

Cloud Based Baseball Analytics Framework

Michael Kresca
Computer Science Department
University of Illinois
Urbana-Champaign USA
mkresca3@illinois.edu

Nischol Antao
Computer Science Department
University of Illinois
Urbana-Champaign USA
antao@illinois.edu

Robert Rupprath
Computer Science Department
University of Illinois
Urbana-Champaign USA
rkr2@illinois.edu

Abstract—This paper describes a research project that involves building a cloud based baseball analytics framework. The specific purpose of the framework is to provide users with a web based interface that will allow for the easy analysis of Historical Baseball Statistics between the year 1887 and 2016. The paper discusses the challenges associated with storing, analyzing and visualizing a large set of statistical data. It also discusses the challenges associated with implementing and scaling a platform that facilitates large scale access to the data set, and easy and repeatable analysis of it. To implement this framework, Cloud based infrastructure services such as Amazon EC2 and S3 were utilized. Cloud based platform services such as Hortonworks Data Cloud, and its associated underlying software packages, such as Spark SQL were heavily leveraged as well.

Keywords—Baseball, Statistics, Big Data, Cloud, Infrastructure as a Service, Platform as a Service, Analytics Framework, Data Analysis, Data Processing, Data Visualization

I. INTRODUCTION

Baseball has long been Americas pastime. As a sport, it has lent itself to the rigorous recording of Statistics. Despite the multitude of baseball statistics that can be obtained freely over the internet, there is no existing web based framework that easily allows users to generate their own queries about baseball statistics. Our research project aims to address this gap. We intend to build a cloud based, data analytics framework for baseball statistics. This framework will provide users with the ability to generate complex queries on a large, and extensible dataset. It will also allow them to easily visualize the results of these queries. The project will utilize cloud based infrastructure and distributed computing data processing engines to develop, host and scale this framework.

II. BACKGROUND

When researching data sets to implement our course research project, we came across a large database of baseball player, manager and team historical statistics [1]. We were unable to find any publicly available resource, that allowed us to easily run advanced queries on this, and other similar baseball statistical data sets [2] [3]. These statistical datasets provide web based database front end interfaces that sort data by pre-defined categories, and data slices. A user can explore these pre-defined categories/slices; however, they cannot define their own category/data slice. Additionally, these Front ends do not allow users to pose their own advanced analytical queries, and visualize them.

To address this gap, we decided to build our own cloud based framework, that would allow a user to query, analyze and visualize Historical (1887-2016) Baseball Statistics. We believe that a framework that delivers the functionality detailed in the items listed below, would serve to advance knowledge, and benefit society.

- 1) Cloud Based Storage of a Large, Frequently Queried Extensible Dataset
- 2) An Analytics Engine that allows users to define their own custom queries on a large data set
- 3) A Visualization Engine that allows users to easily visualize answers to their queries.
- 4) Cloud Based hosting of the Framework, to allow for load balancing and scalability

III. MOTIVATION

The primary purpose of our research project is to build system that has Intellectual Merit and serves to provide a Broader Impact

The framework we intend to build, will serve to advance intellectual merit, by providing users with the ability to pose and answer their own questions, about the sport of baseball. These questions could serve to explain how the sport has evolved, and help to uncover interesting trends about the sport of baseball and its rich history. We intend to provide examples on how to query the framework (detailed in the Implementation section), that would help to advance a user's knowledge of Big Data Processing Technologies.

The framework will provide a broader impact to society by providing a free, easy to use, baseball analytics service that is not in existence today. The service provided by this framework will allow users to find the answers to advanced queries, that are not necessarily easily queried using a search engine, or a web based database front end. An example of such a query would be to examine the impact that Steroids had on the longevity and success of a player's career. For example, A user could study the performance of all players over the age of 35, over different time periods (1887-2016), and use the results to determine if there is a period of time where it is evident that older players performed significantly better. This could be correlated with the period of time that steroids were prevalent in the game, to examine the impact Steroids had on the game of Baseball.

IV. TOOLS & IMPLEMENTATION

To build the Analytics framework, we intend to leverage cloud based infrastructure and platform services. Amazon S3 storage [4] will be used as the cloud storage location for the Baseball Statistical Data. Local copies of these database will be instantiated into faster Instance memory on Amazon Elastic Containers (EC-2 servers) [4], which will be used to host the Data Analytics Engine. The Web Front End for this framework will be load balanced by Amazon Elastic Load Balancing Servers [4]. The Analytics Engine will be powered by Cloud Based Platform services that provide an eco-system to read, analyze, interpret and visualize big data. We will specifically be using Hortonworks Data Cloud (HDP) [5] services to build our framework. This platform will provide us with an Apache Spark [6] data processing engine, Apache Zeppelin [7] Notebooks, for Data Exploration and Visualization and Apache YARN [8] for managing our distributed server clusters. We intend to implement queries and data analysis using a combination of the Spark SQL and Python programming languages. Apache Zookeeper [9] will be utilized to maintain eventual consistency between database transactions, that are required to be persistent.

