Reasoning

1. Why ORC in courses?

Since there's only one fact table in our data warehouse, we chose the dimension table that will have the biggest number of rows and will be often used in queries (if the school would function in real life). ORC makes our table more compressed and allow for faster data read.

2. Why partitioning was used in Courses?

We often run queries grouped by subject, so partitioning the Courses by subject makes sense. Partitioning allows Hive to scan only the specific partition, making our queries faster.

3. Why Parquet in Monitoring?

The monitoring table server as fact table in our data warehouse. It will have huge amount of rows and used in almost each query. So to optimize query performance and faster reads, we decided to choose Parquet as a data storing type, which will be a better fit than a text file

4. Why Bucketing was used in Monitoring?

We bucketed the Monitoring table by student_id because this makes joins with the Students table (on student_id) much faster. Aggregations based on student_id, like calculating attendance or grades, are also more efficient since rows for the same student_id are stored together in the same bucket.

5. Why is Monitoring an external table?

In all of public high schools in Gdansk, teachers use Gdańska Platforma Edykacyjna (GPE) for daily student-centered activities (e.g. checking attendance, adding grades). That’s why we decided to treat monitoring table as external, because all information will be extracted from Gdańsk Education Software and formatted in the way to fit the structure of monitoring table.

6. Why partitioning was used in Classes (static approach)?

We partitioned the Classes table by year, because we'll pretty often use queries only for classes/students of the specific year, to check their performance, etc.

Schools usually organise data by the academic year (e.g. distinct folders: each for specific year). Since we assumed that the data is already divided, we used static approach to insert it

Example Scenarios:

1. Teachers and administrators frequently need to assess student performance in different subjects over the course of the academic year.

2. Teachers need to check and report attendance for individual students on a daily basis.

3. New courses, students, and classes are introduced every academic year, and need to be added to the system without affecting the existing data.

4. School administration wants to check if the students improved in their exams compared to previous years

5. School administration wants to check if there are any specific factors that affect exam results. By finding those correlations, schools can make well-thought decisions to improve student results

6. School administration wants to check the best scoring students and the teachers that teaches the best to award them at the end of the year