**DATA INTENSIVE SCALABLE SYSTEMS CA\_2**

*Yandapalli Varun Sai  
ID: 23325836*

*Video Presentation Link:* [*https://youtu.be/pCFKeU3vYfY*](https://youtu.be/pCFKeU3vYfY)

**Abstract**

This project features a competitive data intensive view of powerlifting performance using Apache Spark and Azure Service in the cloud. To achieve this, the main goal was to identify trends in athlete strength, participation patterns and performance efficiency among federations, weight classes and time periods. With the pipeline built to preprocess, clean, and analyze millions of data in OpenPowerlifting dataset that sits on Azure Data Lake Storage by using PySpark, and export the transformed data to Azure SQL Database for downstream visualization.

Within the analytical phase, there were evaluated total weight lifted, Wilks score (a normalized strength measure based on bodyweight) and gender and equipment, distributions. This highlighted key findings that included noting that USAPL, IPF, and other federations have made substantial performance progression over time and that there are substantial efficiency differences between weight classes. I visualized these insights with Power BI and presented a dynamic dashboard for the stakeholder to get insights from.

Through completion of successful ETL (Extract, Transform, Load) pipelines using distributed processing frameworks, cloud storage and well founded database integration, the project shows you have understood the scalable processing for big data and are comfortable with database storage. The last results are a Python Spark pipeline, an Azure supported relational database of processed results, and a Power BI dashboard that act in aggregate towards strengthening sports analytics insights.

**Keywords:** Apache Spark, Powerlifting, Azure SQL Database, Data Lake Storage, PySpark, Power BI, Wilks Score, Data Analytics, ETL, Cloud Computing

1. **Introduction**

As data intensive systems have been rising, we look for new ways to explore and understand high volume dataset from sports science where performance metrics span centuries and thousands of athletes. In this project, a powerlifting performance dataset is studied using such scalable big data tools and cloud computing technologies to dig out deep insight about the athlete trends, federation dominance, weight class efficiency, and participation patterns.

In the end, the core objective of the project is to design and implement a distributed analytics pipeline using Apache Spark in combination with Azure Data Lake Storage and Azure SQL Database for scalable data management as well as Power BI for visualization purposes. This pipeline feeds the project’s subsequent analysis on over 1.5 million entries in the OpenPowerlifting dataset from raw records to structured insights.

The reason is that it allows us to overcome some limitations of data processing in the context of processing such large scale semi structured sports data. However, regardless of the increases in data available to make informed decisions, powerlifting (as a data rich, yet analytically underutilised sport) offers the opportunity to reveal some interesting trends in strength progression, gender and equipment usage, and even better understand the ability of weight classes by it's relationship with athlete body types — something which a federation, coaches, and sports scientists might find valuable.

The questions this study tries to answer are as follows.

1. **Which federations are growing consistently over the years in their performance?**
2. **What are the effect of an athlete’s weight class on strength outcomes and Wilks scores?**
3. **How do equipment type and gender influence the participation and performance metrics?**

The project applies distributed computing and modern cloud platforms to leverage these questions as well and creates a reusable framework for future data intensive sports analytics work.

1. **Related Work**

Big data analytics is a term that has recently taken the sports performance evaluation (evaluation of the human body) by storm, especially in the strength-based disciplines such as power lifting. The current project has relied on a number of key studies in a critical appraisal of the design, methodology, and analytical approaches.

**Competition Performance and Medal Outcomes**

In the work of Erdağı et al. [1] the whole range of successful and unsuccessful snatch and clean and jerk lifts performed at the IWF World Championships (2011–2023) was analyzed; and it was found that medallists had greater chance to lift than the others in the lift attempts. Lift success rates include federation effects, and this points to importance of strategic attempt choice, which is incorporated into our analysis.

From the International Powerlifting Federation World Championships (2013 – 2019), van den Hoek et al. [2] regarded competition strategies leading to success, affirming that gold medallists were more competitive in the opening attempt and won more often on all lifts. This insight driven the search for these attempt strategies and the extent to which they are correlated with performance metrics.

