**Tiêu đề:** Giới thiệu Amazon Keyspaces CDC streams

Với [Amazon Keyspaces (for Apache Cassandra)](https://aws.amazon.com/keyspaces/), bạn có thể vận hành workload Cassandra trên AWS bằng dịch vụ cơ sở dữ liệu serverless, được quản lý hoàn toàn. Bạn sẽ có khả năng mở rộng throughput và bộ nhớ gần như vô hạn trong khi vẫn duy trì độ trễ mili‑giây. Do tương thích với Cassandra, bạn có thể sử dụng mã nguồn và công cụ hiện có mà không cần phải thay đổi nhiều. AWS đảm nhiệm các nhiệm vụ phức tạp như cấp phát server, vá lỗi và sao lưu, để bạn có thể tập trung vào phát triển ứng dụng.

Tuần trước, AWS đã công bố **CDC streams** (Change Data Capture streams) cho Amazon Keyspaces — tính năng giúp ghi nhận ngay lập tức các thay đổi dữ liệu (insert, update, delete) trên bảng. Bạn có thể phản ứng theo thời gian thực, mở ra nhiều khả năng tích hợp ứng dụng. Dữ liệu thay đổi có thể được đẩy tới Amazon OpenSearch Service để tìm kiếm nâng cao, phục vụ các scenai generative AI, đồng bộ vào kho dữ liệu để phân tích báo cáo, hoặc xây dựng kiến trúc event-driven hiện đại. Tất cả được thực hiện trong môi trường fully managed, đảm bảo khả năng mở rộng và độ tin cậy cao.

**Tổng quan giải pháp**

**Amazon Keyspaces CDC streams** cung cấp một giải pháp mạnh mẽ để ghi nhận các thay đổi ở cấp độ hàng trong bảng, lưu trữ chúng dưới dạng một chuỗi sự kiện có thứ tự trong log trong vòng 24 giờ. Mỗi khi một hàng được chèn, cập nhật hoặc xóa, luồng CDC sẽ tạo ra một bản ghi chi tiết bao gồm thông tin khóa chính cùng với trạng thái trước và sau của hàng dữ liệu đó. Điều này cho phép các ứng dụng tiêu thụ và xử lý các thay đổi này gần như theo thời gian thực.

Khi bạn bật luồng (streams) cho một bảng, các bản ghi sẽ được tổ chức thành các **shard** dựa trên phạm vi khóa chính. Trong mỗi shard, bản ghi được gán số thứ tự duy nhất tăng dần (monotonically increasing) để đảm bảo thứ tự nghiêm ngặt. Amazon Keyspaces sẽ **tự động quản lý vòng đời của shard**, tách hoặc gộp chúng theo lưu lượng truy cập thực tế. Cơ chế quản lý linh hoạt này giúp tối ưu hiệu suất đồng thời vẫn duy trì khả năng theo dõi lịch sử bản ghi qua các shard.

*Sơ đồ minh họa cho giải pháp này được trình bày bên dưới (trong bài viết gốc).*

A diagram of a row and row

AI-generated content may be incorrect.

Hệ thống cung cấp hai khả năng then chốt:

1. Mỗi bản ghi thay đổi (mutation) chỉ xuất hiện đúng một lần trong luồng CDC.
2. Các bản ghi được giữ đúng thứ tự như thứ tự thay đổi thực tế dựa trên khóa chính.

Điều này đảm bảo tính nhất quán dữ liệu và thứ tự chính xác cho các ứng dụng phía sau. Luồng CDC giữ bản ghi trong **24 giờ** và đây là khoảng thời gian giữ cố định, kể cả khi tính năng CDC bị tắt sau đó.

Các thao tác theo lô (batch operations) sẽ được tự động phân tách thành từng bản ghi thay đổi ở cấp hàng nhưng vẫn giữ đúng thứ tự. Với **các cột tĩnh** (static columns), vốn chia sẻ giá trị giữa tất cả các hàng trong một phân vùng, các thay đổi sẽ được ghi nhận thành các bản ghi riêng biệt để phản ánh chính xác mô hình dữ liệu.

