#### 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 №3

по дисциплине

"Бизнес анализ данных"

Выполнил:

ст. гр. ПРИм-124

Парахин К.В.

Приняла:

доц. Кафедры ИСПИ

Озерова М.И.

Владимир, 2025 г.

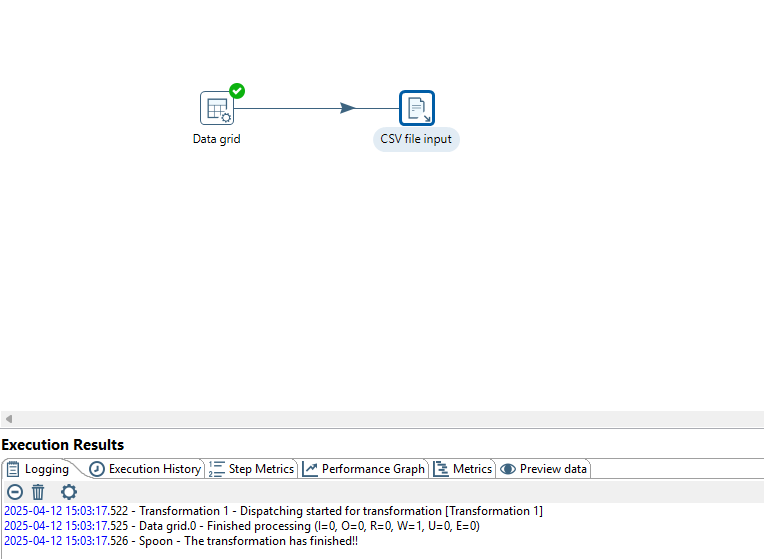
Цель работы:

Научиться пользоваться специализированными средствами для выполнения ETL – процесса (Pentaho Data Integration)

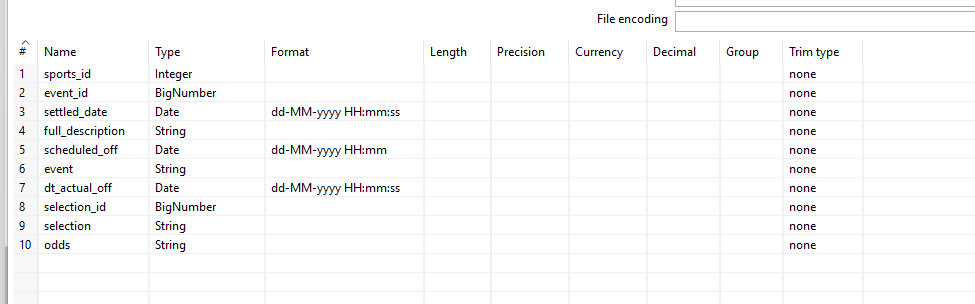
Выполнение работы:

Предполагает разработку механизма загрузки данных из CSV файла в реляционную БД (MySQL, PostgreSQL) с помощью библиотеки встроенных компонентов Pentaho Data Integration (Kettle). В качестве датасета и целевой БД должны использоваться исходные данные и хранилище, спроектированное в рамках лабораторной работы №2.

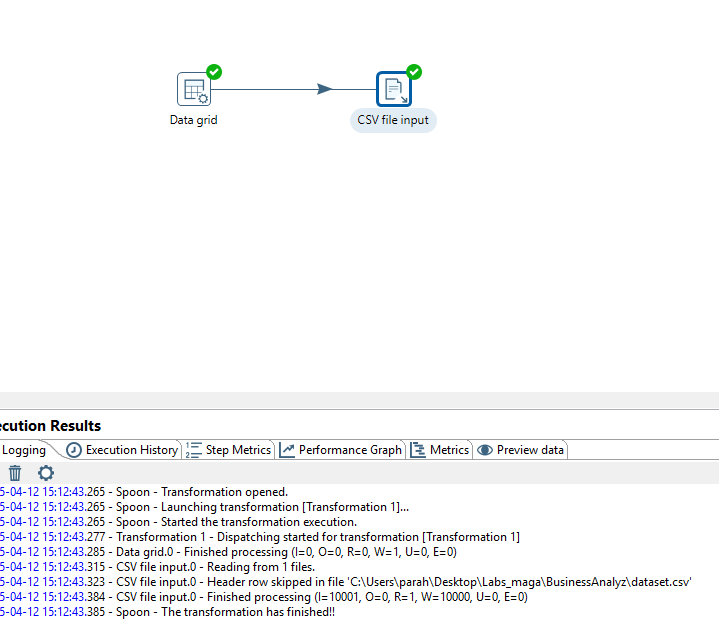
Для начала открываем программу Pentaho и создаем компонент data grid – в который указываем через filename путь к .csv файлу  
Затем добавляем компонент CSV file input



