**CS-498 Cloud Computing Applications Final Project Proposal**

**Team Members:** Nischol Antao, Mike Kresca, Rob Rupprath

**Background**: Baseball has long been Americas pastime. As a sport, it has lent itself to the rigorous recording of statistics, most of which are freely available over the Internet. During our research into datasets to choose for our final project, we came across a large database of baseball player, manager and team statistics, that spanned from 1871-2016 : <http://www.seanlahman.com/baseball-archive/statistics/> . Being baseball fans, we decided to use distributed computing to help us analyze this large dataset, to see if we could answer specific questions we had, about how the game has evolved. We were also curious to see if we could identify in definite trends in the data, that could be used to answer questions about how the sport has changed, or prove/disprove any commonly held conceptions about the game today.

**Proposal:** We intend to analyze a database of Historical baseball records, that provides player, team and manager data from 1871-2016, to answer some of the specific questions listed below. We will also attempt to Visualize our Analysis, using D3/Vega/Tableau to make our analysis more visually appealing, and allow consumers to interact with our analyzed data set.

1) How has the Global Representation of Baseball Players changed over time? What countries produce the most baseball players in number and per capita.

2) Does money buy Championships? How have the Highest Spending Teams performed versus the Lowest Spending Teams over Time.

3) At What Age to Players provide most Value? After how many years in the league are players most productive, and when do their skills start to decline? Can the Steroid era be uniquely identified in time, by looking at these data trends

4) Who has performed better, Left Handed or Right Handed Pitchers? Has this trend changed over time?

5) How do teams perform on the road versus at home? Which Teams have the best Home field advantage and which teams have the worst? How has this changed over time.

6) Is there a correlation between Travel Distance and Performance? Do Teams typically perform better in the second road game, as opposed to the first?

As a consequence of our dataset spanning several years, and spanning several players, Distributed computing is an ideal means to be able to analyze it quickly and effectively. Slicing the data by Year, or by Player, or Team, naturally lends itself to operations that should be distributed. Additionally several of the aggregation operations on each of these slices, would be iterative, which would serve as an ideal use case for an engine such as Apache Spark. The intellectual merit criterion of the project would be served by providing visuals of answers to the questions we have posed, thereby educating other fans of the game, on the evolution of the sport, and of definite trends that exist in the data. To meet the Broader impact criterion, we will aim to establish a framework, that makes it easy to pose other similar questions, in the future, and to easily be able to extract answers to these questions from the dataset. The framework will also allow for users to extend the dataset to include data from future years.