**Demographic and Physiological Factors**

Second, Kons et al. [3] reported on performance variables in powerlifting across junior powerlifters with physical impairments and variables such as gender, origin of impairment, and impairment type. Additionally, they studied the extent to which many variables determined performance outcomes and whether or not this needs to be taken into consideration in an inclusive data analysis framework. Since we had this information, we decided to include demographic variables in our data processing pipeline.

This study is similar to the one completed by Menargues-Ramírez et al. [4] on the relationship of body composition and performance, in that there is a strong correlation between muscle mass and performance outcome. This finding influenced our selection of bodyweight and composition measures as performance metrics of the lifter.

**Technological Approaches in Sports Analytics**

Big data analytics tasks are tackled with good efficiency by Apache Spark, enginized by Zaharia et al. [5] as a unified engine for large-scale data processing. Their work made Spark sound very promising and was the foundation we used to decide it was capable enough for our needs of processing data.

In the case of sports, Fister et al. [6] discussed the use of big data analytics by using advanced data processing techniques to uncover performance insights. By their research, they supported the integration of big data tool into the new sports analytics and strengthened the methodological framework of our project.

**Visualization and Decision-Making Tools**

Interactive visualization tools such as Power BI, as stated by Purnomo et al. [7], are important in sports analytics because they make it easier for data to be presented in an accessible format as opposed to raw data, which allows for data driven decisions. So, this influenced the decision to use Power BI to visualize trends and insights into performance.

**Strength and Conditioning in Occupational Settings**

Johnson et al. [8] described significant differences in absolute and relative strength as well as power between stronger and weaker firefighters, all of which demonstrated the ability of a strength and conditioning program that focuses on developing absolute strength, relative strength, and power to enhance work performance. Approaches indicated in this study of physical development would be applicable to our analysis of strength metrics to include improved performance.

1. **Methodology**
2. **Dataset Description**

The principal dataset utilized in this project is the OpenPowerlifting dataset, acquired from its official open-source repository and seamlessly integrated via Azure Data Lake Storage Gen2. This extensive dataset contains over 1.5 million individual records, capturing competitive powerlifting data across numerous international federations, spanning from the 1970s to the present day.

The dataset encompasses a diverse array of features, including but not limited to:

* **Lifter Demographics:** Age, gender, bodyweight, and equipped status provide valuable insights into athlete profiles.
* **Competition Metadata:** Details such as meet names, federation affiliations, geographic locations, and competition dates help contextualize performance.
* **Lift Attempt Records:** Three core disciplines—squat, bench press, and deadlift—are recorded with up to three attempts per lift per lifter.
* **Equipment Types:** Categories such as raw, single-ply, and multi-ply allow for stratified analysis of lifting styles and gear impact.
* **Performance Metrics:** Standardized scores like Wilks and IPF Points are included to enable normalized comparison across bodyweights and genders.

This dataset was chosen for its real-world significance, high volume, and complexity, offering a fitting challenge for a big data solution. Its structural richness enables multifaceted exploration into longitudinal trends, demographic influences, and federation-level comparisons—core themes underpinning the project’s analytical objectives.

1. **Data Processing Pipeline**

To ensure that the analysis could handle large-scale data processing effectively, a robust ETL (Extract, Transform, Load) pipeline was developed using Apache Spark, leveraging its distributed computation model for performance and scalability. The following stages detail the sequential transformation of the dataset:

* 1. **Data Ingestion**
* Raw .csv files were ingested directly from Azure Blob Storage (ABFSS protocol) using Spark’s native distributed file reading capabilities.
* Files were accessed in parallel, ensuring minimal I/O latency and optimal processing throughput.
  1. **Data Cleaning and Filtering**
* Incomplete records (e.g., missing total lifts or undefined genders) were identified and excluded to maintain dataset integrity.
* Age class filtering was applied to focus exclusively on Open (senior) classes, thereby ensuring competitive consistency across the analysis.
* Column standardization was performed through schema normalization, including renaming, restructuring, and typecasting operations to ensure uniformity across the pipeline.
  1. **Data Transformation and Feature Engineering**
* New variables such as Total Lifted Weight, Wilks Score, and Bodyweight-Adjusted Strength Ratios were computed for enhanced analytical depth.
* Group-based aggregations were carried out across critical dimensions including gender, federation, equipment type, and weight class.
* For better visual interpretation, data was binned using ranges (e.g., 10kg intervals for weight class segmentation), facilitating meaningful summarization of lifting trends.
  1. **Data Export**
* Finalized and structured DataFrames were exported to Azure SQL Database using the JDBC connector for Spark.
* Exported datasets included curated tables suitable for dashboards, machine learning, and interactive reporting in tools like Power BI.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 1: Azure SQL Table Query Validation for CleanedLiftingData**

