Data Engineering Assessment

Buffalo Sabres

November 4, 2021

1 Database Design for Hockey Analytics

1.1 Problem Set Up

Elite Hockey Prospects (EHP) is an online resource for hockey statistics, present and historical. They maintain an API for data access, which requires a credentials for private access. See the documentation here for more details: https://app.swaggerhub.com/apis-docs/esmg/Eliteprospects/1.0

Suppose that you had credentials for full access to the EHP API.

For analysis purposes, it can be useful to link the present-day statistics (including those collected internally, those provided by the NHL, and those that we create using third-party data) to historical player statistics (e.g. pre-draft and/or junior hockey statistics) and biographical information (e.g. height, weight, nationality).

Each of these data sources has its own player identification system, and each has its own standard for player and team name conventions. For example, some data sources encode text (e.g. for player names) differently than others. Similarly, standards for dealing with Russian names (e.g. Evgeni vs. Yevgeni) differ across data sources. Finally, common nicknames are used in some data sources, but not others (e.g. Vyachslav vs. Slava, Alexander vs. Alex, etc).

Moreover, as every new season begins, and as each new game is played, this information is constantly updating and evolving. New players arrive in the NHL as they develop and progress through junior leagues, college, European professional hockey, North American minor leagues, etc; and new draft-eligible players arrive in their junior leagues (and, consequently, our databases) every year.

1.2 Your Task

Describe how you would design a system to coalesce all of this information into a single Buffalo Sabres statistical system. Specifically:

1. How would you gather, store, and update information from the EHP API?
2. What data tables do you think would be useful for hockey analytics purposes? Pick 2-5 potential data tables that you would recommend including in an internal version of the EHP database. Describe the information in each table. Describe your proposed schema for the tables in this database.
3. How would you deal with the name-standardization and data linkage issues (so that all player information can be quickly linked across data sources)?
4. How would you design this system to handle the streaming nature of the data (i.e. updating databases to include the most recent player statistics, handling new players in each underlying data source, etc)?
5. What other issues do you expect to encounter along the way here? What measures would you put in place to prepare for these potential obstacles (and/or other unexpected issues), to ensure that there no downstream issues in our hockey research systems?

We encourage you to be specific about both the tools/software/systems/methods/etc that you would use for each piece of this problem.

2 Cloud vs. On Premise

Would you recommend housing our internal databases (and data analysis systems) in the cloud or on premise? What are the pros and cons of each approach?

3 Data Science Discussion Questions

1. What is the difference between questions of the form “who performed best in the past” vs. “who do you predict will perform best in the future”?
How does your method for answering these questions change, if at all?

2. Suppose you had 20 years of goaltender statistics. Describe how would you answer the following question: *Does a goalie's ability change over time?*

3. Last season, the NHL forwards with highest rates of giveaways (per 60 minutes) include names like Jack Hughes, Evgeni Malkin, Johnny Gaudreau, Leon Draisaitl, Sebastian Aho, and Mitch Marner – players typically thought of to be good with the puck on their stick. On the surface, this doesn't make sense: Why are some of the league's most talented players among the worst in the league when it comes to giving the puck away?

   What are the issues with a metric like "giveaways per 60"? How would you design a better metric to capture how responsible a player is with the puck?