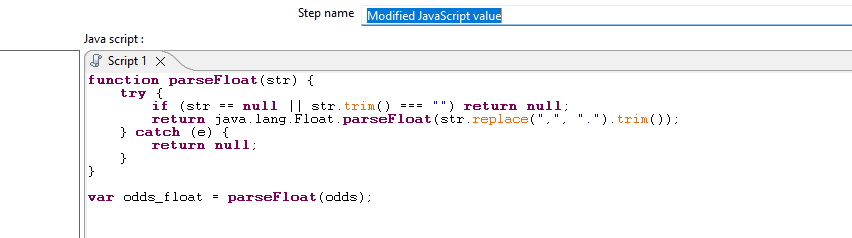
В рамках этого компонента настраиваем импорт конкретных колонок из csv файла, для них можно типизировать тип и формат



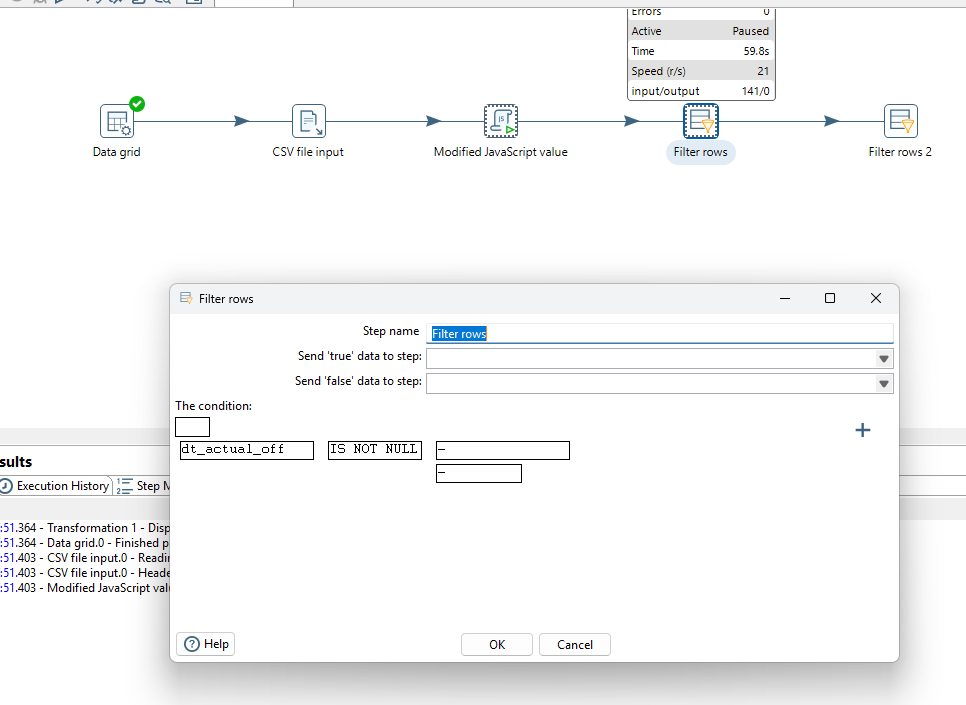
Как видно, импорт данных проходит успешно и теперь с ними можно манипулировать далее.



Но возникает проблема, что некоторые типы данных стандартным csv импортером не получается корректно импортировать (из-за особого формата) – поэтому с помощью отдельного компонента Modified JS value можно осуществить создание новой колонки нужного типа с помощью JS скрипта



Затем также можно провести фильтрацию данных по нескольким колонкам – условие для фильтрации связано с тем, что некоторые поля с датами представлены в dataset некорректно и поэтому невозможно их провалидировать – так что такие строчки отбрасываются



Далее уже можно проводить конкретные действия по вставке данных в OLTP модель Postgres SQL СУБД

