USA Baseball Upgrades Data Infrastructure With Scalable AWS Data Lake

Executive Summary

USA Baseball is the national governing body for amateur baseball overseeing more than 15 million players across the United States. The organization has served as a resource to member amateur baseball organizations since 1978. One of USA Baseball's most exciting initiatives is its Prospect Development Pipeline (PDP), which is a collaborative effort between Major League Baseball (MLB) and USA Baseball to establish a player development pipeline leading to the MLB Draft for amateur baseball players in the United States and involves significant data collection and presentation.

Previously, the PDP pathway relied on inefficient and manual data management practices, limiting what scouts and players could do with the information. USA Baseball decided to upgrade its data infrastructure and move forward with a data lake project on the cloud. Due to the organization's small in-house development team, USA Baseball chose to bring in ClearScale for engineering support.

ClearScale Provides Data Driven Player Analysis for USA Baseball

Russell Hartford
Senior Director of Technology

The Challenge

USA Baseball wanted to develop a better system for managing and sharing data collected through its joint PDP initiative with MLB. The program is a large-scale effort that involves conducting amateur player assessments throughout the country and collecting various metrics about athletes using sensors, real-time observation, and other tactics. Since 2017, USA Baseball has gathered data on thousands of players to help them improve and to assist MLB scouts in their work.

Some PDP data is collected automatically, and some is entered manually. All data ultimately ends up in csv files or pdfs in a Dropbox folder that eventually gets shared with MLB scouts. A set of R scripts produces detailed 26-page pdf reports, which are all loaded to a 3D issue Flipbook that players can use to review their performances at specific events.

While this process worked, it left a lot of opportunity on the table. USA Baseball envisioned being able to populate and publish more detailed dashboards that would provide better insights into player performance. The organization also wanted to see if all data could be stored in a scalable way that would also enable advanced analytics. That way, scouts and players could identify trends and discover relationships between key data points that would otherwise be hard to find.

The challenge was that USA Baseball didn’t have the resources in-house to develop an MVP. The organization’s engineering team is small and focuses more on SQL and R programming. That’s why USA Baseball chose to bring in an outside cloud consulting expert and AWS Premier Consulting Partner in ClearScale. ClearScale has proven big data management expertise and deep experience with setting up scalable data lakes on Amazon Web Services (AWS) cloud.

> “With ClearScale we now have a working data lake that allows us to put our data into a more structured cloud environment. This is going to allow us to improve our reporting processes, improve our methods of sharing data with various stakeholders, and put us into a better position to gain new insights into the data, run analytics on our data, and help players develop, while also helping the MLB scouting community better evaluate the top talent in the nation.”
>
> **Russell Hartford,** Senior Director, Technology, USA Baseball

The ClearScale Solution

USA Baseball charged ClearScale with the following requirements:

- Set up a more structured data storage environment
- Enable data lake analyses using standard SQL queries and data science IDEs
- Write new Athena SQL scripts to generate sample reports
- Develop a pipeline to load historical data from Dropbox into the new data lake
- Develop a pipeline to load incremental data from Excel (csv) files

ClearScale fulfilled all of these requirements and provided USA Baseball with a secure data lake for managing and storing all information collected through the PDP initiative. The organization now has the data management infrastructure it needs to handle its ever-growing data volumes in a highly efficient manner.

The Benefits

With its new data lake, USA Baseball has been able to improve its data-sharing and reporting processes, both with athletes and with the 30 MLB clubs. The organization can now run advanced analytics on its data, as it's now structured more appropriately for downstream use. USA Baseball can provide data to the 30 Clubs more efficiently and players have more information to leverage in determining their development plans.

USA Baseball is now planning a second phase to the project by which it will pull in other data sources within a centralized portal that players can use to review their performance histories and trends across multiple events, as well as provide MLB scouts with the ability to conduct more in-depth analyses without having to work with raw data.