نظرة عامة على المشروع

هذا المشروع يهدف إلى تنفيذ **Slowly Changing Dimension Type 2 (SCD Type 2)**فيHive ، وهي تقنية تُستخدم لتتبع تاريخ التغييرات في بيانات العملاء مع الحفاظ على السجلات القديمة والجديدة معًا. الفكرة الأساسية هي أن كل تغيير في بيانات العميل يُنشئ سجلاً جديدًا بدلاً من تحديث السجل القديم.

هيكل الجداول المستخدمة

1. **الجدول الداخلي(customer_int)**

هذا جدول داخلي بالكامل تحت إدارة .Hive يتم استخدامه لتخزين البيانات الأولية للعملاء.

الأعمدة:

- CustomerID: معرف فريد للعميل
- Name: اسم العميل
- Emailالبريد الإلكتروني
- Phone_Number: رقم الهاتف
- Address: العنوان (قد يحتوي على فواصل داخلية)
- JOIN_Date: تاريخ انضمام العميل
- Start_Date: تاريخ بدء صلاحية السجل
- End_Date: تاريخ انتهاء صلاحية السجل
- Is_Current: يشير إلى ما إذا كان السجل هو الإصدار الحالي ('1') أم تاريخي('0')

ملاحظة مهمة :عند حذف هذا الجدول، يتم حذف البيانات نهائيًا لأنه جدول.managed

2. **الجدول الخارجي(customer_ext)**

جدول خارجي يشير إلى موقع محدد في.(HDFS (/user/itversity/customer_external_data

الميزة الرئيسية :عند حذف هذا الجدول، تبقى البيانات الفعلية في HDFS ولا تُحذف، مما يوفر طبقة أمان للبيانات.

3. **جدول الأبعاد(customer_dim)**

هذا هو الجدول الأساسي الذي يتم تطبيق منطق SCD Type 2 عليه. يحتوي على نفس أعمدة الجدول الداخلي مع إضافة التواريخ التي تحدد فترة صلاحية كل إصدار.

4. جدول المرحلة(customer_stage)

جدول مؤقت يُستخدم لاستقبال بيانات التحديثات الجديدة من ملف customer_updated.csv يحتوي على أعمدة أقل من الجدول الرئيسي لأنه لا يحتاج إلى تواريخ الصلاحية والحالة الحالية.

🔧معالجة مشكلة الفواصل في عنوان العميل

البيانات تحتوي على عنوان العميل (Address) الذي قد يتضمن فواصل داخل النص، مثل "Main St, 123 Apt 4B, New York".

الحل :تم استخدام **OpenCSVSerde**مع تحديد:

- separatorChar = ",": الفاصلة هي الفاصل بين الأعمدة
- quoteChar = "\"": علامة التنصيص تُستخدم لتجميع الحقول التي تحتوي على فواصل
- escapeChar = "\\": حرف الهروب للتعامل مع علامات التنصيص الداخلية

هذا يسمح لـ Hive بقراءة الملف بشكل صحيح دون الحاجة إلى تغيير الفاصل الأساسي.

📂تحميل البيانات

البيانات الأساسية

يتم تحميل ملف customer_scd2_mixed.csvإلى ثلاثة جداول:

- customer_int: الجدول الداخلي
- customer_ext: الجدول الخارجي
- customer_dim: جدول الأبعاد (كنقطة بداية)

بيانات التحديثات

يتم تحميل ملف customer_updated.csv إلى جدول customer_stage هذا الملف يحتوي على البيانات الجديدة التي قد تحتوي على تغييرات لبعض العملاء.

منطق (SCD Type 2)الجزء الأهم

بما أن Hive لا يدعم أوامر UPDATE أو DELETE بشكل مباشر، تم استخدام طريقة INSERT OVERWRITEلإعادة بناء الجدول بالكامل. الاستعلام يتكون من 4 أجزاء يتم دمجها باستخدام :UNION ALL

الجزء الأول: السجلات التاريخية

يتم الاحتفاظ بجميع السجلات التي هي تاريخية(Is_Current = '0') كما هي دون تغيير.

الجزء الثاني: السجلات الحالية التي لم تتغير

هذه السجلات لم يطرأ عليها أي تغيير في بيانات customer_stage يتم الاحتفاظ بها كسجلات حالية (Is_Current = '1').

الجزء الثالث: إغلاق السجلات التي تغيرت

عندما يتغير عميل ما، يتم أخذ سجله الحالي وتحويله إلى سجل تاريخي عن طريق:

- تعيين End_Date إلى تاريخ اليوم
- تغيير Is_Current من '1'إلى '0'

الجزء الرابع: إضافة السجلات الجديدة

بعد إغلاق السجل القديم، يتم إنشاء سجل جديد بنفس بيانات العميل ولكن مع:

- Start_Date= تاريخ اليوم
- End_date = (Null) لأنه السجل الحالي
- Is_Current = '1'

الشرط :يتم إضافة سجل جديد فقط إذا كان العميل غير موجود في customer_dim أو إذا كانت بياناته مختلفة عن السجل الحالي.

استعلامات التحقق

بعد تنفيذ عملية التحديث، تم إنشاء عدة استعلامات للتحقق من صحة البيانات:

1.عرض العملاء ذوي الإصدارات المتعددة

هذا الاستعلام يُظهر العملاء الذين لديهم أكثر من إصدار (سجل تاريخي + سجل حالي)، مما يثبت أن SCD Type 2يعمل بشكل صحيح.

2.إحصائيات عامة

يُظهر هذا الاستعلام:

- إجمالي السجلات في الجدول
- عدد السجلات الحالية(Is_Current = '1')
- عدد السجلات التاريخية(Is_Current = '0')
- عدد العملاء الفريدين

3.عرض السجلات الحالية

يُظهر أول 10 عملاء نشطين حاليًا.

4.عرض السجلات التاريخية

يُظهر أول 10 سجلات تم استبدالها بسبب التغييرات.

النتائج النهائية

بعد تنفيذ جميع الأكواد، تم الحصول على النتائج التالية:

المقياس

إجمالي السجلات

السجلات الحالية(Is_Current = '1')

السجلات التاريخية(Is_Current = '0')

العملاء الفريدين

هذه النتائج تُثبت نجاح تنفيذ:SCD Type 2

- العملاء الذين تغيرت بياناتهم لديهم إصداران (سجل تاريخي + سجل حالي)
- السجلات التاريخية تحمل تاريخ انتهاء محددًا
- السجلات الحالية تحمل End_Date = NULL

ملاحظات مهمة

1. **الفرق بين الجدول الداخلي والخارجي :**الجدول الداخلي يفقد بياناته عند الحذف، بينما الجدول الخارجي يحتفظ بالبيانات في.HDFS
2. **التعامل مع الفواصل :**استخدام OpenCSVSerde يحل مشكلة الفواصل داخل حقل العنوان دون الحاجة لتغيير الفاصل الأساسي.
3. **بديلUPDATE/DELETE :** بما أن Hive لا يدعم هذه الأوامر، تم استخدام INSERT OVERWRITE لإعادة بناء الجدول بالكامل.
4. **تتبع التاريخ :**كل تغيير في بيانات العميل يُنشئ سجلاً جديدًا مع الحفاظ على السجل القديم، مما يتيح تتبع تاريخ العميل بالكامل.

🎯خلاصة

هذا المشروع يُعد حلاً متكاملاً لإدارة أبعاد العملاء المتغيرة ببطء (SCD Type 2) في بيئةHive ، مع مراعاة القيود الخاصة بـ) Hive عدم دعم (UPDATE/DELETE والبيانات المعقدة (وجود فواصل داخل الحقول).