Dữ liệu CDC được **mã hóa tự động khi lưu trữ**, sử dụng cùng khóa mã hóa với bảng gốc. Trong môi trường triển khai đa vùng (multi-Region), các luồng CDC hoạt động **độc lập** tại từng vùng AWS. Điều này có nghĩa là mặc dù mỗi vùng duy trì luồng CDC nhất quán riêng biệt, **thứ tự sự kiện có thể khác nhau giữa các vùng** do tính chất bất đồng bộ của việc sao chép đa vùng và quá trình xử lý xung đột.

Trong các phần tiếp theo, chúng ta sẽ khám phá cách bật CDC streams và truy cập chúng để phục vụ cho các mục đích sử dụng phía sau.

**Truy cập các endpoint**

Khi một luồng CDC được tạo cho bảng, **Amazon Keyspaces** bắt đầu thu thập thông tin về các thay đổi trong bảng đó. Luồng CDC được xác định bằng một **Amazon Resource Name (ARN)** theo định dạng sau:

arn:<AWS\_PARTITION>:cassandra:<REGION>:<AWS\_ACCOUNT\_ID>:/keyspace/<KEYSPACE\_NAME>/table/<TABLE\_NAME>/stream/<STREAM\_LABEL>

Khi kích hoạt luồng CDC, bạn có thể chọn loại thông tin (hay **view type**) mà luồng sẽ ghi nhận trong từng bản ghi. Lưu ý rằng bạn **không thể thay đổi view type sau khi đã bật** luồng CDC. Amazon Keyspaces hỗ trợ các loại view sau:

* **NEW\_AND\_OLD\_IMAGES (mặc định)** – Ghi lại phiên bản của hàng trước và sau khi thay đổi
* **NEW\_IMAGE** – Ghi lại phiên bản của hàng sau khi thay đổi
* **OLD\_IMAGE** – Ghi lại phiên bản của hàng trước khi thay đổi
* **KEYS\_ONLY** – Ghi lại khóa phân vùng và khóa sắp xếp của hàng đã thay đổi

Bạn có thể tiêu thụ dữ liệu từ luồng CDC thông qua [**Amazon Keyspaces Streams API**](https://docs.aws.amazon.com/keyspaces/latest/StreamsAPIReference/Welcome.html)hoặc sử dụng [**Kinesis Client Library (KCL)**](https://docs.aws.amazon.com/streams/latest/dev/kcl.html)**.**Mỗi phương thức có ưu điểm riêng, tuy nhiên **KCL hỗ trợ sẵn các tính năng như quản lý shard tự động, chịu lỗi (fault tolerance), và cân bằng tải cho worker**, rất hữu ích khi xử lý dữ liệu phức tạp ở quy mô lớn. Trong khi đó, cách dùng API trực tiếp cung cấp quyền kiểm soát chi tiết hơn và phù hợp với các tình huống đơn giản, tùy biến logic xử lý luồng.

**Điều kiện tiên quyết**

Bạn cần có một **tài khoản AWS** với các quyền thích hợp để tiếp tục. Vui lòng tham khảo tài liệu chính thức để biết chi tiết về các quyền cần thiết.

**Bật tính năng CDC streams**

Hãy xét một tình huống sử dụng thực tế: một **nền tảng truyền thông kỹ thuật số** cần lưu trữ và quản lý nội dung do người sáng tạo tải lên. Mỗi nội dung (video, hình ảnh hoặc âm thanh) sẽ được gán một **content\_id duy nhất** và được theo dõi bằng các thông tin metadata như **thời điểm tải lên** và **trạng thái xử lý**.

