[DelimitedText] Limitation with Multiline CSV source in Dataflow

Last updated by | Jackie Huang | Jan 4, 2022 at 12:24 AM PST

Contents

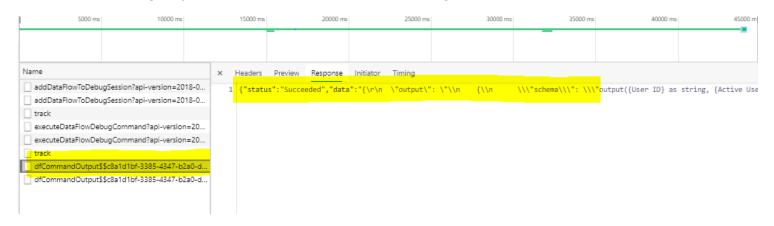- Issue
- How to Troubleshoot
- Recommendation

Issue

In the Mapping data flow, currently, the multiline CSV source does not work with the \r\n as row delimiter. Sometimes has extra lines at carriage returns, breaking source values. This is a limitation in the underlying Spark engine used. This result in failures like

- The last column coming out to be null or missing
- Schema validation at source fails.
- Schema import fails to show correctly in the UX as well as the last column has a new line character in the name.
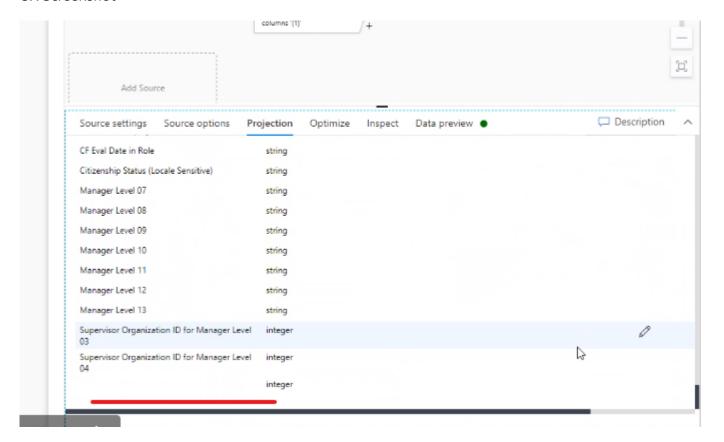
How to Troubleshoot