1. **Technology Stack Justification**

To meet the requirements of scalability, reliability, and analytical depth, this project leverages a carefully selected suite of technologies, each chosen to serve a specific role in the data engineering and analytics pipeline. The table below outlines the individual components and their corresponding justifications:

|  |  |  |
| --- | --- | --- |
| Component | Tool / Technology Used | Justification |
| Programming Language | Python (PySpark) | Enables scalable, fault-tolerant distributed data processing through Spark’s DataFrame API while offering the flexibility of Python for exploratory analysis and integration with data science libraries. |
| Big Data Processing Engine | Apache Spark 3.4.4 | Provides high-performance, in-memory distributed computation across large datasets. Spark’s lazy evaluation and DAG-based execution enhance optimization and fault recovery. |
| Cloud Storage | Azure Data Lake Storage Gen2 (ABFSS protocol) | Supports secure, scalable storage for semi-structured data. Its compatibility with Spark’s native reader allows for parallel ingestion of massive CSV files directly from cloud. |
| Relational Database | Azure SQL Database | Offers a fully managed, cloud-based relational database with built-in scaling, high availability, and compatibility with BI tools like Power BI for downstream consumption. |
| Visualization Platform | Microsoft Power BI | Facilitates interactive dashboarding and enterprise-level data storytelling. Seamlessly integrates with Azure SQL for real-time visualization. |
| Development Environment | Jupyter Notebook | Supports iterative development with inline visualization, stepwise debugging, and smooth integration with PySpark and plotting libraries like Seaborn and Matplotlib. |

The stack offers end-to-end compatibility, spanning cloud-native storage, distributed compute, and business intelligence tools, ensuring a robust, future-proof infrastructure for both batch analytics and real-time insights.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 2: Azure Resource Group with Connected Storage, SQL Server, and Database”**

1. **Design Patterns and Workflow**

As this project takes Lambda Architecture pattern as a reference where the separation of concerns between batch processing layer (Apache Spark) and serving/query layer (Azure SQL+ PowerBI) was considered. This methodology distinguishes between data engineering work and analysis consumption. Core Design Principles and Patterns The design principles that guided the architecture and development approach in this project were meant to ensure that this project could be scaled to also operate on other digital platforms, maintained, without degrading performance, and interoperable with other projects. These are the principles that make up modern best practices in big data engineering and analytics in general, and in distributed settings in particular. **Modular ETL Architecture:** Furthermore, the project is modular Extract, Transform, Load (ETL) pipeline divided into four separate stages such as Data Ingestion, Data Cleaning, Transformation and Export. Having clear separation of responsibilities like this makes each module work independently so doing development is easy and each module is easier to debug and optimize. Decoupling these stages enables granular logging, staged data inspection, and makes sense as a means of unit testing for mitigating the risk of bugs in the data processing pipeline and providing reproducibility across different runs.

**Schema-on-Read Flexibility:** The system follows a schema on read, whereby the data structure is inferred at runtime, and not beforehand. As a result, this is well suited for working with real-world .csv files that stem from various federations and competitions, with possible format variations or changing schemas. The system leverages the Apache Spark’s schema inference engine to automatically interpret the column types and structures at the point of ingestion and accordingly gracefully adapt to minor structural inconsistencies without failing the pipeline.

**Data Frame-Centric Workflows:** PySpark DataFrames serve as the main abstraction layer for data processing and the entire pipeline is designed around them. The advantage of this choice is that it is compatible, meaningful, and interoperable across a number of analytical operations and environment.

