# About this exercise

Hill View High School (fictional school) has been using three different systems to track attendance (“C”, “K” and “S”) all with varying degrees of success. We need to help Hill View create a clean attendance record to load into a new sole attendance system.

There are five tasks to complete in this exercise:

1. Create a clean data set which accomplishes the following
   1. Provide a complete attendance record for each day (Unique Identifier, Attendance Status, Date)
   2. Create a flexible structure so if business rules/assumptions change the work is easily reusable and adaptable.
   3. Ensure the solution is able to ingest new data (upsert/merge functionality)
   4. Ensure the solution is idempotent
2. Create an exception data set that show places where you needed to apply business logic to clean up the data
   1. Use your best judgement in making decisions for how to clean up that data.
   2. Please provide a description for any conflict resolution methods you applied.
   3. If the data is cleaned up we should be able to re-run the conflict data set to update the clean data set with fixed values.
3. Run a couple of simple analyses on the data.
   1. What day of the year had the highest number of absences?
   2. Are specific days of the week more likely to result in absences/tardies?
4. Propose any recommendations to the school administrator to help simplify/ease the administration of attendance.
5. Provide the code used, and tools selected, and be able to explain your process.

## About the Data Sets

**List of School Days –** This lists all of the days for which attendance should be taken

**Attendance\_c.csv –** This contains a data export from the “c” system it has three columns

* External ID – Identifies the person for whom the event is logged
* Eventdate – When the event was logged
* Eventtype – The type of logged event

**Attendance\_s.csv –** This contains a data export from the “s” system it has 8 columns

* ID – Internal user ID for the “s” system
* CardSourceID – This specifies the event type.
  + CardSourceID 73 are attendance events
* CardHolderID – External user ID for the “s” system
* DateIn – Time when an event is recorded
* DateOut – Time when an event is recorded
* HoursWorked – Difference between DateIn and DateOut

**Attendance\_k.csv**

* This contains a data export from the “k” system and has 4 columns
* External ID – Identifies the person for whom an event is logged
* Behavior – The event
* Behavior Category – The event type
* Behavior Date – Date when the event occurred

## Business Logic

This section will outline the rules for inferring meaning in **Attendance\_s.csv**

* If cardsourceid = 73 and there is a value for datein then the event should be labeled “tardy”
* If cardsourceid = 73 and there is a value for dateout then the event should be labeled “early dismissal”

**If you have questions or concerns about this exercise please email –** [**anthony@innovateedunyc.org**](mailto:anthony@innovateedunyc.org)