УДК 004.6

***О.В. Литвинова, Л.В. Аршинский***

*Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация*

# АНАЛИЗ ПРОБЛЕМ СЛИЯНИЯ БАЗ ДАННЫХ АВИАПЕРЕВОЗОК С РАЗЛИЧНЫМИ ИЗМЕРЕНИЯМИ

***Аннотация***. В работе рассматриваются проблемы и решения, связанные со слиянием баз данных в сфере авиационных перевозок с различными измерениями. Основное внимание уделяется техническим и методологическим аспектам интеграции данных, таким как несовместимость форматов, преобразование данных, нормализация и стандартизация. Применение методов ETL-процессов, интеграционных платформ, и OLAP-кубов позволяет эффективно управлять данными, обеспечивать их согласованность и актуальность. Представлены примеры использования OLAP-анализа для интеграции данных, что позволяет улучшить аналитические возможности, оптимизировать маршруты и повысить качество обслуживания. В результате, применение предложенных методов способствует более обоснованным управленческим решениям и повышению эффективности бизнес-процессов в авиационной отрасли.

***Ключевые слова:*** авиационные перевозки, слияние баз данных, ETL-процессы, интеграционные платформы, нормализация данных, стандартизация, OLAP-анализ, многомерный анализ данных, управление данными, аналитические возможности.

***O.V. Litvinova, L.V. Arshinskiy***

*Irkutsk State Transport University, Irkutsk, Russian Federation*

# ANALYSIS OF DATABASE INTEGRATION ISSUES IN AIRWAYS TRANSPORTATION WITH DIFFERENT DIMENSIONS: TECHNOLOGICAL AND METODOLOGICAL ASPECTS

***Abstract.*** This project addresses the challenges and solutions related to merging databases with different dimensions in the field of airways transport. The focus is on the technical and methodological aspects of data integration, including format incompatibility, data transformation, normalization, and standardization. The application of ETL processes, integration platforms, and OLAP cubes facilitates effective data management, ensuring consistency and relevance. Examples of OLAP analysis for data integration are demonstrating improvements in analytical capabilities, route optimization, and service quality enhancement. The application of these methods contributes to more informed managerial decisions and increased efficiency in business processes within the aviation industry.

***Keywords*:** airways, flights, database integration, ETL processes, integration platforms, data normalization, standardization, OLAP analysis, OLAP-cubes, multidimensional data analysis, data management.

**Введение.** В современном мире авиационные перевозки играют ключевую роль в глобальной логистике и транспортной инфраструктуре [1]. Управление данными в этой сфере стало одной из наиболее сложных и критичных задач, особенно когда речь идет о слиянии различных баз данных. В частности, слияние баз данных авиационных перевозок из разных стран с различными измерениями представляет собой значительный вызов, поскольку эти данные часто хранятся в разных форматах, используют различные схемы и включают множество аспектов, таких как время, маршруты, язык, различные типы баз данных и систем, а также типы грузов и прочее [2].

Эффективная интеграция этих данных необходима для улучшения аналитических возможностей, оптимизации процессов и принятия обоснованных решений. Однако процесс слияния сталкивается с целым рядом технических и методологических проблем. Технические аспекты включают несовместимость форматов данных, проблемы с качеством данных и необходимость их преобразования. Методологические аспекты охватывают вопросы моделирования данных, согласования бизнес-процессов и построения интеграционных стратегий [3].

**Методы слияние баз данных с разными измерениями.** Слияние баз данных, имеющих разные измерения, представляет собой комплексную задачу, требующую не только

технической компетенции, но и стратегического подхода к интеграции информации. В работе использованы несколько ключевых методов, которые помогли эффективно справляться с этой задачей и обеспечить целостность и согласованность данных.

