

УДК 004.052.42:01

Аналіз платформ зберігання даних для створення інформаційного сховища медичних досліджень

DOI 10.36994/2788-5518-2021-01-01-11¹¹

Скулиш М.А., д.т.н.,с.н.с., Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського». mskulysh@gmail.com

Перебаєва К.С., студент, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського». kperebaieva@gmail.com

Анотація. У статті розглядається аналіз типів нереляційних баз даних для зберігання медичних досліджень. Розроблена інформаційна модель для зберігання медичних карт. Обрано платформу для зберігання медичних даних Mongo DB на основі документо-орієнтованої бази даних. Бази даних, орієнтовані на документи, забезпечують механізм виконання запитів до колекцій та отримання документів, які задовольняють вимоги до атрибутів. Документи групуються разом, утворюючи структуру, яка називається колекцією. Може бути кілька колекцій баз даних. Ця структура подібна до функціонування таблиці, яка присутня в реляційній базі даних. Сімейство нереляційних моделей даних поділяють на чотири види: ключ-значення, документо-орієнтовані, стовпчикові і графові бази даних. Проаналізувавши кожний вид, для медичних карт було обрано документо-орієнтовану базу даних. Запропоновано обрати документо-орієнтовану систему керування базами даних – MongoDB.

Ключові слова: база даних, медична карта, колекція, кореляційна модель, множина вимірювань, інформаційне сховище, концепція MongoDB

ANALYSIS OF THE STORAGE PLATFORM FOR CREATING AN INFORMATION REPOSITORY OF MEDICAL RESEARCH

Mariia Skulysh, Dr.habil. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. mskulysh@gmail.com

Kateryna Perebaieva. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine kperebaieva@gmail.com

Abstract. This article discusses the analysis of types of non-relational databases, namely these document-oriented databases, "key values" databases, columnar databases, graphic databases for storing medical research. To ensure an understanding of the structure of medical data storage, an information model for the storage of medical records was created. After a detailed analysis of the advantages and disadvantages of non-relational database types, this paper identifies the features of the document-oriented database type and created this model for further analysis of general optical information analysis for medical research. Due to the choice of a document-oriented type of database, a platform for storing MongoDB medical data based on a document-oriented database was

¹¹ Скулиш М.А., Перебаєва К.С.

chosen. MongoDB is a cross-platform open source document management (DBMS) system. The choice of the document-oriented DBMS is based on the fact that, unlike relational analogues, there are no restrictions on the set of fields in the documents. MongoDB meets all the requirements described in these article, as well as these documents - focused DBMS. MongoDB was chosen precisely because of the popularity, similarity of the approach to data storage with relational databases and BSON data storage format. To summarize the entire study, it should be noted the following research results: 1. Created and described an information model for the storage of medical research. 2. On the basis of the qualitative analysis of four types of non-relational models of databases, designed by a documentary-oriented model of databases to create an information analysis of medical research. 3. Based on the selected type of databases, it is proposed to choose the document-oriented database management system - MongoDB Big Data Platform for medical data storage.

Keywords: *dang database, medical card, collection, correlation model, set of measurements Information repository, MongoDB concept.*

АНАЛИЗ ПЛАТФОРМ ХРАНЕНИЯ ДАННЫХ ДЛЯ СОЗДАНИЯ ИНФОРМАЦИОННОГО ХРАНИЛИЩА МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Скулиш М.А., д.т.н., с.н.с. . Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского»
mskulysh@gmail.com

Перебаева Е.С., Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского»
kperebaieva@gmail.com

Аннотация. В статье рассматривается анализ типов корреляционных баз данных для хранения медицинских исследований. Разработана информационная модель для хранения медицинских карт. Избрано платформу для хранения медицинских данных Mongo DB на основе документо-ориентированной базы данных. Базы данных, ориентированные на документы, обеспечивают механизм выполнения запросов к коллекциям и получения документов, удовлетворяющих требованиям к атрибутам. Документы группируются вместе, образуя структуру, называемую коллекцией. Может быть несколько коллекций баз данных. Эта структура подобна функционированию таблицы, которая присутствует в реляционной базе данных. Семейство нереляционных моделей данных разделяют на четыре вида: ключ-значения, документо-ориентированные, колонки и графы базы данных. Проанализировав каждый вид, для медицинских карт была избрана документо-ориентированная база данных. Предложено выбрать документо-ориентированную систему управления базами данных - MongoDB.

