Analyse sur la criminalité à Chicago

By Akram ZAHRY

Cadre Logiciel Pour Le Big Data

Département Informatique

01/01/2025

Table des matières

1 Sujet

L'objectif de votre travail consiste à analyser des données de criminalité afin d'extraire des informations pertinentes sur divers aspects tels que :

1. **Le taux d'arrestation par zone communautaire :**
   Identifier les zones avec les taux d'arrestation les plus élevés.
2. **Transformation des données :**
   Nettoyer et préparer les données pour qu'elles soient exploitables, notamment en traitant les colonnes inutiles, les valeurs manquantes et les formats de dates.
3. **Heures critiques pour les crimes :**
   Déterminer les heures où le nombre de crimes est le plus élevé.
4. **Hotspots de kidnapping et vol :**
   Identifier les emplacements où ces crimes spécifiques sont les plus fréquents.
5. **Distribution des types de crimes par district :**
   Étudier la répartition des crimes par type dans chaque district.

2 Problématique

L'analyse des données de criminalité pose plusieurs défis en raison de leur volume et de leur hétérogénéité. Ces défis peuvent être résumés en trois grandes questions :

1. **Comment structurer et nettoyer les données efficacement ?** Les données peuvent contenir des valeurs manquantes, des doublons, ou des colonnes inutiles qui rendent leur exploitation difficile.
2. **Quels outils utiliser pour effectuer des analyses à grande échelle ?** L'utilisation d'outils adaptés, tels que Spark et Hive, est essentielle pour traiter un volume important de données tout en assurant la reproductibilité des résultats.
3. **Comment extraire des insights pertinents à partir des données transformées ?** L'objectif final est d'identifier les zones et moments les plus à risque afin d'aider à la prise de décision pour la sécurité publique.

3 Technologies

Dans ce projet, plusieurs technologies ont été utilisées pour traiter et analyser les données de criminalité. Les principales technologies sont les suivantes :

1. **Google Cloud** : Plateforme de cloud computing utilisée pour le stockage des données dans Google Storage et le traitement des données via un cluster Dataproc.
2. **Dataproc** : Service de traitement de données massives de Google Cloud, utilisé pour gérer un cluster avec un nœud master et deux workers pour l'exécution des tâches de calcul.
3. **Looker** : Outil de visualisation de données qui permet de créer des rapports interactifs pour explorer et analyser les données de manière intuitive.
4. **Spark** : Framework de traitement de données massives utilisé pour analyser et transformer de grandes quantités de données de manière distribuée.
5. **Scala** : Langage de programmation utilisé avec Spark pour le développement d'analyses distribuées et efficaces.

4 Infrastructure

L'infrastructure Dataproc utilisée repose sur un cluster créé avec le système managé Dataproc, composé d'un nœud maître et de deux nœuds workers. Voici les détails de l'infrastructure :

1. **Zone** us-east1-c
2. **Cluster Name** cluster1
3. **Cluster UUID** fe0b7e1b-0f92-4e3b-bdf9-fc957163eb5c
4. **Type de machine** - Nœud maître : n1-standard-2 - Nœuds workers : n1-standard-2
5. **Disques** - Nœud maître : 1000 Go (disque de démarrage, type pd-standard) - Nœuds workers : 100 Go (disque de démarrage, type pd-standard)
6. **Image utilisée** dataproc-2-2-deb12-20241026-165100-rc02 (Debian 12)
7. **Préemptibilité** - Nœud maître : Non préemptible - Nœuds workers : Non préemptibles
8. **Configuration réseau** - Accès interne uniquement - Réseau privé par défaut - Zone réseau : us-east1-c - Comptes de service autorisés : cloud-platform
9. **Configuration des ressources** - **YARN** - Nombre de cœurs par nœud : 2 cœurs - Mémoire totale allouée : 12288 MB - Mémoire disponible : 12288 MB - Cœurs CPU alloués : 4 cœurs - **Hadoop** - Nombre de réplicas par bloc : 2 - Capacité totale HDFS : 210 Go - Capacité utilisée HDFS : 22 Mo - **Spark** - Mémoire pour le daemon Spark : 1920 MB - Nombre d'exécuteurs : 2 - Mémoire par exécuteur : 2688 MB - Nombre de cœurs par exécuteur : 1 - Taille maximale des résultats du driver : 960 MB
10. **Statut du cluster** RUNNING
11. **Historique du statut** - Création : 2024-12-22 - Passage à l'état RUNNING : 2024-12-22 - Arrêts et redémarrages fréquents avec des horaires détaillés.