**Использование ETL-процессов**. ETL (Extract, Transform, Load) — это метод для извлечения данных из различных источников, преобразовывать их в нужный формат и загрузить в центральное хранилище на Google Cloud. Процесс извлечения данных включал выбор и сбор информации из различных источников. Преобразование данных в согласованный формат устранило различия в измерениях и структурах данных. Наконец, загрузка данных в единую систему обеспечило их доступность для дальнейшего использования и анализа. [3]. Этот процесс повысил эффективность управлениями данными о рейсах, пассажирах и бронированиях, собранными из различных источников, таких как глобальные системы дистрибуции (GDS) и внутренние системы авиакомпаний. В нашем случае, для извлечения даннных использовались API, а также прямые интеграции с внешними базами данных. Данные, извлеченные из разных систем, отличались по структуре и форматам (например, разные форматы дат, валюты или расстояния). После преобразования данные загружались в центральное хранилище данных, размещённое на Google Cloud. Это хранилище данных позволило легко и быстро получать доступ к актуальной информации для анализа, оптимизации тарифов, управления бронированиями и предоставления персонализированных предложений клиентам.

Такой подход авоматизирует обработку больших объёмов данных и обеспечивает их точность и актуальность, помогает принимать обоснованные решения и повышать качество обслуживания клиентов.

**Интеграционные платформы и middleware**. Для слияния баз данных с разными измерениями использованы интеграционные платформы и middleware решения. Эти инструменты предоставляют наборы функций для синхронизации и консолидации данных в реальном времени. Они помогли объединить данные из различных источников, обеспечивая их согласованность и актуальность. В нашем случае, компания применяет интеграционные платформы и middleware решения для объединения данных из различных систем и баз данных. Эти технологии позволяют эффективно синхронизировать и консолидировать данные из систем управления полетами (Flight Management Systems), систем бронирования, CRM, а также внешних партнерских сервисов, таких как глобальные системы дистрибуции (GDS) и аэропортовые информационные системы.

**Нормализация и стандартизация данных**. Важный шаг в процессе слияния данных – это их приведение к единому формату. Нормализация включает преобразование данных в единые единицы измерения и стандартизацию форматов даты и времени [3]. Этот этап важен для устранения различий, которые могут возникнуть из-за несоответствий в форматах или способах хранения данных в разных системах. Приведение данных к единому виду помогает избежать ошибок и путаницы в процессе их объединения [8].

В рассматриваемом случае возникла проблема при слиянии баз данных различных авиакомпаний из разных стран, которые используют различные системы для управления своими данными (например, Amadeus, Abacus, Galileo, E-tern и т.д.). Часть авиакомпаний хранит информацию о рейсах в формате "DD/MM/YYYY" для даты и использует часы в 12-часовом формате (например, американские и канадские), в то время как другая часть авиакомпаний (например, азиатские авиакомпании) использует формат "YYYY-MM-DD" и 24-часовой формат времени, европейские и российские авиакомпании используют 24-часовой формат "DD-MM-YYYY".

Другая проблема заключалась в том, что часть баз данных хранила информацию о расстоянии в милях, а другая — в километрах.

Для приведения данных к единому формату все значения расстояний преобразованы в одну единицу измерения. Для стандартизации форматов даты, времени и расстояния использовалось преобразование даты из различных форматов в формат "YYYY-MM-DD" при

помощи кодов на MySQL. Например, дата "15/08/2024" преобразована в "2024-08-15". Аналогично, время в 12-часовом формате "02:30 PM" - в 24-часовой формат "14:30".

**Проверка и очистка данных**. После преобразования данных, проведена их проверка на наличие несоответствий и ошибок и очистка. При этом использовались седующие методы: сравнение с исходными данными, проверка диапазонов значений и кросс-валидация, проверка на дубликаты, проверка форматов данных, валидация целостности данных и проверка на пропущенные значения, сравнение с внешними источниками и анализ выбросов. Регулярная проверка и очистка данных от ошибок и устаревших записей являются критически важными для поддержания качества интегрированных данных. Это включает выявление и исправление некорректных или неполных данных, что помогает избежать проблем при их использовании. Проверка данных также может включать проверку на соответствие стандартам и требованиям бизнеса.

