ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2020 Управление, вычислительная техника и информатика № 50

УДК 004.6
DOI: 10.17223/19988605/50/3

А.В. Воробьев, Г.Р. Воробьева

ПОДХОД К ПОВЫШЕНИЮ ПРОИЗВОДИТЕЛЬНОСТИ ПРОГРАММНЫХ ПРОЦЕССОВ ОБРАБОТКИ И ХРАНЕНИЯ БОЛЬШИХ ОБЪЕМОВ ГЕОМАГНИТНЫХ ДАННЫХ

Работа выполнена при поддержке гранта РФФИ № 20-07-00011-а.

Обсуждаются вопросы повышения вычислительной скорости процессов аналитической обработки больших объемов геомагнитных данных, являющихся результатом непрерывного наблюдения за параметрами геомагнитного поля распределенными магнитными станциями и обсерваториями. Предложена гибридная архитектура, сочетающая особенности реляционной, иерархической и колоночной моделей данных, использующая правила ссылочной целостности и POSIX-структуру адресации компонентов. Проводится анализ эффективности предложенного подхода на основе оценки вычислительных затрат на хранение и обработку геомагнитных данных.

Ключевые слова: геомагнитные данные; реактивность программного обеспечения; аналитическая обработка; большие данные.

Одним из основных источников знаний о характере и закономерностях пространственно-временного распределения параметров магнитного поля Земли и его вариаций являются геомагнитные данные, регистрируемые магнитными станциями и обсерваториями в режиме реального времени. При этом специализированное программное обеспечение для хранения и обработки геомагнитных данных к настоящему времени не разработано, а анализ данных выполняется отдельными исследователями посредством загрузки результатов наблюдений, хранящихся в репозиториях геомагнитных данных, техническое сопровождение которых осуществляется мировыми и региональными центрами геомагнитных данных [1. С. 390; 2. С. 2].

Общепринятым способом представления геомагнитных данных является формат IAGA2002, развиваемый Международной ассоциацией геомагнетизма и аэрономии [3. С. 5]. В структуре документа выделены: служебный заголовок, экспликация геомагнитных данных, значения параметров геомагнитного поля с соответствующими временными метками. Значения параметров и их временные метки заданы в ASCI-кодировке и разделены равным числом пробелов. Такое описание данных обеспечивает возможность использования формата для представления значений на длительном временном интервале – от нескольких секунд до многих месяцев.

Посуточное распределение результатов наблюдений параметров геомагнитного поля и его вариаций по отдельным файлам, низкоскоростные протоколы передачи данных, отсутствие веб-сервисов и API – далеко не полный перечень проблем, с которыми сталкивается разработчик программных средств для обработки геомагнитных данных формата IAGA2002. При этом наибольшую сложность с технической точки зрения представляет производительность программного продукта. Кроме того, локальное сохранение загруженных из репозиториев геомагнитных данных сопряжено с существенными затратами дискового пространства: например, годовой архив минутных значений результатов наблюдений параметров геомагнитного поля и его вариаций занимает в среднем объем в 40 МБ. На сегодняшний день в общей сложности доступны результаты более чем десятилетних наблюдений почти 300 магнитных станций и обсерваторий, что пропорционально увеличивает такие аппаратные затраты. Вместе с тем технические возможности научных организаций, занимающихся

исследованиями геомагнитного поля и его вариаций, зачастую ограничены, что не позволяет хранить подобные архивы наблюдений полностью и тем более выполнять их масштабную аналитическую обработку и визуализацию. Большие объемы геомагнитных данных и производительность программных средств их обработки напрямую связаны: к примеру, выполнение однопредикатного запроса к годовому архиву геомагнитных наблюдений одной магнитной обсерватории занимает в среднем 70 с при условии локального размещения обрабатываемых данных. Очевидно, что увеличение объемов обрабатываемых данных и сложности запросов к ним, а также использование, например, низкоскоростных протоколов для обращения к удаленным репозиториям в разы снизит производительность программного обеспечения.