Ключевые слова: база Данг, медицинская карта, коллекция, корреляционная модель, множество вимірювань Інформаційне сховище, концепція MongoDB:

Вступ

Інформаційні технології оточують нас всюди. Дані обробляються для отримання інформації, яка згодом може бути використана для різних цілей. Однак із зростанням соціальних медіа та розумних пристроїв, дані вже не є простим набором даних, з яким можуть впоратися традиційні інструменти та технології.

Сприяє оцифрування та зростання популярності сучасних технологій. Більше того, ці дані не просто великі за обсягом, але вони також включають дані різних видів, що генеруються періодично. Ми управляємо, ділимося та зберігаємо всі наші аспекти життя онлайн. У кожного з нас є акаунт в соціальних мережах. Крім того, ми ще робимо покупки онлайн, користуємося пошуком, завантажуюмо музику, тобто самостійно продукуємо гігабайти інформації.

Разом зі стрімким накопиченням інформації швидкими темпами розвиваються і технології аналізу даних. Якщо ще нещодавно було можливо лише сегментувати клієнтів на групи зі схожими перевагами, то тепер можливо будувати моделі для кожного клієнта в режимі реального часу, аналізуючи, як приклад, його переміщення по мережі Інтернет для пошуку конкретного товару.

Найбільшою проблемою у вирішенні цієї "великої проблеми даних" є те, що нинішні або традиційні системи не можуть зберігати та обробляти дані такого роду. Тому це породило необхідність масштабованих систем, які можуть зберігати різноманітні форми даних і обробляти їх для створення корисних аналітичних рішень.

Існуючі рішення

Зберігання

великих

даних

[1

<https://researchbank.swinburne.edu.au/file/af6e8e88-e9cc-45e7-ae5d-883151450be1/1/PDF%20%28Published%20version%29.pdf>] - загальний термін, що використовується для опису інфраструктур зберігання, призначених для зберігання, управління та пошуку даних, як правило, великих за обсягом, високих швидкостей і різноманітних за різноманітністю. В таких інфраструктурах, дані зберігаються таким чином, що їх використання, обробка та доступ стають простішими. Більше того, такі інфраструктури можуть масштабуватися відповідно до вимог програми або послуги.

Основним завданням зберігання великих даних є підтримка операцій введення та виведення збережених даних на додаток до зберігання великої кількості файлів та об'єктів. Зазвичай архітектури, що використовуються для зберігання великих даних, включають а кластер підключеного до мережі сховища, пули безпосередньо підключеного сховища або сховища на основі формату зберігання об'єктів [1].

Різні дослідницькі роботи класифікують доступні рішення NoSQL на основі підтримуваних моделей даних. Однак найбільш прийнята класифікація класифікує рішення NoSQL на основі з чотирьох моделей даних, а саме орієнтованих на документо-орієнтованих, графових, «ключ-значення» та стовпчикових.

Наукова стаття «Рішення для зберігання даних для систем великих даних: Якісне дослідження та Порівняння» (Storage Solutions for Big Data Systems: A Qualitative Study and Comparison):

- Забезпечує класифікацію рішень NoSQL на основі підтримуваної моделі даних, яка може бути документо-орієнтованою, графовою, «ключ-значення» або стовпчиковою.

- Проводить аналіз функцій 80 рішень NoSQL з урахуванням критеріїв вибору технологій для систем великих даних разом із сукупною оцінкою відповідних та невідповідних випадків використання для кожної моделі даних.

- Класифікує формати файлів великих даних за п'ятьма категоріями, а саме на основі тексту, рядків, стовпців, в пам'яті та послуги зберігання даних.

- Порівнює доступні формати файлів даних, аналізуючи переваги та недоліки та випадки використання кожного файлу даних формат.

- Оцінює проблеми, пов'язані із переходом зберігання великих даних наступного покоління в бік децентралізованого зберігання та технології блокчейн.

