Requirement Gathering Summary – Project 3 (Athlete Wearable Tech Team)

Requirement Area	Response Summary
Valuable Data Types	Performance metrics (e.g., heart rate, cadence, speed), health metrics (e.g., HRV, sleep quality), training summaries, environmental data (e.g., temperature), derived metrics (e.g., fatigue index), time-series data.
Data Format	CSV and XLSX for current data; future API integrations (e.g., Strava, Garmin) may bring in JSON.
Data Structure	Structured data in tabular format, suitable for relational databases; future support for semi-structured (JSON) may be needed.
Data Challenges	Limited dataset size, missing values, inconsistent device formats, timestamp misalignment, and unit standardization across sports and devices.
Data Cleaning Required	Yes – tasks include outlier removal, missing value imputation, timestamp alignment, normalization, duplication removal, and validation.
Data Access Frequency	Multiple times daily by coaches, athletes, and analysts; real-time access during training and competitions.
Storage Estimate	~250–750MB daily; ~7.5–22.5GB monthly; estimated growth to 500GB+ annually, with 5x increase expected in 3 years.
Reporting & Visualization	Needs customizable dashboards, interactive visualizations, comparative views, mobile-friendly formats, export options, real-time alerts, predictive insights, and athlete-specific views. BI tools like Power BI or Tableau preferred.
Unique Identifiers/Timestamps	Yes – includes Athlete ID, Session ID, and timestamps for time-series analysis and syncing.
Daily/Weekly/Monthly Data Volume	~300MB daily average; peaks up to 1.125GB/day. Weekly: up to 7.875GB. Monthly: up to 33.75GB in peak periods.
Expected Growth Rate	Projected 5x increase over 3 years due to athlete expansion, more metrics, and increased session frequency.
Current Data Storage	Local (CSV/XLSX in GitHub repo), experimental data in Jupyter Notebooks. Planned integration with Strava/Garmin APIs. GitHub storage becoming insufficient as data scales.
Security Requirements	Must comply with Australian Privacy Act, APPs, NDB Scheme.

	Implement RBAC, 2FA, encryption, audit logging, and special protection for under-18 athletes. Data should be stored in Australia.
Querying/Analysis Frequency	Real-time (during sessions), daily batch (summaries), weekly and monthly analysis (trends), quarterly strategic reviews. System must support both real-time and batch querying.
Processing Needs	Yes – includes aggregation (daily/monthly summaries), filtering (invalid/outliers), transformation (unit standardization, derived metrics). Best handled at warehouse level and automated via ETL pipelines.
Expected Insights	Training load/recovery balance, performance trends, injury risk, sport-specific efficiency, environmental effects, recovery patterns, training zones, long-term trajectories, predictive models (fatigue, injury, performance forecasting). Used by coaches, athletes, and analysts to personalize and enhance performance strategies.

Notes for Planning:

- This data strongly supports the need for:
    - Structured relational storage (e.g., PostgreSQL or SQLite on-prem).
    - ETL pipelines for cleaning & transformation (can be Python scripts or tools like Airflow).
    - Real-time and batch processing.
    - Local BI visualization (Metabase or Superset as open-source options).
- JSON support should be future-proofed but isn't urgent in Sprint 1.
- This data could potentially be reused/shared with other projects (e.g., if Project 4 also uses session-based metrics), supporting combined solutions with logical data silos.