FEMA Analysis Report

Group 3
Siddharth Bahekar || Kiran Kesarapu || Vikas Yellapu || Sriram Sai Bokka
College of Engineering
Northeastern University
Toronto, ON

Team Contributions

The project was executed collaboratively by the team, with equal distribution of work across data engineering, analytics, and documentation activities. Each team member contributed to specific components of the pipeline and played an active role in the successful completion of the project, including the report and presentation deliverables.

Siddharth Bahekar

- Identified and selected the **public FEMA dataset** as the data source, ensuring relevance and usability for cloud ingestion and analytics.
- Designed and developed **Python notebook** for pre-validation of data and performing cleaning.
- Contributed extensively to the **project report**
- Created the **pipeline architecture diagram** and designed key visual elements for the presentation.

Kiran Kesarapu

- Configured and implemented **Azure Data Factory** for ingesting data from HTTP endpoints into Azure Data Lake Gen2 (Raw).
- Ensured secure and efficient **storage and retrieval** of ingested raw data from the data factory pipelines.
- Drafted and presented the **Storage Management & Integration** slides for the final presentation.
- Contributed to the **report sections** related to storage setup and data governance.

Vikas Yellapu

- Developed scalable **data transformation workflows** using **Azure Databricks**, handling data cleansing
- Maintained Databricks **notebooks** for ETL processing, with integration into the cloud pipeline.
- Documented the **Data Processing & Transformation** aspects of the project
- Created corresponding slides and visualizations for **Databricks processing** in the presentation.

Sriram Sai Bokka

- Configured **Azure Synapse Analytics** to query transformed data stored in Data Lake Gen2.
- Built and published interactive **Power BI dashboards** for data visualization and reporting.
- Drafted the **Data Analytics & Visualization** sections in the report
- Presented insights and dashboard walkthroughs during the **final project presentation**.

GitHub Link - https://github.com/yvikasofficial/data_arch_final_project

Abstract

This project presents the design and implementation of a cloud-based data pipeline leveraging the Microsoft Azure ecosystem to ingest, process, analyze, and visualize publicly available disaster recovery data. The primary data source is the *FEMA Public Assistance Funded Projects Details* dataset, accessed via FEMA Open Data Portal. The project architecture, as illustrated in the final pipeline diagram, represents a scalable and modular solution suitable for large-scale data processing and analytics.

The pipeline begins with Azure Data Factory, which automates the ingestion of raw data from a public HTTP endpoint into Azure Data Lake Storage Gen2. The raw data is then processed and transformed using Azure Databricks, where Python scripts are used for cleansing, normalization, and enrichment. Transformed data is stored back in a structured format in the data lake. Azure Synapse Analytics is used to perform advanced querying and prepare the data for analysis. Finally, Power BI connects to Synapse to visualize key trends and insights through interactive dashboards.

This end-to-end solution demonstrates the effective use of cloud-native services for building a robust data infrastructure. It emphasizes automation, scalability, and real-time insights, with potential applications in disaster recovery analysis, public transparency, and data-driven decision-making.

Dataset Description and Justification

For this project, we utilized the FEMA Public Assistance Funded Projects – Details v1 dataset, which is publicly available via the FEMA OpenFEMA Data Portal at https://www.fema.gov/openfema-data-page/public-assistance-funded-projects-details-v1. This dataset provides comprehensive and granular information about federally funded public assistance projects that support state, local, tribal, and territorial governments as well as certain private non-profits in responding to and recovering from major disasters and emergencies.

Dataset Overview

The FEMA Public Assistance dataset includes detailed records of disaster recovery projects funded under the Robert T. Stafford Disaster Relief and Emergency Assistance Act. Key attributes within the dataset include:

- Disaster Number – A unique identifier for the disaster event.
- Project Title and Description – Descriptive information about the recovery activity.
- Recipient and Subrecipient Details – Entities receiving the funding (e.g., state governments, local agencies, hospitals).
- Funding Amounts – Obligated and federal share funding in USD.
- Damage Category – Categories such as roads, bridges, water control facilities, buildings, utilities, etc.
- Geographical Information – State, county, and place names where the projects were carried out.
- Project Status and Dates – Status of the funding approval and project lifecycle timestamps.

The dataset is updated regularly, ensuring that it reflects recent disasters and newly approved funding projects.

Justification for Selection

This dataset is particularly suitable and interesting for a cloud-based data engineering and analytics project for several key reasons:

1. Volume and Complexity: The dataset presents a realistic challenge in terms of large-scale data ingestion, storage, transformation, and querying. It provides the opportunity to build and test scalable cloud infrastructure.
2. Diverse Data Types: The dataset contains a variety of data types including text, categorical variables, dates, and numeric fields, which allows for complex transformation logic and meaningful visual analytics.
3. Real-World Relevance: Disaster recovery is a socially impactful and high-stakes domain. Analyzing how federal funds are distributed across regions, disaster types, and damage categories allows for insightful public policy evaluation and transparency reporting.
4. Public Accessibility and Openness: Since the dataset is publicly available via an HTTP endpoint, it was ideal for implementing an end-to-end automated ingestion pipeline using Azure Data Factory, demonstrating integration with open government data sources.
5. Supports Rich Visualization: The dataset enables the creation of powerful Power BI dashboards that can display trends such as funding distribution over time, geographic impact of disasters, and project types by state or county. These visualizations are not only technically impressive but also informative for stakeholders.

