Technologies Cloud-native et Big Data : les clés pour tirer le meilleur parti de l'intelligence géospatiale

Une analyse Big Data plus agile

Pour beaucoup d'entreprises, la mise en œuvre du Big Data n'est pas synonyme de rentabilité. Nombre d'entre elles peinent à tirer des informations stratégiques pertinentes de leurs vastes ressources de données.

Le traitement géospatial vient cependant changer la donne. Grâce à l'enrichissement des données et à la création de zonage automatique, l'intelligence géospatiale Big Data et Cloud-native leur permet de transformer d'importants volumes de données en informations exploitables. Elles peuvent exécuter des opérations spatiales au sein d'environnements natifs et Big Data, puis utiliser les résultats de l'intelligence géospatiale pour des applications d'IA liées au machine learning et à l'analyse augmentée.

Precisely offre une approche unique en intégrant la technologie de localisation aux processus de conteneurs et Big Data. En préférant l'intégration à la connexion, vous interprétez les données transactionnelles plus rapidement et pouvez résoudre vos problématiques stratégiques en bénéficiant de la clarté nécessaire.

Le défi des environnements Big Data et Cloud-native

S'appuyant sur les environnements Big Data et Cloud-native, les entreprises stockent et traitent des jeux de données colossaux : appels clients, transactions financières, flux de réseaux sociaux, etc. Cependant, il leur est souvent difficile de générer des informations stratégiques pertinentes à partir de ces données. Alors que le volume et la vélocité des données ne cessent d'augmenter, les indicateurs de performance clés restent flous.

Le défi consiste à connecter les données non seulement au sein des jeux de données mais également entre les jeux de données, de manière à :

- Assurer leur exactitude et leur précision
- Permettre l'enrichissement sans perturbation des processus existants
- S'adapter à la vitesse et aux volumes extrêmement élevés des données

Une technologie vous permettant de générer des informations plus riches et d'accélérer votre RSI

- Évolutivité élevée, traitement des données à haute vitesse
- Géo-enrichissement
- Partitionnement des données au niveau cluster
- Traitement des données au niveau nœud

De précieuses applications

Grâce à l'approche intégrée, les entreprises peuvent formuler des problématiques métier et les résoudre au sein d'un environnement Big Data/Cloud-native. Par exemple :

Les entreprises de télécommunications traitent en permanence un volume colossal d'enregistrements d'appels. Ceux-ci peuvent être classifiés et présentés dans des cartes de couverture extrêmement précises, quasiment en temps réel. Ce type d'analyse visuelle aide les entreprises à améliorer leur service client, à réduire les résiliations, à cibler leur clientèle plus efficacement et à gagner des parts de marché.

Les cabinets de services financiers traitent en continu un grand nombre de transactions. Chacune d'elles peut être géo-enrichie avec des centaines d'attributs axés sur les biens et les zones environnantes, et des modèles opérationnels permettent de mieux prédire les valeurs d'emprunt ou des biens. Ces établissements peuvent tirer parti de ce processus pour mieux estimer les valeurs d'emprunt et maximiser la valeur monétaire pour l'entreprise.

L'avantage de l'intelligence géospatiale

L'intelligence géospatiale apporte un éclairage nouveau sur l'analyse des données. Le Big Data contient souvent des informations de localisation : adresse d'un client, signal GPS de téléphone mobile, emplacement d'un distributeur de billets, transaction en magasin, localisation sur un réseau social, etc.

Grâce au géo-enrichissement, un processus ajoutant un contexte aux données métier, les organisations peuvent enrichir leurs enregistrements avec des attributs de tiers et les résultats de requêtes géospatiales.

Grâce à ces informations intégrées et enrichies :

- Des workflows basés sur des règles peuvent exploiter ces données ajoutées pour automatiser les décisions stratégiques.
- L'agrégation spatiale peut condenser les volumes de données pour en simplifier la gestion.
- Les données peuvent être généralisées dans un contexte spatial, pour des résultats plus faciles à modéliser et à visualiser.
- Les organisations peuvent obtenir de nouvelles perspectives sur les enjeux opérationnels et les réponses à mettre en place.

