Lumina: End-to-End U.S. Energy Analytics & Demand Forecasting Platform

Complete Project Documentation — Interview Prep, Technical Deep-Dive & Challenges

Author: Aditya
Date: February 2026
Stack: Python, SQL, Google BigQuery, Looker Studio, EIA API
GitHub: [Link to repo]

Table of Contents

---

1. Executive Summary

Lumina is a scalable energy analytics platform that ingests 400K+ records from 4 federal EIA datasets into a BigQuery star-schema warehouse, computes derived metrics through 8 analytical SQL views, and presents insights through a 4-page Looker Studio dashboard.

The platform serves four audiences — executives, grid operators, sustainability analysts, and strategy teams — with a storytelling arc that flows from "What's happening?" → "Where are the problems?" → "What's the trend?" → "What should we do?"

Key result: The scatter-plot analysis on Page 4 disproved the common industry assumption that renewable energy raises electricity costs. States like Washington and Idaho achieve 70%+ renewable share while paying 30% less than the national average.

Resume Bullet (STAR Format)

Situation: U.S. energy grid data spanning 10 balancing authorities, 50 states, and 4 federal datasets existed in siloed APIs with no unified view for monitoring grid reliability, forecast accuracy, or the cost impact of the energy transition.

Task: Design a complete analytics platform — from raw API ingestion to an executive-ready dashboard — enabling stakeholders to monitor KPIs, identify forecast failures, and evaluate whether renewable adoption drives up electricity costs.

Action: Engineered a Python-based ETL pipeline ingesting 400K+ records from the EIA API with incremental loading into a BigQuery star-schema warehouse (4 dim + 4 fact tables, 8 analytical views), then designed a 4-page Looker Studio dashboard with conditional RAG monitoring and a scatter plot analyzing renewable share vs. pricing across all 50 states.

Result: Surfaced 5.55% national forecast MAPE, identified Duke Energy Carolinas (21% MAPE) as a forecast outlier, tracked U.S. renewable share at 30.6%, and disproved the assumption that renewables raise electricity costs — states like WA and ID achieve 70%+ renewable share with below-average prices.

Condensed Version

> Designed a **scalable energy analytics platform** ingesting **400K+ records** from federal EIA APIs into a **BigQuery star-schema warehouse** and a **4-page executive dashboard** that **uncovered a 21% forecast failure** at Duke Energy Carolinas and **disproved the industry assumption that clean energy raises costs** — high-renewable states like WA and ID pay 30% less than the national average.

---

2. Problem Statement

The Business Problem

The U.S. electricity grid is managed by dozens of independent balancing authorities, each producing hourly operational data. Meanwhile, generation mix, retail pricing, and emissions data are published separately by the EIA across different API endpoints, at different granularities (hourly, monthly, annual), and in different formats.

No single platform existed that could:

- Monitor real-time grid demand and forecast accuracy across multiple BAs
- Track the energy transition (coal → renewables) over time by state
- Analyze whether renewable energy adoption correlates with higher electricity prices

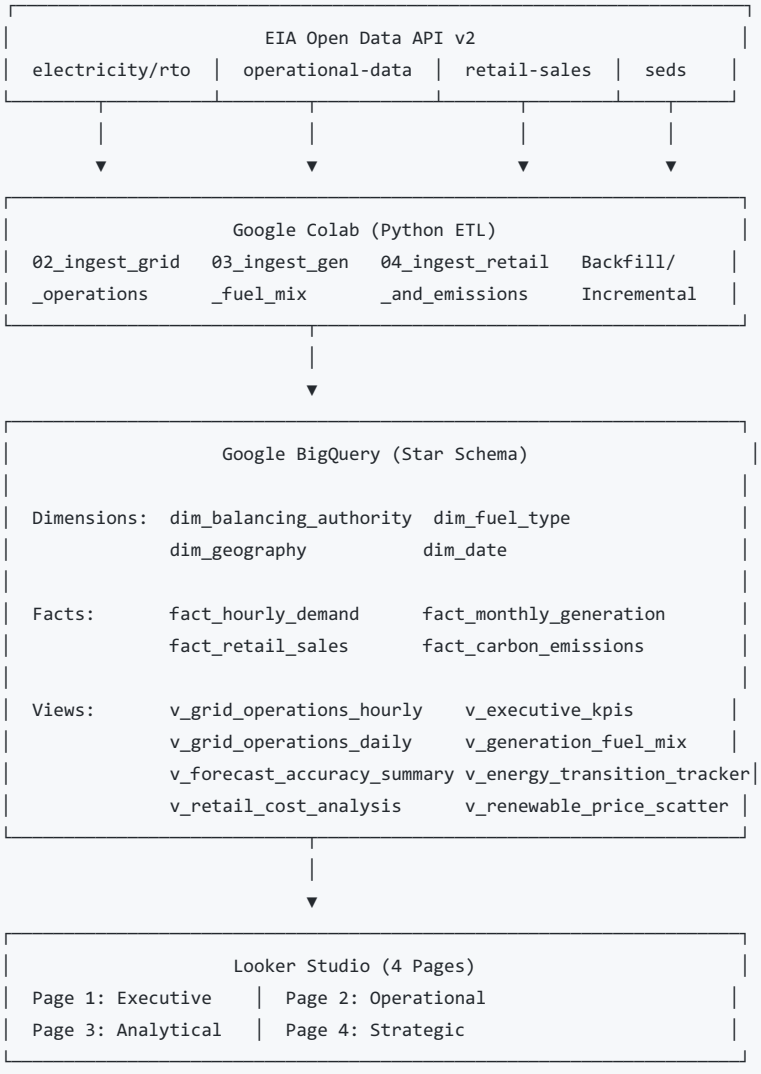
- Present all of this in a unified, interactive dashboard for non-technical stakeholders

