

AETHERSTREAM ANALYTICS ENGINEERING PRACTICE SET

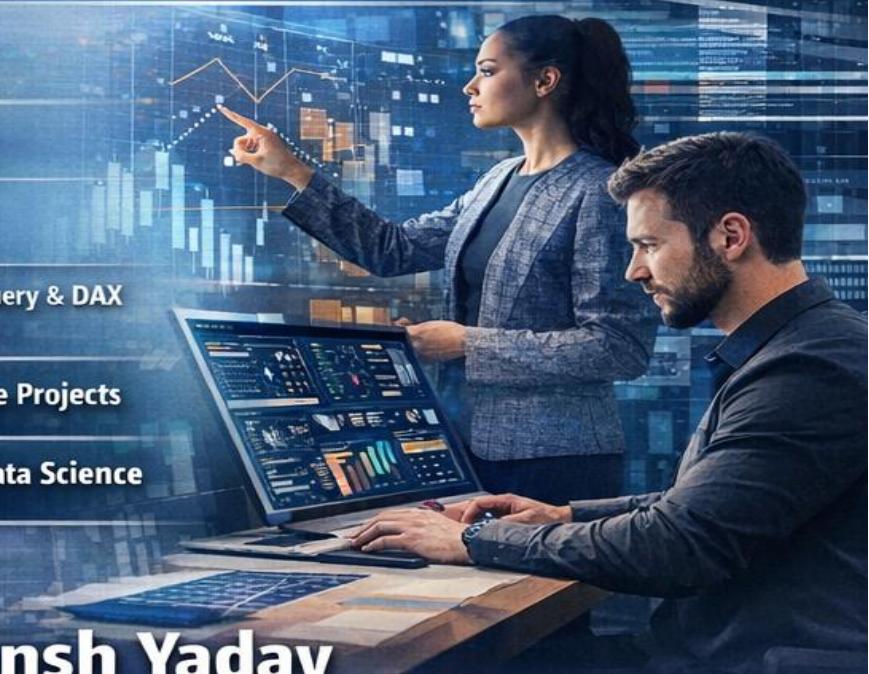
From Raw Data to Executive Intelligence

This practice set covers:

- | | |
|--|--|
|  Power Query (High-Scale ETL) |  DAX |
|  Power Pivot (Star Schema Modeling) |  |
|  Advanced DAX |  |
|  Statistical Analysis & Forecasting |  |
|  Automation (VBA + Python) |  |
|  Professional Dashboards | |

WHO THIS PRACTICE SET IS FOR:

- | |
|---|
|  Aspiring Data Analysts |
|  BI / Power BI Professionals |
|  Analytics Engineers |
|  Excel Users moving into Power Query & DAX |
|  Students Building Portfolio-Grade Projects |
|  Professionals Transitioning into Data Science |



By Shivansh Yadav

Hands-On, Industry-Grade Analytics Framework for Real-World Enterprise Challenges.

ABOUT THE AUTHOR

Shivansh Yadav is a dedicated learner in Data Science with strong hands-on experience in Data Analytics and Data Visualization. He has worked extensively with Microsoft Excel (Advanced), Power Query, Power Pivot, DAX, SQL, Python, and professional dashboard engineering.

This practice framework represents a structured learning journey built through real analytical workflows, practical experimentation, and deep exploration of modern BI concepts — from spreadsheet foundations to enterprise-grade modeling, automation, and predictive analytics.

The objective is to develop analytical thinking and system-level understanding before transitioning into full-scale Data Science.

For assignments, projects, and industry-focused learning resources across Python, SQL, Power BI, Machine Learning, DSA, and Analytics Engineering:

🔗 GitHub: <https://github.com/Venom-Shivu>

Connect professionally and follow ongoing technical work:

🔗 LinkedIn: <https://www.linkedin.com/in/the-venom/>

INTELLECTUAL PROPERTY NOTICE

© Shivansh Yadav — All Rights Reserved.

This practice set, curriculum structure, module organization, and associated materials are the intellectual property of Shivansh Yadav.

Unauthorized reproduction, redistribution, modification, or commercial use — in whole or in part — is strictly prohibited without explicit written permission from the author.

This material is intended solely for educational and professional development purposes.

⌚ Module 1: High-Scale Data Ingestion & ETL (Power Query)

Objective: Ingest and normalize fragmented, "dirty" data from global CDN logs and SAP exports.