Les avantages d'une approche native et optimisée

De nombreux fournisseurs de technologies géospatiales proposent des solutions qui se connectent aux plateformes Big Data et Cloud-native, puis transfèrent les données de ces plateformes distribuées vers leur propre technologie SIG basée sur serveur. Cette approche de type « connecteur » présente cependant un inconvénient : elle ne tire pas parti de la puissance de traitement de la plateforme distribuée (EMR, Cloudera ou Spark, par exemple.).

En fait, les opérations géospatiales elles-mêmes sont effectuées sur un serveur individuel ou sur un petit cluster de serveurs, ce qui limite votre capacité à traiter des jeux de données volumineux.

Precisely adopte une approche différente ; une approche distribuée, élastique et éphémère. Les opérations géospatiales peuvent être exécutées de façon native sur de nombreux types de plateformes distribuées, optimisant les capacités des environnements de traitement distribués.

Les sections qui suivent présentent la technologie que nous proposons, puis les étapes de traitement permettant de l'utiliser de façon native dans un environnement Big Data.

Une technologie innovante et modulaire

La solution Spectrum Location Intelligence for Big Data comprend des kits de développement logiciel (SDK) dédiés à la technologie de localisation. Ils permettent aux entreprises d'enrichir leurs propres informations clients et d'assurer l'agrégation spatiale des résultats afin de concentrer les données sous une forme exploitable. Cette solution inclut également des API et des données.

- Nos SDK Java peuvent être transférés vers n'importe quel environnement Big Data, tel que Kubernetes ou Spark. Dans un environnement dynamique et en perpétuelle évolution, les entreprises peuvent ainsi librement choisir leur technologie.
- Nous proposons plus de 350 jeux de données pouvant être utilisés pour ajouter un contexte spatial et servir de conteneur pour l'agrégation. Ils peuvent être analysés et visualisés à l'aide de nos technologies de cartographie basées sur le Web.

Exemples de performances atteintes grâce à la stratégie technologique Cloud-native et optimisée Precisely

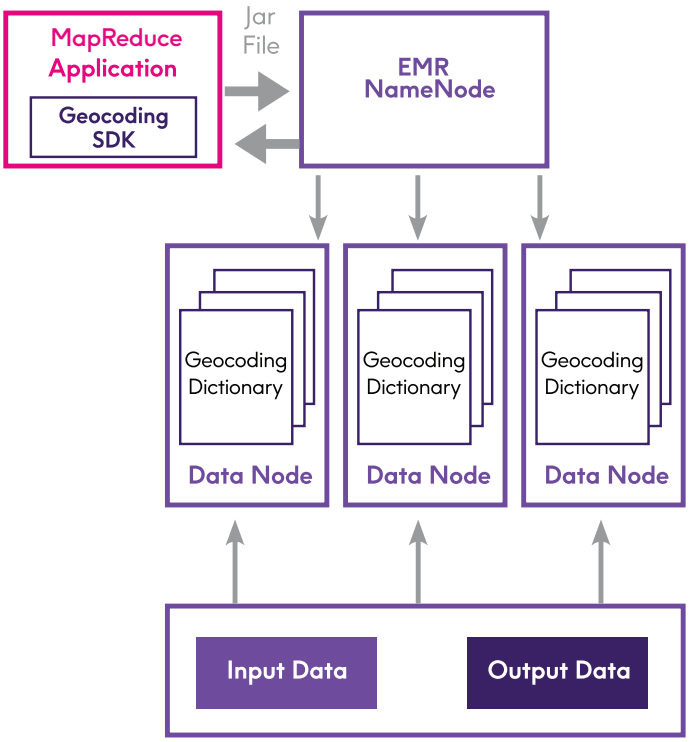
Il s'agit de chiffres réellement atteints par nos clients. Nous améliorons notre technologie en permanence et pouvons atteindre des performances encore bien supérieures avec des technologies plus récentes telles que Spark.

Géocodage	Géocodage au centroïde de la parcelle de 106 millions d'adresses aux États-Unis	**30 minutes**	Cluster Hadoop à 5 nœuds
Jointure spatiale « Rechercher l'entité la plus proche »	Jointure spatiale entre 1 milliard de points mobiles et 12 millions de points d'intérêt	**36 minutes**	EMR à 20 nœuds sur AWS
Traitement de type « point dans un polygone »	Agrégation de 19 milliards d'enregistrements d'appels mobiles dans 950 millions de polygones	**30 minutes**	Cluster Hadoop à 56 nœuds

Notre technologie à l'action

Dans le schéma ci-dessous, le SDK de géocodage est utilisé pour illustrer cette architecture et la façon dont elle peut être mise en œuvre dans une variété de processus liés au Big Data. Il montre comment Precisely intègre des capacités de géocodage à Hadoop de façon native. Le SDK de géocodage et les fichiers de dictionnaires de géocodage sont les éléments clés de la solution.