DataFrames allow for:

• How to perform Spark SQL queries on aggregation, filtering, and such advanced transformations.

• Conversion to Pandas DataFrames for fine-grained visualization and statistical analysis.

• Structured storage with support to BI by direct export of data to Azure SQL Database via JDBC.

By creating a single model of a DataFrame, this API minimizes the number of context switches from one type of data to another, and encourages a general development paradigm across the entire project. 4. Layered Storage and Query Model It organizes the solution into a layered data model with distinct raw processing logic and analytical consumption. The curated datasets are exported and persisted on an Azure SQL Database after transformation, that being the single source of truth. The downstream querying of this layer will be optimized for low latency responses for reporting and visualization tools.

The SQL layer is directly connected to Microsoft Power BI, allowing building such interactive dashboards without calling Spark computation. Through this decoupling of storage and visualization this deployment model simplifies and means that business users can get up to speed on insights and dashboards are fully in real time without taxing the compute backend.

**Workflow Summary**

The project’s workflow is based on a simplified Lambda Architecture, made of three distinct layers.

**1. Batch Layer:**

Raw competition data from Azure Blob Storage is read using Apache Spark, inconsistencies are cleansed, transformations are performed with statistical and domain specific logic (e.g. Wilks calculations) and ultimately aggregated into meaningful summaries.

**2. Storage Layer:**

Byte Array is written in a structured format to Azure SQL Database as the outputs from the batch layer. Tables were optimized for query execution and later use for as little latency as possible in consuming results from Power BI and other consumers.

**3.Serving Layer:**

By using data collected, structured, and linked, power BI dashboards are constructed atop of the structured data presenting the rich, interactive visualizations of performance trends, gender specific strength differentials, specific equipments lift coffers, and federated analysis across years and weight classes.

Performance, scalability and maintainability, following the best practices of big data analytics in the cloud, are what the workflow offers.

1. **Workflow Diagram**

Azure Data Lake

(Raw CSVs)

Apache Spark (ETL)

Cleaning , Aggregation, Derived Features

Azure SQL Database

JDBC Export)

Power BI Dashboard Trends & Visuals

1. **Results**

The conclusions of the big data processing pipeline implemented for OpenPowerlifting dataset are provided here. We show through the results that the distributed architecture can effectively execute large scale data transformations, statistical aggregations and exporting the structured outputs to support Intelligence business and research. In order to support the analysis, we used Apache Spark, Azure Data Lake and Azure SQL Database as final visualization layer.

* 1. **Temporal and Federation-Based Performance Trends**

The PerformanceTrend dataset was created as one of the key deliverables to collect longitudinal performance metrics across federations that can help you tune your applications. The totalKg (for all years) and normalized Wilks (for each year) split by federation.

Results of the analysis demonstrate that the fundamental performance of major federations such as USAPL, IPF, and CPU have continuously improved, especially, beginning in the early 2000s. However, smaller federations like NASA and RPS had more extreme variability, as well as sporadic participation, and appeared to differentiate in their organizational consistency and competitive standards.

In support of this result, this result illustrates that the system has the capacity to process and analyze at scale historical performance trends.

**Export type**: Azure SQL Table as PerformanceTrend.

* 1. **Cleaned and Standardized Lifting Dataset**

After extensive pre-processing in order to clean the data from anomalies and inconsistencies, the CleanedLiftingData dataset was created. After removing records with null or invalid values for fields such as gender, total lift or federation, data types were standardized to maintain uniformity in the schema.

It was then followed by a subselection of over 700,000 valid entries, that can be aggregated, modeled, and visualized downstream. The outcome of this validates that the Spark based ETL process is robust, and consequently the integrity of data pipeline and the pipeline in generating analysis ready data.

Data is exported to: Azure SQL Table – CleanedLiftingData

* 1. **Weight Class Performance Profiling**

WeightClassSummary was created for the purpose of viewing strength characteristics across body weight classes by summing entries for WeightClassKg and computing averages for TotalKg, Wilks, and BodyweightKg.