📝 PROBLEM STATEMENTS

1. **[L1] The Multi-Region Log Merger:** Ingest 24 monthly CSVs of CDN traffic. Standardize "Timestamp" columns that alternate between US (MM/DD/YYYY) and UK (DD/MM/YYYY) formats.
2. **[L2] The Subscription Tier Unpivot:** A legacy Excel forecast has "Tiers" (Basic, Pro, Ultra) as column headers. Transform this into a 3-column "Long" table: [Date], [Tier], [Forecasted_Rev].
3. **[L2] The Hardware SKU Masking:** Extract Product IDs and Region Codes from a semi-structured string (e.g., `SKU:9921-ID|LOC:APAC|WH:SG`). Ensure no "Ghost Spaces" remain.
4. **[L3] The Recursive API Paginator:** Connect to the AetherStream Support API. Use `List.Generate` in M-Language to fetch all 10,000 tickets across 100 paginated results.
5. **[L2] The Weekend Hedge Logic:** Merge Sales data with a "Daily FX Rate" table. If a sale occurs on a Sunday (missing rate), use the rate from the previous Friday.
6. **[L1] The Scientific Notation Cleanup:** Telemetry data imported "Voltage" as Scientific Notation (1.2E+02). Force-convert this into a precise Decimal type for sensor analysis.
7. **[L2] The Fuzzy User Merge:** Join a "Marketing Lead" sheet with the "User Database" where names have 15% character variance (e.g., "Shivansh Yadav" vs "Shivansh Y.").
8. **[L3] The Folder-to-JSON Transformer:** Ingest a folder containing 5,000 JSON files of user device metadata. Flatten the "Nested Record" structure into a flat relational table.
9. **[L2] The Logic-Based Filtering:** Filter out internal IP addresses (192.x.x.x) from the "Public Traffic" table without hard-coding, using a dynamic parameter list.

10. [L3] **The Conditional Error Firewall:** Use `try...otherwise` logic in M-Language to replace `#DIV/0!` and `#VALUE!` errors in the "Latency" column with the regional median.

11. [L2] **The Case-Standardization Script:** Ensure all "City" names are converted to `PROPER` case and "Country Codes" are strictly `UPPER` case to prevent duplicate grouping in Pivot Tables.

12. [L3] **The Big Data Sample Engine:** Create a Power Query logic that loads only a 1% "Random Sample" of a 5-million row log file for rapid prototyping.

❖ SOLUTION GUIDELINES (MODULE 1)

- **Q1-Q4:** Utilize 'Get Data > From Folder'. For pagination, use `List.Generate(() => [Result = try Get(1) otherwise null...])`.
- **Q5-Q8:** Use 'Left Outer Join' followed by 'Fill Down' for currency gaps. Use `Json.Document` and the 'Expand' icon for nested data.
- **Q9-Q12:** Use 'Merge Queries' with Fuzzy Matching (0.8 threshold). Implement `Number.FromText` for scientific notation.

Module 2: Enterprise Relational Modeling (Power Pivot)

Objective: Engineer a "Single Source of Truth" using Star Schemas and the VertiPaq Columnar Engine.

PROBLEM STATEMENTS

13. **[L1] The Universal Temporal Spine:** Build a `Dim_Date` table starting from 2020. Include a "Fiscal Month" column where the year starts in July (AetherStream Fiscal Cycle).
14. **[L3] The Multi-Fact Granularity Bridge:** Connect a "Daily Sales" table and a "Monthly Sales Target" table to the same `Dim_Date` table using a Month-Year key.
15. **[L2] The Snowflake Melting Strategy:** You have 4 tables: [Product] -> [SubCat] -> [Category] -> [Division]. Denormalize them into one `Dim_Product` table for speed.
16. **[L3] The Role-Playing Date Architecture:** In a hardware logistics table, map `Order_Date`, `Ship_Date`, and `Return_Date` to one Date table using inactive relationships.
17. **[L2] The SCD Type 2 Trap:** A user changed their location from "India" to "USA." Ensure historical streaming revenue for June stays in "India" while July revenue moves to "USA."
18. **[L3] The Cross-Database Join:** Link a SQL Server "Transaction" table (5M rows) with an Excel "Product Catalog" in the Power Pivot model.
19. **[L2] The Many-to-Many Bridge:** Link "Movies" to "Genres" where one movie can have multiple genres using a "Bridge Table" to avoid circular dependencies.
20. **[L1] The UI Cleanup:** Hide 100% of Surrogate Keys (IDs) from the Pivot Table field list so users only see "Human-Readable" names.
21. **[L2] The Sort-Order Governance:** Fix the "Month Name" column so it sorts by the "Month Number" (Jan, Feb...) instead of alphabetically (Apr, Aug...).

