AMITY UNIVERSITY ONLINE

SOCIAL MEDIA ANALYTICS PROJECT

Data Integration Documentation

Student Name:
Hemant Borana

Course: Data Modeling and Visualization (DMV)

Module:
Module: 5

Date:
December , 2025

Semester: 5

SOCIAL MEDIA ANALYTICS PROJECT

Data Integration Documentation

Student Name: Hemant Borana
Course: Data Modeling and Visualization (DMV)
Module: 5
Semester: 5
Date: December , 2025

Contents

PROJECT OVERVIEW

This project integrates data from three different sources (CSV, JSON, and XML) to create a unified social media analytics dashboard. The goal was to combine structured and semi-structured data to analyze engagement patterns across multiple platforms.

Key Objectives:

- Successfully import data from multiple formats
- Clean and standardize data across sources
- Merge datasets using appropriate join strategies
- Create unified visualizations
- Demonstrate understanding of streaming data concepts

Technologies Used:

- Python 3.x
- Pandas for data manipulation
- JSON and XML libraries for parsing
- Matplotlib and Seaborn for visualization
- Jupyter Notebook for development

DATA SOURCES DESCRIPTION

CSV File: Social_Media_Posts.csv

Format: Structured tabular data
Records: 150 posts
Columns: 9 (post_id, platform, post_type, category, post_date, likes, shares, comments, reach)

Key Characteristics:

- Contains posts from 4 platforms: Facebook, Instagram, Twitter, LinkedIn
- Date range: October 2 - November 29, 2024 (58 days)
- Missing values: 7 missing categories, 6 missing shares
- Primary key: post_id

JSON File: User_Engagement.json

Format: Semi-structured nested data
Records: 539 engagement records from 100 users
Structure: Nested JSON with metadata and user engagement history

Key Characteristics:

- Contains user-level engagement data (likes, comments, shares)
- Includes sentiment analysis for comments
- Timestamps for each engagement action
- Time spent metrics per engagement
- Foreign key: post_id (links to posts)

XML File: Platform_Metrics.xml

Format: Hierarchical semi-structured data
Records: 4 platforms + 32 weekly records
Structure: Nested XML with platform statistics, performance, demographics, and weekly metrics

Key Characteristics:

- Platform-level aggregate metrics
- Demographic information (age groups, gender split, location)
- Performance indicators (engagement rates, conversion rates)
- 8 weeks of historical trends per platform
- Join key: platform name

DATA LOADING PROCESS

CSV Data Loading

Method: pandas.read_csv()

Steps:

1. Loaded CSV file using pd.read_csv()
2. Verified data shape: 150 rows × 9 columns
3. Checked data types (object, int64, float64)
4. Identified post_date as object type requiring conversion
5. Previewed first and last records to understand data range

Challenges:

- Date field stored as string, required conversion to datetime
- Float type for shares indicated presence of missing values

JSON Data Loading

Method: json.load() + pandas DataFrame conversion

Steps:

1. Opened and parsed JSON file using json.load()
2. Explored nested structure (metadata + user_engagement array)
3. Flattened nested engagement_history into individual records
4. Created DataFrame with 539 engagement records
5. Preserved user-level information (join_date, favorite_platform)

Challenges:

- Nested structure required manual flattening
- Variable fields (sentiment only present for comments/like_comment)
- Multiple engagement records per user needed proper handling

XML Data Loading

Method: xml.etree.ElementTree

Steps:

1. Parsed XML file using ET.parse()
2. Navigated hierarchical structure (root → platform → statistics/performance/demographics)
3. Extracted platform-level statistics into dictionary format
4. Separately extracted weekly metrics into time-series format
5. Converted both to pandas DataFrames

Challenges:

- Hierarchical structure required iterative extraction
- Multiple data types (statistics vs weekly trends) needed separate handling
- Text values required type conversion (int, float)

DATA CLEANING & TRANSFORMATION

Posts Data Cleaning (CSV)

Missing Value Handling:

- Filled 7 missing categories with 'Uncategorized'
- Filled 6 missing shares with 0 (assumed no shares recorded)

Data Type Conversions:

- Converted post_date from object to datetime

Feature Engineering:

- Created total_engagement = likes + shares + comments
- Calculated engagement_rate = (total_engagement / reach) × 100
- Extracted time features: post_month, post_day, post_weekday

Result: Clean dataset with 150 rows × 14 columns (no missing values)

Engagement Data Cleaning (JSON)

Missing Value Handling:

- Identified 254 missing sentiments (expected for likes/shares)
- Created has_sentiment flag for analysis

Data Type Conversions:

- Converted engagement_timestamp to datetime
- Converted user_join_date to datetime

Feature Engineering:

- Extracted engagement_hour from timestamp
- Extracted engagement_day (day name)
- Categorized time_spent into buckets: quick (<30s), medium (30-120s), long (>120s)

Result: Enhanced dataset with 539 rows × 13 columns

Platform Metrics Cleaning (XML)

Data Parsing:

- Extracted gender_split text into separate male_percentage and female_percentage columns
- Converted all numeric strings to appropriate types (int, float)

Feature Engineering:

- Calculated engagement_per_post = total_impressions / total_posts

Result:

- Platform stats: 4 rows × 14 columns
- Weekly metrics: 32 rows × 5 columns (clean, no missing values)