Khi dữ liệu nội dung được ghi vào bảng, các ứng dụng phía sau sẽ sử dụng **API của Streams** để giám sát thay đổi và kích hoạt các tiến trình xử lý như:

* Hệ thống đề xuất nội dung,
* Gửi thông báo cho người dùng,
* Cập nhật dashboard phân tích theo thời gian thực.

**Thực hiện các bước sau để bật CDC streams:**

1. Truy cập **Amazon Keyspaces Console**, chọn **Keyspaces** trong bảng điều hướng bên trái.
2. Chọn **Create keyspace** (Tạo keyspace).
3. Tạo một **keyspace tên là media**, nơi bạn sẽ lưu metadata cho bảng media\_content.
4. Chọn **Create keyspace** để hoàn tất.

A screenshot of a computer

AI-generated content may be incorrect.

1. Mở trình soạn thảo CQL (CQL editor) để tạo bảng với lệnh sau:

CREATE TABLE media.media\_content (

content\_id uuid,

title text,

creator\_id uuid,

media\_type text,

upload\_timestamp timestamp,

status text,

PRIMARY KEY (content\_id)

);

1. Chọn Run Command

A screenshot of a computer

AI-generated content may be incorrect.

1. **Tại mục Chi tiết luồng (Stream details),** chọn **Bật luồng (Turn on streams)** để có thể ghi nhận các thay đổi dữ liệu trên bảng với kiểu hiển thị mặc định.
2. Sau đó, chọn **Lưu thay đổi (Save changes)**.

A screenshot of a computer

AI-generated content may be incorrect.

Sau khi luồng được bật, bạn sẽ thấy **trạng thái luồng hiển thị là “on”** và **kiểu hiển thị (view type) là “New and old images”**.

A close up of a letter

AI-generated content may be incorrect.

Bạn đã hoàn tất việc thiết lập để bật luồng dữ liệu (streams) cho bảng **media\_content**. Bây giờ, hãy chèn một số dữ liệu mẫu và xem cách truy cập dữ liệu từ luồng CDC.

**Truy cập luồng CDC**

Trước khi truy cập luồng, hãy chèn một số dữ liệu vào bảng bằng **CQL editor**. Sử dụng các câu lệnh **INSERT** sau:

INSERT INTO media.media\_content (content\_id, title, creator\_id, media\_type, upload\_timestamp, status) VALUES (uuid(), 'Summer Vacation Video', uuid(), 'video', toTimestamp(now()), 'active');

INSERT INTO media.media\_content (content\_id, title, creator\_id, media\_type, upload\_timestamp, status) VALUES (uuid(), 'Birthday Party Photos', uuid(), 'image', toTimestamp(now()), 'processing');

INSERT INTO media.media\_content (content\_id, title, creator\_id, media\_type, upload\_timestamp, status) VALUES (uuid(), 'Podcast Episode 1', uuid(), 'audio', toTimestamp(now()), 'active');

INSERT INTO media.media\_content (content\_id, title, creator\_id, media\_type, upload\_timestamp, status) VALUES (uuid(), 'Wedding Ceremony', uuid(), 'video', toTimestamp(now()), 'archived');

INSERT INTO media.media\_content (content\_id, title, creator\_id, media\_type, upload\_timestamp, status) VALUES (uuid(), 'Concert Recording', uuid(), 'audio', toTimestamp(now()), 'processing');

Giờ đây, bạn có thể **truy xuất các bản ghi thay đổi** của bảng bằng cách sử dụng **AWS CLI** để gọi các API mới của Amazon Keyspaces CDC Streams.