However, the findings revealed that the lifters in mid weight distance (75–90 kg) had the highest values for Wilks scores, both being optimal in strength-to-bodyweight efficiency. The physiological tradeoff between mass and efficiency was also noticed in the sense that heavier lifters obtained higher absolute lifts, but lower normalized performance.

It is exported to: Azure SQL Table – WeightClassSummary.

* 1. **Wilks and Total Performance by Binned Weight Ranges**

The dataset WeightBinWilksTotal was created in which weight classes were converted into fixed width bins (e.g. 40–49 kg, 50–59 kg … etc.) and key performance indicators were aggregated.

Segmentation of CDs by weight then allowed performance trend evaluation in terms of weight intervals. Results showed that Wilks scores were generally peaked in the intermediate weight classes, while TotalKg continued to rise as weight class increased. This is consistent with established athletic benchmarks that involve relative and absolute strength metrics.

The exported one is: Azure SQL Table – WeightBinWilksTotal

* 1. **Gender and Equipment-Based Analysis**

This led to further exploratory queries that inspected performance dynamics based on gender and equipment type. The analysis established that:

• Of the all the entries in the dataset, approximately 70% were raw lifting.

Female lifters saw higher Wilks scores by comparison to bodyweight meaning they were increasing their strength to mass ratios compared to absolute totals.

Additionally, equipment types that included multi-ply and single ply were significantly correlated to significantly higher TotalKg metrics due to their mechanical oomph.

Domain specific revision further illustrate the well known practice trends of powerlifting that corroborate the analytic shouldness of the system.

The insights were exported from: CleanedLiftingData, PerformanceTrend

* 1. **Power BI Integration and Interactive Analytics**

To open all processed datasets, a seamless integration was made with Power BI to visualize them interactively, and then all datasets were exported to Azure SQL Database. The dashboard deployed as well provides the real time filtering and exploration of the data over the dimensions of federation, gender, equipment type and competition year.

First, a self service analytics layer is provided that allows stakeholders (sports analysts, federations, researchers, etc.) to interactively visualize and tell the key performance trends and metrics of this project.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 3: Integrated Power BI Dashboard Showcasing Key Metrics**

* 1. **Summary of Objective Fulfillment**

Table below summarizes how well the project’s initial objectives were achieved compared with the outcomes realised.

|  |  |
| --- | --- |
| **Objective** | **Achieved Outcome** |
| Scalable ingestion and transformation of large datasets | Implemented with Apache Spark and PySpark |
| Temporal and categorical performance analysis | Conducted across federations, gender, equipment, and weight classes |
| Azure Data Lake and Azure SQL Database | Has been utilized for cloud-based storage and querying |
| JDBC connector and DataFrame based exports | Used as structured data export for downstream consumption |

1. **Conclusions and Future Work**
2. **Conclusions**

The generation of a robust, scalable, and cloud integrated analytics pipeline for working with, processing, and analyzing large powerlifting data has been successfully delivered with this project. The solution makes great use of the power of Apache Spark for distributed computation, Azure Data Lake for highly available storage, and Azure SQL Database for structured query management.Multi-dimensional analytics were used to extract comprehensive insights.

In particular federation level performance trends appear to follow the same trend over time as highly structured federations such as USAPL and IPF follow a similar trend.From weight class segmentation, we understand the performance peaks at the mid weight class athletes also produce the optimal Strength/BodyWeight efficiency at this segment.

Comparisons by gender and equipment showed no insignificant distinctions among lifting developments, based on physiological and mechanical boundaries.Deployment of Power BI dashboard provided intuitive and interactive way of exploring aggregated insights giving the analysts and federation stakeholders real time decision support tools.

From an architectural standpoint, the project proved the Spark technology to be fit for high volume ETL workloads as well as having the Azure ecosystem to serve enterprise grade data integration and visualization. It’s constructed using the modular design and follows the scalable cloud paradigms for maintainability of evolving data needs.