5 Data

Les données que nous utilisons concernent les crimes survenus à Chicago depuis 2001 jusqu'à aujourd'hui. Ces données sont fournies sous forme d'un tableau avec 22 colonnes et 8,07 millions de lignes. Les colonnes sont les suivantes :

1. **ID** Identifiant unique pour chaque enregistrement.
2. **Case Number** Le numéro de dossier de la police de Chicago (Numéro du département des archives), qui est unique pour chaque incident.
3. **Date** La date à laquelle l'incident a eu lieu. Celle-ci peut parfois être une estimation.
4. **Block** L'adresse partiellement rédigée de l'incident, positionnée sur le même bloc que l'adresse réelle.
5. **IUCR** Le code de rapport criminel uniforme de l'Illinois (IUCR). Ce code est directement lié au type primaire et à la description.
6. **Primary Type** La description principale du code IUCR.
7. **Description** La description secondaire du code IUCR, une sous-catégorie de la description principale.
8. **Location Description** Description du lieu où l'incident a eu lieu.
9. **Arrest** Indique si une arrestation a été effectuée.
10. **Domestic** Indique si l'incident est lié à la violence domestique telle que définie par la loi sur la violence domestique de l'Illinois.
11. **Beat** Indique la zone géographique où l'incident a eu lieu. Chaque zone est une petite zone géographique de la police, avec une voiture dédiée. Trois à cinq zones forment un secteur de police, et trois secteurs forment un district. La ville de Chicago a 22 districts de police.
12. **District** Indique le district de police où l'incident a eu lieu.
13. **Ward** Indique le quartier (district du Conseil municipal) où l'incident a eu lieu.
14. **Community Area** Indique la zone communautaire où l'incident a eu lieu. Chicago possède 77 zones communautaires.
15. **FBI Code** Indique la classification criminelle selon le système de rapport d'incidents national du FBI (NIBRS).
16. **X Coordinate** La coordonnée x du lieu de l'incident dans la projection State Plane Illinois East NAD 1983. Cette coordonnée est décalée par rapport à l'emplacement réel pour des raisons de confidentialité, mais elle se situe sur le même bloc.
17. **Y Coordinate** La coordonnée y du lieu de l'incident dans la projection State Plane Illinois East NAD 1983. Elle est également décalée pour des raisons de confidentialité mais se situe sur le même bloc.

18. **Year** L'année où l'incident a eu lieu.
19. **Updated On** Date et heure de la dernière mise à jour de l'enregistrement.
20. **Latitude** La latitude de l'endroit où l'incident a eu lieu. Elle est décalée par rapport à l'emplacement réel pour des raisons de confidentialité.
21. **Longitude** La longitude du lieu de l'incident, également décalée pour des raisons de confidentialité.
22. **Location** Le lieu où l'incident a eu lieu, au format permettant la création de cartes. Ce lieu est également décalé pour préserver la confidentialité mais reste dans le même bloc.

Source des données : https ://www.kaggle.com/datasets/middlehigh/los-angeles-crime-data-from-2000 ?resource=download

6 Réalisation

Les différentes étapes réalisées pour l'analyse des données de criminalité sont les suivantes :

1. **Étape 1 : Préparation et transformation des données**
   — Suppression des colonnes inutiles comme *Updated On*, *ID*, etc.
   — Transformation de la colonne *Date* en séparant la date et l'heure, puis en ajoutant des colonnes supplémentaires *AM PM*, *Hour*, etc.
   — Remplacement des valeurs manquantes dans des colonnes comme *Location Description* et *Community Area* par leurs valeurs les plus fréquentes.
2. **Étape 2 : Calcul des statistiques par zone**
   — Groupement des données par zone communautaire pour calculer le total des crimes, des arrestations, et le taux d'arrestation (%).
3. **Étape 3 : Analyse temporelle des crimes**
   — Conversion des heures au format 24h pour une agrégation plus précise.
   — Identification des heures critiques où le taux de criminalité est le plus élevé.
4. **Étape 4 : Hotspots de kidnapping et vol**
   — Filtrage des données pour ne retenir que les crimes de type *KIDNAPPING* ou *ROBBERY*.
   — Groupement par emplacement et description pour identifier les zones où ces crimes sont les plus fréquents.
5. **Étape 5 : Distribution des types de crimes par district**
   — Groupement des données par District et Primary Type pour analyser la distribution des crimes dans chaque district.

7 Conclusion

L'analyse des données de criminalité à partir des incidents recensés à Chicago met en lumière des tendances importantes, tant sur le plan temporel que géographique. La préparation et la transformation des données ont permis de structurer les informations et d'identifier les zones critiques où les taux de criminalité sont les plus élevés. Les statistiques par zone communautaire et les analyses temporelles ont révélé des insights clés, comme les heures et emplacements les plus sujets aux crimes, ainsi que les types d'incidents les plus fréquents dans chaque district.

Les résultats obtenus fournissent une base solide pour la prise de décision, notamment dans la répartition des ressources policières, l'élaboration de politiques publiques et la sensibilisation communautaire. Ils peuvent également être utilisés pour développer des outils prédictifs et des modèles visant à prévenir des crimes spécifiques.

Ce projet démontre la puissance des approches basées sur les données pour comprendre et répondre aux enjeux sociétaux complexes. En exploitant davantage de données et en intégrant des techniques avancées comme l'apprentissage automatique, les analyses futures pourront encore affiner les recommandations et contribuer à une sécurité publique accrue.