Еще одна проблема связана с избыточностью формата IAGA2002. Обилие служебных символов, многократное повторение крайне редко изменяемых метаданных магнитных станций и обсерваторий в каждом суточном файле с результатами наблюдений приводит к тому, что объем полезной информации в IAGA2002-документе составляет менее 30% от его общего объема. При этом большинство разрабатываемых в научных организациях программных средств и систем зачастую ориентированы на использование устаревших технологий, не предназначенных для обработки данных такого большого объема.

Указанные проблемы приводят к необходимости совершенствования формата представления геомагнитных данных для обеспечения возможности создания высокопроизводительных программных средств их обработки и визуализации. Для решения поставленной задачи в настоящей работе предлагается новый гибридный формат долговременного хранения геомагнитных данных, представленный совокупностью трех взаимосвязанных компонент и отличающийся тем, что использует правила ссылочной целостности для объединения реляционной, иерархической и колончатой моделей данных, применяемых для описания метаданных и геомагнитных данных, а также реализует комбинацию текстового и бинарного форматов представления информации с целью повышения реактивности программных средств аналитической обработки геомагнитных данных, с одной стороны, и сокращения затрат требуемого объема физической памяти – с другой. Предлагаемый формат используется для представления данных в гибридном хранилище в составе предложенного авторами единого пространства геомагнитных данных [1. С. 395].

Результаты проведенных сравнительных экспериментов показали, что предложенный формат обеспечивает существенное повышение производительности вычислений, проводимых применительно к наборам разнородных геомагнитных данных, а также позволяет значительно сократить вычислительные затраты, связанные с их физическим хранением.

1. Структура описания метаданных

Служебный заголовок геомагнитных данных содержит признаковое описание магнитной обсерватории / станции, крайне редко изменяется и повторяется в каждом файле со значениями параметров геомагнитного поля, зарегистрированных обсерваторией / станцией. Очевидным шагом оптимизации формата представления геомагнитных данных является устранение избыточности служебного заголовка. Для этого предлагается отделить служебный заголовок и объединить метаданные всех магнитных станций и обсерваторий.

Метаданные магнитной станции / обсерватории, представленные множеством разноформатных объектов и их признаков, могут быть описаны посредством реляционной модели, заданной несколькими сущностями (рис. 1). Родительские сущности представляют собой обобщенные справочники параметров обсерватории, а каждый экземпляр дочерней описывает определенную станцию / обсерваторию посредством набора значений атрибутов. Сущности заданы в нормальной форме Бойса–Кодда и связаны друг с другом отношением типа «один-ко-многим».

Сущность «Observatory» предназначена для представления обобщенных данных о магнитной обсерватории / вариационной станции. Идентификатором каждого ее экземпляра выступает трех-

значный IAGA-код (поле «IAGAcode», текстовый формат, фиксированная размерность в 3 символа), который присваивается каждой станции / обсерватории, зарегистрированной в магнитной сети (независимо от ее принадлежности научной организации). Официальное название обсерватории, представленное в ее технической документации, задается в поле «Name» (текстовый формат, динамическая размерность). Для представления геодезических координат магнитной станции / обсерватории, таких как широта, долгота и высота над уровнем моря, использованы поля «Geodetic Longitude», «Geodetic Latitude», «Elevation» соответственно. Кроме того, в поле «Digital Sampling» задается значение скорости сбора данных с цифровых устройств или оцифровки аналогового сигнала в магнитной обсерватории (число одинарной точности). Также в поле «Data Interval Type» (текстовый формат, фиксированная размерность в 1 символ) предусмотрено хранение данных о временном интервале публикации геомагнитных данных (мгновенные регистрируемые значения или средние значения для интервалов от 1 с).