**Моделирование данных**. Эффективное моделирование данных является важной частью процесса слияния. Создание модели данных, которая учитывает все измерения и их взаимосвязи, помогает обеспечить правильное слияние [7]. Использование OLAP кубов в моделировании данных позволяет организовать данные в структуру, удобную для анализа и интеграции [5]. Эти схемы четко определяют, как различные измерения взаимодействуют между собой и каким образом их возможно объединить.

Для совмещения нескольких баз данных с различными измерениями использованы методы анализа данных и технологии OLAP-кубов. Также учтены влияние внешних факторов, таких как погодные условия, на пунктуальность рейсов и сезонность. OLAP-анализ (Online Analytical Processing) представляет собой методологию обработки данных, ориентированную на анализ многомерных данных для выявления тенденций, паттернов и взаимосвязей. В контексте реляционных баз данных авиаперевозок, OLAP-анализ позволяет исследовать данные в многомерных аспектах, таких как направление, регион, время, авиакомпания, статус полета и другие ключевые факторы [6]. В контексте авиаперевозок, OLAP-куб может быть построен для отражения ключевых измерений, таких как место отправления и прибытия, дата, авиакомпания, регион бронирования и прочие. Атрибуты, такие как авиакомпания, количество пассажиров и определенная услуга, количество обращений в авиакомпанию на одно бронирование, могут служить фактами, которые можно анализировать в различных перспективах.

Для международной компании, которая занимается перевозками пассажиров воздушным транспортом, для поиска скрытых закономерностей в базах данных Big Query и разработки рекомендаций для принятия управленческих решений использован OLAP метод.

В данном проекте использованы три базы данных с разными направлениями измерений:

– Language – язык обращения (двенадцать европейских языков);

– Region – регион бронирования (регион бронирования задаётся, т.к. в базах данных в целях защиты информации не указывалась персональные данные клиентов, такие как город вылета и город прилета, фамилия и имя, пол);

– City – город обращения в контактный центр и язык, на котором обратился клиент.

Для комбинирования трех баз данных с различными направлениями измерений разработан OLAP-куб, содержащий измерения «Language», «Region», «City» (рисунки 1, 2 и таблица 1).