Potential Applications

- Identifying states or regions that receive the most public assistance and for what types of damages.
- Visualizing temporal trends in disaster recovery funding across years.
- Analyzing effectiveness and efficiency in fund allocation and project completion status.
- Supporting academic or governmental research into disaster management and public policy.

Architecture and Design Overview

The objective of this project was to design and implement a robust, scalable, and modular **cloud-based data pipeline** to automate the ingestion, transformation, and visualization of publicly available disaster recovery data. The architecture is built entirely on Microsoft Azure cloud services, incorporating a suite of tools tailored to modern data engineering workflows.

Two representations of the architecture are included in this report:

1. **The original architecture diagram** (Figure 1) – This is a detailed, component-wise breakdown highlighting the roles of each service within a logical data flow.
2. **A simplified high-level pipeline view** (Figure 2) – This version provides a linear and intuitive overview ideal for stakeholder presentation and executive summaries.

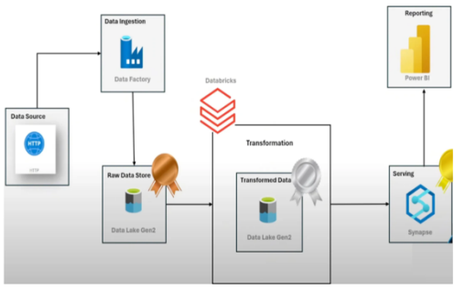


Figure 1

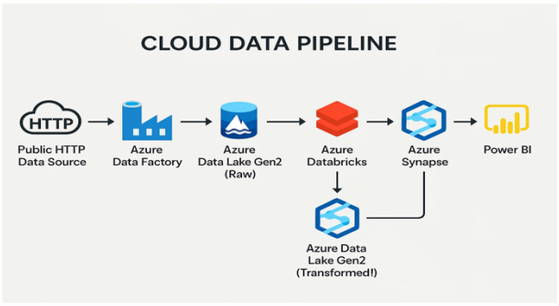


Figure 2

Design Components and Flow

1. Data Source (HTTP Endpoint)

The dataset used originates from FEMA's open data portal, accessible via a **public HTTP endpoint**. This real-world data source enables the development of a repeatable pipeline that demonstrates ingestion from external sources, simulating scenarios such as open government data integration and API-based data feeds.

2. Data Ingestion – Azure Data Factory

Azure Data Factory (ADF) is used as the **data orchestration service**. It is responsible for automating the ingestion of data from the HTTP endpoint into the raw storage layer. ADF pipelines are configured to extract data on a schedule, enabling periodic refresh and near real-time updates.

3. Raw Data Storage – Azure Data Lake Storage Gen2

The ingested data is landed into **Azure Data Lake Storage Gen2 (Raw zone)**. This layer acts as the **persistent storage for unprocessed data**, ensuring immutability and traceability of the original input. Using the bronze-silver-gold data architecture pattern, this raw data is classified as "bronze-tier" storage.

4. Transformation – Azure Databricks

Databricks serves as the **core processing engine** for the pipeline. Using python notebooks, the team developed data transformation logic including:

- Data cleansing and schema alignment
- Handling missing or null values
- Type casting and enrichment
- Deriving new metrics for analysis

The output of this step is written back to the **Data Lake Gen2 (Transformed zone)** – equivalent to the "silver-tier" in the architecture.

5. Serving Layer – Azure Synapse Analytics

Azure Synapse acts as the **serving and query layer**. It connects directly to the transformed data and provides SQL-based access for reporting. Synapse enables scalable and interactive querying, allowing seamless integration with reporting tools.

6. Reporting – Power BI

The final layer is implemented using **Power BI**, which is connected to Synapse to visualize insights from the transformed FEMA data. Dashboards are developed to provide:

- Funding distribution by region and disaster type
- Project status over time
- Category-wise allocation trends

