Consumer Risk & Operational Intelligence: A Full-Stack Data Engineering & Analytics Solution

1. The General View (The 'Why')

The Problem

We're facing a massive volume of consumer feedback. With over **86,000 complaints** in the CFPB database for credit cards, trying to review them manually is impossible. This creates three key business risks:

1. **Regulatory Pressure:** We have to respond within 15 days or risk fines.
2. **Operational Cost:** If we're slow, it means our internal processes are inefficient.
3. **Hidden Risks:** Critical issues like fraud ("identity theft") or systemic product failures are buried in unstructured text, so they don't show up in our standard reports.

The Objective

I built a three-part dashboard to turn this raw data into something we can act on:

- **For Executives:** A high-level view of compliance status and any sudden spikes in complaints.
- **For Investigators:** The ability to drill down into specific products or states to find the root cause of a problem.
- **For Operations:** A way to track our response times and make sure we're meeting our regulatory deadlines.

2. Data Origin & Ingestion

Source

I'm using public data from the **CFPB Consumer Complaint Database**. This is real-world data of complaints filed against financial companies.

Ingestion Step (ELT Strategy)

I chose an ELT (Extract, Load, Transform) approach. The first step is to get the raw CSV file into our database as-is, without changing anything. This way, we have a perfect copy of the source and don't lose any data.

- **Technique:** I used the T-SQL BULK INSERT command because it's the fastest way to load large flat files.
- **Handling Dirty Data:** The raw text fields often contain commas (e.g., "Chicago, IL"), which can break a standard CSV import. I used FORMAT = 'CSV' and FIELDQUOTE = '"' to handle this. I also loaded all columns as VARCHAR first to prevent any data type errors from stopping the entire load process. ‘tablock’ for performance.

3. The SQL Engineering Deep-Dive (Step-by-Step)

Step A: Cleaning & Standardization

With the raw data loaded into a staging table, I created a SQL View named View_Core_Cleaning to clean and structure it.

- **Robust Casting:** Instead of a simple CAST, I used TRY_CAST(Date_received AS DATE). If a date is formatted incorrectly, this function returns NULL instead of erroring out, which keeps the data pipeline from failing.
- **Null Handling:** Some records were missing a State. I used ISNULL(State, 'Unknown') to fill in these gaps, ensuring every complaint can be mapped geographically.
- **Standardization:** To protect privacy and simplify reporting, I standardized location data, like grouping complaints by ZIP code prefixes.

Step B: Text Mining & Feature Engineering

The most valuable information is often in the Consumer_complaint_narrative text field. I built a simple keyword scanner directly in SQL to pull out key risks.

- **Risk Flags:** I used CASE statements with LIKE to search for specific keywords and create new columns.
    - *Identity Theft & Fraud:* Searched for terms like '%identity theft%', '%fraud%', or '%stolen%'.
    - *Credit Reporting:* Searched for '%credit report%' or '%fcra%'.
- **Impact:** This turns unstructured text into simple binary flags (0 or 1). It allows us to count "risk events" without a human having to read thousands of narratives and it saves enormous amounts of time.

Step C: Dimensional Modeling

A single, flat table is slow for BI tools. I remodeled the data into a **Star Schema**.

- **Dimension Tables:** I created separate tables for repeating attributes: Dim_Location, Dim_Product, Dim_Company, and Dim_Issue.
- **Fact Table:** The main Fact_Complaints table now only holds keys (like Location_ID) and numbers (like Days_To_Process).
- **Why this is better:** This model is much more efficient. For example, the database stores "California" once in Dim_Location instead of 10,000 times in the main table. This makes filtering in Power BI incredibly fast.

4. Preparation for Power BI

Staging for Performance
The final step in the SQL backend is to take the logic from my views and store the results in physical tables (the Fact and Dimensions). This means all the heavy-duty processing is done *before* the data ever gets to Power BI.

5. Conclusion & Business Impact

This automated pipeline provides a clear, repeatable way to monitor consumer risk. We can move from being reactive to proactive.

- **Pinpoint Risk:** We can now instantly see if a problem is emerging in a specific state, like **Texas** or **California**.
- **Product Insights:** We can isolate which products, like a specific credit card, are causing the most complaints.
- **Proactive Compliance:** Instead of just reacting to old complaints, we can monitor trends in fraud or service issues as they happen.

This project covers the full data lifecycle ,from ingestion and cleaning to text analysis and creating an optimized model for reporting.

6. Power BI: The Visual Intelligence Layer

A. Data Model Architecture (The Star Schema Strategy)

I used a strict **Star Schema** in Power BI to ensure the report is fast and responsive, even with all 86,741 records.

- **One-to-Many Relationships (1:*):** I connected the smaller dimension tables (Dim_Location, Dim_Product) to the central Fact_Complaints table. Filters flow one way, from the dimensions to the fact table.
- **The Performance Win:** This is how Power BI's VertiPaq engine works best. When a user filters by "Company Name," Power BI filters the very small Dim_Company table first and then uses the Company_ID to instantly filter the huge Fact_Complaints table. This makes slicers and visuals respond almost instantly.

B. Interface Design & UX Strategy

I designed the report as an **Executive Command Center** for active monitoring, not just a static report.

- **"Dark Mode" Theme:** I used a high-contrast dark theme to make it easier for analysts to look at the screen for long periods and reduce eye strain.

- **Visual Hierarchy:** I used color intentionally. Teal shows good performance ("On Target"), while Orange/Amber draws the eye immediately to problems like "Disputes" and "Risk."

7. The DAX Measurement Framework (Analytical Logic)

While SQL did the heavy lifting, I used DAX to create dynamic measures that respond to user filters.

