his is a csv file with 1.9 million rows and 4 columns (provider\_id, detail, source, event\_time).

This is the data of provider activity status on the platform for Sep 2017. The objective is to compute hour wise provider wise no of seconds online.

Rules :

1. Each Day measuring period is between 8 AM and 7 PM.
2. Events from 7 PM to midnight can be ignored
3. Events from midnight to 8 AM are important. (eg : Online at 730 AM can be considered as online at 8 AM. Please note that order of events is important)
4. If detail contains 'True' or source contains 'Action on Job' then 'online'
5. If detail contains 'False' then 'offline'
6. Any thing other an 4 or 5 is error state, can be ignored
7. If an event is both online and offline then treat it as online.
8. It is possible that there may be continuous online or offline events.

Please share the code  and result.

Please share the code  and result.

Please solve this without using any loops.

Sample scenario for a pro p , sep 10 only 3 events ( 1015 : online , 1130 online, 1345 offfine) .The output I expect is this

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **provider\_id** | **date** | **Hour startime** | **Hour end time** | **Seconds online** |
| p | 10-Sep | 8 | 9 | 0 |
| p | 10-Sep | 9 | 10 | 0 |
| p | 10-Sep | 10 | 11 | 2700 |
| p | 10-Sep | 11 | 12 | 3600 |
| p | 10-Sep | 12 | 13 | 3600 |
| p | 10-Sep | 13 | 14 | 2700 |
| p | 10-Sep | 14 | 15 | 0 |
| p | 10-Sep | 15 | 16 | 0 |
| p | 10-Sep | 16 | 17 | 0 |
| p | 10-Sep | 17 | 18 | 0 |
| p | 10-Sep | 18 | 19 | 0 |

Please also fill the following metrics while sharing the output :

1. No of Online Events
2. No of Offline Events
3. No of Providers left after removing a) error rows b) events on or after 7 PM
4. No of Rows in final output data frame (No of Providers \* 30 \* 11)
5. No of Rows in final output where seconds online > 0