

Change Data Capture (CDC)

Xeratic Internship

Research and Development CDC

Research and Development (RND) pada CDC (Change Data Capture) mencakup pengembangan dan penelitian terkait metode, teknologi, dan algoritma yang terkait dengan CDC. RND CDC bertujuan untuk meningkatkan kinerja, efisiensi, dan kemampuan CDC dalam mereplikasi perubahan data dari sumber data ke sistem atau aplikasi tujuan.

Beberapa aspek RND yang terkait dengan CDC meliputi:

1. Algoritma dan Metode Pencocokan Data: RND dapat fokus pada pengembangan algoritma dan metode pencocokan data yang lebih baik dan lebih efisien. Ini melibatkan penelitian tentang teknik pencarian, pemetaan, dan identifikasi perubahan data yang lebih akurat dan cepat.
2. Pengoptimalan Kinerja: RND CDC juga dapat berfokus pada pengoptimalan kinerja CDC secara keseluruhan. Hal ini meliputi penelitian tentang cara mengurangi beban pada sumber data, mempercepat proses replikasi, dan mengoptimalkan penggunaan sumber daya.
3. Skema Penyimpanan dan Transfer Data: RND CDC dapat melibatkan penelitian tentang skema penyimpanan dan transfer data yang efisien. Hal ini meliputi pengembangan format data yang kompak dan efisien, metode kompresi data, teknik transfer data yang cepat dan aman, serta manajemen penyimpanan data yang optimal.
4. Integrasi dengan Teknologi Terkini: RND CDC juga melibatkan eksplorasi integrasi dengan teknologi terkini, seperti cloud computing, big data, dan teknologi real-time processing. Tujuannya adalah meningkatkan skalabilitas, fleksibilitas, dan ketahanan CDC dalam menghadapi lingkungan yang terus berubah.
5. Keamanan dan Keandalan: Aspek keamanan dan keandalan juga menjadi perhatian dalam RND CDC. RND dapat melibatkan penelitian tentang metode enkripsi data, mekanisme keamanan akses, pemulihan setelah kegagalan, dan pemantauan kualitas data.

Melalui upaya RND yang berkelanjutan, CDC dapat terus meningkatkan performa dan kemampuannya dalam mereplikasi perubahan data dengan akurat, cepat, dan efisien. Hal ini penting untuk mendukung kebutuhan bisnis yang semakin kompleks dan meningkatkan efektivitas penggunaan data dalam organisasi.

Defini CDC

Change Data Capture (CDC) adalah proses untuk mengidentifikasi dan merekam perubahan data yang terjadi pada sebuah sistem atau sumber data. CDC mencatat perubahan data yang dilakukan pada basis data dan memungkinkan sistem lain atau aplikasi untuk memperoleh informasi mengenai perubahan tersebut secara real-time atau dalam waktu nyata.

CDC digunakan dalam lingkungan yang melibatkan replikasi data, integrasi data, analisis data real-time, dan pemantauan data yang akurat. Dengan CDC, perubahan data yang terjadi pada sumber data dapat dikirimkan secara efisien dan akurat kepada sistem lain yang memerlukannya.

Proses CDC bekerja dengan melacak operasi perubahan data, seperti operasi penambahan (insert), pembaruan (update), atau penghapusan (delete) yang terjadi pada sumber data. CDC mencatat perubahan tersebut dalam log transaksi atau jurnal transaksi, dan kemudian meneruskan informasi tersebut kepada sistem atau aplikasi yang membutuhkan data yang diperbarui.

CDC memiliki banyak manfaat, antara lain:

1. Integrasi data real-time: CDC memungkinkan sistem atau aplikasi lain untuk memperoleh data yang diperbarui secara real-time, sehingga informasi yang digunakan selalu mutakhir.
2. Replikasi data: CDC memfasilitasi replikasi data antara sumber data dan target data dengan akurasi yang tinggi, sehingga sistem atau aplikasi lain dapat memiliki salinan data yang konsisten dengan sumber data.
3. Analisis data real-time: CDC memungkinkan sistem analisis data untuk menerima perubahan data secara cepat dan menghasilkan informasi analisis real-time berdasarkan perubahan tersebut.
4. Pemantauan data: CDC memungkinkan pemantauan perubahan data secara real-time, sehingga masalah atau anomali pada data dapat terdeteksi dan ditangani dengan cepat.

CDC digunakan dalam berbagai teknologi dan platform, seperti basis data relasional, sistem manajemen basis data (DBMS), dan alat-alat ETL (Extract, Transform, Load) untuk memfasilitasi aliran data yang efisien antara sistem atau aplikasi yang berbeda.