22. **[L3] The Filter Direction Protocol:** Set the relationship between `Dim_User` and `Fact_Streams` to "Single" to prevent "Double Counting" in complex DAX measures.

23. **[L2] The Binary Optimization:** Save the final model as `*.xlsb` (Binary) and measure the file size reduction compared to `*.xlsx`.

24. **[L3] The Vertical Partitioning:** Remove 15 unused columns from the "Fact_Streams" table in Power Query to optimize the VertiPaq memory footprint.

SOLUTION GUIDELINES (MODULE 2)

- **Q13-Q16:** Use `CALENDARAUTO()` and `USERELATIONSHIP()` DAX for inactive roles.
- **Q17-Q20:** Implement "Surrogate Keys" in Power Query. Use 'Diagram View' to link tables; Right-click > 'Hide in Client Tools'.
- **Q21-Q24:** Select column > 'Sort by Column'. Monitor memory via 'File > Info'.

Module 3: Advanced Computational Engineering (DAX Mastery)

Objective: Programmatically define "North Star" metrics (Churn, Retention, LTV).

PROBLEM STATEMENTS

25. **[L2] The MTU (Monthly Total Users):** Count unique users who had at least 1 streaming session in the last 30 days, regardless of the date selected.
26. **[L3] The Retention Cohort Matrix:** Calculate the % of users who signed up in "Month 0" and are still active in "Month 1," "Month 2," etc.
27. **[L2] The YoY Growth (Fiscal):** Calculate "Year-over-Year" growth based on the AetherStream July-June fiscal cycle.
28. **[L3] The "Basket" Analysis:** Find the top 3 hardware accessories most frequently purchased together with the "AetherGlass" VR headset.
29. **[L2] The Lapsed User Flag:** Identify users who haven't streamed in 60 days but haven't canceled their subscription.
30. **[L3] The Dynamic Top-N Ranking:** Create a measure that sums sales for the "Top N" countries, where "N" is a value selected by a user via a slicer.
31. **[L2] The Semi-Additive Cash Balance:** Calculate "End of Month" cash balance which must not be summed across dates, only the last known value.
32. **[L3] The Global Benchmark Variance:** Calculate a region's revenue performance as a `% Variance` against the "Total Company Average."
33. **[L2] The Currency Switcher Logic:** Build a measure that switches the entire report from USD to EUR or INR based on a slicer selection.
34. **[L3] The Cumulative "Run Rate":** Calculate the "Year-to-Date" running total of revenue and project the "End of Year" total based on current velocity.
35. **[L2] The Rolling 7-Day Latency:** Calculate a moving average of server latency to smooth out daily spikes for executive reporting.

36. [L3] **The Churn Rate (Complex):** `(Users at Start - Users at End) / Users at Start`. Handle cases where "Users at Start" might be zero.

❖ SOLUTION GUIDELINES (MODULE 3)

- **Q25-Q28:** Use `CALCULATE(DISTINCTCOUNT(...), DATESINPERIOD(...))`. For cohorts, use `VAR` to capture Signup Month.
- **Q29-Q32:** Use `EXCEPT()` for lapsed users. Use `RANKX()` for Top-N. Use `ALL()` or `ALLEXCEPT()` for benchmarking.
- **Q33-Q36:** Use `LASTNONBLANK()` for semi-additive metrics. Use `SELECTEDVALUE()` for currency switching.

Module 4: Predictive Analytics & Risk Modeling

Objective: Quantify uncertainty using Statistical Process Control (SPC) and Monte Carlo.