- Le SDK de géocodage est un ensemble de fichiers Jar pouvant être utilisés pour l'écriture d'applications Java EMR ou Spark.
- Les fichiers de données des dictionnaires de géocodage peuvent être pré-installés sur tous les nœuds de données EMR ou distribués dynamiquement sur le cluster avant utilisation.



Spectrum Spatial pour le Big Data

Spectrum Location Intelligence for Big Data (SDK)

- Prend les primitives spatiales (points, lignes et polygones) et applique une fonction géométrique (« contient », « combine », « intersecte », etc.) à l'aide de données spatiales et non spatiales supplémentaires.
- Permet de créer une requête spatiale, par exemple pour agréger des points de données au sein d'un polygone ou trouver le point le plus proche d'une ligne.
- Peut être utilisé pour géo-enrichir un jeu de données en ajoutant des attributs supplémentaires à partir de données de clients, de données de tiers ou de l'un des jeux de données du catalogue de données mondiales Precisely (plus de 350 jeux de données proposés).
- Les clients aux États-Unis peuvent également utiliser les ressources MLD (Master Location Database) pré-enrichies pour les adresses postales afin d'accroître les vitesses de traitement et d'améliorer la précision de la localisation pour les workflows opérationnels.

Spectrum Geocoding for Big Data (SDK)

- Le géocodage transforme une adresse, un emplacement ou un point d'intérêt en paire de coordonnées (latitude, longitude).
- Le géocodage inversé retourne une adresse ou une limite administrative à partir de coordonnées.

Spectrum Routing for Big Data (SDK)

- Prend un emplacement connu (tel qu'un magasin) et utilise le réseau routier pour en tirer des informations telles que les temps de conduite équivalents (isochrones) autour de ce point ou le trajet le plus court jusqu'à ce point.

Exécution du géocodage dans le Big Data

Voici un exemple de la façon dont Precisely peut exécuter le géocodage sous forme de tâche EMR par lots dans une ligne de commande. Notez que l'utilisateur peut configurer différents paramètres de géocodage dans le fichier config.xml, tels que le dictionnaire à utiliser et les champs à retourner. Le géocodage avant et le géocodage inversé sont pris en charge.

```
[jun@osboxes ~]$: hadoop jar Geocoding_Hadoop.jar com.
pb.mr.GeocodingDriver -input /addressdatafolder -output /
geocoderesult -appConfig Geocode_config.xml
```

Vers une meilleure accessibilité

Les utilisateurs familiarisés avec le data engineering utiliseront efficacement les tâches EMR par lots. Cependant, elles sont peu conviviales pour les autres analystes de données. Pour rendre plus accessible le géocodage dans EMR, nous avons développé une fonction de géocodage définie par l'utilisateur, permettant à toute personne familiarisée avec SQL de l'utiliser dans EMR. La plupart des capacités d'intelligence géospatiale Precisely peuvent être déployées dans EMR ou Spark via une approche similaire à l'exemple de géocodage plus haut.

```
HIVE> select geocode (street, city, state, zip, 'USA') from
customersAddTable;
```

Optimisation du traitement des données géospatiales dans EMR et Spark

Le traitement des données géospatiales représente une étape fondamentale dans la quasi-totalité des applications Big Data faisant intervenir la localisation. Par exemple, pour analyser les enregistrements des traces mobiles d'utilisateurs contenant leurs localisations GPS, des données auxiliaires sont ajoutées afin d'apporter un contexte (adresse de l'individu ou point d'intérêt à proximité, par exemple). Pour assurer la mise en œuvre de ces processus de géo-enrichissement à grande échelle, un ensemble de processus géospatiaux extrêmement performants (recherches de site de type « trouver l'entité la plus proche » ou « point dans un polygone ») est nécessaire. Pour tirer parti de la puissance de calcul à grande échelle, ces processus doivent être optimisés pour les technologies Big Data comme EMR ou Spark.