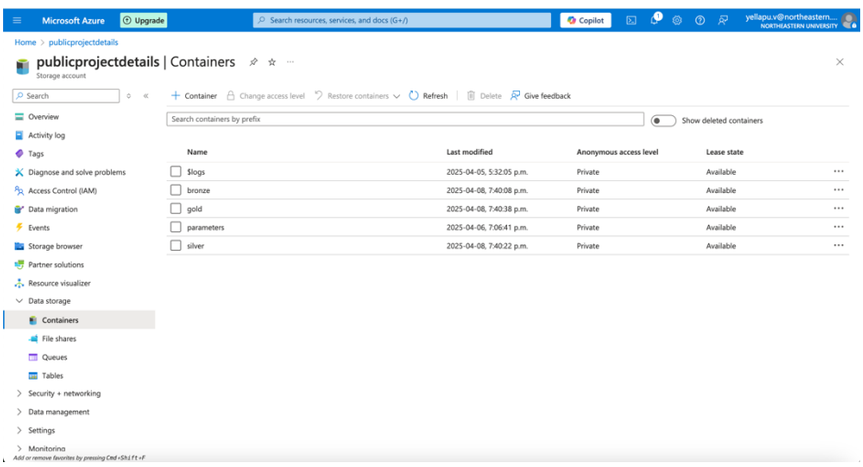
These insights allow users (e.g., public administrators or policymakers) to understand how federal assistance is utilized across the country.

Architectural Principles and Considerations

- **Scalability**: Each component in the architecture is designed to scale independently, ensuring the pipeline remains performant as data volume increases.
- **Modularity**: The use of discrete zones (raw, transformed) allows clear separation of concerns and easier debugging or enhancement.
- **Maintainability**: Notebooks in Databricks and pipeline definitions in ADF promote version control and reusability.
- **Security and Compliance**: Azure-native services ensure that data is encrypted in transit and at rest, with role-based access control available at each stage.
- **Cost-efficiency**: Serverless and pay-as-you-go components (e.g., Data Factory, Synapse Serverless) help control operational costs.

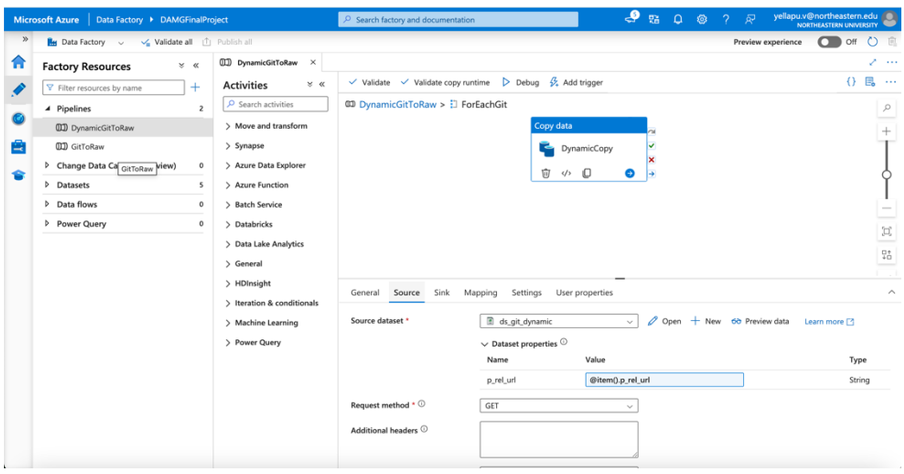
Cloud Implementation Using Microsoft Azure

To process and analyze the FEMA Public Assistance dataset, this project implements a complete cloud-based data pipeline using Microsoft Azure. The primary goal of the pipeline is to automate the extraction of structured data from a public HTTP source, perform necessary transformations, and make the data available for analytics and visualization. The implementation follows a layered medallion architecture and integrates several Azure services including Data Factory, Data Lake Storage Gen2, Databricks, Synapse Analytics, and Power BI.



Data Ingestion Using Azure Data Factory

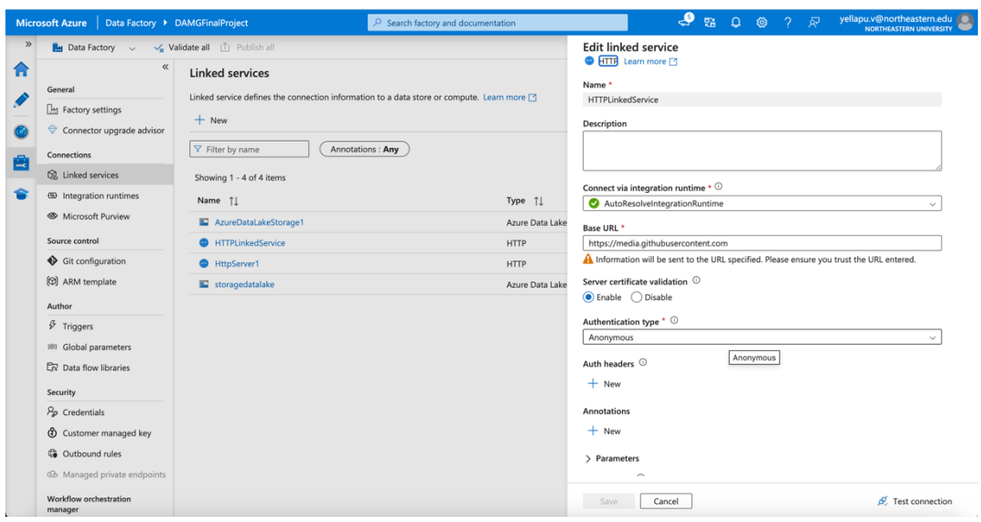
The pipeline begins with Azure Data Factory, which orchestrates the data ingestion process. A public dataset hosted by FEMA on GitHub is accessed via HTTP using a parameterized and dynamic pipeline. Two main pipelines were created within Azure Data Factory: GitToRaw and DynamicGitToRaw. The DynamicGitToRaw pipeline is configured to accept multiple relative URLs via a Lookup activity. These URLs are then iteratively processed within a ForEach loop. For each iteration, the Copy Data activity retrieves data from the HTTP endpoint and writes it to Azure Data Lake Storage.