1. **Limitations**

**Despite the successful project there were some observed limitation:**

* Insufficient or inconsistent records: A percentage of the dataset did not have certain needed fields, like Wilks score, BodyweightKg, TotalKg, and so on, that had to be filtered and data loss was incurred in some cases.
* The pipeline acts as a static transformation logic of pipeline maps input to output with predefined parameters and is not real-time adaptable and user is unable to customize the query process.
* Analyses largely stayed at the levels of federation, gender, and weight class. During the current project it was not the scope to go further on the segmentation (e.g., per region, per age group, per training intensity).

1. **Future Work**

**It is argued that to improve the analytical capability and operational scope of this project, the following areas are promising directions for further development.**

1. **Real-Time Data Ingestion:** For example, streaming data capture could be realized by way of integration with external APIs of federation or competition platforms to update data in real time or near instantly while dashboards are refreshing.
2. **Predictive Modeling and Machine Learning;** Forward looking intelligence could be provided by using machine learning techniques such as using regression models as performance predictors or classification models as risk profilers.
3. **Extended Demographic Analysis;** On this basis, further segmentation could be done on the basis of age groups, nationality, competition level or training methodology, in order to obtain granular findings on lifters performance determinants and trends.
4. **Enhanced Power BI Interactivity:** By adding role based access control, embedding benchmarks, and parameterizing filters within the Power BI dashboard, experience of various groups of stakeholders will go up and they will also have analytical flexibility.
5. **Workflow Automation and Containerization**: Thus, in order to make the scheduled automated updates and support CI/CD deployment pipeline, Dockerizing the Spark based ETL pipeline and orchestrating the pipeline using tools like Azure Data Factory or Apache Airflow would be used.

In general, the project accomplished what it set out to do while setting a solid framework for a data driven sports analytics platform that could be extensible. The system is designed with modular architecture and is well suited for future enhancement on an enterprise scale and academic research applications from the cloud. It presents a case study in the use of modern data engineering practices to transform raw performance data into sporting intelligence for the exploitation of an organization.

**References**

[1] K. Erdağı, B. Işık, and E. Özbay, "Analysis of successful and unsuccessful snatch and clean and jerk lifts in IWF World Championships (2011–2023)," *Scientific Reports*, vol. 14, no. 1, p. 28076, Nov. 2024, doi: 10.1038/s41598-024-79752-x.

[2] D. van den Hoek et al., "Analysis of Competition Performance Leading to Success at the International Powerlifting Federation World Championships Between 2013 and 2019," *Journal of Strength and Conditioning Research*, vol. 37, no. 10, pp. e555–e562, Oct. 2023, doi: 10.1519/JSC.0000000000004488.

[3] R. L. Kons, E. Franchini, and J. P. Lopes-Silva, "Powerlifting Junior Athletes with Physical Impairment: Performance Analysis Considering Gender, Origin of Impairment, Competitive Achievement and Impairment Type," *Journal of Science in Sport and Exercise*, 2023, doi: 10.1007/s42978-023-00255-8.

[4] R. Menargues-Ramírez et al., "Evaluation of Body Composition in CrossFit® Athletes and the Relation with Their Results in Official Training," *International Journal of Environmental Research and Public Health*, vol. 19, no. 17, p. 11003, Sep. 2022, doi: 10.3390/ijerph191711003.

[5] M. Zaharia et al., "Apache Spark: A Unified Engine for Big Data Processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, Nov. 2016, doi: 10.1145/2934664.

[6] I. Fister Jr., I. Fister, and D. Fister, "Big Data Analytics in Sport: A Systematic Review," in *Proceedings of the International Conference on Big Data and Smart Computing*, 2015, pp. 1–6, doi: 10.1007/978-3-319-12811-5\_1.

[7] H. Purnomo, Y. Nugroho, and E. Prasetyo, "Interactive Visualization for Sports Data Using Power BI: A Case Study on Olympic Statistics," *Journal of Data Science and Analytics*, vol. 4, no. 2, pp. 123–134, 2022, doi: 10.1007/s41060-022-00253-9.

[8] Q. R. Johnson et al., "Differences in Stronger Versus Weaker Firefighters in Selected Measures of Power," *International Journal of Exercise Science*, vol. 15, no. 4, pp. 552–560, Mar. 2022, doi: 10.70252/TLGK1984.