NFL Player Load Data Exploration

Using NFL Big Data Bowl Tracking Data to Explore Pipelines

Overview

I have worked with tracking data before, but I wanted to add a small project to my portfolio to demonstrate how I might design and implement a system to analyze tracking data from a sports science perspective. Future iterations should include automated reporting or governance checks, but for now this is just setting up the pipeline and (selfishly) looking at some metrics.

For the project, I used 2021 NFL Big Data Bowl data as my source as it has rich tracking data (albeit only pass plays) from the 2018 season. I wanted to quickly explore how I would approach gathering, cleaning, and analyzing this data, but it is important to note that the data is all from one source. Thus cleaning and creating the system in this case is much easier than what one might encounter in practice.

The system ingests 17 million raw 10Hz optical tracking frames across 253 games and 1,303 players, transforms them through a medallion architecture into analyst-ready metrics, and surfaces them through an interactive dashboard and reproducible monitoring notebook.

System Architecture

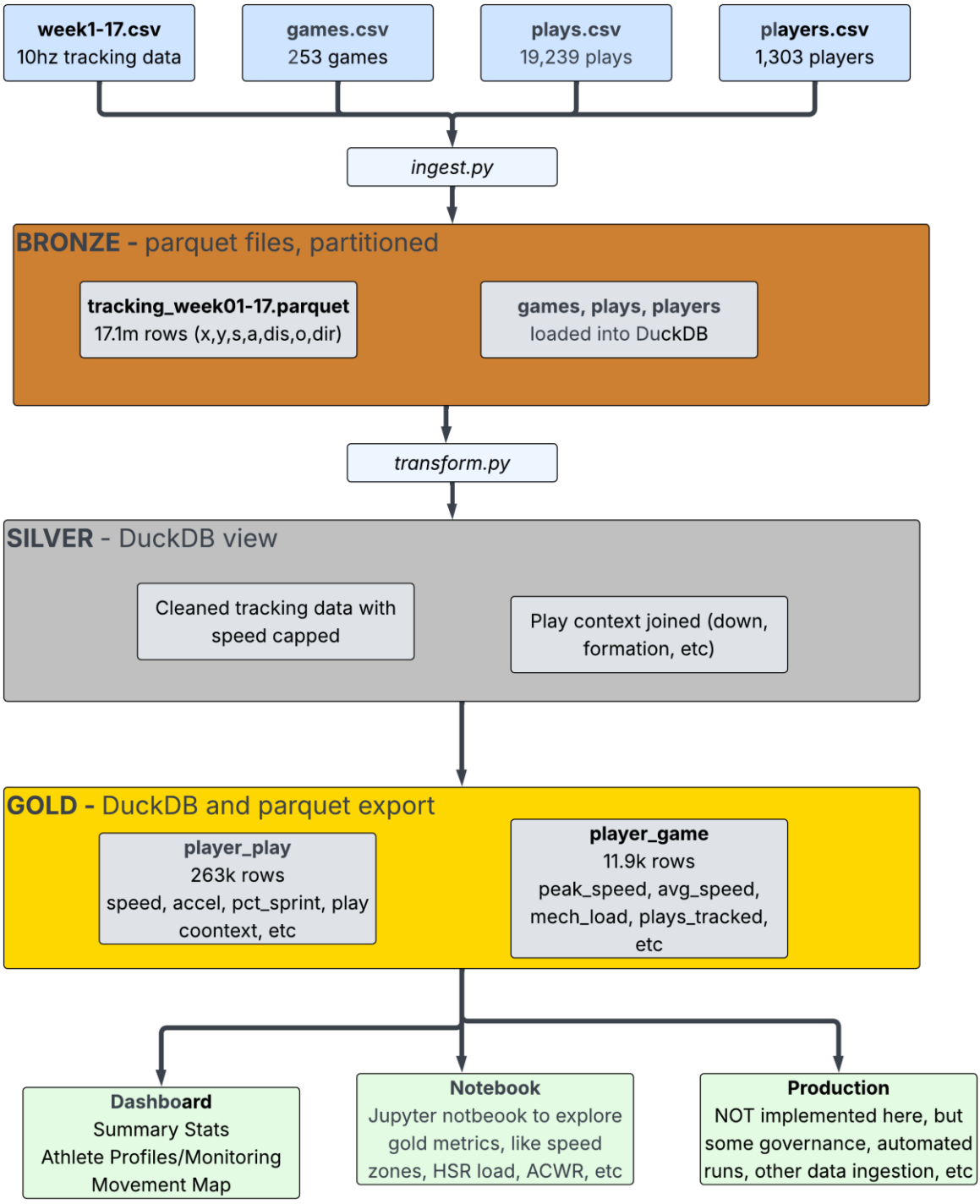
The pipeline follows a Bronze/Silver/Gold medallion architecture that is common in lakehouse environments. Each layer serves a distinct purpose:

