Baseball Systems Data Engineer Exercise

Thank you for your interest in the Baseball Systems Data Engineer position with the Toronto Blue Jays. We would like you to complete a small technical exercise as the next step in the application process.

We do not expect this exercise to take more than a couple of hours so please limit your answers to an appropriate level of detail and complexity. It's more important for us to see how you interpret and think through the questions than to have highly detailed answers or elaborate code.

This MLB StatsAPI endpoint returns JSON data for MLB games that happened on one date in 2023: https://statsapi.mlb.com/api/v1/schedule?sportId=1&date=2023-07-28&hydrate=linescore,gameInfo

Use the data returned to answer the following prompts:

1. Design a relational database schema to store the **game** and **linescore** data returned from the link. Include a diagram showing the relationships between the tables. You do not need to write comprehensive table definitions, but should list what you think are the most important columns and may include notes or annotate the diagram with any comments you think might be useful (for example "*links to the venue table*" or "*include further game info columns*").

   You may also find it helpful to look at the `/team` and `/venue` endpoints (paths are available in the `/schedule` data) but you do not need to use that data and may assume those tables already exist in the database schema.

   You can use draw.io (drawio.com) or Excalidraw (excalidraw.com) to create the diagram but including an image created in other software or using Word's drawing tools is also fine.

2. Write a Python command line program to download the data for a single date and parse it into flat CSV files that correspond to your schema's tables. It should be possible to execute your code from the command line with arguments for the date and sport parameters. Provide requirements and example syntax for how to run your program.

   As before – you do not need to include every possible column in your output, just the most relevant game and linescore columns.

   Explain briefly how you would use your program to download a full season of game data and load it into a database. Describe any edge cases or issues you anticipate may occur that would cause your process to fail.

3. Write SQL queries (any dialect) which you could theoretically run on your schema to answer the following questions about the 2023 season. You do not need to create a database or have access

to all the data to write these queries. You may include partial queries or notes if it helps explain your thought process.

a. How many games were played on each day?
b. What is the average number of runs that the Toronto Blue Jays scored in the 8th inning?
c. What was the gamePk and number of runs of the highest scoring game(s) on each date?
d. What were the final win-loss standings of the American League East division?
e. How often, league wide, did teams win when leading by three runs or fewer entering the 9th inning?
f. What was the largest blown lead in 2023? (extra challenge - this question is optional)

Please return your responses in a zip archive containing:

- Schema diagram
- Python source code
- Written notes and SQL queries
- Example CSV output files for one day of games