**Workflow Overview**

1. **Ingestion (Bronze Layer)**
   * Pull Fingrid hydro power API data (JSON/CSV) using Spark notebook.
   * Store directly in Fabric Lakehouse (bronze tables).
2. **Cleaning & Transformation (Silver Layer)**
   * Normalize timestamps, units, missing values.
   * Join with **hydro plant metadata** (capacity, type).
3. **Analytics (Gold Layer)**
   * Aggregations: hourly/daily averages, regional summaries.
   * Calculate load factor = actual output / capacity.
   * Store as clean fact tables for dashboards.
4. **Streaming Simulation**
   * Use Fabric Eventstream (or Python generator) to push real-time power readings.
   * Spark Structured Streaming job consumes → stores into Lakehouse.
5. **CI/CD**
   * GitHub repo with all notebooks & SQL scripts.
   * GitHub Actions workflow runs unit tests (Spark job validation, SQL checks).
   * Deploy updated pipelines automatically into Fabric workspace.
6. **Visualization (Power BI)**
   * Simple dashboard with:
     + Hydro production trends (hourly, daily)
     + Regional breakdowns
     + Anomaly detection (e.g., sudden dips in output)

| **Dataset** | **What It Tells** | **Time Dimension** | **Granularity** | **Purpose** |
| --- | --- | --- | --- | --- |
| **Fingrid API** | Actual real-time generation (MW) | Minutes | National total | Observed performance |
| **Zenodo** | Modeled long-term capacity factors | Hourly (1981–2010) | Country/bidding zone | Historical baseline |
| **Metadata** | Installed plant capacity, type, location | Static | Plant-level | Reference and scaling factor |

| **Layer** | **Dataset** | **Granularity** | **What It Brings** | **How It Connects** |
| --- | --- | --- | --- | --- |
| **Silver** | silver\_fingrid\_hourly | Hourly | Real generation (MW) | Used as numerator in Observed CF |
| **Silver** | silver\_meta\_capacity\_fi | Static | Installed capacity (MW) by type | Denominator for Observed CF |
| **Silver** | silver\_zenodo\_fi\_cf | Hourly (1981–2010) | Historical capacity factors by hour/month | Baseline to compare observed CF |
| **Gold** | — | Hourly (aligned) | Joins all three → Observed CF, Historical CF, and Deviation | Final KPI output |

A black text on a white background

AI-generated content may be incorrect.

┌────────────────────────┐

│ Fingrid (3-min API) │

│ → Hourly Averages │

└────────────┬──────────┘

│

▼

silver\_fingrid\_hourly

│

┌─────────────────┼──────────────────┐

▼ ▼

┌──────────────────┐ ┌────────────────────┐

│ JRC Metadata (CSV) │ │ Zenodo JingHydro.csv │

│ Installed MW by type│ │ 1981–2010 CF by hour │

└──────────┬──────┘ └───────────┬───────┘

│ │

▼ ▼

silver\_meta\_capacity\_fi silver\_zenodo\_fi\_cf

└────────────┬──────────────────────────┘

▼

┌────────────────────┐

│ GOLD LAYER │

│ Observed CF = │

│ Fingrid / JRC │

│ Historical CF = │

│ Zenodo avg\_cf │

│ Deviation = Δ │

└────────────────────┘

**Hydropower Performance Analysis — From Data Pipeline to Power BI Dashboard**

This project demonstrates the end-to-end development of a modern data-engineering pipeline and analytics model for assessing Finland’s hydropower efficiency against long-term climatic baselines. The objective was to integrate real-time operational data with historical hydrological benchmarks, transforming raw feeds into actionable insights through Microsoft Fabric and Power BI.

The process began with data ingestion at the **Bronze layer**, where three complementary sources were collected.

1. **Fingrid API** provided real-time electricity generation from Finnish hydro power plants at 3-minute intervals. These records reflect the country’s live production output in megawatts (MW).
2. **Zenodo’s JingHydro dataset** offered nationally aggregated hourly capacity-factor data for 30 climatic years (1981-2010), serving as a long-term baseline of hydro potential under historical conditions.
3. **The JRC Hydro-Power-Plant Database** supplied metadata on installed capacities and plant typologies (run-of-river, storage, and pumped-storage), giving crucial context for calculating efficiency metrics.

At the **Silver layer**, each dataset was cleaned and standardized. Fingrid data were resampled from 3-minute intervals to hourly averages and stored as silver\_fingrid\_hourly. Zenodo’s European-wide data were filtered for Finland and separated into run-of-river and storage components (silver\_zenodo\_fi\_cf). Metadata were aggregated to produce Finland’s total installed hydro capacity (silver\_meta\_capacity\_fi).

The **Gold layer** combined these curated sources into analytics-ready tables. Observed capacity factors were computed by dividing actual generation (Fingrid) by total installed capacity (metadata). Historical capacity factors were drawn from Zenodo’s long-term averages. A deviation metric captured how current operational performance diverged from climatological expectations. Monthly aggregates produced two key tables:

* gold\_cf\_deviation\_monthly — observed CF, historical CF, and deviation.
* gold\_zenodo\_baseline\_monthly — 30-year mean and standard deviation of historical CFs.

Finally, results were presented through **Power BI visualizations** connected directly to the Fabric Lakehouse in Direct Lake mode:

1. **Line Chart:** plotted avg\_observed\_cf, avg\_historical\_cf, and baseline\_avg\_cf to compare current October performance with the 30-year baseline.
2. **Column Chart:** visualized avg\_deviation by month, highlighting whether production was above or below expectations (−3.2 % for October 2025).
3. **Area Band Chart:** used baseline\_avg\_cf ± baseline\_std\_cf to illustrate natural historical variability, forming a shaded confidence band around the baseline trend.

Although the pilot currently covers ten days of Fingrid data (October 2025), the architecture is fully scalable: as new months are ingested, Power BI visuals will automatically extend. The project showcases the integration of real-time cloud ingestion, PySpark transformations, and business-ready analytics in Microsoft Fabric. It bridges energy-sector domain data and modern engineering practices—turning raw hydropower measurements into intuitive, evidence-based operational intelligence.

Next, **ETL → analytics → AI augmentation:**