1. While importing projection, press F12 and check the tracelogs of dfcommandoupt.

2. Please check the last column in below DSL.

{"status":"Succeeded","data":"{\r\n \"output\": \"\\n {\\n \\\\"schema\\\\": \\\\"output(UPI as integer, {Date of Birth} as string, Age as short, Gender as string, {Marital Status} as string, Ethnicity as string, Race as string, {Worker Type} as string, {Time Type} as string, {Worker Subtype} as string, FTE as double, FLSA as string, {Is International Assignee} as boolean, {Employment Status} as string, {Global Grade} as string, {Job Level} as string, {Compensation Grade Profile} as string, {I-Grade} as string, {General Salary Plan} as boolean, {Pay Group} as string, {Pay Rate Type} as string, {Pay Frequency - Payroll} as string, {Hourly Rate - Amount} as boolean, {Weekly Hours} as short, {Scheduled Weekly Hours} as short, {Base Salary Months Paid} as boolean, {Allowance Plans} as string, {Bonus Plans} as string, {Stock Plans} as string, {Merit Plans} as string, {Total Base Pay - Amount for Global Salary Study} as boolean, {Total_Base_Pay_Annualized_-_Amount} as boolean, Currency as string, {Pay Frequency} as string, {Primary Compensation Basis - Local Currency} as boolean, {Compensation Range - Minimum} as boolean, {Compensation Range - Midpoint} as boolean, {Compensation Range - Maximum} as boolean, {Compa-Ratio} as boolean, {Last Base Pay Increase - Amount} as boolean, {Last Base Pay Increase - Currency} as string, {Last Base Pay Increase - Percent} as boolean, {Base Pay Change Date} as string, Contract_Pay_Rate as double, Contract_Pay_Rate_Currency as string, Contract_Pay_Rate_Frequency as string, {Last Hire Date} as string, {Continuous Service Date} as string, {Original Hire Date} as string, {Years of Service - Orig} as boolean, EEO as string, AAP as string, {Job Family Group} as string, {Job Family} as string, {Job Profile} as string, {Job Code} as string, {Job Title} as string, {Position ID} as string, {Job Date} as string, {Time in Position} as string, {Has Direct Reports} as boolean, {Management Level} as string, {Manager Level 01} as string, {Manager Level 02} as string, {Manager Level 03} as string, {Manager Level 04} as string, {Manager Level 05} as string, {Manager UPI} as integer, {Manager Name} as string, {Legacy Ledger} as string, {Legacy Dept} as string, {Cost Center ID} as integer, {Cost Center Desc} as string, {Profit Center ID} as integer, {Profit Center} as string, {Legal\\\\Company Name} as string, Division as string, {Finance Division} as string, {Line of Business} as string, Location as string, Loc_Code as string, {Work Address - City} as string, {Work Country} as string, Region as string, {Primary Nationality} as string, {Last Name} as string, {First Name} as string, {Middle Name} as string, {Home Address - City} as string, {Location Hierarchy - For Global Salary Study} as string, ADLOGON as string, {Work Email} as string, {Supervisory Org} as string, {Work Address State Code} as string, {Work Address State} as string, {Home Address State Code} as string, {Home Address State} as string, PERF_UNIT_CODE as integer, PERF_UNIT_NAME as string, PERF_UNIT_HIER_NAME_12 as string, PERF_UNIT_HIER_NAME_11 as string, PERF_UNIT_HIER_NAME_10 as string, PERF_UNIT_HIER_NAME_9 as string, PERF_UNIT_HIER_NAME_8 as string, PERF_UNIT_HIER_NAME_7 as string, PERF_UNIT_HIER_NAME_6 as string, PERF_UNIT_HIER_NAME_5 as string, PERF_UNIT_HIER_NAME_4 as string, PERF_UNIT_HIER_NAME_3 as string, PERF_UNIT_HIER_NAME_2 as string, PERF_UNIT_HIER_NAME_1 as string, GEO_COUNTRY_NAME as string, GEO_SUB_REGION_NAME as string, GEO_REGION_NAME as string, FUNC_AREA_CODE as short, FUNC_AREA_NAME as string, FUNC_AREA_GRP_NAME as string, ENT_CODE as short, ENT_NAME as string, {Actual FTE} as boolean, {Minority Participation (US Only)} as string, {Sales Flag} as boolean, {Hire Type} as string, {Hire Reason} as string, {Hire Completed Date} as string, {Business Process Effective Date} as string, {Staffing Event Effective Date} as string, {Returning Worker} as boolean, {CF-Worker-Benefit-Program-Indicator-LRV} as string, CF_Benefit_Program_Indicator___BLUE as boolean, CF_Benefit_Program_Indicator___GREEN as boolean, All_Managers_in_Management_Chain as string, CF_Benefit_Program_Indicator_Text as string, {Supervisory Organization Name} as string, {Organization Type} as string, {Organization Subtype} as string, {Hispanic or Latino} as boolean, {CF - Worker - LookupHierarchy - Level06FromTop} as string, {CF - Worker - LookupHierarchy - Level07FromTop} as string, {CNV-HCM-Employee Hire Reason} as string, {CF Eval Date in Role} as string, {Citizenship Status (Locale Sensitive)} as string, {Manager Level 07} as string, {Manager Level 08} as string, {Manager Level 09} as string, {Manager Level 10} as string, {Manager Level 11} as string, {Manager Level 12} as string, {Manager Level 13} as string, {Supervisor Organization ID for Manager Level 03} as integer, {Supervisor Organization ID for Manager Level 04} as integer, <mark>{Supervisor Organization ID for Manager Level 05\\\\r} as integer}</mark>\\\" .\\n \\\\"data\\\\": []\\n }\\n \"\r\n}"

3. This caused the last column to be drifted.

UX Screenshot

Recommendation

- If customer CSV files don't have multi-line rows (column value includes the row delimiter), then please uncheck the "multiline rows" option. This is always recommended if column values don't include the row delimiter since it will help the performance as well.

- If the customer CSV file indeed has multi-line rows, dataflow only works well if the row delimiter is '\n', if customer data has a different row delimiter, please suggest customer to either generate the file at the source with \n as row delimiter rather than \r\n. Or use Copy Activity to convert the file from \r\n to \n as a row delimiter.

NOTE: If the customer has multi-char column delimiter + multi-line rows + row delimiter is not '\n' at the same time, this should be very rare case and please do confirm with customer it is really true and suggest testing with the workaround above first. But it is true, we do not have a proper workaround and the PG team has planned efforts to solve the issue before 2021-07, and no exact date before that as this is a complex bug.

Please leverage the AVA thread in case if you have any questions on this.

Additional Information:

- **Icm References:** ICM: [204467268](#) ⬈,[218126956](#) ⬈

How good have you found this content?

😊 🙁