The source dataset is parameterized to allow for dynamic ingestion, referencing the URL values extracted by the Lookup activity. The connection to the GitHub-based HTTP source is established through a linked service (HTTPLinkedService) using an anonymous authentication method, with the base URL set to https://media.githubusercontent.com.

Storage in Azure Data Lake Gen2 (Bronze Layer)

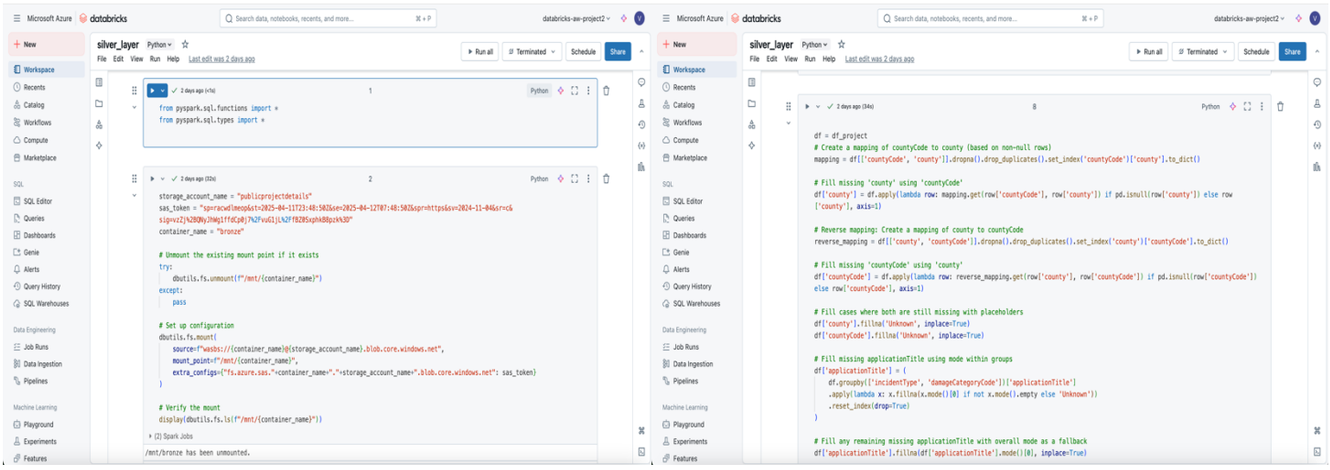
Once the data is ingested, it is stored in Azure Data Lake Gen2 under a container named bronze, representing the raw data storage layer in the medallion architecture.



The data lake is structured to support a progressive refinement of the dataset, with three distinct containers: bronze for raw data, silver for cleaned and enriched data, and gold for curated datasets optimized for analytics. The publicprojectdetails storage account holds these containers, allowing modular access to datasets at various stages of transformation.

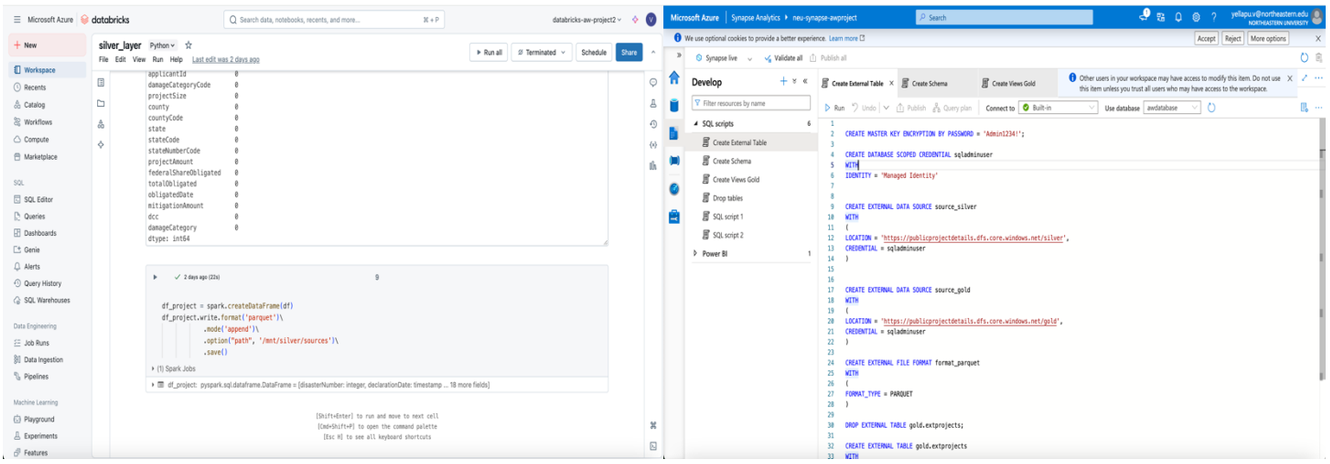
Data Transformation Using Azure Databricks