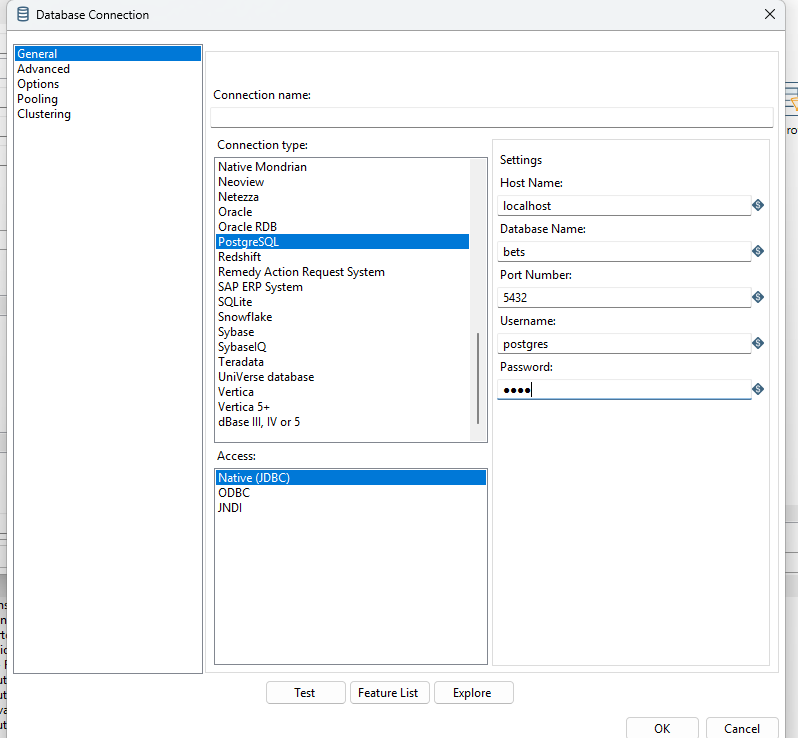
Так как необходимо вставлять данные о sport, event и selection с пропуском дубликатов первичных ключей этих таблиц – то приходится использовать отдельный компонент Execute SQL скрипт – в который вводится скрипт наподобие:

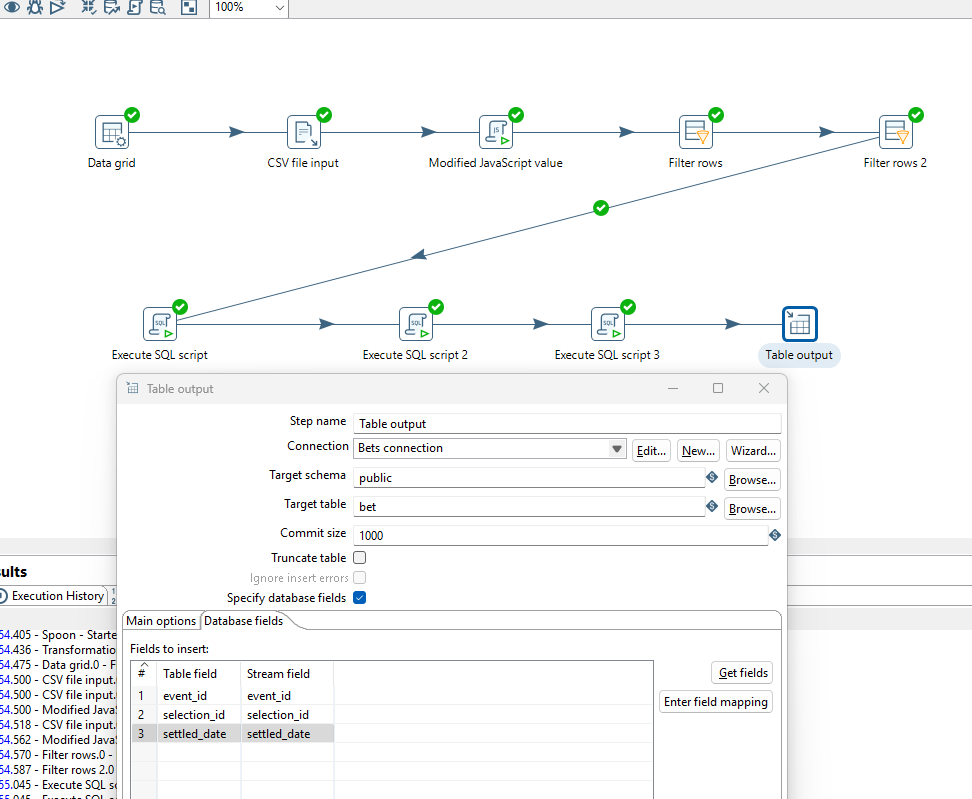
INSERT INTO event (event\_id, sport\_id, full\_description, event, scheduled\_off, dt\_actual\_off)

VALUES (?, ?, ?, ?, ?, ?)

