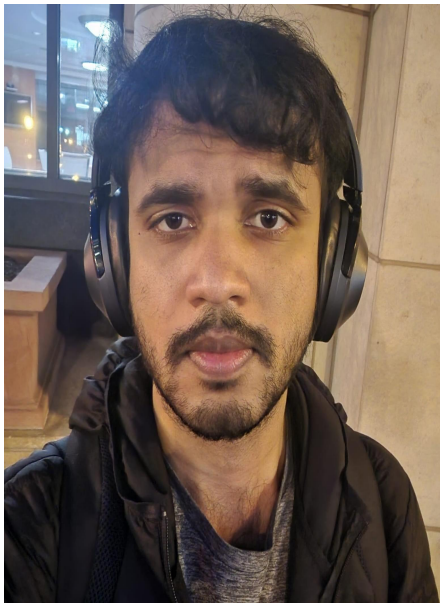
Project Baseball Team Final Report

Lukas Ross, Rishab Haltore, Rishabh Singh, Sobha Garapati



Introduction

The advent of big data has fundamentally transformed the sport of baseball. What began as a data-driven movement spearheaded by amateur statisticians to quantify the value of players and strategies has evolved into a full-fledged analytics revolution embraced across professional baseball organizations. This rapid penetration of data-centric processes in baseball has been enabled by advances in sensor technologies, computation, and database infrastructure, as elucidated extensively in the Related Work section.

The era of data-driven decision-making in baseball, propelled by sources like Statcast, a platform that aggregates a vast array of incredibly specific stats (e.g. launch angles on batted balls, time spent in pitching windups, etc.) offers a granular view of the game. Every pitch, hit, and run is quantified, creating vast data repositories. Extracting meaningful insights from this data ocean demands not just advanced analytics but also a robust and scalable system.

Related Work

Several papers discuss how high-resolution optical recognition systems and radar technologies like Statcast now capture millions of data points per game on factors ranging from pitch speed and trajectory to fielder positioning and sprint velocity (Bai & Bai, 2021; Mizels et al., 2022). Integrated with traditional stats and injury indicators, these vast datasets are leveraged using sophisticated machine learning algorithms to extract nuanced performance insights and predict outcomes like career longevity or injury susceptibility (Dmonte & Dmello, 2017). Healey (2017) was also using Statcast data from an earlier season to prove that past results classified by specific weights based on criteria such as launch angle of hits were more predictive of future success for offensive performers than standard outcome-based statistics.

These analytic techniques have demonstrated tangible impacts on team strategy, player evaluations, roster optimization, and fan engagement initiatives (Watanabe et al., 2021). By quantifying the multidimensional attributes of players, managers can precisely assess talent and optimize lineups to extract maximum tactical advantage - seen in Oakland A's strategic roster building approach highlighted in Moneyball (Dmonte & Dmello, 2017). Granular physical motion indicators allow targeted biomechanical interventions and training adjustments to improve technique and reduce injury risk (Mizels et al., 2022). Descriptive analytics reports even enhance fan experience by providing detailed interactive infographics on player performance. Oh, Han, and Kim (2021) further discussed fan engagement techniques, looking to find correlations between fan attendance and enjoyment with the stadium and gameday experience.

However, ethical questions around usage transparency, predictive model interpretability, and data privacy safeguards have also surfaced, requiring careful regulatory oversight (Watanabe et al., 2021). Overall, the literature extensively covers how baseball has vigorously embraced analytics, demonstrating the transformative potential when cutting-edge technology, statistical rigor, and strategic decision-making intersect. But optimally harnessing big data innovations to elevate performance, engagement, and safety while mitigating risks remains an ongoing pursuit for baseball stakeholders.

“Data analytics and performance: The moderating role of intuition-based HR management in major league baseball” by Kim et al (2021) had the specific objective of evaluating data analytics and performance in the light of the moderating role of intuition-based Human Resource (HR) management in Major League Baseball (MLB). Kim et al (2021, p.204) has identified trends in decreased spectrum of social simplicity or available strategies and diminishing mechanisms of value creation or causal clarity that are directly caused by proportionate increase in reliance on data-driven decisions in highly competitive and specialized industries which have contributed into corresponding directly proportionate diminishing of

positive effects of social capital due to increasing use of data-driven decision making during deployment of human resources.