To prove out our framework and provide examples for other people on how to use it, we intend to use the Historical

Baseball Statistical data set [1] in combination with general information about major league baseball teams [2] [3], to answer the questions listed below

- 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?

V. CHANGES

We have found the dataset to contain almost all the information needed to answer our questions, though some external calculations will need to be performed, in order to complete the analysis. For example, in order to measure the impact of travel on a team, we must calculate the distance between two cities, calculate any time zone difference, and then make an estimate on the travel time that was taken.

VI. CHALLENGES

Unlike most other sports, there is a huge amount of baseball statistical data available, which covers nearly every aspect of the game. This is good in the sense that it enables interesting queries to be performed, which can provide insights, that otherwise might not be available. In another sense, it makes collecting and storing the data more of a challenge. While there is a large amount of data available, it is stored across different databases, in varying formats. One of our challenges will be finding the right balance of dataset size and the insights (queries) which can be learned from the dataset provided. In the best interest of time, we will need to minimize the amount of time spent on the complexities involved in combining multiple data sets and sources. However, at the same time we need to provide a useful dataset which enables some level of interesting queries that a user can perform.

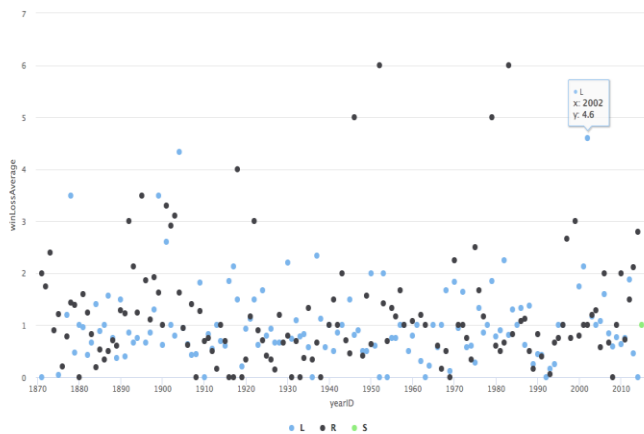
VII. TIMETABLES

Given the remaining time to complete the project, the following timetables will be used:

- | | |
|------------------------|---|
| 1) April 9 – 13 | Data analysis and visualization |
| 2) April 16 – 20 | Continued analysis, cloud framework build |
| 3) April 23 – 27 | Final adjustments and edits |
| 4) May 2 nd | Project Submission |

IX. PRELIMINARY RESULTS

The dataset is well organized and contains twenty-eight different tables for various statistical reporting, with a master table to link each specific dataset. For example, our preliminary analysis of the performance of right-handed pitchers vs. left-handed pitchers used the pitching dataset, while removing outliers who hadn't played enough games for a meaningful performance evaluation, and then adding a win/loss ratio for each player. An example visualization of the right-handed vs left-handed pitchers is shown below:



X. RELATED WORKS

Google and the NCAA recently partnered [10] to perform data analytics in the sport of basketball at the collegiate level. The NCAA has collected a large amount of data over 80 years of the sport's history. The Google cloud has enabled the NCAA to analyze the data in ways which were not possible before.

The insights the NCAA have gained is helping them improve the NCAA tournament team selection. It is also being used to learn about the medical care being provided to the players, and its impact on their performance.

Another related work is the case study of Curtis Granderson [11], a New York Met's outfielder, who was struggling at the plate with a low batting average in 2017. Even though his batting average was low, the Mets kept playing him. They did this because based on plate discipline statistics (only swinging at balls in the strike zone), as well as the speed the ball was leaving his bat, they expected him to break out of his slump soon. Their bet paid off. In the following months, Granderson's batting percentage improved from 0.122 to 0.299. His slugging percentage (total bases divided by number of at bats) shot up to be in the top three in national league. If the associated data analytics had not been performed, the Met's most likely would have benched Granderson, as they would not have expected him to break out of his slump, when he did.

REFERENCES

- [1] Sean Lahman, Baseball Database Website, <http://www.seanlahman.com/baseball-archive/statistics/>
- [2] Major League Baseball, MLB.com website <http://mlb.mlb.com/stats/sortable.jsp>
- [3] Baseball Reference, Baseball Reference website <http://baseball-reference.com>
- [4] Amazon Web Services, Amazon Web Services website <http://aws.amazon.com>
- [5] Horton Works, Hortonworks website <https://hortonworks.com/>
- [6] Apache Spark, Apache Spark website <https://spark.apache.org/>
- [7] Apache Zeppelin, Apache Zeppelin website <https://zeppelin.apache.org/>
- [8] Apache YARN, Apache YARN website <https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html>
- [9] Apache Zookeeper, Apache Zookeeper website <https://zookeeper.apache.org/>
- [10] Google cloud + NCAA documentary <https://cloud.withgoogle.com/ncaa/articles/ncaa-documentary>
- [11] Baseball Analytics – Curtis Granderson Case Study <https://biztechmagazine.com/article/2017/07/baseball-bringing-sports-analytics-forefront>