#### Министерство науки и высшего образования Российской Федерации

#### Федеральное государственное бюджетное образовательное учреждение

#### высшего образования

#### «Владимирский государственный университет

#### имени Александра Григорьевича и Николая Григорьевича Столетовых»

**(ВлГУ)**

**Кафедра информационных систем и программной инженерии**

Лабораторная работа №2.1

по дисциплине

"Бизнес анализ данных"

Разработка многомерного хранилища. Технологии OLAP

Выполнил:

ст. гр. ПРИм-124

Парахин К.В.

Приняла:

Преподаватель

кафедры ИСПИ

Тимофеев А.А.

Владимир, 2025 г.

Цель работы

Ознакомиться с принципами проектирования OLTP и OLAP систем.

Задание

Необходимо выбрать dataset данных в формате .csv, в котором хотя бы 1 млн записей, определить решаемую задачу аналитической подсистемы и описать предметную область, а также выделить 4 аналитических запроса/сценария использования.

Далее построить UML диаграмму классов OLTP системы, за ней построить схему OLAP куба, затем ER-диаграмму и скрипт реальной БД, попробовать выполнить один из сценариев на тестовом наборе данных

И выполнить сравнение полученных схем данных для OLTP и OLAP систем.

Выполнение работы

1. В качеству dataset была выбрана выборка данных о спортивных событиях букмекерской конторы Betfair – расположенный по ссылке: <https://www.kaggle.com/datasets/zygmunt/betfair-sports>

Он представлен в формате .csv, содержит чуть более 1 млн записей  
В нем представлены следующие колонки: sport\_id, event\_id, settled\_date, full\_description, scheduled\_off, event, dt actual\_off, selection\_id, selection, odds.

1. Данный датасет содержит информацию о ставках, сделанных на спортивные события через букмекерскую контору Betfair. Включает данные о различных видах спорта (sport\_id), событиях (event\_id), времени начала матчей (scheduled\_off, dt actual\_off), описании событий (full\_description, event), исходах (selection) и коэффициентах (odds). Также указаны даты завершения ставок (settled\_date) и идентификаторы исходов (selection\_id).

Решаемая задача в области анализа данных - анализ поведения пользователей и эффективности ставок на основе исторических данных

1. Основные сценарии использования аналитической подсистемы

* Анализ популярности видов спорта (запрос на топ-5, топ-10 самых популярных и, соответственно, прибыльных видов спорта среди ставщиков)
* Выявление аномалий в коэффициентах ставок (можно проанализировать медианные значения коэффициентов исходов событий конкретных видов спорта – и найти те подозрительные события, в которых отклонения достаточно большие)
* Анализ эффективности конкретных видов ставок (для определенного спорта вывести самые популярные виды ставок – и, соответственно, исходов, на которые игроки ставили чаще всего)
* Анализ времени отклонения начала событий (можно найти среднее отклонение времени запланированного и реального старта спортивного события – что можем потенциально использоваться системой принятия ставок для внутренних оптимизаций)

1. Далее необходимо промоделировать транзакционную систему данных, на основе вычислений и операций в которой создавался данный dataset

Для этого построим UML Class Diagram с помощью описательного языка PlantUML (содержит сущности Sport, Event, Selection, Bet):

@startuml

class Sport {

+sport\_id: int

+name: string

}

class Event {

+event\_id: int

+sport\_id: int

+full\_description: string

+event: string

+scheduled\_off: datetime

+dt\_actual\_off: datetime

}

class Selection {

+selection\_id: int

+event\_id: int

+selection: string

+odds: float

}

class Bet {

+bet\_id: int

+event\_id: int

+selection\_id: int

+settled\_date: datetime

}

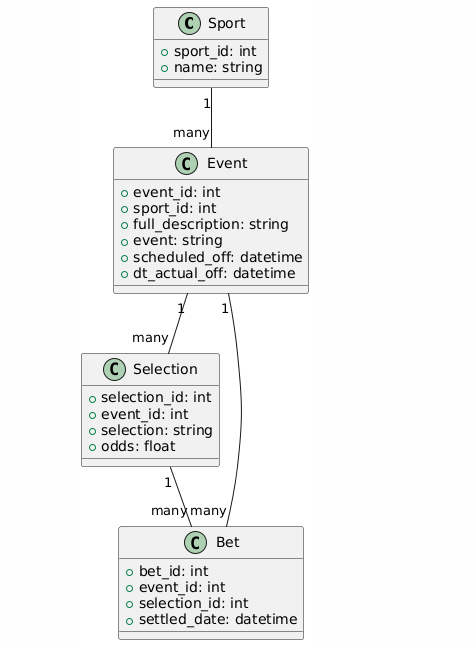
Sport "1" -- "many" Event

Event "1" -- "many" Selection

Event "1" -- "many" Bet

Selection "1" -- "many" Bet

@enduml



1. Далее представим, что мы пытаемся получить схему аналитической системы для этих же данных. Такие данные удобно можно представить в рамках OLAP-куба, который состоит из единственной таблицы фактов: Fact\_Bets, которая содержит следующие 3 меры: total\_bets (общее количество ставок), avg\_odds (средний коэффициент ставки), total\_events (общее число событий)

В качестве измерений можно взять такие: time\_dimension (временное измерение: шкала времени в определенных единицах измерения), sport\_dimension (иерархическое измерение: каждое событие относится к определенному виду спорта), bet\_dimension (тип ставки), selection\_dimension (вариант исхода ставки)

1. Далее составим физическую схему базы данных – в виде ER – диаграммы на описательном языке Plant UML