Обсяг цієї наукової роботи обмежений порівнянням та аналізом доступних рішень для зберігання для систем великих даних. Існує два аспекти інфраструктури зберігання в контексті великих даних. Перша грань стосується зберігання даних та управління за допомогою виділеної інфраструктури зберігання. З огляду на конкретні проблеми зберігання, поставлені Великими даними, такі як масштабованість, доступність, інтеграція та безпека, традиційні системи вважаються нездатними обробляти існуючий сценарій даних. Автори статті доходять до висновку, що NoSQL є життєздатним рішенням проблеми великих даних. [2 Storage Solutions for Big Data Systems: A Qualitative Study and Comparison]

Більшість існуючих досліджень пропонують або порівняльне дослідження між реляційними базами даних та NoSQL базами даних та/або як перетворити реляційні дані у дані NoSQL або як трансформувати концептуальну модель у конкретну базу даних NoSQL; в той час як дуже мало досліджень пропонують критерії вибору фізичної моделі та імплантаційні платформи.

В деяких дослідженнях застосовується модельована архітектура (MDA) для перетворення класу діаграми в БД NoSQL. [3 Abdelhedi, F., Brahim, A.A., Atigui, F., Zurfluh, G.: MDA-based Approach for NoSQL Databases Modelling. In: DAWAK'17. pp. 88–102. Springer (2017)] перетворює схему класів у базу даних NoSQL. Автори представляють загальну логічну модель, яка описує чотири родини NoSQL баз даних. Потім ця логічна модель трансформується у фізичні моделі, що стосуються чотирьох сімей.

У роботі [3 Daniel, G., Gomez, A., Cabot, J.: UmltoNoSQL: Mapping conceptual schemas to heterogeneous datastores. In: 2019 13th International Conference on Research Challenges in Information Science (RCIS). pp. 1–13 (2019)] автори пропонують UMLtoNoSQL, заснований на MDA структурі для відображення концептуальних моделей у декількох рішеннях для зберігання даних. Їх підхід поєднує в собі моделювання методів відображення та набір правил для перекладу OCL обмеження на різні мови запитів.

У статті [4 ModelDrivenGuide: An Approach for Implementing NoSQL Schemas] запропонована ModelDrivenGuide, підхід MDA до моделювання та реалізації Information Systems (IS), який направляє користувача від концептуальної моделі до технічної можливості платформи. Основою підходу є загальна логічна метамодель, що описує 5 сімейств моделей. Представлено QVT правила уточнення, особливо для злиття рядків концептів. Ці правила створюють оптимізовану схему і створення потенційних цільових моделей. Це поєднання даних моделювання та оптимізації перетворюються в підхід, спрямований на пошук ефективної цільової моделі серед 5 сімейств моделей даних. Евристика скорочує пошук простору і націлена на відповідні моделі, оскільки вони покладаються на варіант використання IS.

Опис інформаційної моделі

Результат кожного обстеження – це заповнений медичним працівником протокол дослідження. Всі дані можна розділити на три категорії:

- множина вимірювань;
- інформація про візит до лікаря;
- множина мультимедійних матеріалів (графіки, схеми, різні зображення).

Множина вимірювань та інформація про візит до лікаря призначені для опису досліджуваного об'єкта - анатомії. Однак наша сутність анатомії має ширше значення, ніж анатомія в біологічному сенсі. Вона включає в себе частини тіла (голова, шия і т.п.), органи (печінка, серце і т.п.), тканини, продукти життєдіяльності організму (сеча, кал), кістки, рідини, що є діяльністю будь-яких органів (грудне молоко, секрет передміхурової залози), різні новоутворення, слизові оболонки, мікроорганізми та інше. З одного боку, всі об'єкти входять в різні системи - функціональний підхід. З іншого боку, можна розглянути дані в структурному зрізі. Саме так виникає ієрархія анатомії – досліджуваний об'єкт.

Інформація про візит до лікаря – це набір текстових фраз, що включають у себе скарги пацієнта, стан досліджуваного об'єкта, діагноз, рекомендації до лікування. Найголовніше, що інформація про візит дозволяє висловити свою думку лікаря для надання більш повної картини іншим колегам і пацієнту.

Множина вимірів, наприклад, лабораторні дослідження, кожне з яких можуть мати текстове значення та числове (діапазони значень), оскільки багато параметрів (наприклад, розміри органів, частота серцевих скорочень, кількість наявних речовин) залежить від таких даних як вік, стать, ріст, вага, фаза циклу, вагітність, куріння, час взяття аналізу. Для числового значення використовується референтний інтервал. Референтний інтервал – це межі норми.