Metode CDC

Ada beberapa metode Change Data Capture (CDC) yang umum digunakan, antara lain:

1. Log-based CDC: Metode ini melibatkan perekaman dan analisis log transaksi atau jurnal transaksi yang dihasilkan oleh sistem basis data. Log transaksi merekam setiap operasi perubahan data yang terjadi pada sumber data. Dengan menganalisis log transaksi, perubahan data dapat diidentifikasi, direplikasi, dan ditransfer ke sistem atau aplikasi tujuan.
2. Capture-based CDC: Metode ini melibatkan penangkapan (capture) perubahan data secara langsung dari sumber data. Pada level yang paling dasar, proses capture akan memantau tabel atau objek yang ditentukan untuk mencari perubahan data. Perubahan data kemudian ditangkap dan direplikasi ke sistem atau aplikasi tujuan.
3. Query-based CDC: Metode ini melibatkan eksekusi query periodik pada sumber data untuk mencari perubahan data. Query dieksekusi secara berkala untuk mengidentifikasi data yang telah berubah sejak eksekusi query terakhir. Data yang ditemukan kemudian direplikasi dan ditransfer ke sistem atau aplikasi tujuan.
4. Hybrid CDC: Metode ini merupakan kombinasi dari dua atau lebih metode CDC. Biasanya, kombinasi log-based CDC dengan capture-based atau query-based CDC digunakan untuk memanfaatkan kelebihan masing-masing metode. Misalnya, log-based CDC digunakan untuk mereplikasi perubahan data secara real-time, sementara capture-based atau query-based CDC digunakan sebagai pendekatan yang lebih hemat sumber daya untuk mereplikasi perubahan data yang lebih lama.

Setiap metode CDC memiliki kelebihan, keterbatasan, dan situasi penggunaan yang berbeda. Pemilihan metode yang tepat tergantung pada kebutuhan bisnis, karakteristik sumber data, dan infrastruktur teknologi yang tersedia.

Log-based CDC : Kelebihan dan Kekurangan

Kelebihan Log-based CDC:

1. **Real-time Data Capture:** Log-based CDC memungkinkan perekaman perubahan data secara real-time. Setiap operasi perubahan data yang terjadi pada sumber data direkam dalam log transaksi, sehingga perubahan dapat segera direplikasi dan ditransfer ke sistem atau aplikasi tujuan.
2. **Akurasi dan Integritas Data:** Log transaksi merupakan sumber informasi yang sangat akurat dan terpercaya. Dengan menggunakan log transaksi, log-based CDC dapat mereplikasi perubahan data dengan tingkat akurasi dan integritas yang tinggi. Data yang direplikasi cenderung konsisten dengan sumber data asli.
3. **Fleksibilitas Skema Basis Data:** Log-based CDC mampu menangani perubahan skema basis data yang terjadi pada sumber data. Jika ada perubahan pada struktur tabel atau objek basis data, CDC masih dapat mengenali dan mereplikasi perubahan data dengan benar.
4. **Pengaruh Minimal pada Kinerja Sumber Data:** Log-based CDC tidak mempengaruhi kinerja sumber data secara signifikan karena perekaman perubahan data dilakukan melalui log transaksi yang sudah ada. Proses CDC berjalan secara terpisah dari operasi utama pada sumber data, sehingga tidak menyebabkan beban ekstra pada sumber data.

Kekurangan Log-based CDC:

1. **Ketergantungan pada Log Transaksi:** Log-based CDC membutuhkan log transaksi yang tersedia dan diaktifkan pada sumber data. Jika log transaksi tidak tersedia atau tidak didukung, metode ini tidak dapat digunakan. Beberapa sistem basis data mungkin tidak mendukung log transaksi atau hanya menyediakan log dengan informasi terbatas.
2. **Konfigurasi dan Administrasi yang Kompleks:** Implementasi log-based CDC dapat melibatkan konfigurasi yang kompleks dan memerlukan pengetahuan yang mendalam tentang sistem basis data dan log transaksi. Proses administrasi, pemeliharaan, dan pemantauan log-based CDC juga dapat membutuhkan keahlian teknis yang tinggi.
3. **Ukuran Log Transaksi yang Besar:** Log transaksi dapat menjadi sangat besar pada sistem dengan beban kerja yang tinggi atau sumber data yang besar. Menyimpan, memindahkan, dan mengolah log transaksi yang besar dapat memerlukan sumber daya yang signifikan.
4. **Tidak Cocok untuk Data Historis:** Log-based CDC biasanya fokus pada merekam perubahan data yang terjadi setelah penerapan CDC. Jika Anda memerlukan akses ke data historis, metode ini mungkin tidak cocok karena log transaksi hanya menyimpan informasi perubahan terkini.

