# **Week 5 Report: Collection & Integration Phase**

During the fifth week of the InsightNation project, our focus shifted to one of the most crucial backend activities: collecting and integrating citizen feedback data. This week’s activities laid the groundwork for the analytical components of the platform, ensuring the data pipeline is robust, clean, and aligned with future modeling and visualization tasks. The work involved setting up reliable data sources, automating data collection pipelines, maintaining quality, and integrating datasets for seamless analysis.

**1. Setting Up Data Sources**

The first step involved identifying viable sources of citizen feedback data relevant to public services. Three primary categories were finalized:

* **Online Surveys**: Public satisfaction surveys from open government data repositories.
* **Official Government Portals**: Sites offering structured datasets related to public amenities like sanitation, transport, and parks.

Permissions, formats, and access methods were reviewed to ensure smooth and legal collection. API credentials and public access links were gathered for all sources.

**2. Developing Data Collection Pipelines**

Following source setup, we developed automated pipelines to systematically collect and process data. Python scripts using requests, Tweepy, and Pandas were created to:

* Extract real-time tweets filtered by location and civic-related keywords.
* Download and periodically update CSV/JSON data from open government portals.
* Merge survey responses stored in local CSV format into the main data stream.

Data ingestion was modularized to accommodate future expansion to more sources.

**3. Ensuring Data Quality and Consistency**

Significant effort was placed on preprocessing incoming data to ensure it met quality standards. Common data quality issues addressed included:

* **Missing Values**: Filled using context-aware imputation techniques where appropriate or flagged for exclusion.
* **Duplicates**: Removed based on unique identifiers or text similarity.
* **Formatting Issues**: Standardized column names, timestamp formats, and categorical variables across datasets.

Custom data validation scripts were written to automate this process and ensure reusability.

**4. Data Integration**

After individual datasets were cleaned, they were merged into a unified master dataset. A schema was defined that aligned key fields like date, location, service type, and feedback text. Integration involved:

* Normalizing variables from different sources to a common format.
* Resolving conflicts in overlapping data (e.g., differing sentiment labels from surveys and tweets).
* Consolidating feedback into a structured format suitable for both ML modeling and dashboard visualization.

The integrated dataset now provides a holistic view of citizen sentiment across services and platforms.

**5. Data Validation**

Validation procedures were carried out to ensure the dataset accurately represented public opinion. This included:

* Random sampling and manual verification of tweet sentiment and topic relevance.
* Comparing survey results with external benchmarks to check plausibility.
* Ensuring all services (e.g., transport, sanitation, parks) were adequately represented across cities and demographics.

Discrepancies identified during validation were documented, and corrective actions were taken.

**6. Documentation and Team Coordination**

All steps of the collection and integration process were carefully documented, including:

* Source details and access methods
* Pipeline architecture and code repositories
* Data dictionaries and schema definitions

Regular meetings were held with team members to align on progress, share issues, and review data quality checks. Feedback from the team helped refine the structure of the final integrated dataset.

**Conclusion**:  
Week 5 marked a major technical milestone by transforming scattered public feedback into a unified, high-quality dataset. The pipelines and structures established this week form the backbone for the upcoming analytics and modeling phases. The team will now proceed to exploratory data analysis (EDA) and NLP modeling in Week 6 with a reliable data foundation in place.