Множина мультимедійних матеріалів займає значний обсяг в медичних інформаційних системах та є важливим джерелом даних при діагностиці, оцінці і плануванні терапії. Наприклад, комп'ютерна томографія (КТ), магнітно-резонансна томографія (МРТ), рентген, молекулярна візуалізація,

ультразвукове дослідження, Фотоакустична томографія, рентгеноскопія, позитронно-емісійна томографія з комп'ютерною томографією.

Інформаційна модель



Рис.1. Інформаційна модель

Документо-орієнтована БД

Сімейство нереляційних моделей даних поділяють на чотири види: ключ-значення, документо-орієнтовані, стовпчикові і графові бази даних. Проаналізувавши кожний вид, для медичних карт було обрано документо-орієнтовану базу даних.

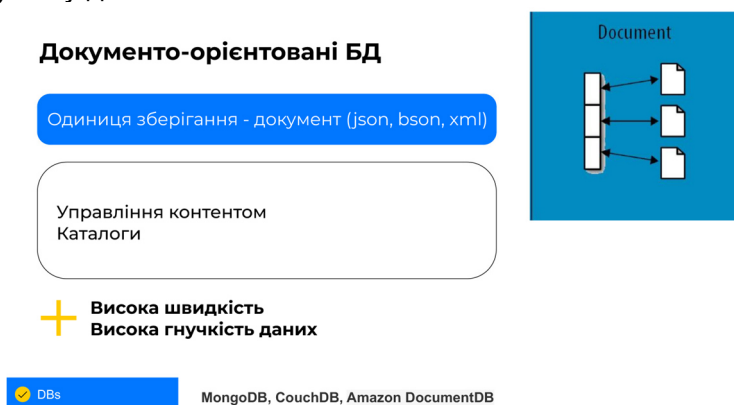


Рис. 2. Структура документно-орієнтованої бази даних

База даних NoSQL, яка використовує документи для зберігання та пошуку даних, називається документо-орієнтованою базою даних[5 Soransso, R. A. S. N., & Cavalcanti, M. C. (2018, April). Data modeling for analytical queries on document-oriented DBMS. In Proceedings of the 33rd Annual ACM Symposium on

Applied Computing (pp. 541-548). ACM.]. Інші назви цієї моделі даних NoSQL - це база даних документів [6 Sun, J., & Milani, C. V. B. (2018). U.S. Patent Application No. 15/406,643] та сховище документів [7 Geissinger, S. (2018). U.S. Patent Application No. 15/354,921.]. Вона насамперед використовується для управління напівструктурованими даними через гнучкість та підтримку змінної схеми. Документо-орієнтовані бази даних дуже схожі на ключ-значення у деяких із областей їх застосування. Але у них одиницею є документ.

Кожен документ може зберігати в собі, як правило, XML, JSON або BSON - бінарному-зберігається JSON. Але зараз практично завжди JSON або BSON.

У будь-який документ, в будь-який JSON ви можете записати абсолютно будь-який набір даних. І вони дуже часто застосовуються - наприклад, коли потрібно зробити який-небудь каталог і коли кожен продукт в каталозі може мати різні характеристики.

Наприклад, профілі користувачів. Хтось вказав свій улюблений фільм, хтось - улюблену їжу. Щоб не засовувати все в одне поле, яке буде зберігати незрозуміло що, ми можемо все записати в JSON документо-орієнтованих бази даних.

Документи групуються разом, утворюючи структуру, яка називається колекцією [5]. Може бути кілька колекцій баз даних. Ця структура подібна до функціонування таблиці, яка присутня в реляційній базі даних.

Бази даних, орієнтовані на документи, забезпечують механізм виконання запитів до колекцій та отримання документів, які задовольняють вимоги до атрибутів. Існує кілька переваг використання цього підходу, які включають:

1. Більша частина зростаючих даних надходить з пристроїв IoT, соціальних медіа та Інтернету. Орієнтовані на документи бази даних пропонують гнучке моделювання даних [5], на відміну від реляційних баз даних, які змушують програми вбудовувати дані в існуючі моделі незалежно від їх потреби.