Si l'on prend pour exemple l'utilisation de l'analyse « point dans un polygone » dans EMR, différents types de stratégies peuvent être identifiés. Ils dépendent des cas d'utilisation et des données à analyser.

Dans de nombreux cas d'utilisation, le nombre de polygones à évaluer est faible. Il peut alors être suffisant d'utiliser un Broadcaster pour l'évaluation, par exemple pour déterminer si les enregistrements de points sont inclus dans des polygones représentant des limites administratives. L'agrégation et l'analyse spatiales classiques s'inscrivent le plus souvent dans ces types de cas d'utilisation.

Dans le contexte de l'IoT (Internet des objets), cette approche simpliste basée sur les Broadcasters échoue dans de très nombreux cas d'utilisation. En effet, le volume de polygones à diffuser vers chaque nœud et à conserver en mémoire est alors bien trop important et les processus spatiaux deviennent excessivement lents. Une autre approche est donc essentielle. Precisely rend ces requêtes Big Data plus agiles, accélère et optimise les résultats, répondant aux besoins du marché.

L'approche Precisely en détail

La préparation des données est une étape critique pour assurer un processus spatial hautement performant. Elle implique d'optimiser non seulement le partitionnement des données spatiales au niveau cluster, mais également le traitement de ces données au niveau nœud.

Le partitionnement des données au niveau cluster détermine dans quelle mesure les grands jeux de données doivent être divisés pour être traités efficacement sur un nœud unique.

Le traitement des données au niveau nœud optimise l'indexation et le traitement spatiaux de petites portions du sous-ensemble de données dans un nœud local pour accélérer le traitement joint des requêtes. Le cas d'utilisation d'une analyse « point dans un polygone » à grande échelle ci-après illustre chacun de ces aspects.

Cas d'utilisation : « Point dans un polygone »

L'objectif : Associer des points de connexion mobile avec des informations GPS à des polygones de limite de magasin pour déterminer les modèles de visite de magasin des utilisateurs mobiles.

Le défi : Les deux jeux de données sont trop volumineux pour être importés sur une même machine (plusieurs téraoctets de points et plusieurs gigaoctets de polygones).

La solution : Une stratégie de partition et l'algorithme correspondant.

Partitionnement des données au niveau cluster

Le partitionnement des données au niveau cluster s'effectue en deux grandes étapes de traitement : le pré-partitionnement et la mise en correspondance.

- 01. Pré-partitionnement
- 02. Processus de mise en correspondance

01. Pré-partitionnement

Le pré-partitionnement utilise les attributs spatiaux inclus dans les données pour organiser les jeux de données dans un système de fichiers Big Data (par exemple, HDFS) avant l'exécution de l'application. Il permet d'interroger ou de traiter les données rapidement lorsque l'application est exécutée. La nature des données est d'abord examinée pour déterminer la meilleure approche de pré-partitionnement.

Cas d'utilisation

Dans ce cas d'utilisation, les données des connexions des mobiles sont diffusées en streaming chaque jour vers un stockage Big Data. Les données de limite de magasin sont mises à disposition par des fournisseurs de données comme Precisely et mises à jour chaque trimestre. Il est plus efficace de pré-partitionner les données de limite de magasin que les données d'utilisateurs mobiles et d'actualiser cette préparation une fois par trimestre, au moment où les données de limite sont mises à jour.

Plusieurs algorithmes permettent de partitionner les données de limite, par exemple les algorithmes orientés espace comme la grille Geohash, ou encore les algorithmes orientés données comme les arbres R. Les algorithmes orientés espace conviennent généralement mieux à un traitement parallèle et sont donc privilégiés au niveau cluster.

La grille régulière est l'algorithme le plus utilisé dans le secteur. **Figure 1 :** Exemple d'utilisation d'une grille régulière pour partitionner un grand jeu de données de polygones.