***OLAP Cube***: Language-Region-City.

***Dimensions***: Language, Region, City.

***Идентификатор Language*** принимает значения:

1 English

2 French

3 Spanish и т.д. (всего 12 языков).

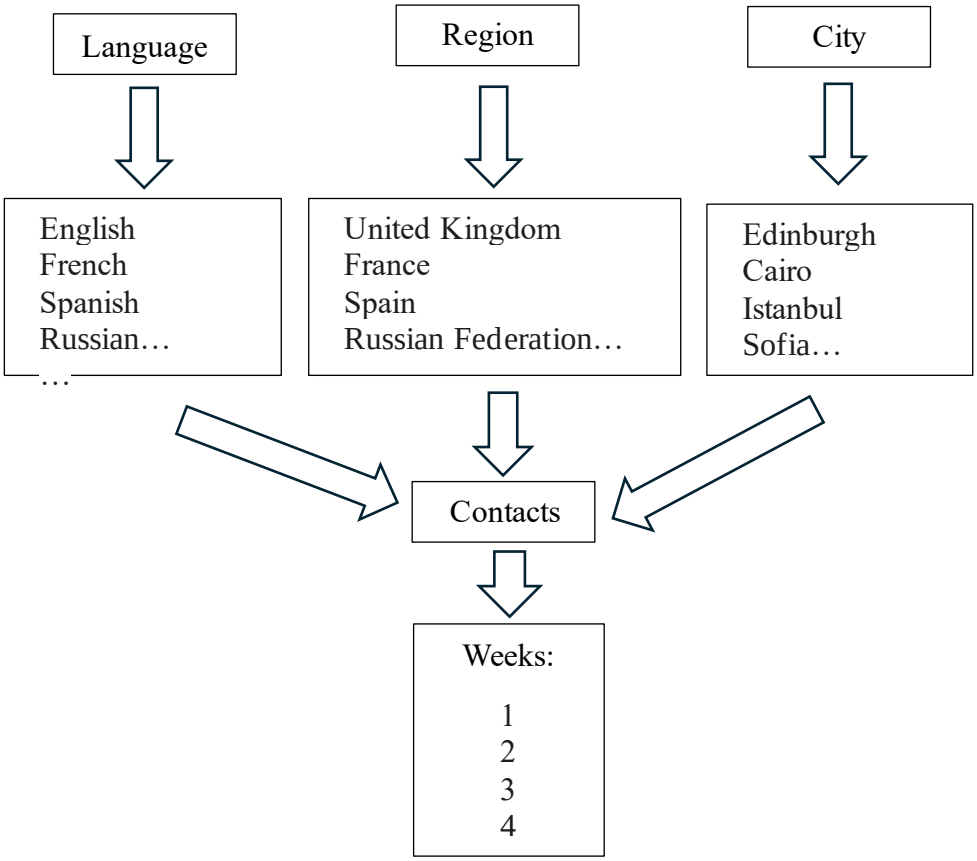


Рис. 1. Схема OLAP-куба

***Идентификатор Region***:
A United Kingdome
B Russia
C France и т.д., все страны Европейского региона
***Идентификатор City***:
X Edinburgh
Y Cairo
Z Istanbul и т.д.
***Measures:*** количество обращений (contacts).

Код для извлечения данных написан на GoogleSQL, а код для комбинирования трех баз данных – на языке Python.

**Таблица 1 *OLAP Cube*: Language-Region-City**

| Language | Region | City | Week | Contacts |
|---|---|---|---|---|
| English | United Kingdom | Edinburgh | 1 | 500,000 |
| French | France | Cairo | 2 | 150,000 |
| Spanish | Spain | Istanbul | 3 | 210,000 |
| Russian | Russian Federation | Sofia | 4 | 560,000 |

Размер баз данных за 2023 год составлял около трёх миллионов строк.

Разработанный OLAP-куб автоматизировал проведение корреляционно-регрессионного анализа удовлетворенности клиентов при внедрении новой услуги «автоматическая регистрация на рейс» по трем измерениям одновременно, что обеспечило более гибкий и мощный способ работы с многомерными данными, включая язык обращения, страну обращения, регион бронирования и город обращения.

Для проведения корреляционно-регрессионного анализа программный код разрабатывался на языке программирования Python с использованием библиотеки pandas для работы с данными и statsmodels при построении регрессионной модели.

В результате корреляционно-регрессионного анализа выявлено, что частота обращений пассажиров при внедрении новой услуги «автоматическая регистрация на рейс» сильно влияет на количество контактов в целом, а также на все ключевые показатели удовлетворенности клиентов, в том числе является одним из факторов, увеличивающим активность пользователей при использовании данной услуги. Это снижает финансовые затраты при ее применении и позволяет использовать данный метод для повышения эффективности пассажирских авиаперевозок.

В результате использования OLAP-анализа и проектирования OLAP-кубов в сочетании с корреляционно-регрессионным анализом в контексте авиаперевозок, предприятия могут выявлять закономерности, оптимизировать маршруты, улучшать качество обслуживания и принимать более обоснованные управленческие решения на основе фактических данных.