2. Ефективність записування баз даних, орієнтованих на документи, краща за звичайні системи. Щоб зробити систему доступною для написання, база даних може також порушити консистенцію даних. Отже, навіть якщо система виходить з ладу і реплікація займає більше часу, ніж очікувалося, операція запису буде швидкою.

3. Відомо, що функції індексації та механізми запитів баз даних, доступні в цій категорії, є швидкими та ефективними [5]. Тому вони пропонують більш швидку продуктивність запитів.

Необхідні завдання властивості БД і відповідно її СУБД приведені у вигляді операцій, які СУБД повинна вміти робити над даними, що зберігаються в БД:

- Медична картка може мати різну кількість полів, оскільки кожна людина матиме різний набір метаданих.
- Має бути можливість додавання нових "для користувача" полів в медичну картку. Тобто має бути присутня можливість простого розширення набору метаданих випробуваного.

У якості СУБД вирішено вибрати MongoDB. MongoDB - це крос-платформова документо-орієнтована система управління базами даних (СУБД) з відкритим кодом.

Вибір саме документо-орієнтованої СУБД заснований на тому, що на відміну від реляційних аналогів тут не накладаються обмеження на набір полів у документа. Так як документо-орієнтовані бази даних ще не так поширені як реляційні, але їхня популярність із кожним роком все збільшується, доцільно буде роз'яснити деякі нюанси термінології документо-орієнтованих баз даних.

У короткому підручнику з MongoDB *The Little MongoDB Book* від автора Karl Seguin [8 <https://www.pvsm.ru/download/mongodb-ru.pdf>] наведені шість основних концепцій MongoDB:

1. MongoDB - концептуально те ж саме, що звичайна, звична нам база даних (або в термінології Oracle - схема). У середині MongoDB може бути нуль або більше баз даних, кожна з яких є контейнером для інших сутностей.

2. База даних може мати нуль або більше «колекцій». Колекція є аналогом таблиці з реляційних БД. Колекції складаються з нуля або більше «Документів». Документ є аналогом записи в таблиці (рядки).

Документ складається з одного або більше «полів», поля, в свою чергу, є аналогом "колонки".

3. «Індекси» в MongoDB майже ідентичні таким в реляційних базах даних.

4. «Курсор» відрізняються від попередніх п'яти концепцій, але вони дуже важливі

Важливо розуміти, що коли ми запитуємо у MongoDB будь-які дані, то вона повертає курсор, з яким ми можемо робити все що завгодно - підраховувати, пропускати певну кількість попередніх записів - при цьому не завантажуючи самі дані.

Основна ж відмінність документо-орієнтованих баз даних від реляційних є те, що в реляційних базах даних «колонки» визначаються на рівні «таблиці», тоді як документо-орієнтованих бази даних визначають «поля» на рівні «Документа». Тобто будь-який документ всередині колекції може мати свій власний унікальний набір полів, що і є ключовим аспектом, що робить документо-орієнтовані бази даних найбільш підходящими для даного завдання.

Вибір припав саме на MongoDB через її популярність та наявність хорошої і повної документації. Основними плюсами MongoDB для даного проекту, крім її документо-орієнтованості, стали:

- схожий з реляційними СУБД підхід до зберігання даних: базаданих колекція-документ-поле. Наприклад, у основного конкурента MongoDB, CouchDB, відсутнє поняття будь-якого поділу документів на групи і всі документи є рівнозначними, а для поділу документів, наприклад, по типу, необхідно додавати спеціальні поля.

- формат зберігання даних. Дані в MongoDB зберігаються в форматі BSON.

BSON - це практично те ж саме, що і JSON (поширений формат представлення даних в web), але в бінарному вигляді. Той факт, що дані зберігаються саме в бінарному вигляді, значно прискорює час їх обробки.

- існує інструмент GridFS, що дозволяє зберігати в БД файли будь-якого розміру. Варто зазначити, що на даному етапі розробки, файли великих розмірів, такі, наприклад, як DICOM-файли, пропонується зберігати в іншій, більш захищеній і призначеній для цього БД, під керуванням спеціалізованої СУБД, наприклад Oracle. Наявність такого інструменту є величезним плюсом.