@startuml

hide circle

skinparam linetype ortho

' Сущности

entity Sport {

\*sport\_id : integer

--

\*name : varchar

}

entity Event {

\*event\_id : integer

--

\*sport\_id : integer

\*full\_description : text

\*event : varchar

\*scheduled\_off : timestamp

\*dt\_actual\_off : timestamp

}

entity Selection {

\*selection\_id : integer

--

\*event\_id : integer

\*selection : varchar

\*odds : float

}

entity Bet {

\*bet\_id : integer

--

\*event\_id : integer

\*selection\_id : integer

\*settled\_date : timestamp

}

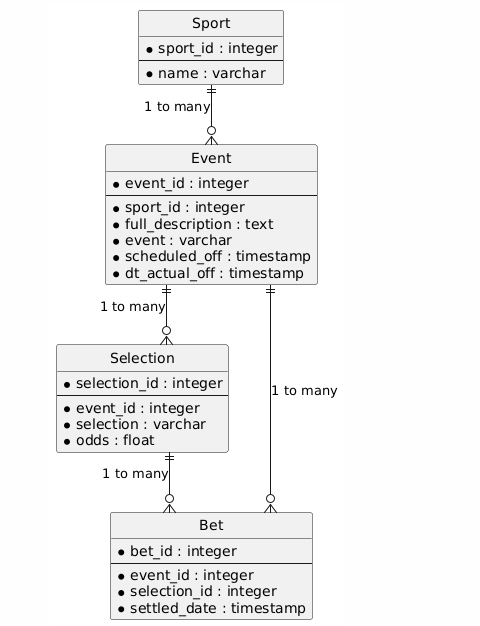
Sport ||--o{ Event : "1 to many"

Event ||--o{ Selection : "1 to many"

Event ||--o{ Bet : "1 to many"

Selection ||--o{ Bet : "1 to many"

@enduml



И на примере реляционной СУБД Postgre SQL представим скрипт БД:

CREATE TABLE Sport (

sport\_id SERIAL PRIMARY KEY,

name VARCHAR(100) NOT NULL

);

CREATE TABLE Event (

event\_id SERIAL PRIMARY KEY,

sport\_id INT REFERENCES Sport(sport\_id),

full\_description TEXT,

event VARCHAR(200),

scheduled\_off TIMESTAMP,

dt\_actual\_off TIMESTAMP

);

CREATE TABLE Selection (

selection\_id SERIAL PRIMARY KEY,

event\_id INT REFERENCES Event(event\_id),

selection VARCHAR(200),

odds FLOAT

);

CREATE TABLE Bet (

bet\_id SERIAL PRIMARY KEY,

event\_id INT REFERENCES Event(event\_id),

selection\_id INT REFERENCES Selection(selection\_id),

settled\_date TIMESTAMP

);

1. Далее я заполнил созданную БД тестовыми данными из 100 первых столбцов исходного dataset) и попробовал использовать предложенный ранее аналитический запрос – а именно вывод топ 5 популярных видов спорта среди введенных данных  
     
   Скрипт:

SELECT

S.name AS sport\_name,

COUNT(B.bet\_id) AS total\_bets

FROM

Bet B

JOIN Event E ON B.event\_id = E.event\_id

JOIN Sport S ON E.sport\_id = S.sport\_id

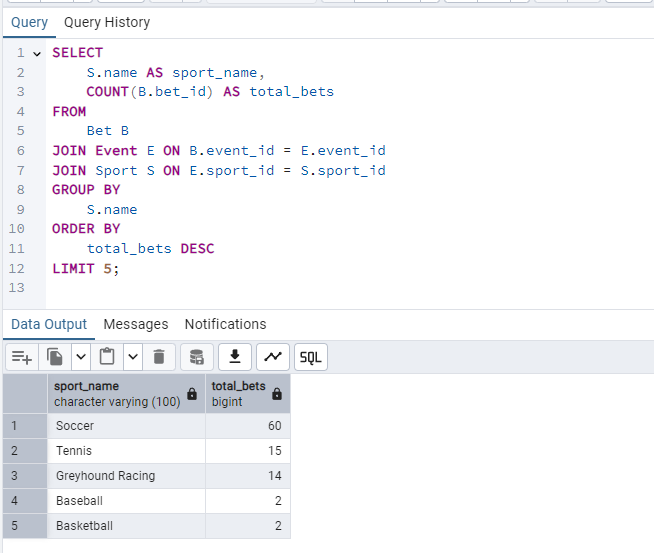
GROUP BY

S.name

ORDER BY

total\_bets DESC

LIMIT 5;



1. По итогу, получилось 2 схемы данных о ставках, выполненные в рамках OLTP и OLAP подхода

Каждая из схем имеет свои преимущества/недостатки – и вполне может быть использована в конкретных ситуациях

OLTP схема имеет преимущества: минимизирует избыточность данных (нормализует данные), предоставляет возможность проведения транзакционных операций записи/обновления/удаления записей из БД, благодаря использования внешних ключей обеспечивает ссылочную целостность данных. Но в недостатках есть следующее: достаточно сложно использовать для комплексных аналитических запросов (типа запросов на события с аномальными коэффициентами) – удобно для несложных запросов, типа самых популярных видов спортов

OLAP схема предоставляет богатый инструмент для аналитических исследований, может содержать и оперировать большими массивами данных, поддерживает многомерные анализы и сложные сценарии. Но ее сложно модифицировать, особенно что то обновлять или удалять, она не работает в параллельных сценариях, во многом дублирует некоторые данные и создает избыточность.

Вывод

В результате выполнения работы были изучены принципы проектирования OLTP и OLAP систем.