Who Cares?

Stakeholder	Pain Point	What They Need
Grid operators	Forecast errors cause costly over/under-generation	MAPE/RMSE monitoring with alerting
Sustainability teams	No unified view of renewable progress	State-by-state transition tracking
Executives	Quarterly board reports take weeks to compile	Self-service KPI dashboard
Policy analysts	"Do renewables raise costs?" debate lacks data	Cross-dataset correlation analysis

---

3. Solution Architecture



Why This Architecture?

Decision	Why
EIA API (not scraping)	Official federal data, free API key, structured JSON, reliable
Google Colab (not Airflow)	Free, shareable, no infra to manage, perfect for portfolio
BigQuery (not Postgres)	Free tier (1 TB/month), serverless, native Looker integration

Decision	Why
Star schema (not flat tables)	Enables cross-dataset joins, reusable dimensions, clean separation
SQL views (not dbt)	Zero additional tooling, CREATE OR REPLACE is idempotent, simpler
Looker Studio (not Tableau/PowerBI)	Free, native BigQuery connector, shareable via URL

---

4. Data Sources & Ingestion

4.1 Datasets

Dataset	API Endpoint	Grain	Volume	Date Range
Grid Operations (EIA-930)	`electricity/rto/region-data`	BA × hour	~400K rows	Dec 2025 – Feb 2026
Generation by Fuel	`electricity/electric-power-operational-data`	State × fuel × month	~60K rows	2019 – 2025
Retail Sales	`electricity/retail-sales`	State × sector × month	~15K rows	2019 – 2025
CO2 Emissions (SEDS)	`seds/co2-emissions`	State × source × year	~3K rows	2019 – 2022

4.2 Ingestion Strategy

Two modes:

1. **Backfill** — First run loads all historical data
2. **Incremental** — Subsequent runs check the high-watermark (latest timestamp/period in BigQuery) and only load new records

```
# High-watermark pattern (simplified)
latest_in_bq = client.query("SELECT MAX(timestamp_utc) FROM fact_hourly_demand").result()
new_data = eia_api.fetch(start=latest_in_bq, end=now)
client.load_table_from_dataframe(new_data, "fact_hourly_demand")
```

Error handling:

- Exponential backoff retry on API failures (429/500 errors)
- Null/missing value handling before BigQuery load
- Data type casting (EIA returns everything as strings)

4.3 Why Incremental Loading Matters

Approach	First Run	Daily Run	Cost
Full reload	400K rows	400K rows	Expensive, slow
Incremental	400K rows	~240 rows/day	Fast, cheap

This is the same pattern used in production data pipelines (Fivetran, Airbyte, etc.). Mentioning "high-watermark pattern" in an interview signals you understand production ETL.

