influential Nodes in Complex Networks Based on AHP / T. Bian, J. Hu, Y. Deng // Physica A: Statistical Mechanics and Its Applications. 2017. Vol. 479. Pp. 422–436. DOI: 10.1016/j.physa.2017.02.085.

18. Identifying Important Nodes in Complex Networks Based on Multiattribute Evaluation / H. Xui, J. Zhang, J. Yang, L. Lun // Mathematical Problems in Engineering. 2018. Art. No. 8268436. 11 p. DOI: 10.1155/2018/8268436.

19. Бочков, А. П. Модель формирования кластеров информативных узлов интегрированной и распределенной обработки данных в вычислительной сети / А. П. Бочков, А. Д. Хомоненко, А. М. Барановский // Наукоемкие технологии в космических исследованиях Земли. 2021. Т. 13, № 1. С. 44–57. DOI: 10.36724/2409-5419-2021-13-1-44-57.

DOI: 10.24412/2413-2527-2023-234-27-33

Integrated and Distributed Data Storage and Processing with Clustering Included

A. A. Bryzgalov, E. S. Kutyeva, E. A. Petrova, Grand PhD A. D. Khomonenko
Emperor Alexander I St. Petersburg State Transport University
Saint Petersburg, Russia
shyrik777888@gmail.com, res19.01@gmail.com, katya26021984@rambler.ru, khomon@mail.ru

Abstract. **The widespread integration of information technology poses a number of complex tasks related to the storage, processing and transmission of data. One of the areas of development here is the development and improvement of integrated and distributed data processing systems. The features and prospects for the development of integrated and distributed data processing systems are considered. A reference security architecture is presented, which is proposed to be considered as the basic architecture of a scalable data storage and processing system for possible improvement of its performance and/or reliability characteristics based on the clustering of its network structure. The choice of the model and method used to identify the influential nodes of the network structure is substantiated.**

Keywords: **clustering, scalable architecture, data processing, integrated processing, distributed processing.**

REFERENCES

1. Kazmina I. V. The Specificity of Improving the Production Processes at Implementing the Reengineering Based on Contemporary Information Technologies [Spetsifika sovershenstvovaniya proizvodstvennykh protsessov pri provedenii reinzhiniringa na osnove sovremennykh informatsionnykh tekhnologiy], *Ekonominfo*, 2016, No. 25, Pp. 52–56.

2. Portnov M. S., Rechnov A. V., Filippov V. P. Potential of Application of Modern Information Technologies in Business Analytics [Potentsial primeneniya sovremennykh informatsionnykh tekhnologiy v biznes-analitike], *Vestnik of the Russian University of Cooperation [Vestnik Rossiyskogo universiteta kooperatsii]*, 2020, No. 2 (40), Pp. 87–92.

3. Foss A. H., Markatou M., Ray B. Distance Metrics and Clustering Methods for Mixed-Type Data, *International Statistical Review*, 2019, Vol. 87, Is. 1, Pp. 80–109.
DOI: 10.1111/insr.12274.

4. Finogeev A. A., Finogeev A. G., Nefedova I. S. Distributed Data Processing in Wireless Sensor Networks Based on Multi-Agent Approach and Fog Computing [Raspredelennaya obrabotka dannykh v besprovodnykh sensornykh setyakh na osnove multiagentnogo podkhoda i tumannykh vychisleniy], *Reliability and Quality: Proceedings of the International Symposium [Nadezhnost i kachestvo: Trudy Mezhdunarodnogo simpoziuma], Penza, Russia, May 23–31, 2016. Volume 1.* Penza, Penza State University, 2016, Pp. 258–260.

5. Dudarev V. A., Temkin I. O., Kornyushko V. F. Integration Methods Analysis for the Development of Information-Analytical Systems on Inorganic Substances Properties [Analiz metodov integratsii dlya razrabotki informatsionno-analiticheskikh sistem po svoystvam neorganicheskikh soedineniy], *Software and Systems [Programmnye produkty i sistemy]*, 2020, Vol. 33, No. 2, Pp. 283–296.
DOI: 10.15827/0236-235X.130.283-296.

6. Bolodurina I. P., Parfenov D. I. The Approaches to Identification of Network Flows and Organization of Traffic Routes in a Virtual Data Processing Center Based on a Neural Network [Podkhody k identifikatsii setevykh potokov i organizatsii marshrutov trafika v virtualnom tsentre obrabotki dannykh na baze neyronnoy seti], *Software and Systems [Programmnye produkty i sistemy]*, 2018, Vol. 31, No. 3, Pp. 507–513.
DOI: 10.15827/0236-235X.123.507-513.