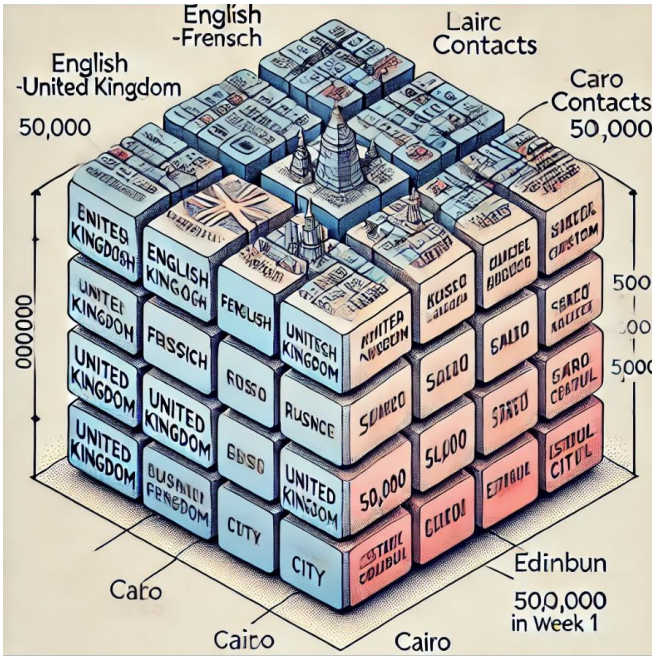

Рис. 2. OLAP-куб

**Заключение.** В заключение следует отметить, что слияние данных из различных баз данных в сфере авиационных перевозок требует комплексного подхода, включающего использование ETL-процессов, интеграционных платформ, нормализацию данных и проверку их качества. Эффективная интеграция и стандартизация данных, такие как преобразование форматов дат и времени, а также единиц измерения, позволяют обеспечить согласованность и актуальность информации. Использование OLAP-кубов для многомерного анализа данных улучшает аналитические возможности и помогает выявлять закономерности, оптимизировать маршруты и повышать качество обслуживания. Этот подход способствует более обоснованным управленческим решениям и повышению эффективности бизнес-процессов в авиационной отрасли.

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Винокуров А. Эволюция и значение воздушных грузовых перевозок. – URL: https://vinocenter.ru/novosti/evolyuciya-i-znachenie-vozdushnyx-gruzovyx-perevozok.html (дата обращения: 16.09.2024)

2. Как повысить эффективность обработки данных в сфере авиации. Кейс Xiamen Airlines. – URL: https://habr.com/ru/companies/glowbyte/articles/720254/ (дата обращения: 16.09.2024)

3. T. Zurek, M. Sinnwell, «Data Warehousing Has More Colours Than Just Black and White», Proc. 25th Int'l Conf. Very Large Databases, Morgan Kaufmann, San Mateo, Calif., 1999

4. Барышков Кирилл Васильевич использование больших данны для повышения эффективности go-to-market стратегий // Практический маркетинг. 2024. №5. URL: https://cyberleninka.ru/article/n/ispolzovanie-bolshih-dannyh-dlya-povysheniyaeffektivnosti-go-to-market-strategiy-1 (дата обращения: 02.09.2024).

5. H.-J. Lenz, A. Shoshani, «Summarizability in OLAP and Statistical Data Bases», Proc. 9th Int'l Conf. Scientific and Statistical Database Management, IEEE CS Press, Los Alamitos, Calif., 1997. – URL: https://www.semanticscholar.org/paper/Summarizability-in-OLAP-and-statistical-data-bases-Lenz-Shoshani/72d5b3fec9a116f119a213633a2a75c96567d5ea (дата обращения: 16.09.2024)

6. C. Ordonez, Z. Chen. Exploration and Visualization of OLAP Cubes with Statistical Tests. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Workshop on Visual Analytics and Knowledge Discovery. 2009,46–55. – URL: https://www2.cs.uh.edu/~ordonez/pdfwww/w-2011-DOLAP-cubevis.pdf (дата обращения: 15.09.2024)

7. Горелов Б. А., Горелов Б. Б. Разработка модели данных для целей оперативной аналитической обработки финансовой информации университета // Унив. управление. 2002. №4(23). С. 33–46. 2. НоженковаЛ. Ф., ШайдуровВ. В. OLAP-технологии оперативной информационно-аналитической поддержки организационного управления // Информ. технологии и вычислит. системы. 2010. № 2. С. 15–27.