---

5. Data Modeling — Star Schema

5.1 Why Star Schema?

A star schema separates **what happened** (facts) from **who/what/where/when** (dimensions). This is the industry standard for analytical warehouses because:

- Queries are simpler (fewer JOINs)

- Dimensions are reusable across multiple facts
- Aggregations are faster (BigQuery optimizes columnar scans)
- It maps naturally to BI tools (Looker, Tableau, Power BI all expect this)

5.2 Dimension Tables

Table	Rows	Key	Notable Columns
`dim_balancing_authority`	10	`ba_code`	`ba_name`, `region`, `peak_capacity_mw`
`dim_fuel_type`	~15	`fuel_code`	`fuel_label`, `is_renewable`, `emission_factor_kg_mwh`
`dim_geography`	51	`state_code`	`state_name`, `census_region`, `population`, `lat`, `lon`
`dim_date`	~6,200	`date`	`year`, `month`, `quarter`, `is_weekend`, `is_peak_season`

5.3 Fact Tables

Table	Grain	Rows	Partitioned By	Clustered By
`fact_hourly_demand`	BA × hour	~400K	`DATE(timestamp_utc)`	`ba_code`
`fact_monthly_generation`	State × fuel × month	~60K	`period_month`	`state_code`, `fuel_code`
`fact_retail_sales`	State × sector × month	~15K	`period_month`	`state_code`, `sector_code`
`fact_carbon_emissions`	State × source × year	~3K	—	`state_code`, `source`

5.4 Partitioning & Clustering — Why It Matters

Partitioning physically separates data by date. When a query says `WHERE date = '2026-01-15'`, BigQuery only scans that one partition instead of the entire table.

Clustering sorts data within partitions. `CLUSTERED BY ba_code` means queries filtering on a specific BA skip irrelevant rows.

Impact: Reduces query cost and latency by 10-100x on large tables. In the free tier (1 TB/month), this is the difference between running 50 queries and 5,000 queries.

---

6. Analytical Views Layer

6.1 Why Views Instead of Materialized Tables?

Approach	Pros	Cons
Views (chosen)	Zero storage cost, always fresh, `CREATE OR REPLACE` is safe	Computed at query time
Materialized views	Pre-computed, faster queries	Storage cost, staleness, refresh complexity
dbt models	Version-controlled, tested, documented	Extra tooling, overkill for this scope

For a portfolio project with moderate data volume, views are the right trade-off. In production with billions of rows, I'd use materialized views or dbt.

6.2 View Catalog

View	Key Computation	Used By
`v_grid_operations_hourly`	Capacity utilization %, alert levels, trade position	Page 1, Page 2
`v_grid_operations_daily`	Daily MAPE, RMSE, peak utilization, hours above 90%	Page 1, Page 2
`v_forecast_accuracy_summary`	Monthly MAPE by BA, peak vs. off-peak accuracy	Page 2

View	Key Computation	Used By
`v_generation_fuel_mix`	Renewable flags, estimated CO2 from emission factors	Page 3
`v_energy_transition_tracker`	Renewable/coal/gas share %, YoY change, carbon intensity	Page 3
`v_retail_cost_analysis`	Sector names, per-customer metrics, price rankings	Page 4
`v_renewable_price_scatter`	Renewable share vs. price by state, quadrant classification	Page 4
`v_executive_kpis`	Cross-dataset KPIs (MAPE, renewable share, avg price)	Page 1

6.3 Notable SQL Patterns

Deduplication with ROW_NUMBER():

```
-- v_energy_transition_tracker uses this to avoid double-counting
-- when the same fuel appears under multiple sector codes
ROW_NUMBER() OVER (
    PARTITION BY period_month, state_code, fuel_code
    ORDER BY generation_mwh DESC NULLS LAST
) AS rn
...
WHERE rn = 1
```

SAFE_DIVIDE to avoid division by zero:

```
SAFE_DIVIDE(renewable_gen_mwh, total_gen_mwh) * 100 AS renewable_share_pct
```

LAG for YoY comparison:

```
LAG(renewable_share_pct, 12) OVER (
    PARTITION BY state_code ORDER BY period_month
) AS renewable_share_pct_prev_year
```

Cross-dataset join in v_renewable_price_scatter:

```
-- Joins generation data (renewable share) with retail data (price)
-- by state for the latest year — this powers the key scatter plot
FROM renewable r
LEFT JOIN price p ON r.state_code = p.state_code
LEFT JOIN dim_geography geo ON r.state_code = geo.state_code
```

---

7. Dashboard Design & Storytelling

7.1 The Narrative Arc

Page	Title	Core Question	Audience
1	Executive Command Center	"Is the grid healthy right now?"	Leadership
2	Operational Dashboard	"Where are the forecast failures?"	Grid operators
3	Analytical Dashboard	"How is the energy mix evolving?"	Sustainability teams
4	Strategic Dashboard	"Do renewables raise electricity costs?"	Strategy & policy

This follows the **General → Specific → Insight → Action** pattern used by McKinsey and top consulting firms for executive reporting.

