Milestone 1

Data Requirements & Sources

GenAI - MGTA 495
April 17, 2025

Group 11: Eleanor Amidei, Carolina Reta, Erin Yabsley

Primary Data Sources and Required Formats

The system would require structured data from both public APIs and internal artist workflows. Potential data types include:

- Streaming data from platforms such as Spotify and YouTube to analyze fan geographies and forecast attendance.
- Venue information, including capacity, genre compatibility, and availability, which could be pulled from online directories or ticketing platforms.
- Offer sheets and performance contracts, which are typically exchanged in PDF or Word format.
- Travel and accommodation quotes, used to support logistics planning, sourced from platforms such as Expedia or Kayak.
- Promotional engagement data from Instagram, TikTok, and marketing tools to monitor campaign effectiveness and target demographics.
- Artist financial and legal documents, such as W-9 tax forms and invoices.

The ideal data formats for these sources would include JSON, XML, CSV for structured data and PDF/DOCX (processed via OCR) for legal and contract-related documents.

Data Collection and Simulation Methods

As part of the system's core architecture, automated data collection would be implemented through third-party APIs to ensure timely and scalable access to relevant tour-planning information.

- Public APIs, including Spotify for Artists, OpenWeatherMap, Ticketmaster, and Meta/Instagram Graph API.
- Web scraping, used selectively when APIs are unavailable (e.g., venue metadata or festival calendars).
- Manual input, where managers or artists provide subjective preferences, availability, or branding guidance.
- Simulated data generation, used in development and prototyping phases when real data is inaccessible or restricted.

Data Privacy and Security Considerations

Given the sensitivity of some data handled by artist teams—particularly legal agreements, financial information, and personal contact details—privacy and security would be critical. The system should:

- Encrypt all sensitive data at rest and in transit.
- Implement role-based access control so only authorized personnel can view or modify confidential information.
- Adhere to relevant regulations.
- Obtain explicit user consent for data access.

Data Quality Requirements and Validation

AI-driven automation would rely on clean, complete, and timely data. To ensure accuracy:

- Fields such as artist location, venue capacity, and legal terms would need to be validated against known standards or schemas.
- Cross-platform consistency checks could prevent duplication or error (e.g., comparing venue names and addresses).
- Validation could flag missing or non-standard clauses in contracts.

Failure to ensure quality could lead to misbooking, miscommunication, or legal risks, especially in contract and logistical planning stages.

Access Requirements

Access to external and internal data would require various forms of credentialed authentication:

- API keys and OAuth tokens for external services (e.g., Spotify, Expedia, OpenWeather).
- User authentication for internal roles (e.g., artist manager, promoter, legal team).
- Secure data storage systems (e.g., AWS or Firebase with encryption and audit logging).