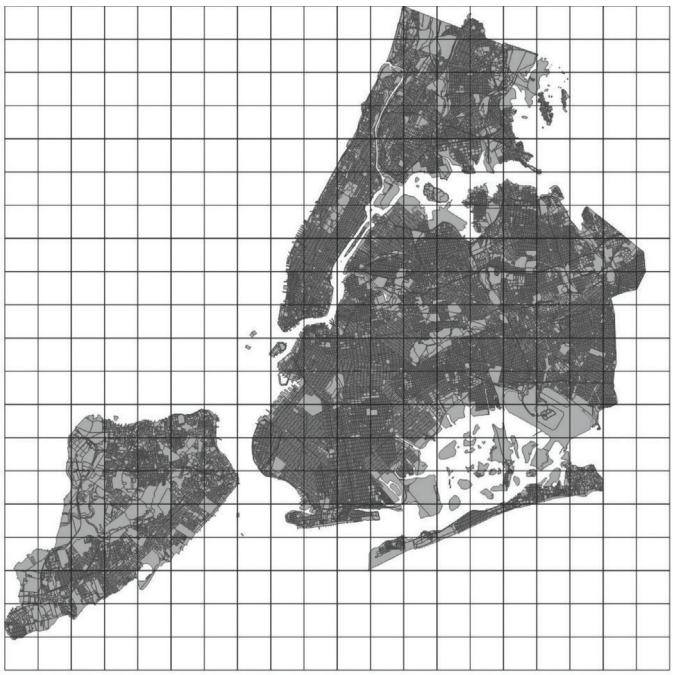


Figure 1

Équilibrage de la charge de données

La méthode basée sur la grille présente cependant un inconvénient majeur : la distribution des données spatiales est souvent faussée.

Cas d'utilisation

Il se peut qu'en une journée, des milliers d'enregistrements d'utilisateurs de données mobiles soient générés à la gare Grand Central de New York et que pas un seul ne soit généré dans le désert de l'Arizona.

La méthode basée sur la grille créera probablement des partitions avec des tuiles de données à haute densité qui, à leur tour, entraîneront des problèmes d'équilibre de charge dans un environnement Big Data de type cluster.

Pour résoudre ce problème, Precisely a développé deux algorithmes :

- Algorithme de grille bissectrice
- Algorithme de tuilage adaptatif

La **Figure 2** compare l'équilibre de charge de données avec l'approche basée sur la grille régulière et les deux nouveaux algorithmes. Plus la distribution des données est plate, moins les problèmes d'équilibrage de charge sont nombreux et meilleures sont les performances dans le cluster EMR. On observe une distribution beaucoup plus plate avec l'algorithme de tuilage adaptatif. Dans le cadre de tests d'analyse « point dans un polygone » avec de gros volumes de points et polygones, l'algorithme de tuilage adaptatif est plus de 20 fois plus performant que la méthode basée sur la grille régulière.