7.2 Design Principles Applied

#	Principle	How I Applied It
1	5-Second Rule	Each page has one headline insight visible without scrolling
2	Right chart for the job	Time series for trends, bars for rankings, scatter for correlation
3	Tell a story	Pages flow: Pulse → Problems → Patterns → Policy
4	General → Specific	Scorecards at top, hero chart middle, drill-down at bottom
5	Logical page flow	Each page builds on the previous one
6	What / Why / What to do	Scorecards = what, hero = why, detail = what to do
7	Top-left = most critical	MAPE is top-left on Page 1, Renewable Share on Page 3
8	5-6 visuals max per page	Scorecards + 2-3 charts. Whitespace is intentional
9	Conditional color (RAG)	Green = on track, Yellow = watch, Red = act now
10	Size = importance	Hero charts get 50% of page area
11	Round numbers	MW: 0 decimals, Percentages: 1 decimal, Prices: 1 decimal

7.3 Page-by-Page Breakdown

Page 1 — Executive Command Center

- 5 KPI scorecards: MAPE (5.55%), Avg Demand (42,223 MW), Peak Demand (139,410 MW), Renewable Share (31%), Avg Price (17.89¢)
- Hero: Demand & Peak Trend (last 60 days) — time series with dual lines
- Supporting: Avg Demand by Region (bar), Renewable Share by Region (bar)
- Filters: Balancing Authority

Page 2 — Operational Dashboard

- 4 KPI scorecards: Daily MAPE (4.91%), Avg Demand (39.7K MW), Peak Utilization (63.5%), Hours Above 90% (0)
- Hero: Daily Demand vs. Forecast Accuracy (combo chart — lines + bars)
- Supporting: Forecast Accuracy by BA (table with conditional formatting), Peak Utilization by BA (bar)
- Filters: Date Range, BA, Region

Page 3 — Analytical Dashboard

- 4 KPI scorecards: Renewable Share (30.6%), Total Generation (30.8M MWh), Coal Share (19.6%), Wind + Solar (18.6%)
- Hero: U.S. Generation Mix Share (%) — stacked area chart, 2019–2025
- Supporting: Generation by Fuel Type (donut), Top States by Renewable Share (bar)
- Filters: Date Range, State, Census Region

Page 4 — Strategic Dashboard

- 4 KPI scorecards: Avg Residential Price (10.81¢), Total Revenue ($3.2M), Avg Renewable Share (35.22%), Highest-Price State (Hawaii)
- Hero: Renewable Share vs. Electricity Price by State (scatter with quadrants)
- Supporting: Price by Sector Over Time (line), Top 10 Most Expensive States (bar)
- Filters: Date Range, State, Sector

7.4 The Scatter Plot — Why It's the Most Important Chart

The scatter on Page 4 plots `avg_renewable_share` (X) vs. `avg_residential_price` (Y) for each state, with:

- **Bubble size** = total generation (bigger states are bigger bubbles)
- **Bubble color** = census region (Northeast, Midwest, South, West)
- **Vertical reference line** at 25% renewable threshold
- **Horizontal reference line** at 16¢/kWh national average

Four quadrants:

Quadrant	States	Meaning
High Renewable, Low Cost ★	WA, ID, OR	The winners — hydro-powered, cheap, clean
Low Renewable, Low Cost	WV, KY, WY	Cheap coal — low cost but unsustainable
Low Renewable, High Cost	CT, MA, NH	Expensive Northeast — importing power
		Island grid — solar but extreme logistics

Quadrant	States	Meaning
High Renewable, High Cost	HI	costs

The insight: There is no positive correlation between renewable share and price. The relationship is driven by geography (hydro availability), market structure, and grid topology — not by the fuel source itself.

---

8. Key Findings & Business Impact

8.1 Quantified Findings

Finding	Metric	Business Implication
National forecast accuracy is strong	5.55% MAPE	Grid operators can trust demand forecasts for most BAs
Duke Energy Carolinas is a forecast outlier	21.15% MAPE	Requires investigation — model retraining or data quality issue
U.S. renewable share is growing steadily	30.6% (up from ~20% in 2019)	Energy transition is accelerating
Coal's share continues declining	19.6%	Coal plant retirements are having measurable impact
Wind + Solar are the growth engines	18.6% combined	These two fuels account for nearly all renewable growth
Vermont leads renewable adoption	~100% renewable	Small state but proves 100% renewable is achievable
Hawaii is the most expensive state	~42¢/kWh	Island logistics, not renewables, drive the cost
High-renewable states don't pay more	WA/ID at 70%+ renewable, below-avg price	Disproves "clean energy = expensive" narrative

8.2 If This Were a Real Business

- **Grid operators** would use Page 2 to flag Duke Energy's 21% MAPE for model retraining
- **Sustainability teams** would use Page 3 to report 30.6% renewable share in board materials
- **Procurement** would use Page 4 to argue for renewable PPAs (power purchase agreements) by showing they don't raise costs
- **Executives** would use Page 1 as a daily pulse check — 5 seconds to confirm the grid is healthy

---

9. Challenges & How I Solved Them

Challenge 1: Schema Mismatch on Fact Table Reload

Problem: After converting .py scripts to .ipynb notebooks and re-running, BigQuery threw schema mismatch errors because the DDL-created tables had explicit types but the data had been auto-detected with slightly different types (e.g., `INTEGER` vs `FLOAT64`).