The raw data stored in the bronze layer is then processed using Azure Databricks. A Databricks notebook is developed to handle data transformation tasks. Initially, the notebook establishes a connection to the bronze container by mounting it using a SAS (Shared Access Signature) token for secure access. The data is loaded into a PySpark DataFrame, after which a sequence of data cleansing and enrichment operations is applied. These operations include filling in missing values for key attributes such as county and countyCode by creating mapping dictionaries and using reverse lookups. Cases where both values are missing are filled with placeholder strings such as "Unknown" to preserve data completeness. Additionally, the notebook implements a strategy to fill missing values in the applicationTitle column by using the statistical mode within grouped data based on incidentType and damageCategoryCode. The cleaned DataFrame is then saved in the silver layer of the data lake in Parquet format, which offers improved query performance and schema flexibility.



Serving Data with Azure Synapse Analytics

After transformation, the data in the silver and gold layers is exposed to analytics tools using Azure Synapse Analytics. Synapse is configured to access external data stored in Azure Data Lake Gen2 by creating external tables over the Parquet files.

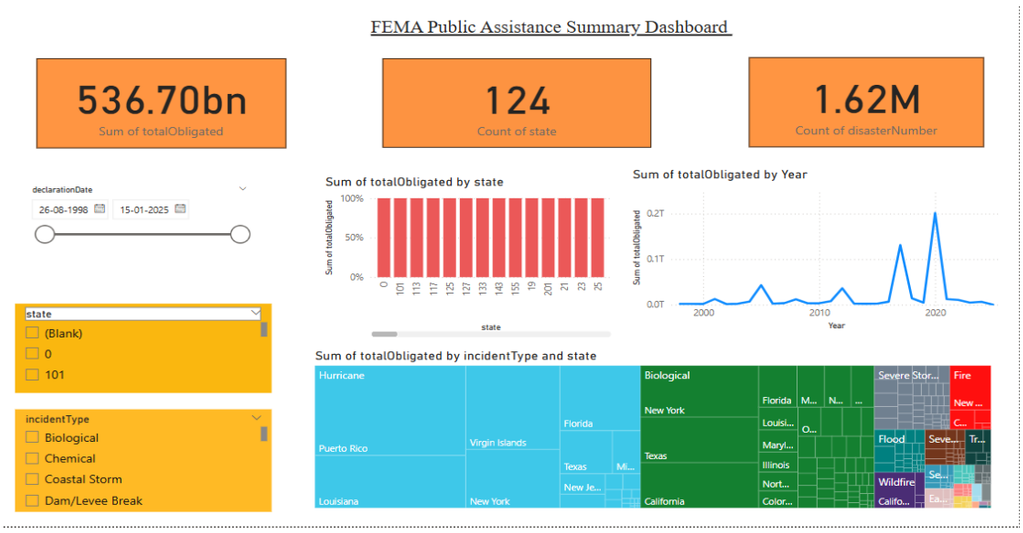


To enable this, a database scoped credential is created using Managed Identity, followed by the definition of external data sources for both silver and gold layers. File formats are specified as Parquet, and external tables are created to represent the processed datasets. This setup allows users to perform T-SQL queries on the transformed data without needing to duplicate or import it into Synapse-managed storage. The data is visualized using Power BI. The BI reports connect directly to Azure Synapse Analytics, leveraging the external tables to perform live queries on the transformed datasets.

Power BI Visualizations and Insights

To understand the analytical scope of the project, two Power BI dashboards and certain charts were developed using transformed FEMA disaster assistance data accessed through Azure Synapse Analytics. These dashboards were designed to provide interactive, visual insights into funding distribution, temporal trends, disaster frequency, and regional impact. The dashboards are connected live to Azure Synapse's external tables, ensuring real-time access to the processed data stored in Azure Data Lake Gen2.

Dashboard 1: FEMA Public Assistance Summary Dashboard
This dashboard presents a high-level overview of FEMA's federal funding for disaster recovery across the United States, with a focus on state-level and incident-type analytics.



