Инструкция по работе с pyspark и чтению файлов с помощью psql

Для начала нам понадобится .csv файл с данными. Я скачала файл с Kaggle.

scp /mnt/d/Student_Mental_Stress_and_Coping_Mechanisms.csv user@91.185.85.179:~/ — копируем файл на машину с GreenPlum

ssh user@91.185.85.179 — подключаемся
Вводим пароль

gpfdist -d ~/ -p 8083 — запускаем gpfdist

Открываем новый терминал
ssh user@91.185.85.179
Вводим пароль
sed -i '1d' Student_Mental_Stress_and_Coping_Mechanisms.csv — удаляем первую строку с заголовками, чтобы не было проблем при создании таблицы.
psql -d idp — подключаемся к базе данных
Пишем запрос для создания External table:

```
CREATE EXTERNAL TABLE student_stress_data (
    student_id VARCHAR(20),
    age INT,
    gender VARCHAR(30),
    academic_performance_gpa FLOAT,
    study_hours_per_week INT,
    social_media_usage_hours_per_day INT,
    sleep_duration_hours_per_night INT,
    physical_exercise_hours_per_week INT,
    family_support INT,
    financial_stress INT,
    peer_pressure INT,
```

```
    relationship_stress INT,
    mental_stress_level INT,
    counseling_attendance VARCHAR(5),
    diet_quality INT,
    stress_coping_mechanisms VARCHAR(100),
    cognitive_distortions INT,
    family_mental_health_history VARCHAR(3),
    medical_condition VARCHAR(3),
    substance_use INT
)
LOCATION
('gpfdist://localhost:8083/Student_Mental_Stress_and_Coping_Mechanisms.csv')
FORMAT 'CSV';
```

Проверяем что таблица успешно создана:

```
SELECT * FROM student_stress_data;
```