

Music Festival Analytics & Risk Prediction Platform

Présentation Théorique du Projet

Décembre 2025

1 Introduction

Ce projet vise à créer une plateforme analytique avancée capable de prédire les incidents de sécurité et d'optimiser les ressources lors de grands festivals de musique. En combinant des données historiques et des flux en temps réel, nous passons d'une gestion de foule réactive à une stratégie de prévention proactive.

2 Sources de Données (Data Sources)

Le projet repose sur l'ingestion de quatre types de données hétérogènes :

- **Foursquare Check-ins** : Données mondiales de fréquentation pour modéliser les flux de mouvement habituels.
- **EMS Emergency Data (Austin, TX)** : Jeu de données historique des appels d'urgence (Kaggle) pour identifier les corrélations entre types d'incidents et événements.
- **OpenWeather API** : Historique et prévisions météo pour analyser l'impact de la chaleur ou des orages sur la santé des festivaliers.
- **Réseaux Sociaux (X/Twitter)** : Analyse de sentiment pour détecter les signes avant-coureurs de mouvements de foule ou de mécontentement.

3 Modélisation : Data Warehouse (Star Schema)

L'architecture de stockage est optimisée pour la performance des requêtes analytiques (OLAP) via un schéma en étoile.

3.1 Table de Faits

La table centrale est `Incident_Fact`. Chaque ligne représente un événement nécessitant une intervention.

Mesures = {Temps de réponse, Nombre de victimes, Gravité}

3.2 Dimensions

- **Dim_Lieu** : Coordonnées géospatiales, zones (VIP, Main Stage, Food Court).
- **Dim_Temps** : Granularité fine (heure, jour, créneau horaire).
- **Dim_Weather** : Température, humidité, vent, alertes météo.
- **Dim_Event_Type** : Genre de musique, démographie estimée de l'artiste.

4 Data Mining et Intelligence Artificielle

Nous appliquons plusieurs modèles pour extraire de la valeur des données :

1. **Règles d'Association** : Identifier les facteurs aggravants.

$$\{\text{Canicule, Artiste EDM}\} \rightarrow \{\uparrow \text{Insulations}\} \quad (\text{Confidence} : 0.85) \quad (1)$$

2. **K-Means Clustering** : Segmentation du site du festival en zones de danger (High-Risk Zones) basées sur la densité et l'historique.
3. **KNN (K-Nearest Neighbors)** : Classification en temps réel du type d'incident probable en fonction du contexte actuel.
4. **Séries Temporelles** : Prévision de la charge de travail des secours (Y_t) :

$$\hat{Y}_{t+h} = f(Y_t, \text{Météo}_t, \text{Affluence}_t) \quad (2)$$

5 Architectures Technologiques

5.1 Option A : The Big Data Engineer (Scalabilité)

- **Streaming** : Apache Kafka pour le flux d'appels 911.
- **Data Lake** : MinIO (S3-compatible) sur Docker.
- **Processing** : Apache Spark & Apache Sedona (pour le traitement spatial).
- **Orchestration** : Apache Airflow.

5.2 Option B : The Smart Hybrid (Agilité)

- **Database** : PostgreSQL avec l'extension **PostGIS** pour les calculs géographiques.
- **Processing** : Spark en mode Local (Processing lourd en container).
- **Dashboard** : Streamlit (Python) pour une visualisation interactive et rapide.

6 Dashboard et Visualisation

La sortie finale est un tableau de bord comprenant :

- Une **Heatmap en temps réel** des risques par zone.
- Un système de **recommandation d'allocation** (ex : "Déployer 2 unités mobiles vers Stage B").
- Un overlay de l'analyse de sentiment des réseaux sociaux.