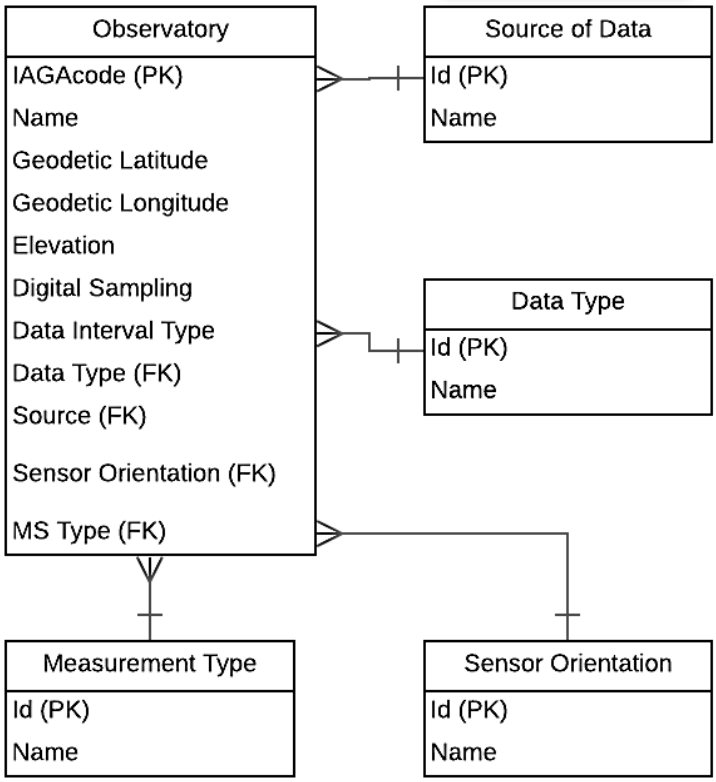

Рис. 1. Реляционная модель для описания метаданных
Fig. 1. Relational Model for Metadata Description

Остальные сущности модели являются независимыми и содержат справочную информацию, используемую при описании магнитных станций / обсерваторий. Так, экземпляры сущности «Measurement Type» (поле «Name», текстовый формат, фиксированная размерность в 4 символа) указывают на наименования регистрируемых станцией параметров геомагнитного поля (допустимые значения: DHIF, DHZF и XYZF). В сущности «Data Type» (поле «Name», текстовый формат, фиксированная размерность в 1 символ) указываются допустимые типы геомагнитных данных (временный (P), окончательный (D), квази-окончательный (Q) или вариационный (V)). Физическая ориентация приборов наблюдения задается в сущности «Sensor Orientation», а курирующая станцию / обсерваторию научная организация – в сущности «Source of Data».

Единый доступ к данным об обсерваториях / станциях позволяет оперативно сформировать набор метаданных по соответствующему IAGA-коду, при этом отсутствует физическое дублирование хранимых данных, присутствующее в применяемом в настоящее время формате представления геомагнитных данных. При этом выделение метаданных магнитных станций позволяет на 80% сокра-

тить затраты памяти, требуемой для физического хранения геомагнитных данных, зарегистрированных обсерваторией за год.

2. Структура описания каталогов данных

Результаты геомагнитных наблюдений физически размещены в иерархической системе директорий, в большинстве решений доступной по протоколу FTP. Структура директорий такова, что корневым элементом является суррогатный каталог с именем, например, магнитной сети, далее он декомпозируется на директории, соответствующие календарным годам наблюдений, каждая из которых делится на каталоги для хранения результатов измерений по месяцам. Такая иерархическая архитектура базируется на принципах построения POSIX-систем с использованием соответствующей адресации.

Древовидная файловая структура может быть описана посредством иерархии элементов формата разметки XML (Extensible Markup Language), где корнем является суррогатный элемент с именем станции / обсерватории, а дочерними по отношению к нему – одноуровневые элементы, соответствующие календарным годам наблюдений. При этом все наблюдения должны быть агрегированы в директорию, где каждой станции / обсерватории соответствует XML-файл с геомагнитными данными. В результате входными параметрами для получения данных являются код магнитной станции / обсерватории и искомый год регистрации наблюдений за параметрами геомагнитного поля и его вариаций. На программном уровне формирование запроса выполняется последовательным применением операций работы с файлами и XPath-запроса непосредственно в теле XML-документа. Централизованное размещение всех геомагнитных данных одной станции / обсерватории позволит существенно повысить производительность программных запросов к ним, поскольку считывание файла и обращение к нему осуществляются единожды, а все последующие действия выполняются со сформированным на его основе виртуальным объектом.