8. Harrison M. Effective Pandas: Patterns for Data Manipulation. Independently published. 2021. 497 p.

**REFERENCES**

1. Vinokurov A. Evolution and Importance of Air Cargo Transportation. – URL: https://vinocenter.ru/novosti/evolyuciya-i-znachenie-vozdushnyx-gruzovyx-perevozok.html (accessed: 16.09.2024)

2. How to Improve Data Processing Efficiency in Aviation. Case Study of Xiamen Airlines. – URL: https://habr.com/ru/companies/glowbyte/articles/720254/ (accessed: 16.09.2024)

3. T. Zurek, M. Sinnwell, "Data Warehousing Has More Colours Than Just Black and White," Proc. 25th Int'l Conf. Very Large Databases, Morgan Kaufmann, San Mateo, Calif., 1999

4. Baryshkov Kirill Vasilievich. Use of Big Data to Improve Go-to-Market Strategy Efficiency // Practical Marketing. 2024. No. 5. URL: https://cyberleninka.ru/article/n/ispolzovanie-bolshih-dannyh-dlya-povysheniyaeffektivnosti-go-to-market-strategiy-1 (accessed: 02.09.2024)

5. H.-J. Lenz, A. Shoshani, "Summarizability in OLAP and Statistical Data Bases," Proc. 9th Int'l Conf. Scientific and Statistical Database Management, IEEE CS Press, Los Alamitos, Calif., 1997. – URL: https://www.semanticscholar.org/paper/Summarizability-in-OLAP-and-statistical-data-bases-Lenz-Shoshani/72d5b3fec9a116f119a213633a2a75c96567d5ea (accessed: 16.09.2024)

6. C. Ordonez, Z. Chen. Exploration and Visualization of OLAP Cubes with Statistical Tests. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Workshop on Visual Analytics and Knowledge Discovery. 2009, 46–55. – URL: https://www2.cs.uh.edu/~ordonez/pdfwww/w-2011-DOLAP-cubevis.pdf (accessed: 15.09.2024)

7. Gorelov B. A., Gorelov B. B. Development of a Data Model for Operational Analytical Processing of University Financial Information // Univ. Management. 2002. No. 4(23). pp. 33–46. 2. Nozhenkova L. F., Shaidurov V. V. OLAP Technologies for Operational Information and Analytical Support of Organizational Management // Information Technologies and Computing Systems. 2010. No. 2. pp. 15–27

8. Harrison M. Effective Pandas: Patterns for Data Manipulation. Independently published. 2021. 497 p.

*Литвинова Оксана Владимировна* – магистрант кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: marino_@mail.ru.

*Аршинский Леонид Вадимович* – д.т.н., доцент, профессор кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: larsh@mail.ru.

**Authors**

*Livinova Oksana Vladimirovna* – master's student of department “Information Systems and Information Security”, Irkutsk State Transport University, Irkutsk, e-mail: marino_@mail.ru.

*Leonid Vadimovich Arshinskiy* – Doctor of Technical Science, professor of department “Information Systems and Information Security”, Irkutsk State Transport University, Irkutsk, e-mail: larsh@mail.ru.

**Для цитирования**

Литвинова О.В., Аршинский Л.В. Анализ проблем слияния баз данных авиаперевозок с различными измерениями // Информационные технологии и математическое моделирование в управлении сложными системами: электрон. науч. журн. 2024. №3. С. 14-20. – Режим доступа: http://ismm-irgups.ru/toma/323-2024, свободный. – Загл. с экрана. – Яз. рус., англ.

**For citations**

O.V. Litvinova and L.V. Arshinskiy, Analysis of database integration issues in airways transportation with different dimensions // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2024. No. 3 pp 14-20.