Висновки

1. Створено та описано інформаційну модель для зберігання медичних карт.
2. На основі проведеного аналізу типів не реляційних моделей баз даних обрано документо-орієнтовану модель баз даних для створення інформаційного сховища медичних досліджень.
3. Запропоновано обрати документо-орієнтовану систему керування базами даних – MongoDB.

Література

1. Min Gu, Qiangping Li, Yaoyu Cao. Optical storage arrays: a perspective for future big data storage
2. Samiya Khana, Qufeng Liub, Syed Arshad Alia, Mansaf Alama. Storage Solutions for Big Data Systems: A Qualitative Study and Comparison
3. Abdelhedi, F., Brahim, A.A., Atigui, F., Zurfluh, G.: MDA-based Approach for NoSQL Databases Modelling. In: DAWAK'17. pp. 88–102. Springer (2017)
4. Jihane Mali, Faten Atigui, Ahmed Azough, Nicolas Travers: ModelDrivenGuide: An Approach for Implementing NoSQL Schemas
5. Soransso, R. A. S. N., & Cavalcanti, M. C. (2018, April). Data modeling for analytical queries on document-oriented DBMS. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing (pp. 541-548). ACM
6. Sun, J., & Milani, C. V. B. (2018). U.S. Patent Application No. 15/406,643
7. Geissinger, S. (2018). U.S. Patent Application No. 15/354,921
8. Karl Seguin (2019). The Little MongoDB Book. [online] Available: <http://github.com/Karlseguin/the-little-mongodb-book>.
9. Скулиш М. А., Суліма С. В. Організація управління в багатоетапних системах масового обслуговування. Вісник Національного технічного університету «Львівська політехніка» Радіоелектроніка та телекомунікації. Львів. 2015. № 818. С. 80–85.
10. Скулиш. М.А. Методи контролю перевантаженнями в процесі багатоетапної обробки заявок. Наукові записки Українського науково-дослідного інституту зв'язку. 2015. №1(35). С.81-86.
11. Скулиш М. А., Суліма С. В. Система управління ресурсами в центрах обробки даних оператора мережі мобільного зв'язку. Visnyk NTUU KPI Seriya – Radiotekhnika Radioaparatabuduvannia, ISSN: 2310-0397, e-ISSN: 2310-0389. 2017. Iss. 68. С. 27-32.

Reference

1. Min Gu, Qiangping Li, Yaoyu Cao. Optical storage arrays: a perspective for future big data storage
2. Samiya Khana, Qufeng Liub, Syed Arshad Alia, Mansaf Alama. Storage Solutions for

Big Data Systems: A Qualitative Study and Comparison

3. Abdelhedi, F., Brahim, A.A., Atigui, F., Zurfluh, G.: MDA-based Approach for NoSQL Databases Modelling. In: DAWAK'17. pp. 88–102. Springer (2017)
4. Jihane Mali, Faten Atigui, Ahmed Azough, Nicolas Travers: ModelDrivenGuide: An Approach for Implementing NoSQL Schemas
5. Soransso, R. A. S. N., & Cavalcanti, M. C. (2018, April). Data modeling for analytical queries on document-oriented DBMS. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing (pp. 541-548). ACM
6. Sun, J., & Milani, C. V. B. (2018). U.S. Patent Application No. 15/406,643
7. Geissinger, S. (2018). U.S. Patent Application No. 15/354,921
8. Karl Seguin (2019). The Little MongoDB Book. [online] Available: <http://github.com/Karlseguin/the-little-mongodb-book>.
9. Skulish M. A., Sulima C. V. Organizaciya upravlinnya v bagatoetapnih sistemah masovogo obslugovuvannya. Visnik Nacional'nogo tehnicnogo universistemu «L'vivs'ka politehnika» Radioelektronika ta telekomunikacij. L'viv. 2015. № 818. S. 80-85.
10. Skulish. M.A. Metodi kontrolyu perevantazhennyami v procesi bagatoetapnoi obrobki zayavok. Naukovi zapiski Ukraïns'kogo naukovo-doslidnogo institutu zv'yazku. 2015. №1(35). S.81-86.
11. Skulish M. A., Sulima C.. V. Sistema upravlinnya resursami v centrah obrobki danih operatora merezhi mobil'nogo zv'yazku. Visnyk NTUU KPI Seriya - Radiotekhnika Radioaparatobuduvannya, ISSN: 2310-0397, e-ISSN