Capture-based CDC : Kelebihan dan Kekurangan

Kelebihan Capture-based CDC:

1. Real-time atau Near Real-time Data Capture: Capture-based CDC memungkinkan perekaman perubahan data secara real-time atau mendekati real-time. Data yang berubah pada sumber data dapat segera ditangkap dan direplikasi ke sistem atau aplikasi tujuan dengan sedikit jeda waktu.
2. Independen dari Log Transaksi: Capture-based CDC tidak bergantung pada log transaksi atau jurnal transaksi yang ada pada sumber data. Ini membuatnya dapat digunakan pada sumber data yang tidak memiliki log transaksi atau tidak mendukung CDC berbasis log.
3. Pemantauan Tabel atau Objek Tertentu: Capture-based CDC memungkinkan Anda untuk memantau tabel atau objek yang spesifik dalam sumber data. Anda dapat mengatur pemantauan hanya pada data yang relevan atau penting, sehingga menghemat sumber daya dan meningkatkan efisiensi.
4. Kemampuan Memfilter Data: Capture-based CDC memungkinkan Anda untuk menerapkan filter pada data yang direplikasi. Anda dapat menentukan kriteria tertentu dalam pemantauan, seperti kolom atau nilai tertentu, untuk memilih data yang akan direplikasi. Ini memberikan kontrol yang lebih besar terhadap data yang dikirimkan.

Kekurangan Capture-based CDC:

1. Dampak pada Kinerja Sumber Data: Proses pemantauan dan penangkapan perubahan data dalam capture-based CDC dapat mempengaruhi kinerja sumber data. Terutama pada sumber data dengan beban kerja tinggi, pemantauan yang intensif dapat menyebabkan penurunan kinerja.
2. Konfigurasi yang Rumit: Implementasi capture-based CDC dapat melibatkan konfigurasi yang rumit dan memerlukan pemahaman yang mendalam tentang sumber data dan sistem yang terlibat. Konfigurasi yang tidak tepat dapat menyebabkan kesalahan atau ketidakcocokan dalam proses pemantauan dan penangkapan.
3. Resiko Kehilangan Data: Capture-based CDC mengandalkan pemantauan dan penangkapan perubahan data secara langsung dari sumber data. Jika ada kegagalan dalam proses penangkapan, ada risiko kehilangan data yang tidak direplikasi atau data yang tidak konsisten.
4. Tidak Cocok untuk Perubahan Skema Basis Data: Jika terjadi perubahan skema basis data pada sumber data, capture-based CDC mungkin memerlukan konfigurasi ulang atau penyesuaian. Proses ini dapat memakan waktu dan mempengaruhi kontinuitas replikasi data.

Query-based CDC : Kelebihan dan Kekurangan

Kelebihan Query-based CDC:

1. Tidak Bergantung pada Log Transaksi atau Pemantauan Aktif: Query-based CDC tidak memerlukan log transaksi atau pemantauan aktif pada sumber data. Prosesnya melibatkan eksekusi query periodik untuk mengidentifikasi perubahan data. Ini membuatnya dapat diterapkan pada sumber data yang tidak memiliki log transaksi atau tidak mendukung pemantauan langsung.
2. Fleksibilitas dalam Memilih Data yang Direplikasi: Dengan query-based CDC, Anda memiliki fleksibilitas untuk merancang query yang memilih data yang akan direplikasi. Anda dapat menentukan kriteria spesifik, seperti kolom, nilai, atau kondisi tertentu, untuk mengambil hanya data yang relevan. Ini memungkinkan kontrol yang lebih besar terhadap data yang direplikasi.
3. Efisiensi Penggunaan Sumber Daya: Query-based CDC menggunakan query periodik untuk mencari perubahan data. Ini dapat lebih efisien dalam penggunaan sumber daya dibandingkan metode CDC yang memantau atau menangkap perubahan secara terus-menerus. Pemantauan data hanya dilakukan pada interval waktu tertentu sesuai dengan kebutuhan bisnis.
4. Tidak Mengganggu Kinerja Sumber Data: Query-based CDC tidak memiliki dampak yang signifikan pada kinerja sumber data karena eksekusi query dilakukan pada interval waktu tertentu. Ini memungkinkan operasi utama pada sumber data berjalan tanpa gangguan, terutama pada sumber data dengan beban kerja yang tinggi.

