**Big Data Project Proposal**

**Dataset Description**

NHL Game Datasets

This data source contains 10 datasets, which basically records all the official metrics measured for each game in the NHL in the past 6 years. In the Game.csv have 16 columns including the information about a game timestamp, game id, season, etc.. And also has table to record the player’s information, team info and the performance of each game. The total size of these datasets is about 1GB.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**MapReduce Analysis**

With these 10 datasets, I can do lots of analysis by using MapReduce methods. Here are some basic ones:

* The averge goals number of each team every year by ascending or descending order.

Key: team\_id /name, year value: goal number --🡪averge.

[game\_skater\_stat.csv]

* The winning rate(win/lose) for each team every year.

Key: team\_id/name, year value: win times / lost times --🡪 rate

* The average age of players for each team

Key: team\_id/name, players’age. ---🡪 average

2 file join (player info.csv, & game\_,,\_,,.csv )

Player\_info.csv: key --🡪player\_id, value🡪date of birth

Game\_skater\_stats.csv: foreign key🡪 player\_id, value-🡪team\_id

* The ranking of best shooter players for each team and then get their basic information like name, nationality, birthdate etc.
* The country or city that has most Goalie or Shooters.

To achieve these goals, I may use multiple methods including basic MapReduce aggregate methods as well as secondary sort, join patterns or even more.