DATA INTEGRATION STRATEGY

Master Dataset Integration

Approach: Multi-step merging strategy

Step 1: Posts + Engagement

- Join type: LEFT JOIN
- Join key: post_id
- Reason: Keep all posts even if no engagement data exists
- Result: 542 rows (some posts have multiple engagement records)

Step 2: Add Platform Metrics

- Join type: LEFT JOIN
- Join key: platform
- Reason: Enrich with platform-level statistics
- Result: 542 rows × 41 columns

Validation:

- Verified all 150 unique posts preserved
- Confirmed 539 engagement records maintained
- Identified 3 posts without any engagement

Aggregated Summaries

Post-Level Summary:

- Aggregated engagement counts per post
- Calculated average time spent per post
- Result: 150 rows (one per post) × 17 columns

Platform-Level Summary:

- Aggregated metrics by platform
- Combined with XML platform metadata
- Result: 4 rows (one per platform) × 21 columns

Daily Trends:

- Time-series aggregation by date and platform
- Result: 164 rows × 5 columns

User Behavior Summary:

- Aggregated patterns by user
- Identified most active platform per user
- Result: 100 rows × 6 columns

Join Type Justification

LEFT JOIN used because:

- Preserved all posts (primary dataset)
- Some posts legitimately have no user engagement yet
- Platform metrics should apply to all posts from that platform

Alternative considered:

- INNER JOIN would lose posts without engagement (not desired)
- RIGHT JOIN would prioritize engagement over posts (incorrect priority)

VALIDATION & QUALITY CHECKS

Data Integrity Checks

Duplicate Check:

- Verified no duplicate post_ids in post summary: ✓ PASS (0 duplicates)

Platform Consistency:

- Master data platforms: 4 unique ✓ PASS
- Platform summary: 4 unique ✓ PASS
- All platforms match expected values

Date Range Validation:

- Posts: Oct 2 - Nov 29, 2024 ✓ PASS
- Engagement: Oct 2 - Nov 30, 2024 ✓ PASS
- Engagement dates overlap with post dates (expected)

Record Count Validation:

- Original engagement records: 539
- Engagements in master: 539 ✓ PASS
- No data loss during merging

Data Quality Metrics

Completeness:

- Posts: 100% complete after cleaning
- Engagement: 47% have sentiment (expected for comment types only)
- Platform metrics: 100% complete

Consistency:

- All platform names standardized across sources
- Date formats consistent after conversion
- Numeric ranges realistic and validated

Accuracy:

- Engagement rates calculated correctly
- Aggregations verified against source data
- Time calculations validated

FINAL OUTPUTS

Integrated Datasets

Master Dataset:

- Dimensions: 542 rows × 41 columns
- Contents: Posts + Engagement + Platform metrics
- Use case: Detailed record-level analysis

Post Summary:

- Dimensions: 150 rows × 17 columns
- Contents: Aggregated metrics per post
- Use case: Post performance analysis

Platform Summary:

- Dimensions: 4 rows × 21 columns
- Contents: Platform-level KPIs
- Use case: Cross-platform comparison

Daily Trends:

- Dimensions: 164 rows × 5 columns
- Contents: Time-series engagement data
- Use case: Trend analysis and forecasting

User Summary:

- Dimensions: 100 rows × 6 columns
- Contents: User behavior patterns
- Use case: Audience segmentation

Unified Dashboard

9 Comprehensive Visualizations:

1. Total Reach by Platform (bar chart)
2. Average Engagement Rate by Platform (horizontal bar)
3. Content Distribution by Category (pie chart)
4. Weekly Engagement Trends by Platform (line chart)
5. Average Engagement by Post Type (grouped bar)
6. User Engagement Type Distribution (bar chart)
7. Average Time Spent by Sentiment (horizontal bar)
8. Followers vs Engagement Rate (bubble chart)
9. Weekly Engagement Heatmap (heatmap)

Dashboard Features:

- Platform filters implicit in visualizations
- Comparative analysis across platforms
- Time-based trend identification
- Engagement pattern insights

Streaming Data Simulation

Demonstration:

- Simulated 5 real-time posts with engagement data
- Showed batch vs stream processing concepts
- Explained hybrid integration approach
- Provided production implementation guidelines

CONCLUSION

This project successfully integrated three diverse data sources into a unified analytics system. The key achievements include:

- **100% data source coverage:** All three formats (CSV, JSON, XML) successfully parsed
- **Zero data loss:** All 150 posts and 539 engagements preserved through integration
- **Comprehensive cleaning:** Missing values handled, data types standardized
- **Effective merging:** Appropriate join strategies maintained data integrity
- **Rich visualizations:** 9-chart dashboard provides actionable insights
- **Scalability consideration:** Streaming data concepts demonstrated for real-time extension

The integration approach is production-ready and can be extended to handle larger datasets and real-time streaming data.

AMITY UNIVERSITY ONLINE

AMITY UNIVERSITY ONLINE

THANK YOU

FOR YOUR TIME AND ATTENTION

CONTACT INFORMATION

Phone: +91 9284494154
Email: hemantpb123@gmail.com
LinkedIn: www.linkedin.com/in/hemant-parasmal-borana

Submitted by:
Hemant Borana

Date:
December 2025