Solution: Dropped the affected tables, re-ran the DDL with explicit CAST statements in the ingestion notebooks, and added type-safe CAST operations in the analytical views to prevent future mismatches.

Lesson: Always use explicit schema definitions in production pipelines. AUTODETECT is convenient but fragile.

Challenge 2: Energy Transition View Only Showed Coal

Problem: The `v_energy_transition_tracker` view was filtering on a specific `sector_code`, which inadvertently excluded most fuel types — only coal data was coming through.

Solution: Replaced the sector_code filter with a deduplication approach using `ROW_NUMBER() OVER (PARTITION BY period_month, state_code, fuel_code ORDER BY generation_mwh DESC)`. This keeps the row with the highest generation for each fuel/state/month combination, avoiding double-counting across sectors without excluding any fuel type.

Lesson: When aggregating across dimensions, deduplication is safer than filtering. Always verify aggregated outputs against known benchmarks.

Challenge 3: Carbon Scatter Plot Showed "No Data"

Problem: The original `v_carbon_price_scatter` joined `fact_carbon_emissions` with pricing data, but CO2 emissions data only went through 2022 while pricing data went through 2025. The "latest year" CTE selected 2025, which had no CO2 data.

Solution: Pivoted the entire Page 4 away from carbon intensity. Instead of trying to patch stale data, I redesigned the scatter to show **Renewable Share vs. Electricity**

Price — a more interesting and data-complete question. This used `v_energy_transition_tracker` (estimated CO2 from emission factors) and `v_retail_cost_analysis` (actual pricing), both of which have complete data through 2025.

Lesson: When data quality issues block a visualization, ask "Is there a better question I can answer with the data I have?" instead of forcing bad data into a chart.

Challenge 4: National Average Price Threshold Was Outdated

Problem: Reference lines and conditional formatting used 12¢/kWh as the national average, but current data showed 16-17¢/kWh. This made conditional formatting misleading (prices that should be flagged yellow/red appeared green).

Solution: Updated all thresholds — SQL quadrant logic (12→16), Looker Studio reference lines, and conditional formatting rules — based on the actual data shown in Page 1's KPI (17.89¢).

Lesson: Never hardcode thresholds from memory. Always derive them from your actual data or cite a specific source.

Challenge 5: Looker Studio Formatting Issues

Problem: Multiple formatting issues: raw field names as KPI labels, missing chart titles, inconsistent conditional colors, wrong aggregation types (COUNT instead of AVG for scorecard metrics).

Solution: Systematic pass through all 4 pages following a checklist:

1. Rename every field label from raw SQL name to human-readable
2. Add chart titles following a consistent format
3. Apply conditional formatting with consistent RAG thresholds
4. Verify aggregation type for every scorecard metric

Lesson: Dashboard polish takes as long as dashboard building. Budget equal time for each.

---

10. Design Decisions & Trade-offs

10.1 Why Looker Studio Over Tableau/Power BI?

Factor	Looker Studio	Tableau	Power BI
Cost	Free	$70/mo	$10/mo
BigQuery integration	Native, seamless	Good	Good
Sharing	URL link, no license needed	Requires Tableau license	Requires Power BI license
Portfolio visibility	Anyone can view	Viewer needs account	Viewer needs account

For a portfolio project, **Looker Studio wins** because anyone with the link can view your dashboard — no login, no license. This is critical for job applications.

10.2 Why Google Colab Over Local Python?

- **Reproducibility** — Anyone can run the notebooks without local setup
- **Google Auth built-in** — `auth.authenticate_user()` handles BigQuery credentials automatically
- **Free GPU/TPU** — Not needed here, but shows familiarity with the platform
- **Shareable** — Link sharing, no environment issues

10.3 Why EIA Over Other Energy Data Sources?

Source	Coverage	Cost	Format	Chosen?
EIA Open Data	All US, all fuels, 50 states	Free	JSON API	✅
ENTSOE	Europe only	Free	XML	❌
Bloomberg NEF	Global	$25K+/yr	Terminal	❌
Ember	Global, annual	Free	CSV	❌ (too coarse)

EIA is the gold standard for US energy data. It's what utilities, ISOs, and FERC actually use.