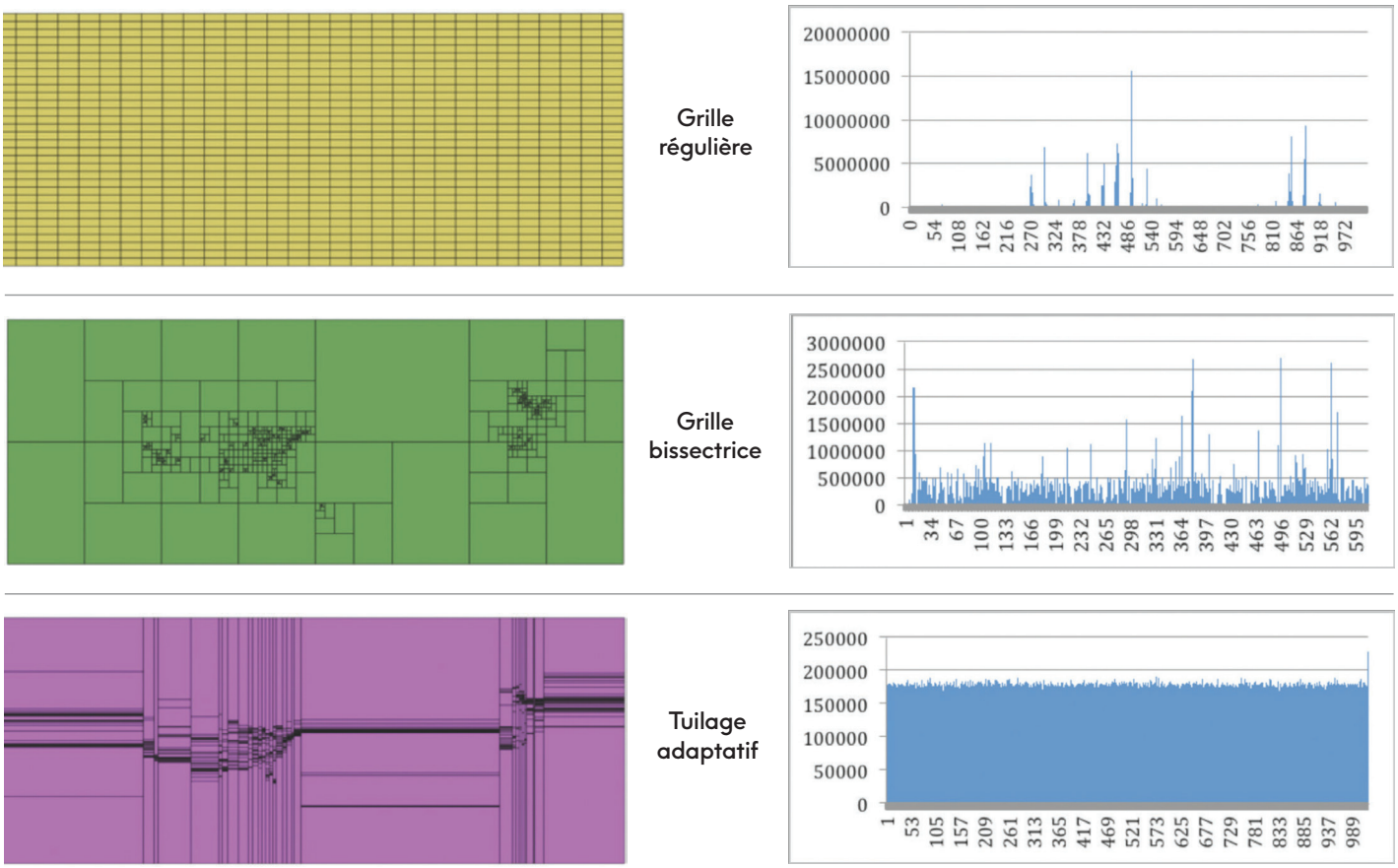


Figure 2

02. Processus de mise en correspondance

Après le pré-partitionnement des données de limite de magasin, les processus de requête ou de jointure spatiale peuvent être élaborés, par exemple à l'aide d'une application Big Data, comme illustré dans la Figure 3 :

- Toutes les limites de magasin pré-partitionnées seront d'abord chargées dans chaque partition avec une clé de partition.
- Les données de points mobiles seront ensuite chargées et mises en correspondance avec une clé de partition.
- Le processus de mise en correspondance est similaire à un simple processus de géohachage et peut s'effectuer rapidement.
- Les paires de données mises en correspondance sont alors importées dans le réducteur en vue de la jointure spatiale au niveau local.

Dans l'exemple ci-dessous, les enregistrements de données de points mobiles n'ont pas été pré-traités, même si un pré-traitement aurait été possible. Par exemple, si le processus avait requis des interrogations spatiales répétées ou une jointure spatiale, une étape de pré-partitionnement supplémentaire aurait pu être ajoutée pour pré-partitionner ce jeu de données de points avec les résultats de partition du jeu de données de limite de magasin.

Une étape de codage spatial pourrait également être exécutée : un algorithme de type géohachage serait appliqué aux champs de latitude/longitude dans chaque enregistrement de donnée de point entrant lors du processus de streaming ou d'importation de données. Une clé basée des grilles variables serait générée et ajoutée à ces enregistrements. Cette clé pourrait ensuite être utilisée dans une base de données HDFS ou NoSQL en vue de l'indexation ou de la partition du stockage de données, permettant, par la suite, d'effectuer des jointures ou interrogations spatiales rapides.