3. Структура описания геомагнитных данных

Постоянно растущий объем геомагнитных данных снижает целесообразность применения текстового формата их хранения в плане как затрат физической памяти, так и производительности выполняемых при этом вычислений. Так, обработка однопредикатного запроса к годовым геомагнитным данным в условиях применения персонального компьютера со средней производительностью (процессор с частотой 1,6 ГГц, 2 ядра, оперативная память 4 Гб, скорость интернет-соединения 342,7 Мбит/с) занимает около 8 с, что существенно превышает общепринятое (с точки зрения эргономики программного обеспечения) время отклика, составляющее 3 с. Отметим, что параметры сетевого соединения здесь имеют принципиальное значение, поскольку в соответствии с концепцией единого пространства геомагнитных данных [1. С. 398] результаты геомагнитных измерений хранятся на сервере, обращение к которому осуществляется по протоколу HTTP(s).

Предварительно целесообразно отметить ряд параметров, которые представляются избыточными с точки зрения необходимости их физического хранения. Прежде всего к ним относится порядковый номер дня в году – параметр, который может быть оперативно вычислен с помощью библиотечных функций на основании календарной даты. Физическое хранение даты и времени регистрации параметра геомагнитного поля в каждой строке суточного файла наблюдений неэффективно, но эта проблема решается применением формата XML в описании геомагнитных данных обсерватории (поэтому в качестве временной метки выбран не порядковый номер дня в году, а дата, что обеспечивает уникальность элемента в составе описания магнитной станции). Остальные параметры, заданные в структуре геомагнитных данных, представляют собой непосредственно результаты измерений, заданные в формате разделенной пробелами строки.

Особенность аналитической обработки геомагнитных данных связана с тем, что наибольшая вычислительная нагрузка приходится на большие выборки записей, зачастую с группированием и агрегированием. При этом количество операций записи не так велико, а добавление новых записей

обычно осуществляется крупными блоками. Небольшое количество столбцов, громоздкие и частые операции выборок, редкие и крупные обновления данных – признаки, указывающие на целесообразность организации хранения геомагнитных данных с помощью колоночных СУБД (имеется в виду именно модель данных, поскольку на физическом уровне колоночное представление обычно используется в архитектуре хранилищ данных). Такие СУБД обеспечивают высокую скорость и гибкость выполнения сложных запросов при сохранении преимуществ использования структурированного языка SQL, а также соответствуют обязательным требованиям ACID.

Колоночная организация хранения геомагнитных данных позволит существенно повысить производительность операций их обработки. Это связано в первую очередь с тем, что при построчной записи чтение с диска происходит более линейно. Более предсказуемое чтение файла при построчной записи позволяет операционной системе эффективнее использовать дисковый кэш.

На сегодняшний день широкое распространение получил колоночно-ориентированный формат представления данных Apache Parquet, отличительной особенностью которого является возможность программного управления механизмом сжатия данных в столбцах. Кроме того, Parquet реализован с использованием алгоритма измельчения и сборки записей, вмещающих сложные структуры данных, которые также можно использовать для их хранения.

Еще одним важным преимуществом Parquet является его бинарный формат, обеспечивающий хранение данных в том виде, в котором они представляются компьютеру в процессе работы программы. Поэтому при чтении файла не выполняются дополнительные преобразования, что существенно повышает скорость работы с данными, что и требуется для повышения производительности программной обработки геомагнитных данных. Колоночный Parquet на программном уровне позволяет не считывать все данные при выполнении запросов, извлекая только значения определенных столбцов, что также повышает производительность обработки данных. Сжатие по столбцам позволяет существенно сэкономить место при физическом хранении геомагнитных данных.