PROBLEM STATEMENTS

37. **[L3] The Monte Carlo "Site Crash" Simulation:** Run 5,000 simulations of "Peak Concurrent Users" to find the probability of hitting the 1.2M server limit.
38. **[L3] The OLS Regression for Revenue:** Find the 'Slope' and 'Intercept' between "Marketing Spend" and "New Signups." Predict revenue for a \$5M spend.
39. **[L2] The A/B Test Significance:** AetherStream changed its "Buy Now" button color. Use a T-Test to prove if the 0.5% conversion lift is statistically significant ($P < 0.05$).
40. **[L2] The Hardware Failure Outliers:** Use Z-Scores to flag "AetherPods" with battery life that is 3 standard deviations away from the mean.
41. **[L3] The Demand Forecast (Holt-Winters):** Predict streaming hours for Q4 2026 using exponential smoothing that accounts for "Seasonality" (Winter peaks).
42. **[L2] The Price Elasticity Matrix:** Create a "What-If" table showing how a 10% price hike affects total LTV if churn increases by 2%.
43. **[L3] The Support Ticket Sentiment Score:** Use a "Weighted Keyword" logic in DAX to score tickets from -1 (Angry) to +1 (Happy).
44. **[L2] The Correlation Mapping:** Use the 'CORREL' function to find the link between "Internet Speed" (Mbps) and "Video Quality" (4K Adoption).
45. **[L3] The Solver Logistics Optimizer:** Find the cheapest shipping route from 5 global warehouses to 50 countries while meeting a 2-day delivery constraint.
46. **[L3] The Churn Propensity Z-Model:** Combine 5 behavioral metrics into a single "Risk Score" using a weighted non-linear formula.

47. [L2] **The Goal Seek for Break-Even:** Find the exact "Unit Price" needed for the AetherGlass VR to break even after \$200M in R&D.

48. [L3] **The SPC Control Chart:** Calculate the Upper (UCL) and Lower (LCL) control limits for "Server Response Time" using $\$3\sigma$ logic.

❖ SOLUTION GUIDELINES (MODULE 4)

- **Q37-Q40:** Use `NORM.INV(RAND(), Mean, SD)`. Use `LINEST` for regression. Use `T.TEST` for A/B testing.
- **Q41-Q44:** Use `FORECAST.ETS`. Build a 2-Variable Data Table for Sensitivity.
- **Q45-Q48:** Use the `Solver Add-in`. For SPC, calculate `Mean + (3 * STDEV.P)`.

Module 5: Strategic Visual Design & Automation

Objective: Design an Executive-grade "North Star" Dashboard following the F-Pattern.

PROBLEM STATEMENTS

49. **[L3] The CEO "Hero" Header:** Position the 3 most critical KPIs in the top-left "F-Pattern" zone. Use "Traffic Light" (Red/Green) indicators for target variance.
50. **[L2] The Sparkline Narrative:** Integrate 12-month trendline sparklines into a "Product Performance" table to show "Directional Velocity."
51. **[L2] The Dynamic Narrative Label:** Create a text box using formulas: "Our top region is " & [Top_Region] & " with " & [Growth_%] & " growth."
52. **[L3] The Drill-Through Deep-Dive:** Configure a map visual so clicking "Germany" opens a detailed "Berlin Infrastructure" report.
53. **[L2] The High Data-Ink Ratio Cleanup:** Take a default chart and remove gridlines, Y-axis, and legends. Use "Data Labels" only on peaks.
54. **[L3] The Scenario Toggle:** Build a "Best Case / Worst Case" toggle that updates all 10 charts on the dashboard simultaneously.
55. **[L2] The Accessibility Check:** Ensure all charts use "Color-Blind Safe" palettes (Blue/Orange) and have Alt-Text descriptions.
56. **[L3] The Tooltip Overlay:** Design a "Hidden" chart that appears only when a user hovers over a "Region" bar in the main chart.
57. **[L1] The Slicer Synchronization:** Ensure a single "Date Slicer" controls both the "Financials" tab and the "Marketing" tab.
58. **[L3] The VBA "Report Generator":** Write a macro that generates 50 individual PDFs (one for each regional manager) and saves them to a folder.
59. **[L3] The Python Correlation Heatmap:** Use the `matplotlib` library inside an Excel cell to plot a heatmap of 10 different business variables.

60. [L3] **The Enterprise Hand-off:** Publish the Excel Data Model to Power BI and set up a "Scheduled Refresh" via a SQL Gateway.

❖ SOLUTION GUIDELINES (MODULE 5)

- **Q49-Q52:** Follow Edward Tufte's design principles. Use 'Insert > Text Box > Formula Bar' for dynamic text.
- **Q53-Q56:** Use 'Sync Slicers'. Use 'Bookmarks' (Power BI) or 'Grouped Shapes' (Excel) for toggles.
- **Q57-Q60:** Use 'VBA (ActiveWorkbook.ExportAsFixedFormat)'. Use `=PY()` for Python integration.

Final Project Submission Checklist

- Star Schema: Is the model using a Star Schema with a central Fact table?
- Formatting: Are all currencies to 0 decimals and percentages to 1 decimal?
- Performance: Does the Pivot Table refresh in under 2 seconds?
- Governance: Is there a "DOCS" sheet explaining the logic of every measure?