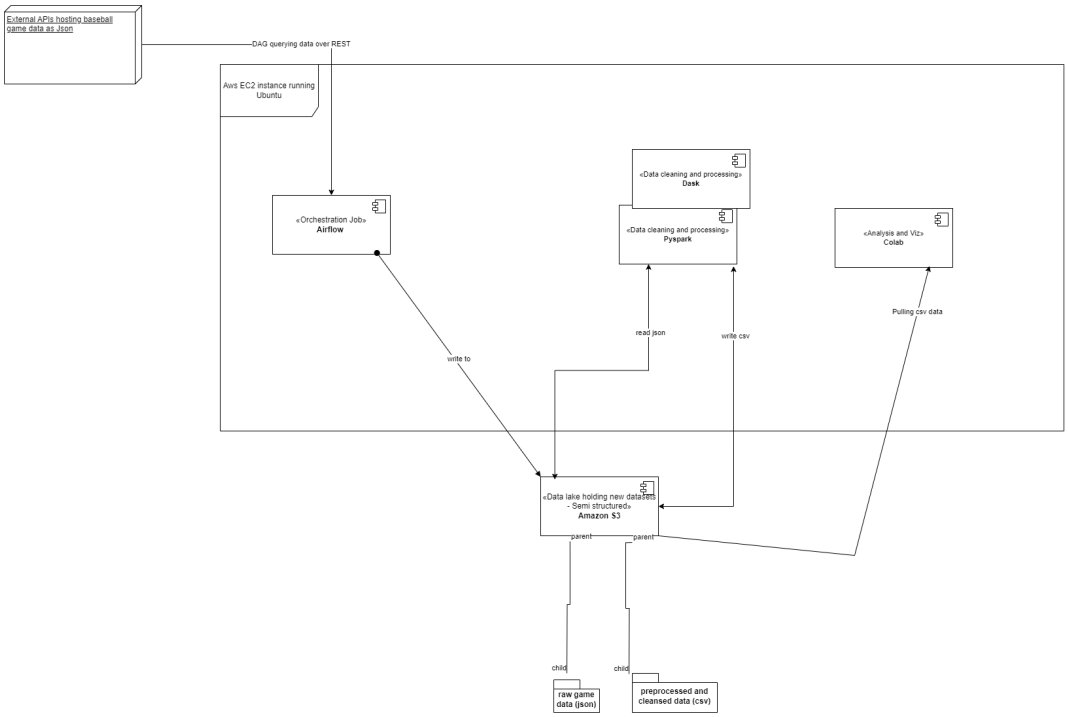
"A multidisciplinary perspective on publicly available sports data in the era of big data: a scoping review of the literature on Major League Baseball" by Huang and Hsu, published in 2021 had specific objective of applying multidisciplinary perspective on publicly available sports data to understand development of data-driven baseball research that satisfies application domains of big data maturity model. The research paper can provide insights into suitability of multidisciplinary perspectives, and especially biomechanical data in MLB.

Design

Source Data Description

Our two large datasets were MLB schedule data and MLB game data. Both of these were returned as semi-structured JSON data with many nested and subnested attributes. The data stretched back into the earlier 1900s, and we had to account for missing values by using "N/A" for these fields, which was more prevalent in earlier data.

Component Diagram



Design Elements

- External API : MLB (Major League Baseball)-sourced data from historical game results and player performances, offered up as JSON (semi-structured data)
- EC2 Instance : we used EC2 to ensure we had a collaborative environment available for development between all team members, and to also represent a “real world” environment in which we are able to vertically scale resources on the box when performing operations that need more or less memory/CPU/etc.
- Airflow : Apache’s open-source data pipeline management tool. Used for the initial extraction of the data from the API and the transfer to S3 as JSON
- S3 : Amazon’s unstructured/semi-structured storage solution. Used to store our un-processed JSON and our processed CSVs.

