Read CSV with newlines

how can i read a csv file with newlines in the column headers with spark? the field are quoted

When reading a CSV file in Spark with newlines in the column headers and quoted fields, you need to ensure that Spark properly handles multi-line headers and quoted fields. Here's how you can do it:

1. Use `spark.read.option("header", "true")`

- Spark will recognize the first row as headers.

2. Enable `multiLine` mode

- This allows Spark to correctly parse headers that span multiple lines.

3. Ensure `quote` is set correctly

- If your fields are quoted (e.g., `"column\nname"`), use `quote="\""`.

Example Code:

```
python

from pyspark.sql import SparkSession

# Initialize Spark session
spark = SparkSession.builder.appName("CSVReader").getOrCreate()

# Read CSV with proper options
df = spark.read.option("header", "true") \
               .option("multiLine", "true") \
               .option("quote", "\"") \
               .option("escape", "\"") \  # Handle quotes within values
               .csv("path/to/your_file.csv")

df.show()
```

Explanation:

- `header="true"`: Treats the first row as column names.
- `multiLine="true"`: Handles newlines inside quoted fields (including headers).
- `quote="\""`: Defines `"` as the quote character.
- `escape="\""`: Escapes embedded quotes inside quoted fields.

If your headers contain actual newlines, they may still appear broken in the resulting DataFrame. Consider preprocessing the file to clean up headers before loading into Spark. 🚀