Alternatives to handle EVENTS.csv.gz

- Keep the file as it is (gzipped) and work with it in memory (if you have enough memory).
- Upload it to a table in GCP BigQuery, extract your views and work with smaller files.
- Break up this file in smaller files. For example, you can split the rows by SUBJECT_ID and store a subset of these subjects in distinct files.
- You can work with a sample of these subjects.
- You can use aggregation to reduce the number of lines per SUBJECT_ID. For example, transform items into columns and use max, avg and min or one of those to represent the corresponding aggregate. If the item is urine collection, for example, calculate for the SUBJECT_ID the avg, max and min for this item. This (or these) values(s) will become (a) new column(s) for this patient. You should find a good way of representing values for patients that do not present that item). If you choose to represent the data in a an aggregated way, it may make sense to drop the time variable.
- If you choose to keep a timeline of items for each patient you would need to choose a package to handle temporal data. In that case you need to reduce the number of variables (generally, these packages work with a table having columns SUBJECT_ID, ITEM_ID, Value, Timestamp).
- When training, do not forget that you can not use the full timeline.
- Suggestion of label for classification or regression: Length of Stay (LoS), which can be calculated as the time of the last item for a patient minus the first time an item was adiministered to the patient.
- If you want to do something new, convert the table into Prolog facts, code some rules for time precedence and use Aleph to learn LoS (should be a binary variable). Most probably you will need to sample the data in order to be able to learn temporal relations using Aleph.

Tools:

- R: data.table, parallel libraries, h2o
- h2o
- GCP: dataflow (pipelines), spark, python multiprocessing, BigQuery
- DASK
- virtual machines with GPUs and appropriate libraries
- and many others available in python and other languages