Набор геомагнитных данных в формате Parquet представлен двумя разделами. Первый из них является схемой документа и содержит описание структурных и параметрических ограничений представления данных: определяются состав столбцов, их наименования и последовательность, алгоритм сжатия и пр. Второй компонент представляет собой геомагнитные данные – значения параметров геомагнитного поля и его вариаций. Для хранения данных выделено 5 столбцов (колонок): один под временную метку, а оставшиеся – под три компонента и полный вектор геомагнитного поля соответственно. Для упрощения структуры в документе выделена только одна страница, а все столбцы образуют одну группу. Поэтому ко всем составляющим Parquet-документа применен один и тот же алгоритм сжатия (в нашем случае – gzip).

4. Интеграция компонент гибридного формата хранения геомагнитных данных

В общем виде хранение геомагнитных данных подразумевает смешанную логическую и физическую интеграцию предложенных выше компонент (рис. 2). Образуется иерархия структур данных, корневым элементом которой выступает реляционная структура с метаданными магнитных станций и обсерваторий. Результаты геомагнитных наблюдений физически размещаются в едином каталоге, в котором каждой обсерватории выделен XML-документ с именем, содержащим IAGA-код. В составе XML-документа каждый соответствующий году наблюдений элемент содержит блок CDATA, в котором размещается набор геомагнитных данных в бинарном формате Parquet. При этом анализатор запросов, предусмотренный в архитектуре единого пространства геомагнитных данных [1. С. 398], обеспечивает проверку ссылочной целостности как по заданному IAGA-коду, так и по указанным временным меткам.

Взаимодействие с хранилищем данных осуществляется строго в соответствии с иерархической структурой. По IAGA-коду из реляционной структуры выгружаются метаданные.

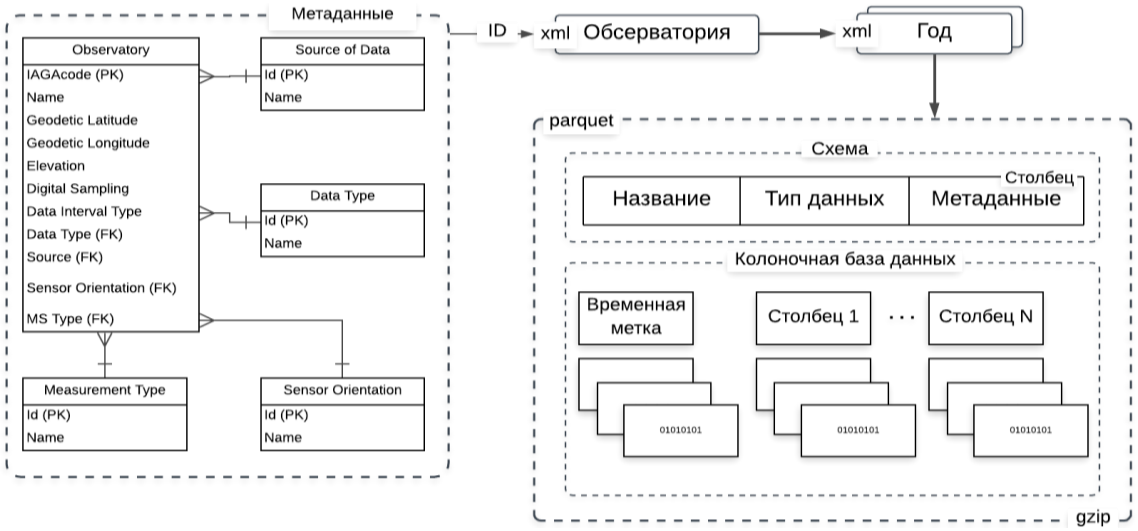

Рис. 2. Гибридная архитектура представления геомагнитных данных
Fig. 2. Hybrid architecture of geomagnetic data presentation

Далее тот же код используется для обращения к XML-файлу станции / обсерватории, а оттуда посредством XPath-запроса выбирается секция CDATA с искомыми геомагнитными данными. При необходимости выполняются фильтрация, группирование и агрегирование результатов наблюдений с использованием языка запросов SQL.