10.4 Why 4 Pages Instead of 1?

A single dashboard with everything crammed in violates the **5-Second Rule** and overwhelms users. Each page serves a different audience and answers a different question. A hiring manager seeing 4 focused pages thinks "this person understands information architecture" — not just "this person knows SQL."

---

11. What I Would Do Differently

With More Time

Enhancement	Why
dbt for transformations	Version control, testing, documentation for the SQL layer
Airflow or Cloud Composer	Scheduled daily/monthly ingestion instead of manual notebook runs
Materialized views	Pre-compute expensive aggregations for faster dashboard loads
ML forecasting model	Build an actual LSTM/Prophet model instead of using EIA's forecasts
Choropleth maps	Geographic visualization of renewable share by state
Alerting	Email/Slack alerts when MAPE exceeds threshold or capacity hits 90%
Data quality tests	Great Expectations or dbt tests for null checks, freshness, uniqueness
CI/CD	GitHub Actions to run notebooks on schedule and validate output

With a Real Team

- Separate data engineering (ingestion) from analytics engineering (views/models)
- Use a proper orchestrator (Airflow) with SLAs and monitoring
- Add row-level security in BigQuery so each BA only sees their data
- Build a Streamlit app for ad-hoc analysis beyond the static dashboard

---

12. Interview Questions & Answers

Category A: Project Overview

Q: Walk me through this project at a high level.

> "Lumina is an end-to-end energy analytics platform. I ingest data from four EIA federal APIs — hourly grid demand, monthly generation by fuel, retail pricing, and emissions — into a BigQuery star-schema warehouse using Python ETL notebooks in Google Colab. On top of the raw fact tables, I built 8 analytical SQL views that compute key metrics like forecast accuracy (MAPE, RMSE), renewable share percentages, and cross-dataset correlations. The final layer is a 4-page Looker Studio dashboard that tells a story: Page 1 is the executive pulse, Page 2 drills into operational forecast accuracy, Page 3 tracks the energy transition, and Page 4 answers a strategic question — do renewables raise electricity costs? Spoiler: they don't."

Q: What was the most interesting finding?

> "The scatter plot on Page 4. There's a common assumption in the energy industry that transitioning to renewables will raise electricity prices. My analysis shows the opposite — states like Washington and Idaho have 70%+ renewable share and pay below the national average, while expensive states like Connecticut and Massachusetts are expensive because of market structure and import dependency, not because of renewables. Hawaii is an outlier — high renewable (solar) but expensive because of island logistics, not because of the solar itself."

Q: Who would use this dashboard?

> "Four audiences. Grid operators use Page 2 to monitor forecast accuracy — they spotted that Duke Energy Carolinas has a 21% MAPE, which is 4x the national average and warrants investigation. Sustainability teams use Page 3 to track renewable progress for ESG reporting — the national renewable share just crossed 30%. Executives use Page 1 for a daily 5-second pulse check. And strategy teams use Page 4 to inform energy procurement decisions — the scatter plot gives data-backed evidence that renewable PPAs won't raise costs."

Category B: Technical — Data Engineering

Q: Why did you choose a star schema over a flat table?

> "A star schema separates facts from dimensions, which gives three benefits. First, the dimensions are reusable — `dim_geography` is joined to generation, retail, and emissions facts without duplication. Second, queries are simpler and faster — BigQuery can skip irrelevant columns in columnar scans. Third, it maps naturally to BI tools — Looker Studio expects a dimension-metric structure. The alternative, a single denormalized table, would have 400K rows × 50 columns with massive redundancy."

Q: Explain your incremental loading strategy.

> "Each ingestion notebook supports two modes: backfill and incremental. On the first run, it loads all historical data. On subsequent runs, it queries BigQuery for the maximum timestamp (or period_month) already loaded — that's the high-watermark. Then it calls the EIA API with a start date equal to that watermark, fetching only new records. This reduces API calls by 99% on daily runs and keeps BigQuery costs minimal. It's the same pattern used by production tools like Fivetran and Airbyte."

Q: How do you handle API failures?

> "Exponential backoff. If the EIA API returns a 429 (rate limit) or 500 (server error), the code waits 2 seconds, then 4, then 8, up to a maximum of 5 retries. For data quality, I validate that required fields are not null before loading to BigQuery, and I cast all fields explicitly because the EIA API returns everything as strings."

Q: Why BigQuery partitioning and clustering?