- Pyspark : Apache's open-source Spark system for distributed data engineering processes. Used in our case for pre-processing and data cleaning at scale
- Dask : Another open-source distributed processing data tool. We used it once we realized Pyspark was too slow in processing the games data specifically
- Docker/Docker-Compose : we used Docker for containerization of the EC2 services, which are how many of them are optimized to be used on a cloud machine, and Docker-Compose for orchestration of these apps speaking to each other.
- Google Colab : For analysis and visualization of the the resultant data

Design Choices And Limitations

We chose PySpark initially for the task of data cleaning and preprocessing because of its ability to flatten each field (with the dataframe.explode command and others) specifically and other first-class parsing capabilities. This was very effective for the 8 GB data set of baseball schedule data. We realized, however, that using PySpark for baseball game data, which was around 160 GBs, was not going to work well with both the size of the data and how nested it was. We instead used the parallel processing platform Dask and the Python Pandas library to perform similar nested operations, and were ultimately successful in getting both of these json data sets cleaned, processed, and converted into Json. We found that using Dask cut the estimated time to process this data by 2/3rds.

For the analysis section, we used Google Colab, but because we were limited by the RAM on the Colab machine, we had to do our analysis in chunks. Another optimization is that during the analysis section, we pulled the data down as .pkl files which are byte streams and are are more space-efficient.

Analytics Section Design

With our analytics section, we used our derived data to answer questions such as some of the following:

- How do engagement scores vary with time of the year & progress of the season? Is there seasonality associated with these scores?
- Does successful teams enjoy higher engagement scores compared to remaining teams? Do higher ranked teams see more engagement from fans?
- How does starting position of a player play a role in driving engagement scores? Are players in certain positions likely to drive higher engagement?

We used an MLB-provided “engagement” metric present on both games and players and applied statistical methodologies against them to answer the questions above and similar questions. See the “Reflections” section for insights into what we learned from this analysis.

Challenges

Design and Conceptual Challenges

Scalability

A major challenge in sports analytics is dealing with the immense scale of data being generated. As discussed in the Watanabe et al. (2021) paper, sports events like the Super Bowl can generate millions of social media posts. Collecting, storing, and analyzing such a high volume of unstructured data presents scalability issues. Traditional data management and analysis methods struggle with this volume. Scaling up the computational infrastructure to handle large volumes in a cost-effective manner is an ongoing challenge.

Heterogeneity

The variety and heterogeneity of data in sports analytics also poses challenges. As highlighted in the Bai and Bai (2021) review, sports data contains diverse modalities like videos, sensor data, text, tabular statistics etc. Integrating such multimodal data and making sense of unstructured formats like video and text adds to the complexity. Lack of standardized formats and semantics makes combining different data sources difficult. Oh, Han, and Kim (2021) similarly discussed the difficulties of combining both manually-collected interviews and unstructured data from social media platforms into a coherent data narrative. Kim et al (2021) also bring up that we discover evidence of significant fan preference heterogeneity across Major League Baseball markets, toward both home and visiting club quality, using a generalized linear mixed model.

Velocity

The speed at which sports data is generated also creates big data challenges. As discussed in the Dmonte and Dmello (2017) paper, real-time IoT sensors and tracking technologies produce streaming data at high velocity during sporting events. Healey (2017) also discussed velocity of data collection via MLB's proprietary sensors, and also mentioned how velocity would get much higher than the 2014 data set used. The analytics has to keep pace with this volume and velocity of generation. Building scalable and flexible data pipelines to ingest and process high-velocity streaming data is an open challenge.

Veracity

Ensuring the quality and veracity of sports data is not easy given the diversity of sources. As the Mizels et al. (2022) review highlights, even basic metrics like errors in baseball have human judgment involved. Ensuring the reliability and accuracy of performance data from

various providers is an ongoing concern. Cleaning noisy multimodal data at scale remains challenging.

Privacy

The Watanabe et al. (2021) paper highlights privacy as an emerging challenge with the proliferation of personal and biometric data. Protecting athlete privacy, preventing leaks of health and performance data, and ensuring ethical use of analytics is of utmost importance. Appropriate de-identification, access control, and governance frameworks need to be implemented.