Kekurangan Query-based CDC:

1. Tidak Real-time: Query-based CDC tidak memungkinkan replikasi data secara real-time. Proses query periodik berarti ada jeda waktu antara identifikasi perubahan data dan replikasi ke sistem tujuan. Jeda ini dapat bervariasi tergantung pada frekuensi eksekusi query.
2. Keterbatasan Performa: Jika sumber data memiliki jumlah data yang besar atau kompleksitas yang tinggi, eksekusi query periodik dapat mempengaruhi performa. Query yang rumit atau pemrosesan data yang intensif dapat memperlambat waktu respons atau mempengaruhi kinerja sumber data.
3. Dibatasi oleh Kemampuan Query Bahasa: Query-based CDC tergantung pada kemampuan dan fitur query bahasa yang tersedia pada sumber data. Jika sumber data memiliki batasan dalam hal ekspresi query atau dukungan bahasa query, mungkin ada keterbatasan dalam mendapatkan data yang diinginkan.
4. Tidak Cocok untuk Data Historis: Query-based CDC biasanya fokus pada perubahan data yang terjadi setelah penerapan CDC. Jika Anda memerlukan akses ke data historis, metode ini mungkin tidak cocok karena hanya melacak perubahan data pada interval waktu tertentu setelah implementasi.

Hybrid CDC : Kelebihan dan Kekurangan

Kelebihan Hybrid CDC:

1. Kombinasi Kelebihan Metode Lain: Hybrid CDC memadukan kelebihan dari berbagai metode CDC, seperti log-based, capture-based, dan query-based CDC. Dengan demikian, ini memungkinkan penggunaan yang lebih fleksibel dan dapat disesuaikan sesuai dengan kebutuhan bisnis dan karakteristik sumber data yang berbeda.
2. Real-time atau Near Real-time Data Capture: Dalam model hybrid CDC, penggunaan metode capture-based atau log-based CDC dapat memberikan kemampuan untuk mereplikasi perubahan data secara real-time atau mendekati real-time. Ini memungkinkan akses cepat ke data yang diperbarui dan meningkatkan responsifitas sistem atau aplikasi tujuan.
3. Selektivitas dan Filter Data: Dalam hybrid CDC, Anda dapat menggabungkan kemampuan query-based CDC untuk memilih dan memfilter data yang akan direplikasi. Ini memberikan kontrol yang lebih besar terhadap jenis data yang ingin Anda sertakan dalam proses replikasi, sehingga menghemat sumber daya dan meningkatkan efisiensi.
4. Ketersediaan Data Historis: Dalam beberapa implementasi hybrid CDC, log-based CDC dapat digunakan untuk mereplikasi data historis, sedangkan metode lainnya digunakan untuk mereplikasi perubahan data saat ini. Ini memungkinkan akses ke data historis yang relevan dan juga memungkinkan pengambilan data yang real-time.

Kekurangan Hybrid CDC:

1. Kompleksitas Konfigurasi dan Administrasi: Implementasi hybrid CDC dapat melibatkan konfigurasi yang lebih kompleks karena melibatkan kombinasi berbagai metode CDC. Pemahaman yang mendalam tentang metode yang digunakan dan integrasi yang tepat diperlukan untuk memastikan konfigurasi yang benar dan pengelolaan yang efektif.
2. Peningkatan Kompleksitas dan Ketergantungan pada Sumber Daya: Dengan menggabungkan metode CDC yang berbeda, dapat terjadi peningkatan kompleksitas dalam hal manajemen dan penggunaan sumber daya. Proses replikasi data dapat memerlukan lebih banyak sumber daya dan pemantauan yang cermat untuk memastikan integritas data.
3. Dapat Memiliki Dampak pada Kinerja Sumber Data: Penggunaan metode CDC yang berbeda dalam hybrid CDC dapat mempengaruhi kinerja sumber data, terutama jika sumber data memiliki beban kerja yang tinggi. Penting untuk memantau kinerja secara terus-menerus dan melakukan penyesuaian jika diperlukan.
4. Konfigurasi yang Tepat untuk Kebutuhan Bisnis: Dalam mengimplementasikan hybrid CDC, penting untuk mengkonfigurasi metode yang sesuai dengan kebutuhan bisnis yang spesifik. Membuat keputusan yang tepat tentang metode yang digunakan dan bagaimana mereka berinteraksi satu sama lain adalah faktor kunci untuk kesuksesan implementasi hybrid CDC.