A. Regulatory & Risk Compliance (SLA Monitoring)

- **% Timely (99.7%):** It's calculated as the percentage of complaints marked as timely, using a simple count of timely cases divided by total complaints. The logic is straightforward and efficient, making it easy to interpret and reliable for compliance tracking.
- **Compliance Failures (226):** A direct count of records where we missed the 15-day deadline. This is an instant audit list for the compliance team.
- **Narrative Inclusion Rate (41%):** Shows what percentage of complaints include a detailed story. A higher number here means we have richer data for finding root causes.

B. Financial Impact & Customer Sentiment

- **% Financial Relief (12.4%):** The percentage of cases where we gave a refund or credit. This is a direct measure of the cost of complaints.
- **% Dispute Rate (14%):** Tracks how often a customer disputes our initial response. This is a key metric for "first-contact resolution" failure.
- **Theft/Fraud Alert (12.5%):** A risk density metric. It's the number of fraud-flagged complaints divided by the total. It warns us if a specific product is being targeted.

C. Operational Throughput & Capacity

- **Daily Volume (237):** The total number of complaints divided by the number of days in the period. This helps managers plan staffing.
- **Avg Processing Hours (30.4):** The average time it takes us to close a complaint, measured in hours.
- **Peak Delay (115 Hours):** Shows the single longest time it took to resolve a complaint. Even if our average is good, this helps us find the outliers that need attention.
- **Outcome Diversity (5):** A count of the different ways we close complaints. If this number is high, it might mean our resolution process isn't standardized.

D. Technical Note: The "Boolean" Advantage

A key decision was to pre-calculate flags in SQL. Rather than writing slow DAX like FILTER(Table, Table[Column] = "Yes"), I had SQL create columns such as Is_Timely using 1/0. When imported into Power BI, these are treated as TRUE/FALSE, allowing DAX measures to run simply as CALCULATE(COUNTROWS(...), Is_Timely = TRUE()) — extremely fast and efficient.

Chapter 8: Findings & Strategic Insights

Finding 1: Credit Bureaus Are a Major Source of Complaint Volume

The "Top 5 Institutions by Complaint Volume" chart shows that while our company (Capital One) is number one, the next three largest sources of complaints are the major credit bureaus: Equifax, TransUnion, and Experian. This strongly indicates that a significant portion of complaints are not about our products directly, but rather about the accuracy of data we report to these third parties, which in turn affects our brand reputation.

Finding 2: 'Closed with Explanation' Is the Largest Resolution Category and a Likely Driver of Disputes

The "Global Resolution Distribution" treemap shows our most frequent outcome is "Closed with explanation," accounting for over 53,000 cases. Given our overall 14% "Consumer Dispute Rate," it is highly probable that this resolution category is the primary driver. This suggests customers may perceive these explanations as insufficient or generic, prompting them to dispute the outcome.

Finding 3: Texas Has More Complaints Than California, Which Is Unusual

According to the "Operational Footprint" table, Texas is our top state for complaint volume (12,348), surpassing California (11,012). This is a typical, as California's larger population usually drives the most volume. The data also shows that over 96% of Texas complaints originate from the web channel, suggesting a specific digital-first issue may be localized to that state, potentially related to regional products, economic conditions, or state-specific regulations.

Finding 4: 'Problem with a purchase' Is the Top Issue by Volume and Poses a High Dispute Risk

The "Dispute Risk vs. Complaint Volume by Issue" scatter plot identifies 'Problem with a purchase shown on your report' as our most critical issue. It is the largest bubble, positioned furthest to the right (indicating the highest complaint volume), and is one of the highest on the vertical axis (indicating a high dispute risk). This issue represents our primary operational challenge.

Finding 5: Relief Is Granted More Often for Store Cards Credit Cards Than for General-Purpose Credit Cards

A detailed analysis of the "Resolution Outcome Profile by Sub-Product" chart reveals a minor but consistent difference in outcomes. Customers with "Store credit cards" are slightly more likely to be granted relief than those with "General-purpose credit cards." While the overall resolution profiles are very similar, the "Relief Granted" bar for store cards is wider by an estimated 2 percentage points. This subtle variation may indicate that issues specific to store cards are marginally more likely to be resolved in the customer's favor, potentially due to the nature of the complaints or specific terms in our retail partnership agreements.

Finding 6: Compliance Failures Are Driven by Process Errors

Our "Max Complaint Response Time" is 115 hours (under 5 days), which is well within the typical 15-day regulatory deadline. This proves that the 226 "Compliance Failures" are not due to slow responses. Instead, they are likely caused by administrative or process errors, such as missing a required legal disclosure or failing to follow a specific communication protocol.

Finding 7: Financial Impact Rate and Fraud Alert Rate Are Two Distinct, High-Level Risks

The dashboards show a "Financial Impact Rate" of 12.4% and a "Fraud Alert Rate" of 12.5%. While numerically similar, these are independent metrics. We cannot conclude that fraud-related complaints are the direct cause of financial payouts. The data simply shows that around 12% of all complaints involve a financial loss to the company and, separately, that around 12% of all complaints have narratives that trigger a fraud alert.

Finding 8: Complaint Volume Follows a Predictable Seasonal Pattern

The "Monthly Complaint & Dispute Trends" chart reveals a clear seasonal pattern: complaint volumes dip in November before spiking in December and January. This post-holiday surge is a predictable trend. While our average of 237 daily complaints is a good baseline for staffing, we must plan for a significant increase in workload during Q1 to maintain service levels.

Key Takeaway: We're fast, but not effective. Our 99.7% timely response rate looks great for compliance reports, but the 14% dispute rate shows customers aren't satisfied with our answers. We're solving the *regulatory* deadline problem, not the *customer's* actual problem.