Model Interpretability

As machine learning is increasingly used for prediction and decision-making in sports, explaining model predictions and ensuring transparency is vital. Sports teams need to trust and interpret the underlying logic. Lack of model interpretability and auditability remains a challenge as highlighted in the Bai and Bai (2021) review. Adoption of explainable AI techniques is an active area of research.Computing and cloud technologies make firms for example MLB's to perceive big data analytics as part of artificial intelligence and machine learning which is consistent with application of biomechanical data on sensing. Researchers put claims that managers of MLB's have potential to apply artificial intelligence which includes scope of current research on biomechanical data which is consistent with application of artificial intelligence identified by Tambe, Cappeli and Yakubovich (2019).

Implementation Challenges

Airbyte and Kafka

Initially, our team was going to use Kafka, a distributed streaming platform, to stream data from the original source, a REST API, and use Airbyte to consume this stream. However, we discovered a blocking bug when deploying Airbyte and Kafka on the AWS EC2 box we used - because we were using Docker and Docker-Compose to run these services in containers, and because of Kafka and Airbyte networking design choices (essentially, Kafka does not expose a publicly available port that can be used by service discovery in some case), we were unable to get Airbyte and Kafka to speak to one another on this box, despite this paradigm working on a locally spun-up instance. This caused us to reconsider our approach and remove Kafka from our design.

Data Cleaning and Processing

The data from the API needed a fair amount of data cleaning as the JSON did not have consistent types in main cases. Within our Pyspark pre-processing and cleaning code, we had to dynamically detect types based on column names. The JSON had many different nested types of data, as well, and as semi-structured data, we had to dynamically detect these nested types while processing the data.

AWS Infrastructure Challenges

We were running all of our code on an EC2 instance on AWS for two reasons: 1) the ability to dynamically scale up and down our resource usage when we needed to depending on the workload of the pipeline 2) to have a shared instance that the whole team could work on in parallel. AWS support removed access to two EC2 instances in a row, and they were unable to share the rationale. Eventually, we were able to get a third EC2 instance spun up and working so that we could complete the pipeline.

Reflections

Processing and cleaning JSON game data from S3 was going to take 11 hours with PySpark at the rate it was moving. This was unacceptably long, so we experimented with using Dask. We found that this caused the final time to be only a third of that. We found that using both PySpark and Dask was a good way to take advantage of using the right tool for the job. However, in the long term for a real system, maintaining two separate distributing processing systems is going to be a challenge, so we would probably consider trying to port all logic into Dask. We would also like to be able to use more custom parallel logic such as with threadpooling, which would enable us to speed up the processing beyond Dask and PySpark's native parallelism.

Additionally, we were hampered by the AWS hurdles we faced. Having to recreate our box twice and not receiving adequate support did not help us be at our most productive. While using an AWS EC2 instance did help our goal of fostering collaboration and also let us scale up as we needed more resources, it may ultimately have been slower in terms of development time than just running everything on a local box given the additional iterations we had to.

For our analysis, using Google Colab we ran specific regressions into fan and player engagement and player performance. Our objective was to use our derived data to help measure engagement versus fan performance, a measurement that isn't necessarily reported on in the media, mostly because a lot of that information is kept proprietary for marketing reasons. We thought that this was a good chance to widen the discourse around the way data is used. Specifically, for big data in baseball, the emphasis is often on just winning games, and how specific decisions can be made to help that. We wanted to look into how player and team performances resonate with fans, and indeed we did see some examples where there were gaps between the most winning team and the teams with the most engagement. We used an MLB-provided measure of fan engagement, and further research could instead rely on

exogenous measures of fan engagement, perhaps including platforms that house deeply engaged fans, like Reddit or X (formerly Twitter). Basically, there is early evidence that winning is important but not the only thing that keeps fans engaged, and finding out what those other factors are is a good place for future research.

Finally, as next steps we would like to build out a more thorough visualization, as being able to visualize the analytics results in a Tableau or similar would make it easier for business stakeholders (or perhaps, the manager of a baseball team!) to understand the results of our analysis and make transformative changes on the field.

Appendix A : Updated Use Case Diagram