**Lấy danh sách các shard trong luồng**

Sử dụng lệnh get-stream để lấy thông tin về các **shard** trong luồng CDC:

aws keyspacesstreams get-stream \

--stream-arn arn:aws:cassandra:<AWS\_REGION>:<account\_id>:/keyspace/media/table/media\_content/stream/2025-05-16T20:23:29.918 \

--endpoint https://cassandra-streams.<AWS\_REGION>.api.aws

Kết quả đầu ra sẽ hiển thị các **shard** trong luồng CDC như sau:

{

"streamArn": "arn:aws:cassandra:<AWS\_REGION>:<account\_id>:/:/keyspace/media/table/media\_content/stream/2025-05-16T20:23:29.918",

"streamLabel": "2025-05-16T20:23:29.918",

"streamStatus": "ENABLED",

"streamViewType": "NEW\_AND\_OLD\_IMAGES",

"creationRequestDateTime": "2025-05-16T15:23:29.918000-05:00",

"keyspaceName": "media",

"tableName": "media\_content",

"shards": [

{

"shardId": "shardId-00000001747427011617-c6dcfa62",

"sequenceNumberRange": {

"startingSequenceNumber": "6500003910786539459926"

},

"parentShardIds": [

null

]

},

{

"shardId": "shardId-00000001747427011721-ee994773",

"sequenceNumberRange": {

"startingSequenceNumber": "6300003358018353308682"

},

"parentShardIds": [

null

]

},

{

"shardId": "shardId-00000001747427011822-286d4fbc",

"sequenceNumberRange": {

"startingSequenceNumber": "6400000689805508842349"

},

"parentShardIds": [

null

]

},

{

"shardId": "shardId-00000001747427011926-6cb2833c",

"sequenceNumberRange": {

"startingSequenceNumber": "6400002768480685596628"

},

"parentShardIds": [

null

]

}

]

}

Để **lấy các bản ghi từ luồng CDC**, trước tiên bạn cần **lấy một "shard iterator"** bằng cách sử dụng API get-shard-iterator.

**Shard iterator** đóng vai trò như **điểm bắt đầu** trong mỗi shard, cho phép bạn truy xuất các bản ghi theo đúng thứ tự.

Trong ví dụ này, chúng ta sử dụng loại iterator **TRIM\_HORIZON**, nhằm truy xuất dữ liệu **từ điểm bắt đầu (hoặc điểm đã được cắt tỉa gần nhất) của shard**:

aws keyspacesstreams get-shard-iterator \

--stream-arn arn:aws:cassandra:<AWS\_REGION>:<account\_id>:/keyspace/media/table/media\_content/stream/2025-05-16T20:23:29.918 \

--endpoint https://cassandra-streams.<AWS\_REGION>.api.aws \

--shard-id 'shardId-00000001747427011617-c6dcfa62' \

--shard-iterator-type TRIM\_HORIZON

Output:

{

"shardIterator": "arn:aws:cassandra:<AWS\_REGION>:<account\_id>:/keyspace/media/table/media\_content/stream/2025-05-16T20:23:29.918|1|"

}

Sử dụng **shard iterator** đã lấy được ở bước trước, bạn có thể truy xuất các bản ghi CDC bằng cách gọi API **get-records** như sau:

aws keyspacesstreams get-records \

--shard-iterator 'arn:aws:cassandra:<AWS\_REGION>:<account\_id>:/:/keyspace/media/table/media\_content/stream/2025-05-16T20:23:29.918|1|' \

--endpoint https://cassandra-streams.<AWS\_REGION>.api.aws

Kết quả trả về sẽ chứa **các bản ghi dữ liệu mà bạn đã chèn vào bảng** trước đó.  
Bạn có thể **lặp lại quy trình tương tự** với các shardId khác mà bạn đã lấy ở **Bước 1** để thu thập toàn bộ dữ liệu từ các shard còn lại.

Như vậy, bạn đã hoàn tất quá trình thiết lập và truy cập **luồng CDC** thông qua các API của Amazon Keyspaces.  
Ngoài ra, bạn cũng có thể **tích hợp CDC stream với các dịch vụ AWS khác** như **AWS Lambda** để xây dựng các hệ thống xử lý dữ liệu theo sự kiện một cách tự động và linh hoạt.