7. Unified Automated System of Act and Claim Work of the Economy of Commercial Work in the Field of Cargo Transportation of a New Generation (EASAPR M NP) [Edinaya avtomatizirovannaya sistema aktovo-pretenzionnoy raboty khozyaystva kommercheskoy raboty v sfere gruzovykh perevozok novogo pokoleniya (EASAPR M NP)], *RZhD Software*. Available at: http://software.rzd.ru/catalog/122 (accessed 31 May 2023).

8. Gibadullin R. F., Novikov A. A., Hevronin N. V., Peruhin M. Yu. Development of a Parallel Module for Generating a Secure Cartographic Database [Razrabotka parallelnogo modulya generatsii zashchishchennoy kartograficheskoy bazy dannykh], *Herald of Technological University [Vestnik tekhnologicheskogo universiteta]*, 2016, Vol. 19, No 10, Pp. 102–105.

9. Polenov M. Yu., Ivanov D. A. Modified Distributed Data Processing Architecture for Geoinformation Systems [Modifitsirovannaya raspredelennaya arkhitektura obrabotki dannykh dlya geoinformatsionnykh sistem], *Izvestiya Southern Federal University. Engineering Sciences [Izvestiya Yuzhnogo federalnogo universiteta. Tekhnicheskie nauki]*, 2020, No. 6 (216), Pp. 99–108.
DOI: 10.18522/2311-3103-2020-6-99-108.

10. Strengholt P. Data Management at Scale: Best Practices for Enterprise Architecture [Masshtabiruemye dannye. Luchshie shablony vysokonagruzhennykh arkhitektur]. Saint Petersburg, Piter Publishing House, 2022, 368 p.

11. Ershov K. S., Romanova T. N. Analysis and Classification of Clustering Algorithms [Analiz i klassifikatsiya algoritmov klasterizatsii], *New Information Technologies in Automated Systems: Proceedings of the Nineteenth Scientific and Practical Seminar [Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh: Materialy Devyatnadtsatogo nauchno-prakticheskogo seminara], Moscow, Russia, April 21, 2016.* Moscow, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, 2016, Pp. 274–279.

12. Michalski R. S. Knowledge Acquisition Through Conceptual Clustering: A Theoretical Framework and an Algorithm for Partitioning Data into Conjunctive Concepts, *International Journal of Policy Analysis and Information Systems*, 1980, Vol. 4, No. 3, Pp. 219–244.

13. Fisher D. H. Knowledge Acquisition Via Incremental Conceptual Clustering, *Machine Learning*, 1987, Vol. 2, Is. 2, Pp. 139–172. DOI: 10.1023/A:1022852608280.

14. Askerova B. G. Optimum Entropy Clustering in Information Systems [Optimalnaya entropiynaya klasterizatsiya v informatsionnykh sistemakh], *Software and Systems [Programmnye produkty i sistemy]*, 2017, Vol. 30, No. 4, Pp. 643–646. DOI: 10.15827/0236-235X.030.4.643-646.

15. Nazarov A. O. Model and Method of Conceptual Clustering of Objects Characterized by Fuzzy Parameters [Model i metod kontseptualnoy klasterizatsii obektov, kharakterizuemykh nechetkimi parametrami], *Fundamental Research [Fundamentalnye issledovaniya]*, 2014, No. 9-5, Pp. 993–997.

16. Zhang J., Wang B., Sheng J., et al. Identifying Influential Nodes in Complex Networks Based on Local Effective Distance, *Information*, 2019, Vol. 10, Is. 10, Art. No. 0311, 15 p. DOI: 10.3390/info10100311.

17. Bian T., Hu J., Deng Y. Identifying Influential Nodes in Complex Networks Based on AHP, *Physica A: Statistical Mechanics and Its Applications*, 2017, Vol. 479, Pp. 422–436. DOI: 10.1016/j.physa.2017.02.085.

18. Hui X., Zhang J., Yang J., Lun L. Identifying Important Nodes in Complex Networks Based on Multiattribute Evaluation, *Mathematical Problems in Engineering*, 2018, Art. No. 8268436, 11 p. DOI: 10.1155/2018/8268436.

19. Bochkov A. P., Khomonenko A. D., Baranovsky A. M. Model of Formation of Clusters of Informative Nodes of Integrated and Distributed Data Processing in a Computer Network [Model formirovaniya klasterov informativnykh uzlov integrirovannoy i raspredelennoy obrabotki dannykh v vychislitelnoy seti], *High Technologies in Earth Space Research [Naukoemkie tekhnologii v kosmicheskikh issledovaniyakh Zemli]*, 2021, Vol. 13, No. 1, Pp. 44–57.
DOI: 10.36724/2409-5419-2021-13-1-44-57.