# **Week 6 Report: Data Sources & Pipeline Development (Collection & Integration Phase)**

During the sixth week of the InsightNation project, we continued our focus on the **Collection and Integration Phase**, with specific emphasis on identifying data sources and building robust data pipelines. These steps are essential to ensure the successful acquisition, quality assurance, and integration of citizen feedback data into the InsightNation analytics platform.

The first step in this week’s workflow involved the **identification and validation of data sources**. We conducted a detailed review to confirm that each data source aligns with the project’s objectives of understanding public sentiment and improving service delivery. We finalized sources including structured survey datasets, public government feedback repositories, and social media content (primarily from Twitter and Reddit). These sources offer a diverse blend of structured and unstructured data, critical for gaining comprehensive insights into public opinions.

Once the sources were finalized, we moved forward with **establishing data pipelines**. Using Python-based scripts, we developed ETL (Extract, Transform, Load) routines to systematically collect and stage incoming data. Survey data was imported in batch processes, while social media content was collected using Twitter’s API and Reddit’s public JSON feeds. Each source had its own collection method, but we designed a unified structure to ingest them into a central storage system.

Parallel to this, **data quality assurance** processes were implemented to maintain integrity and consistency across all incoming datasets. We applied data validation rules to check for null values, type mismatches, and outlier entries. For instance, citizen satisfaction scores outside the expected range (e.g., 1–5) were flagged for review. Deduplication logic was applied to remove repeated entries, especially in social media data, where reposts and retweets could create redundancy.

A significant milestone this week was the **integration of all collected data into a single structured dataset**. We created a master dataset with harmonized field names, consistent data types, and uniform time stamps. This involved careful mapping between fields from various sources, such as aligning “service\_feedback” from surveys with “complaint\_text” from Twitter.

Next, we focused on **developing and testing automated data pipelines**. Cron jobs and scheduled Python scripts were implemented to automate the data fetching and preprocessing steps. This automation ensures the InsightNation platform can continuously and reliably update its datasets with minimal manual intervention.

To further improve performance, we worked on **optimizing the data collection and processing workflows**. This involved batching certain operations, reducing API call frequency, and caching intermediate results. These optimizations are expected to significantly reduce processing time as data volume scales.

Finally, we ensured that every step of this week’s work was thoroughly **documented**. This included maintaining records of all data sources, pipeline configurations, validation rules, and encountered issues. The documentation will serve as a key reference during later phases and for troubleshooting future data processing anomalies.

Overall, Week 6 was pivotal in building the data foundation for the InsightNation platform. With structured pipelines, verified data sources, and quality checks in place, the project is now well-positioned to begin the analytical and modeling phases in the coming weeks.