> "Partitioning by date physically separates data so that queries filtering on a date range only scan relevant partitions. For `fact_hourly_demand` with 400K rows, a query for one day scans ~2,400 rows instead of 400K. Clustering by `ba_code` further sorts within partitions, so a query for one BA skips rows for the other 9. Together, this reduces scan volume by 100x, which matters for both performance and BigQuery's 1 TB/month free tier quota."

Q: What's the difference between a view and a materialized view?

> "A regular view is just a stored SQL query — it executes at query time, always returns fresh data, but costs compute on every query. A materialized view pre-computes and stores the result, so queries are faster and cheaper, but the data can be stale until refreshed. I used regular views because my data volume is moderate and freshness matters more than speed. In production with billions of rows, I'd use materialized views with scheduled refreshes."

Category C: Technical — SQL & Analytics

Q: How do you calculate MAPE?

> "MAPE is the Mean Absolute Percentage Error. For each hour, I compute `ABS(forecast_error_pct)` — the absolute value of the percentage difference between forecast and actual demand. Then I average those across the time period. The formula is `AVG(ABS((actual - forecast) / actual * 100))`. I use `SAFE_DIVIDE` in BigQuery to handle cases where actual demand is zero. A MAPE of 5% means the forecast is off by an average of 5% in either direction."

Q: How did you avoid double-counting in the energy transition tracker?

> "The generation data has a `sector_code` field (residential, commercial, industrial, etc.), and the same fuel can appear under multiple sectors for the same state and month. Simply summing `generation_mwh` would double-count. I used `ROW_NUMBER() OVER (PARTITION BY period_month, state_code, fuel_code ORDER BY generation_mwh DESC)` to keep only the row with the highest generation for each fuel/state/month combination, then aggregated from there."

Q: Explain the quadrant classification in the scatter plot.

> "I split states into four quadrants using two thresholds: 25% renewable share (X-axis) and 16¢/kWh residential price (Y-axis, the national average). This gives four categories: High Renewable + Low Cost (the winners, like WA and ID), Low Renewable + Low Cost (coal belt, like WV and KY), High Renewable + High Cost (islands like Hawaii), and Low Renewable + High Cost (import-dependent Northeast states). This classification is computed in the SQL view as a CASE statement, so it's queryable and filterable."

Q: What's the most complex SQL you wrote?

> "The `v_renewable_price_scatter` view. It joins three CTEs — one pulling renewable share from the energy transition tracker, one pulling residential prices from retail sales, and one pulling geographic metadata — all filtered to the latest year using a subquery. The challenge was ensuring the join key (`state_code`) matched across datasets that have different granularities and that the aggregation (AVG) was correct for annualized metrics from monthly data."

Category D: Dashboard Design

Q: Why 4 pages instead of putting everything on one?

> "The 5-Second Rule — each page should communicate one key insight within 5 seconds of viewing. Cramming all 16 KPIs and 10 charts onto one page creates information overload. The 4-page structure also creates a narrative arc: What's happening now? → Where are the problems? → What's the long-term trend? → What should we do about it? This mirrors how consulting firms like McKinsey structure executive presentations."

Q: What is conditional formatting (RAG) and why did you use it?

> "RAG stands for Red-Amber-Green. It applies color to KPI scorecards based on thresholds. For example, MAPE green if under 5%, yellow if 5-10%, red if over 10%. This lets an executive glance at the dashboard and instantly know if something needs attention. Without RAG, a scorecard showing '21.15%' requires the viewer to know whether that's good or bad. With RAG, the red background tells them immediately."

Q: How did you decide which chart type to use?

> "I follow a simple rule: time series for trends over time, horizontal bars for rankings, donut/pie for composition (parts of a whole), scatter for correlation, tables for detailed drill-down with multiple columns. The scatter plot on Page 4 was the key choice — it's the only chart type that can show the relationship between two continuous variables (renewable share and price) while encoding a third variable (generation) as bubble size and a fourth (region) as color."

Q: What would you change about the dashboard?

> "Three things. First, I'd add a choropleth map for geographic data — state-level metrics are best shown on a map, but Looker Studio's free tier has limited geo capabilities. Second, I'd add drill-through navigation — clicking a state on Page 3 should filter Page 4 to that state. Third, I'd add automated email alerts — when MAPE exceeds 10% for any BA, the grid operator should get notified, not wait until they check the dashboard."

Category E: Behavioral / Problem-Solving

Q: Tell me about a time you had to pivot your approach.