Key Metrics and Filters

The dashboard includes three primary KPI cards:

- Total Obligated Amount: $536.70 billion, representing the cumulative disaster assistance allocated.
- Count of States: 124 unique state and territory identifiers (likely due to multiple naming or coding conventions).
- Count of Disaster Numbers: 1.62 million individual project declarations.

Slicers are embedded to filter the data by:

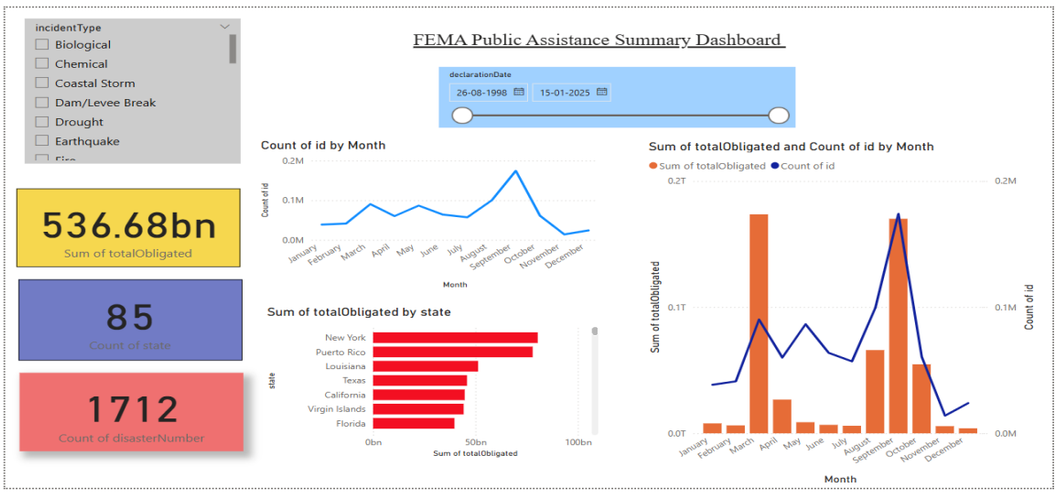
- Declaration Date range (from 1998 to 2025)
- Incident Type (e.g., Hurricane, Fire, Flood)
- State

Visual Components and Interpretation

- Bar Chart (Funding by State): This shows normalized funding distributions, enabling a quick comparison across various state codes. Some numerical labels such as "0", "101" might reflect inconsistencies or legacy formatting, suggesting scope for additional data cleansing.
- Line Chart (Year-wise Total Obligated Funding): Distinct spikes are observed in years 2005, 2012, and 2020, corresponding to Hurricane Katrina, Superstorm Sandy, and the COVID-19 pandemic, respectively. These funding surges underline FEMA's responsive increase in obligations during nationwide crises.
- Treemap (Funding by Incident Type and State): The largest allocations are linked to Hurricanes, particularly in Puerto Rico, Florida, and Louisiana, reflecting their vulnerability to tropical disasters. Other significant event types include biological disasters (e.g., pandemic response in New York and Texas), Floods, and Fires.

Dashboard 2: FEMA Monthly Trends and Funding Patterns

This second dashboard shifts analytical focus from state-level summaries to monthly project activity and seasonal funding patterns. It is designed to support temporal trend analysis, helping policymakers and analysts understand disaster behaviors over time.



KPI Cards

- Total Funding: $536.68 billion (almost identical to Dashboard 1, showing consistency)
- Unique States: 85 (filtered for refined and valid state entries)
- Total Disaster Declarations: 1,712, matching distinct incidents captured post-transformation

Key Visualizations

- Line Chart (Project Volume by Month): Displays count of project IDs each month. Spikes in October and September suggest strong correlation with hurricane season, aligning with FEMA's historical funding cycles.
- Combo Chart (Funding vs. Volume by Month): These visual overlays funding (columns) and project count (line) for every month. A prominent insight is the divergence between volume and funding—for instance, fewer projects in March may have higher associated costs, indicating large-scale or high-impact events.
- Bar Chart (Funding by State): Highlights top-funded regions, reaffirming that New York, Puerto Rico, and Louisiana consistently receive the highest assistance. This repetition across dashboards increases confidence in the data and insights.

Slicers enable filtering by:

- Incident Type (Biological, Earthquake, Hurricane, etc.)
- Date Range (full control over temporal selection)

Insights Derived from Dashboards

1. Temporal Trends

- Peaks in funding align with specific disasters: **2005 (Katrina)**, **2012 (Sandy)**, **2017 (Maria and Harvey)**, and **2020-2021 (COVID-19)**.
- Seasonal trends are evident, with **September–October** showing high project activity due to the **Atlantic hurricane season**.
- The **combo chart** exposes discrepancies where a few high-cost projects can outweigh hundreds of smaller ones.