5. Экспериментальные исследования

Оценка эффективности предложенного гибридного формата хранения геомагнитных данных выполнена на основании сравнительного анализа распространенных форматов данных (рис. 3). По результатам исследования распространенных форматов и архитектур данных [4. С. 18] отобраны следующие: IAGA2002 (он же CSV) [3. С. 5]; реляционная база данных (RDB, relational database, на примере СУБД MS SQL Server 2017); XML [5. С. 1147]; JSON [6. С. 7]; AVRO [7. С. 267]; HDF5 [8; 9. С. 393]; neo4j [10. С. 232; 11. С. 11].

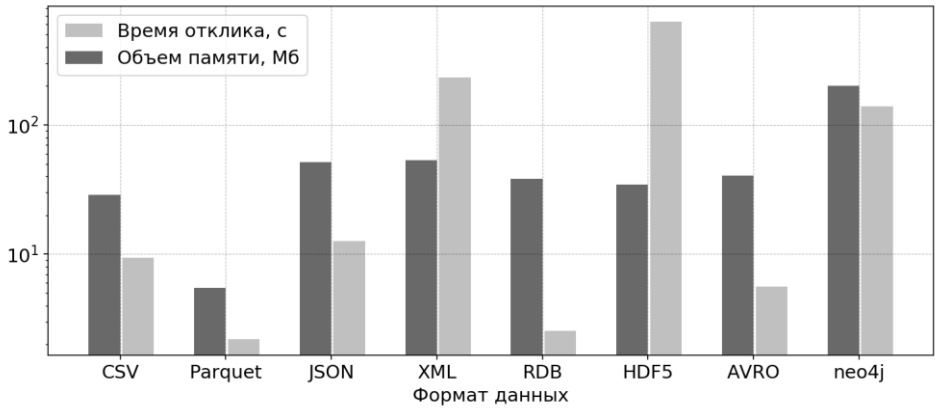

Рис. 3. Результаты сравнительного анализа форматов для хранения геомагнитных данных
Fig. 3. Results of comparative analysis of geomagnetic data formats

Критериями оценки эффективности гибридного формата хранения геомагнитных данных определены реактивность программной обработки данных и объем требуемого для их размещения дискового пространства. Выбор первого из критериев связан с тем, что существующие технологии аналитической обработки геомагнитных данных недостаточно эффективны в плане затрат вычислительных ресурсов на выполнение операций, а также времени на сбор и интеграцию данных на этапе их пред-

варительной обработки. Требуется оценить, насколько предлагаемый формат позволит повысить производительность выполнения операций обработки данных. Второй критерий оценки эффективности является в большей степени вспомогательным, поскольку в современных условиях развития технологий облачных хранилищ проблема занимаемого данными объема дискового пространства теряет свою остроту. Однако в большинстве случаев при анализе изменения значений параметров геомагнитного поля и его вариаций исследователи прибегают к аккумулированию всех необходимых данных на персональном компьютере, что требует больших объемов дискового пространства.

Исследование эффективности гибридного формата выполнено на примере получения выборки из годового архива минутных наблюдений станции с IAGA-кодом BOX за период 01–06.03.2018. Тем самым имеет место двухпредикатный запрос, выполнение которого предполагает обращение к суточному архиву геомагнитных данных по IAGA-коду станции (BOX), формирование набора данных, а выборку данных из соответствующих секций CDATA.

Экспериментальные исследования показали, что минимальное время отклика программного сценария обработки геомагнитных данных достигается при использовании для их хранения формата Parquet (2,2 с), что примерно в 4,3 раза меньше, чем для формата IAGA2002/CSV.

Согласно результатам исследований, применение предложенного формата для хранения геомагнитных данных позволяет минимизировать требования к объему дискового пространства. Так, по сравнению с форматом IAGA2002/CSV, для хранения годового архива геомагнитных наблюдений одной станции требуется примерно в 5,2 раза меньше объема дискового пространства.