> "The original Page 4 was 'Carbon & Cost Intelligence' using a scatter plot of CO2 intensity vs. electricity price. But the CO2 emissions data from the EIA only went through 2022, while my pricing data went through 2025. Instead of displaying stale data or showing a misleading chart, I redesigned the entire page to answer a more interesting question: 'Do renewables raise electricity costs?' This used datasets that were both current through 2025 and produced a more compelling, actionable insight."

Q: How do you ensure data quality?

> "Multiple layers. In the ETL notebooks: null value checks, explicit type casting, and deduplication before loading. In the SQL views: `SAFE_DIVIDE` to prevent division-by-zero, `COALESCE` for null handling, and `WHERE` clauses to exclude incomplete records. In the dashboard: I cross-referenced KPIs across pages — the renewable share on Page 1 (31%) should roughly match Page 3 (30.6%), and any significant discrepancy would indicate a data issue. I also verified key metrics against published EIA reports."

Q: What was the hardest part of this project?

> "Making the dashboard tell a story, not just show data. Anyone can connect BigQuery to Looker Studio and drop charts on a page. The hard part is choosing what NOT to show, structuring the narrative arc across pages, getting the conditional formatting thresholds right (12¢ vs 16¢ makes a huge difference in interpretation), and making each chart earn its space on the page. The technical ETL was straightforward — the storytelling and design took twice as long."

Category F: Domain Knowledge

Q: What is a balancing authority?

> "A balancing authority (BA) is an organization responsible for maintaining the balance between electricity supply and demand in a specific region. In the US, there are about 60 BAs — the major ones include PJM (Mid-Atlantic), MISO (Midwest), ERCOT (Texas), CAISO (California), and others. They operate the grid in real-time, dispatching generators and managing imports/exports. My dashboard tracks 10 of the largest BAs, which together serve the majority of the US population."

Q: What does 5.55% MAPE mean in practical terms?

> "If the forecast says demand will be 40,000 MW, the actual demand will typically be between 37,780 MW and 42,220 MW — plus or minus about 2,200 MW. In the energy industry, a MAPE under 5% is considered excellent, 5-10% is acceptable, and over 10% needs attention. Duke Energy Carolinas at 21% MAPE means their forecasts are off by about 1 in 5 megawatts — that's costly because it means either spinning up expensive peaker plants or curtailing generation unnecessarily."

Q: Why is the renewable share metric important?

> "Renewable share tracks the energy transition — the global shift from fossil fuels to clean energy. At 30.6% nationally, the US is roughly a third of the way to 100% clean electricity. This metric matters for three reasons: regulatory compliance (many states have Renewable Portfolio Standards mandating specific targets), investor reporting (ESG metrics directly affect company valuations), and operational planning (renewable generation is intermittent, which affects how the grid is managed)."

Q: What drives electricity price differences between states?

> "Five main factors, which my scatter plot helps visualize: (1) Fuel mix — coal and hydro are cheap, gas is moderate, imported power is expensive. (2) Market structure — deregulated markets (Texas, PJM) vs. regulated utilities (Southeast) price differently. (3) Geography — island states (Hawaii, Alaska) have extreme logistics costs. (4) Grid topology — states that import power pay transmission costs. (5) Population density — sparse states spread fixed grid costs over fewer customers. Notably, renewable share is NOT a primary driver — that's the key insight from the scatter plot."

---

13. Technical Glossary

Term	Definition
MAPE	Mean Absolute Percentage Error — avg forecast miss as a percentage
RMSE	Root Mean Square Error — penalizes large forecast misses more than small ones
Star schema	Data model with central fact tables surrounded by dimension tables
Partitioning	Physically splitting a table by a column (usually date) for faster queries
Clustering	Sorting data within partitions by specified columns for faster filtering
High-watermark	The latest timestamp/date already loaded — used for incremental ingestion
ETL	Extract, Transform, Load — the data pipeline pattern
RAG	Red, Amber, Green — conditional color coding for KPI status
BA	Balancing Authority — entity responsible for grid balance in a region
EIA	Energy Information Administration — US federal energy data agency
RTO	Regional Transmission Organization — operates the bulk power grid
ISO	Independent System Operator — similar to RTO, manages grid operations
MISO	Midcontinent Independent System Operator (Midwest)
PJM	PJM Interconnection (Mid-Atlantic, largest US BA)
ERCOT	Electric Reliability Council of Texas
CAISO	California Independent System Operator
PPA	Power Purchase Agreement — contract to buy electricity at a fixed price
ESG	Environmental, Social, Governance — sustainability reporting framework
CTE	Common Table Expression — SQL `WITH` clause for readable subqueries

---

Document generated February 2026. Metrics reflect data through February 2026.