Layer	Tooling	Purpose
Bronze	Parquet (partitioned by week)	Raw CSV ingestion converted to columnar format. Preserves source fidelity with no transformations.
Silver	DuckDB view	Cleaned, typed, and enriched tracking data. Speed noise filtered, signed acceleration derived from speed delta, play context joined from games and plays tables.
Gold	DuckDB tables + Parquet export	Aggregated per-player per-play and per-player per-game metrics. Mechanical load, speed zones, high intensity event counts. Ready for dashboard and notebook consumption.

Because I'm doing this locally for a portfolio, I just used DuckDB as the query engine as it will handle the massive data efficiently without a server. A production environment would obviously use something like Databricks and role-based access, automated updates, etc.

Two files are used for the simple pipeline: ingest.py (to turn the raw BDB data into bronze parquet), and the transform.py to create the silver and gold tables (pretty simple since BDB already had a lot of these cleaned and standardized).

Diagram of Workflow: From Tracking Data to Analysis



Load Monitoring Dashboard

I used ClaudeCode to make a quick Streamlit dashboard visualizing the data: one tab is a brief overview and the other is an athlete detail tab. In future versions maybe add a place in the overview or a separate tab to monitor ingestion, data quality, updates, etc.

Figure 1. Population-level load dashboard showing distance leaders, peak speed distribution by position, weekly mechanical load trends, and high intensity event scatter.