Заключение

В результате проведенных исследований предложен гибридный формат хранения геомагнитных данных, который отличается тем, что использует правила ссылочной целостности для объединения реляционной, иерархической и колоночной моделей данных, применяемых для описания метаданных и геомагнитных данных, а также использует комбинацию текстового и бинарного форматов представления информации с целью повышения реактивности программных средств аналитической обработки геомагнитных данных, с одной стороны, и сокращения затрат требуемого объема физической памяти – с другой.

ЛИТЕРАТУРА

1. Воробьев А.В., Воробьева Г.Р., Юсупова Н.И. Концепция единого пространства геомагнитных данных // Тр. СПИИРАН. 2019. Т. 18, № 2. С. 390–415.
2. Geomagnetic Observations and Models / ed. by M. Mandea, M. Korte. Dordrecht : Springer, 2011. P. 149–181. (IAGA Special Sopron Book Series 5). https://link.springer.com/book/10.1007/978-90-481-9858-0 (accessed: 22.05.2019).
3. Intermagnet technical reference manual. Version 4.6 / ed. by Benoît St-Louis. Edinburgh, 2012. 92 p. https://www.intermagnet.org/publications/intermag_4-6.pdf (accessed: 22.05.2019).
4. Carrera D., Rosales J., Blanco G.A.T. Optimizing Binary Serialization with an Independent Data Definition Format // Int. J. of Computer Applications. 2018. V. 180, No. 28. P. 15–18.
5. Yahui Y. Impact data-exchange based on XML // Proc. 7th Int. Conf. Computer Science & Education (ICCSE). 2012. P. 1147–1149.
6. Peng D., Cao L., Xu W. Using JSON for Data bn exchanging in Web Service Applications // J. of Computational Information System. 2011. V. 7 (16). P. 5883–5890.
7. Plase D., Niedrite L., Taranovs R. Comparison of HDFS compact data formats: Avro Versus Parquet // Mokslas-Lietuvos ateitis. 2017. No. 9. P. 267–276.
8. HDF5. URL: https://www.hdfgroup.org/HDF5/ (accessed: 22.05.2019).
9. Emeakaroha V., Healy P. et al. Analysis of Data Interchange Formats for Interoperable and Efficient Data Communication in Clouds // Proc. of the 2013 IEEE/ACM 6th Int. Conf. on Utility and Cloud Computing. P. 393–398.
10. Femy P.F.M., Reshma K.R., Surekha S.M. Outcome analysis using Neo4j graph database // Int. J. on Cybernetics & Informatics (IJCI). 2016. V. 5, No. 2. P. 229–236.
11. Angles R., Gutierrez C. Survey of graph database models // ACM Computing Surveys. 2008. V. 40, No. 1. P. 1–39.

Поступила в редакцию 19 июля 2019 г.

Vorobev A.V., Vorobeva G.R. (2020) APPROACH TO IMPROVING THE PERFORMANCE OF SOFTWARE PROCESSES FOR PROCESSING AND STORING LARGE VOLUMES OF GEOMAGNETIC DATA. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie vychislitelnaja tehnika i informatika* [Tomsk State University Journal of Control and Computer Science]. 50. pp. 23–30

DOI: 10.17223/19988605/50/3

The issues of increasing the computational speed of software processes for the analytical processing of large volumes of geomagnetic data, which are the result of continuous monitoring of the parameters of the geomagnetic field by a great number of distributed ground magnetic stations and observatories, are discussed. A comparative review of the existing geomagnetic data architecture (presented in the framework of the specified IAGA-2002 format provided by International Association of Geomagnetism and Aeronomy), as well as popular data formats is given, and arguments are presented in favor of the need to improve the approach to organizing the results of geomagnetic observations.