2. Geographic Trends

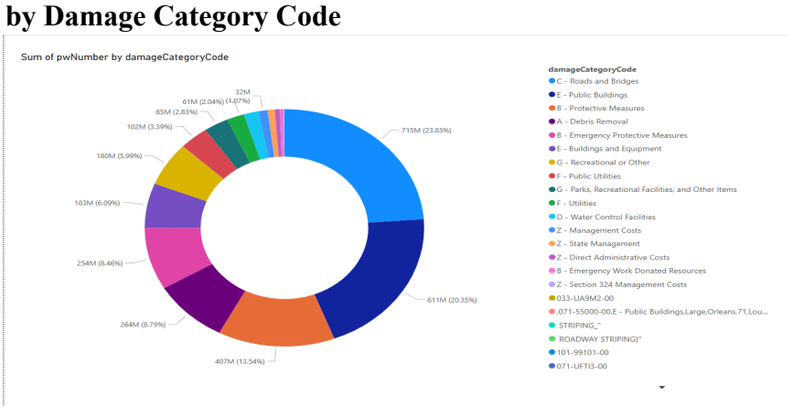
- **New York**, **Puerto Rico**, and **Louisiana** emerge as perennial recipients of high funding—likely due to dense infrastructure and historical disaster exposure.
- **Virgin Islands** and **Florida** also show disproportionately high project counts, possibly reflecting frequent but lower-cost interventions.
- **Data normalization issues** (e.g., state code "101") call for attention to metadata standardization in future pipeline iterations.

3. Disaster-Type Specific Trends

- **Hurricanes** dominate federal spending, both in number and amount.
- **Biological disasters**, particularly **COVID-19**, show sharp, recent peaks—highlighting FEMA's shifting operational focus.
- Less frequent but high-cost incident types (e.g., **Earthquakes, Fires**) appear in smaller numbers but with significant financial impact per project.

Supporting Charts

1. Sum of pwNumber by Damage Category Code



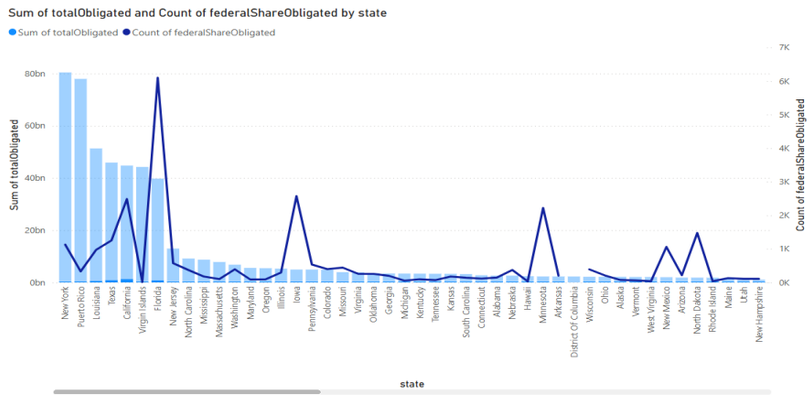
Key Observations:

- The largest share belongs to **Roads and Bridges (C)** at 23.83%, followed by **Public Buildings (E)** at 20.35%.
- Other notable categories include **Protective Measures (B)** and **Debris Removal (A)**.

- The long tail includes administrative costs, parks, water facilities, and some incorrectly formatted records (likely data cleaning issues).

---

2. Total Obligated and Count of Federal Share Obligated by State



This dual-axis chart plots:

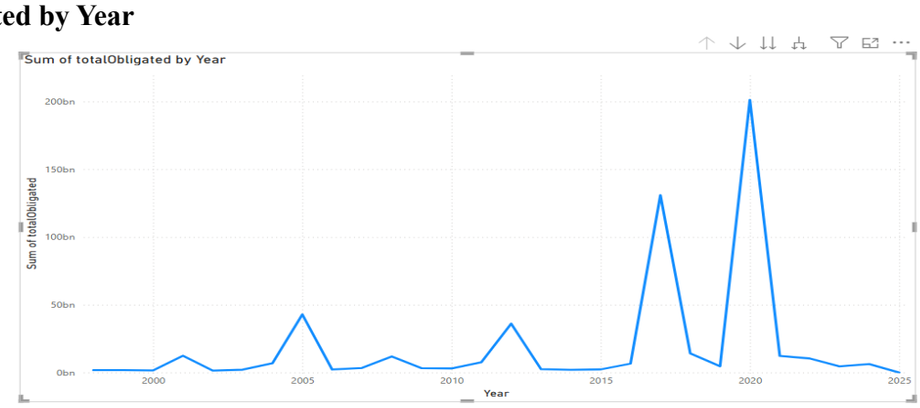
- The **sum of total obligated funds** (bar)
- The **count of federally obligated shares** (line) across all U.S. states and territories.

Key Observations:

- **New York, Puerto Rico, and Louisiana** lead in total federal obligation.
- **Florida** shows the highest project count, indicating many smaller-scale obligations.
- Virgin Islands has a significant project count relative to its total obligation, possibly due to multiple small-scale recovery efforts.