ON CONFLICT DO NOTHING

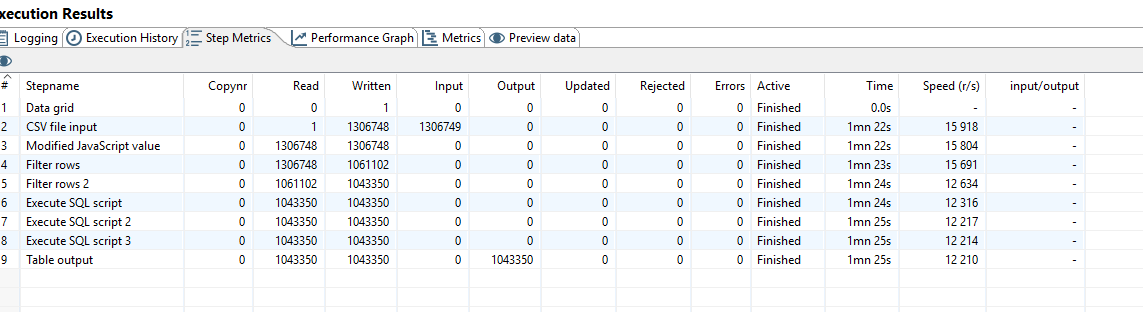
Для финальной таблицы bet уже нет такой проблемы, ведь ключ bet\_id является суррогатным – поэтому возможно использование для вставки компонента Table Output





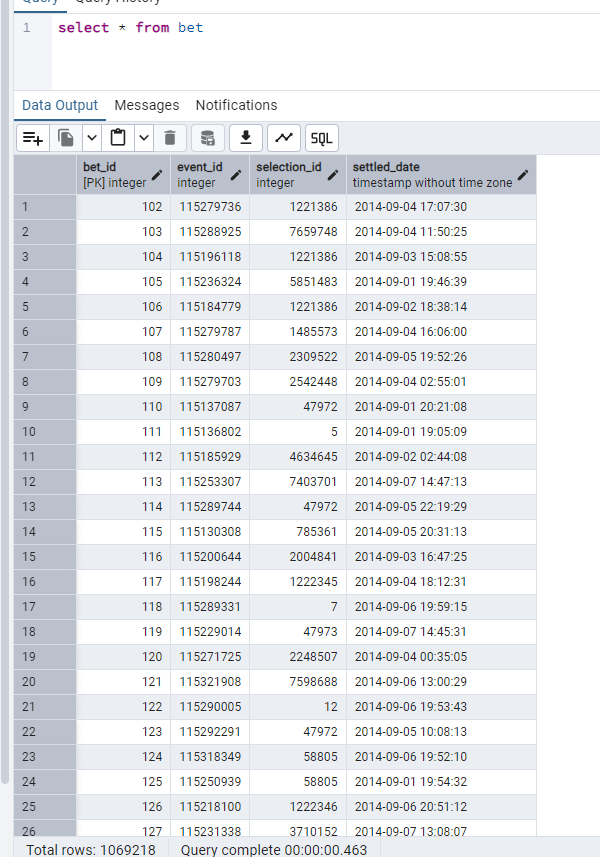
Как видно, весь поток работы проходит успешно – данные транзакционно и последовательно записываются внутрь Postgres SQL базы данных

Попробуем запустить данный ETL процесс для 1 300 000 строк данных внутри .csv файла и чуть упростив условия фильтрации невалидных значений (строки, которые не будут вставлять в колонки БД не будут валидироваться + там где время указано не в общем формате колонк, оно будем округляться)



В итоге видно, что весь ETL-процесс занял около 85 секунд – и его отдельные под-процессы, которые проходили по батчам (равным 50 000 записей) в параллельном режиме – выполнились примерно за одно и то же время

Далее проверим целевую таблицу с данными о ставках в postgre sql:



Если убрать дополнительные шаги модификации уже импортированных из .csv данных перед их экспортом в БД – то потенциально можно чуть оптимизировать общее время работы процесса

Вывод

В результате выполнения работы я освоил использование Pentaho Data Integration инструмента для выполнения ETL-процесса по импорту данных из CSV датасэта в OLTP модель.