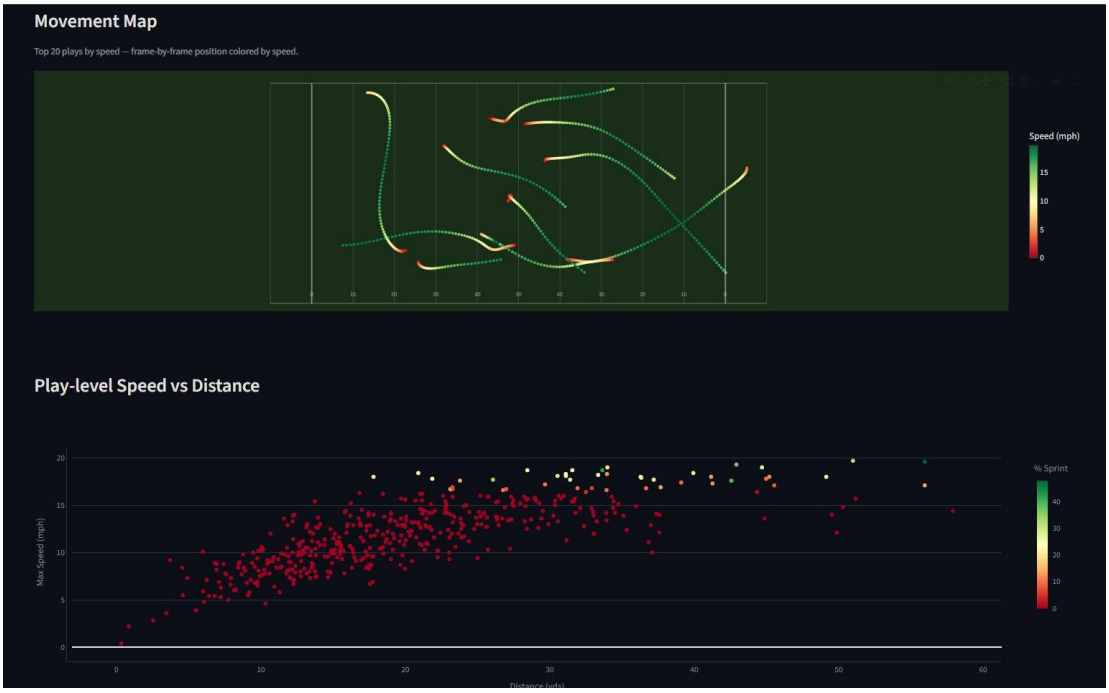


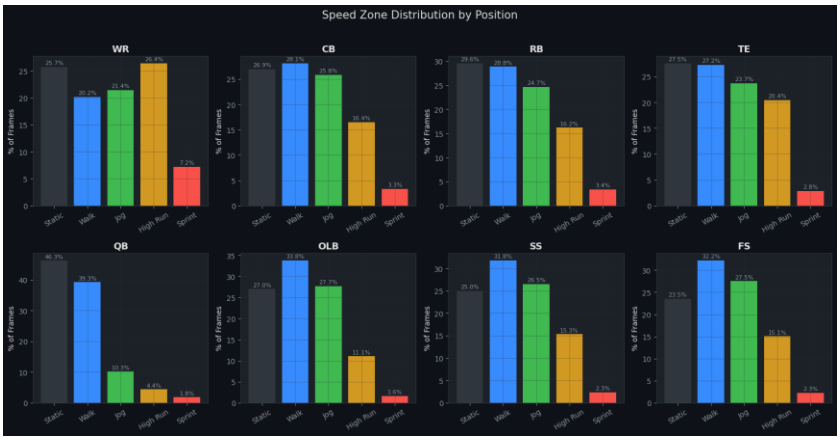
Figure 2. Athlete profile tab showing bio, weekly performance trends, field movement map (top 20 plays by speed colored by intensity), and play-level tables.

Athlete Monitoring Metrics

Next I wrote a Jupyter notebook to demonstrate more of the analytical side of this project. The intent is to show that the data system produces outputs that can be used to answer sports science questions I found in literature.

Speed Zone Profiling

Speed thresholds adapted from Dwyer and Gabbett (2012) define five zones from static to sprint. Positional profiles reveal distinct movement signatures: wide receivers spend 7.2% of frames at sprint intensity, quarterbacks only 1.8%, and outside linebackers show the highest walk fraction (33.8%) reflecting zone coverage demands. These benchmarks form the reference standard for individual deviation monitoring. A note again that the BDB data was just from pass plays this year and the NFL already did their own cleaning of the uploaded data set.