---

3. Total Obligated by Year



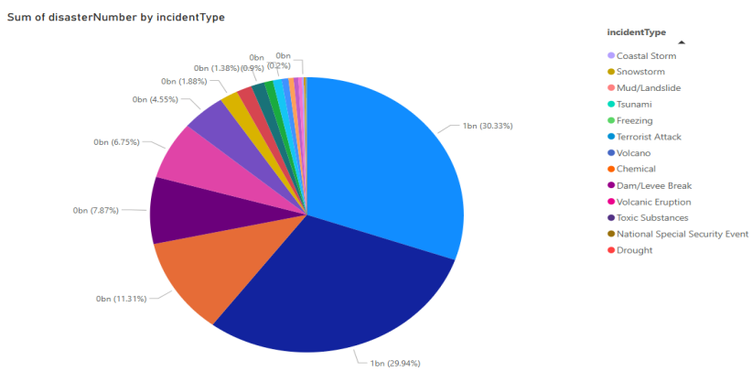
Displays yearly trends of the **sum of total obligated funds** from the late 1990s to 2025.

Key Observations:

- Sharp peaks are seen in **2017**, **2020**, and **2021**—corresponding to major disaster events like hurricanes and the COVID-19 pandemic.
- A large spike in 2017 likely reflects Hurricane Maria and other significant events.

---

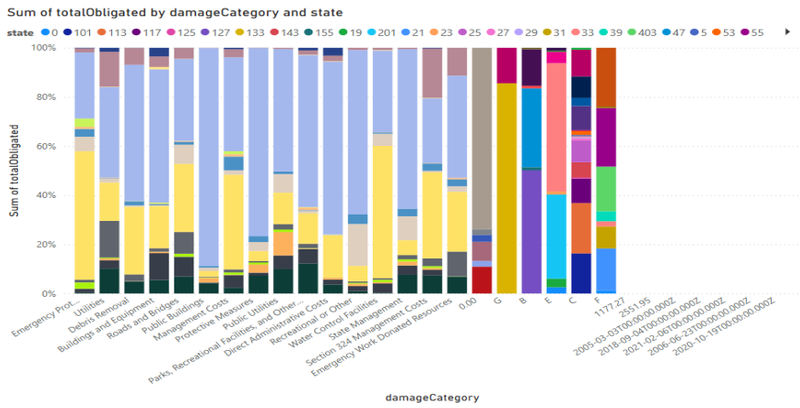
5. Disaster Numbers by Incident Type



Key Observations:

- **Hurricanes and Floods** account for over 60% of disaster events in the dataset.
- Other incidents include **Fires, Snowstorms, and Earthquakes**, but in smaller proportions.
- Rare events like **Volcanic Eruption** or **Terrorist Attacks** also appear.

6. Total Obligated by Damage Category and State



This chart shows how different **damage categories** are funded across various **states**, normalized to percentages.
Key Observations:

- **Emergency Protective Measures and Debris Removal** dominate early states (e.g., Puerto Rico).
- Some states like **California** have higher funding for **Utilities and Buildings**.
- Data variation suggests regional-specific disaster recovery strategies.

Conclusion

This project successfully demonstrates the end-to-end design and implementation of a scalable, cloud-native data pipeline using Microsoft Azure services. By integrating a real-world dataset from the FEMA Public Assistance Program, we built a modular architecture capable of ingesting, transforming, and visualizing large volumes of disaster recovery data.

Each component of the pipeline was carefully selected to serve a specific function within the modern data engineering lifecycle—Azure Data Factory for ingestion, Azure Data Lake Gen2 for structured storage, Azure Databricks for scalable transformations, Azure Synapse for serving and querying, and Power BI for interactive data visualization. The architecture promotes best practices such as data zoning (raw and transformed), metadata tagging, scheduled refreshes, and role-based access control, ensuring the solution is production-ready and extensible.

Our analytical layer, powered by Power BI, surfaced meaningful insights from the FEMA dataset—highlighting patterns in federal funding allocation by disaster type, geography, and damage category. These insights not only validate the accuracy and efficiency of the pipeline but also reveal critical information that can support public sector transparency, disaster preparedness planning, and policy making.

The project serves as a comprehensive example of how cloud technologies can be leveraged for impactful, real-time data-driven decision-making. The system architecture and analytical outputs demonstrate the potential of cloud infrastructure to address real-world problems through automation, scalability, and deep analytics.

References

- Federal Emergency Management Agency. (n.d.). *Public Assistance Funded Projects – Details v1*. FEMA OpenFEMA Data Page. https://www.fema.gov/openfema-data-page/public-assistance-funded-projects-details-v1
- Federal Emergency Management Agency. (n.d.). *OpenFEMA API samples*. GitHub. https://github.com/FEMA/openfema-samples
- DataCamp. (n.d.). *Power BI dashboard tutorial: Learn to create interactive dashboards*. https://www.datacamp.com/tutorial/power-bi-dashboard-tutorial