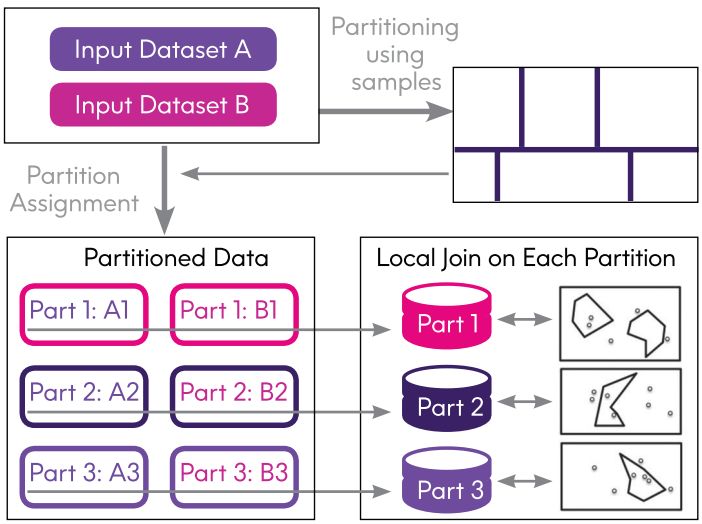


Figure 3

Traitement géospatial au niveau nœud

Le traitement géospatial inclut également deux grandes étapes de traitement : la constitution de l'index spatial local et l'application d'opérations géométriques détaillées.

Cas d'utilisation

Une fois les enregistrements de données de points mobiles et les données de limite de magasin partitionnés, mis en correspondance et envoyés vers les différents nœuds esclaves, l'opération au niveau nœud est très similaire au traitement géospatial sur une machine unique.

Constitution de l'index spatial local

Un index spatial local pour les données en mémoire du nœud est constitué dynamiquement au démarrage de l'application. Un algorithme d'index spatial orienté données (arbre R, par exemple) est généralement utilisé à ce niveau, plutôt qu'un algorithme d'index orienté espace, privilégié au niveau cluster.

Application d'opérations géométriques détaillées

Les opérations géométriques détaillées appropriées sont appliquées entre les données de points et les données de polygones. Différents types d'analyse « point dans un polygone » sont pris en charge, de l'analyse permettant simplement de déterminer si un polygone contient un point spécifique à l'analyse avancée pouvant également retourner la distance entre le point et les côtés du polygone.

Avec ce type d'analyse « point dans un polygone » basée sur les partitions, les utilisateurs peuvent associer des connexions mobiles dynamiques et des limites de magasin plus efficacement qu'avec des plateformes de serveurs classiques. Ils peuvent exécuter plusieurs analyses rapidement, chaque jour, pour générer des informations stratégiques sur les modèles de visite de magasin par les clients.

Ce type d'analyse spatiale haute précision, hautement évolutive et à haute vitesse sur les données mobiles était jusqu'ici très difficile à réaliser. Avec la solution Precisely, l'exécution est rapide et les informations stratégiques sont faciles à assimiler.

Une solution pour aujourd'hui et pour demain

La technologie Cloud-native/Big Data évolue rapidement et en permanence. Aujourd'hui, nous commençons à voir Spark remplacer Hadoop comme la dernière technologie Big Data, Kubernetes, et le rythme de l'innovation se poursuit.

Les cas d'utilisation varient également selon les entreprises, impliquant différentes options technologiques telles que le traitement par lots dans Spark, le streaming en temps réel dans Kubernetes, ou encore l'interrogation spatiale interactive dans des bases de données NoSQL comme HBase.

Dans cet environnement diversifié, en rapide évolution, Precisely adopte une approche agile vous offrant de précieux avantages :

- La possibilité d'appréhender au mieux le traitement spatial et la technologie Big Data dans chaque cas d'utilisation
- L'intégration de capacités de pointe dans la plupart des composants et plateformes Big Data
- Un gain de flexibilité pour répondre aux multiples attentes des utilisateurs
- L'application des capacités à chaque cas d'utilisation d'analyse spatiale avec une grande efficacité
- La maximisation de la nature distribuée des clusters Big Data pour optimiser les applications actuelles, impliquant de grands volumes de données
- En vous appuyant sur la technologie et les capacités Cloud-native/Big Data appropriées, vous pouvez obtenir des informations stratégiques pertinentes et en tirer le meilleur parti.

precisely

À propos de Precisely

Leader mondial en matière d'intégrité des données, Precisely garantit la précision et la cohérence des données de ses 12 000 clients, dont 90 % d'entreprises figurant au classement Fortune 100, dans plus de 100 pays. Les produits d'intégration de données, de qualité des données, d'intelligence géospatiale et d'enrichissement des données de Precisely permettent de prendre les meilleures décisions stratégiques, pour les meilleurs résultats.

Pour plus d'informations, rendez-vous sur precisely.com.