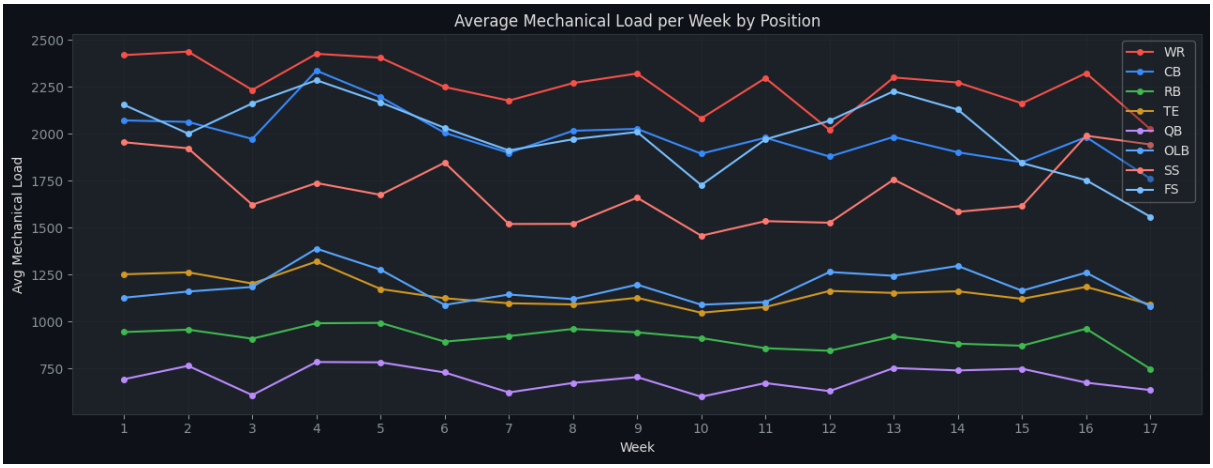


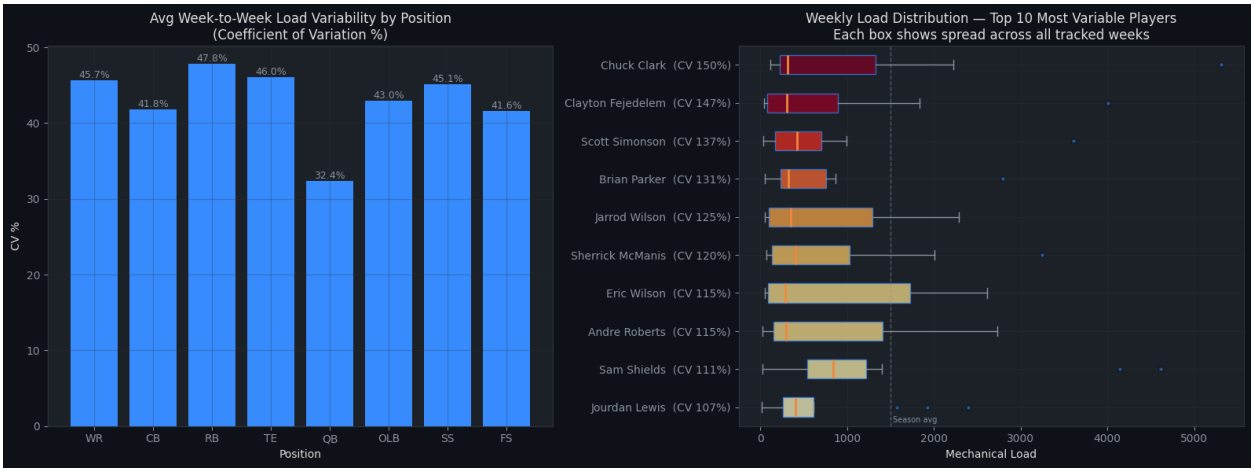
High Speed Running Load

High speed running (HSR) distance above 12.3 mph and sprint distance above 15.7 mph are computed per player per game. Wide receivers accumulate the greatest HSR volume (283 yards/game) and sprint distance (95 yards/game). Free safeties cover the most total distance (592 yards/game) but at lower intensity, demonstrating that raw distance alone is an insufficient load metric across positions.

Week-to-Week Load Variation

Mechanical load (average speed x distance per play, summed across a game) is computed for each player-game. Week-to-week coefficient of variation is calculated per player to identify those with unstable load profiles across the season, a known correlate of injury risk in the GPS literature.