To solve this problem, a new hybrid format for long-term storage of geomagnetic data is presented, represented by a set of three interrelated components and characterized in that it uses the rules of referential integrity to combine relational, hierarchical and columnar data models used to describe metadata and geomagnetic data, and also sets POSIX-component addressing structure and implements a combination of textual and binary formats for presenting information. The main purpose of the proposed architecture is to increase the reactivity of software tools for analytic processing of geomagnetic data, on the one hand, and reducing the cost of the required amount of physical memory, on the other hand.

The results of the comparison of the proposed hybrid format for presenting geomagnetic data with the existing approach to describing geomagnetic observation data (IAGA-2002), as well as other common formats for presenting large volumes of structured and semi-structured data (XML, JSON, Avro, etc.) are presented. In this case, the criteria for evaluating the effectiveness of a hybrid format for storing geomagnetic data determined the reactivity of software data processing and the amount of required disk space for their placement. The results of the experiment showed that the proposed format provides a significant increase in computing performance (about 4 times), conducted in relation to sets of heterogeneous geomagnetic data, and also significantly reduces the computational costs associated with their physical storage (approximately 5 times).

Keywords: geomagnetic data; software reactivity; analytical processing; big data.

VOROBEV Andrei Vladimirovich (Candidate of Technical Sciences, Associate Professor, Ufa State Aviation Technical University, Ufa, Russian Federation).
E-mail: geomagnet@list.ru

VOROBEVA Gulnara Ravilevna (Candidate of Technical Sciences, Associate Professor, Ufa State Aviation Technical University, Ufa, Russian Federation).
E-mail: gulnara.vorobeva@gmail.com

REFERENCES

1. Vorobev, A.V., Vorobeva, G.R. & Yusupova, N.I. (2019). Conception of geomagnetic data integrated space. Tr. SPIIRAN – *SPIIRAS Proceedings*. 18(2). pp. 390–415. DOI: 10.15622/sp.18.2.390-415
2. Mandea, M. & Korte, M. (eds) (2011) *Geomagnetic Observations and Models*. Dordrecht: Springer. pp. 149–181.
3. Trigg, D.F. & coles, R.L. (ed.) (2012) *Intermagnet Technical Reference Manual*. 4.6. Edinburgh: [s.n.]. [Online] Available from: https://www.intermagnet.org/publications/intermag_4-6.pdf (Accessed: 22nd May 2019).
4. Carrera, D., Rosales, J. & Blanco, G.A.T. (2018) Optimizing Binary Serialization with an Independent Data Definition Format. *International Journal of Computer Applications.* 180(28). pp. 15–18. DOI: 10.5120/ijca2018916670
5. Yahui, Y. (2012) Impact data-exchange based on XML. *Proc. 7 th Int. Conf. Computer Science & Education (ICCSE)*. pp. 1147–1149. DOI: 10.1109/ICCSE.2012.6295268
6. Peng, D., Cao, L. & Xu, W. (2011) Using JSON for Data bn exchanging in Web Service Applications. *Journal of Computational Information System*. 7(16). pp. 5883–5890.
7. Plase, D., Niedrite, L. & Taranovs R. (2017) Comparison of HDFS compact data formats: Avro Versus Parquet. *Mokslas – Lietuvos ateitis*. 9. pp. 267–276.
8. *HDF5*. (n.d.) [Online] Availablef rom: https://www.hdfgroup.org/HDF5/ (Accessed: 22nd May 2019).
9. Emeakaroha, V., Healy, P. et al. (2013) Analysis of Data Interchange Formats for Interoperable and Efficient Data Communication in Clouds. *Proc. of the 2013 IEEE/ACM 6th Int. Conf. on Utility and Cloud Computing*. pp. 393–398. DOI: 10.1109/UCC.2013.79
10. Femy, P.F.M., Reshma, K.R. & Surekha, S.M. (2016) Outcome analysis using Neo4j graph database. *International Journal on Cybernetics & Informatics (IJCI).* 5(2). 2016. pp. 229–236. DOI: 10.5121/ijci.2016.5225. 229
11. Angles, R. & Gutierrez, C. (2008) Survey of graph database models. *ACM Computing Surveys*. 40(1). pp. 1–39. DOI: 10.1145/1322432.1322433