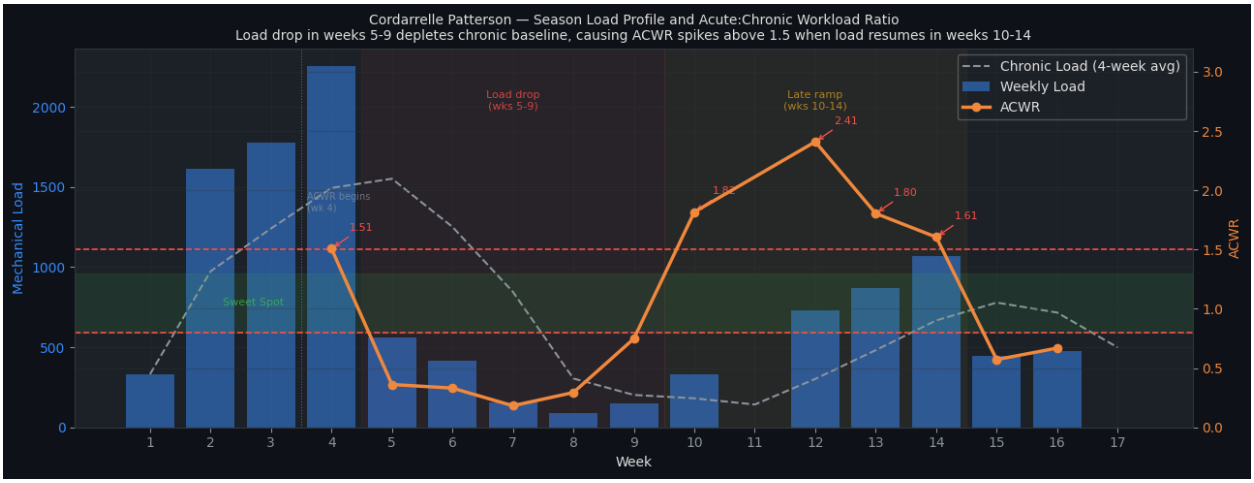
Acute Chronic Workload Ratio

The ACWR (Hulin et al., 2016) compares a player's current week load to their rolling 4-week average. A ratio between 0.8 and 1.3 is the established sweet spot. Ratios above 1.5 flag load spikes are associated with elevated injury risk. The notebook produces individual ACWR profiles across the season and flags weeks outside the sweet spot by position.

Example: Cordarrelle Patterson. I had to dig back in the 2018 logs for this one, it was his only year in New England, and he won a Super Bowl. Firstly, as a utility guy, Patterson has spent his career moving between starting running back duties, to back up roles or primarily returning kicks. He fluctuates where and how often he appears on the field. This unique profile illustrates a critical part of ACWR I'm just learning about: a spike does not require a high-load week. It requires a low chronic baseline

In 2018: Patterson's early-season load was substantial, peaking in Week 4. However, starting in Week 5 his offensive role contracted, dropping to as few as 6 snaps per game. He had a fumble in Week 7 that likely put him in the doghouse and saw his usage decline sharply. By week 9, his 4-week chronic load average had fallen to just 202 mechanical load units. That same week, with the Patriots depleted at running back, Patterson backed up James White in the backfield against Green Bay - recording 61 yards on 11 carries and scoring a touchdown.

From the coaching staff's perspective this was a low-risk deployment of a versatile player, also a move they had to make. From a load monitoring perspective, his body had de-conditioned over the prior four weeks and was now absorbing a significantly elevated stimulus relative to its recent baseline. His ACWR values in Weeks 10-14 (ranging from 1.61 to 2.41) were not driven by exceptionally high loads, but instead just a low chronic baseline from his stretch of low usage. This shows how raw load numbers alone may not have been concerning, but the ACWR shows the return to play was a spike in intensity for Patterson. This is one of many scenarios a systematic athlete monitoring infrastructure is designed to catch



Production Considerations

This project was a quick attempt to use public data (Kaggle) to highlight a generalized approach to working with tracking data for sports science tasks. The following adaptations would be required for a production football environment:

- Cloud data platform (Snowflake or Databricks) replacing DuckDB for multi-user access and compute scaling
- Role-based access controls to enforce athlete data privacy
- Replacing manual script execution for automated weekly pipeline runs post-game
- Real GPS/IMU sensor data could also be used for not just in-game load but also weekly practice logs
- Integration with injury and availability records to enable prospective workload flagging against historical injury context

References

Dwyer, D.B. and Gabbett, T.J. (2012). Global positioning system data analysis: Velocity ranges and a new definition of sprinting for field sport athletes. Journal of Strength and Conditioning Research, 26(3), 818-824.

Hulin, B.T., Gabbett, T.J., Lawson, D.W., Caputi, P. and Sampson, J.A. (2016). The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players. British Journal of Sports Medicine, 50(4), 231-236.

Tierney, P., Young, A., Clarke, N. and Duncan, M. (2016). Match play demands of 11 versus 11 professional football using Global Positioning System tracking: Variations across common playing formations. Human Movement Science, 49, 1-8.

NFL Big Data Bowl (2021). Player tracking data, 2018 NFL season. Kaggle. https://www.kaggle.com/c/nfl-big-data-bowl-2021