**Các lưu ý**

Giả sử bạn nhận được **nhiều MB dữ liệu mỗi giây** từ các thay đổi nội dung truyền thông, việc **quản lý tiêu thụ luồng CDC trực tiếp thông qua API** sẽ trở nên **khó khăn** do phải xử lý các vấn đề như:

* Quản lý shard
* Đánh dấu vị trí đã xử lý (checkpointing)
* Khả năng mở rộng (scaling)

Với những trường hợp có **lưu lượng lớn**, chúng tôi **khuyến nghị sử dụng Kinesis Client Library (KCL)** thay vì gọi API trực tiếp.

**KCL** cung cấp các tính năng sẵn có như:

* Điều phối worker
* Tự động checkpoint
* Tự động scale các trình xử lý dữ liệu

→ Đây là giải pháp lý tưởng để xử lý **khối lượng lớn dữ liệu luồng** trong môi trường sản xuất (production).

**Lợi ích và các tình huống sử dụng phổ biến**

**✅ Tích hợp dữ liệu thời gian thực (Real-time data integration):**

* Đồng bộ liền mạch giữa **Amazon Keyspaces** và **data warehouse**
* Truyền thay đổi dữ liệu sang các nền tảng phân tích để có insight theo thời gian thực
* Xây dựng kiến trúc hướng sự kiện (event-driven) phản hồi tức thời với thay đổi dữ liệu

**🔍 Tìm kiếm và phân tích (Search and analytics):**

* Giữ cho chỉ mục tìm kiếm luôn cập nhật bằng cách truyền thay đổi đến **OpenSearch Service**
* Hỗ trợ tìm kiếm toàn văn (full-text) trên dữ liệu Cassandra
* Cung cấp dashboard phân tích theo thời gian thực với dữ liệu mới nhất

**🛡️ Tuân thủ và nhật ký kiểm tra (Compliance and audit trail):**

* Lưu giữ lịch sử chi tiết về các thay đổi dữ liệu
* Theo dõi **ai**, **đã thay đổi gì** và **khi nào** để đáp ứng yêu cầu tuân thủ
* Hỗ trợ **khôi phục theo thời điểm** và đối chiếu dữ liệu (reconciliation)

Đây là những lợi ích và ứng dụng nổi bật khiến **Amazon Keyspaces CDC Streams** trở thành một công cụ mạnh mẽ trong việc xây dựng **ứng dụng hiện đại dựa trên dữ liệu**, yêu cầu khả năng xử lý và tích hợp theo **thời gian thực**.

**Dọn dẹp tài nguyên (Clean up)**

Khi bạn đã hoàn tất thử nghiệm giải pháp, hãy **xóa các tài nguyên đã tạo** để tránh phát sinh chi phí không cần thiết.

Xóa bảng media\_content:

aws keyspaces delete-table --keyspace-name media \

--table-name media\_content

Xóa keyspace media:

aws keyspaces delete-keyspace --keyspace-name media

**Kết luận**

Trong bài viết này, chúng ta đã tìm hiểu về **Amazon Keyspaces CDC Streams** — một tính năng cho phép **ghi lại các thay đổi dữ liệu** trong bảng Amazon Keyspaces.

Bạn đã được hướng dẫn cách:

* Bật CDC stream cho bảng,
* Chèn dữ liệu và truy xuất các bản ghi thay đổi,
* Tích hợp CDC với các ứng dụng phía sau theo hướng event-driven.

**Việc truyền dữ liệu thay đổi** từ Amazon Keyspaces đến các ứng dụng hoặc nguồn dữ liệu khác giúp:

* Xây dựng hệ thống phản hồi theo sự kiện (event-driven),
* Tích hợp sâu với các hệ thống phân tích hoặc lưu trữ khác,
* Giải quyết những bài toán khó về mô hình hóa dữ liệu theo thời gian thực.

**Thông Tin Về Tác Giả**

**Rajesh Kantamani**

**Rajesh** là một **Chuyên gia Giải pháp Cơ sở dữ liệu cấp cao (Senior Database Specialist Solutions Architect)**.  
Anh hợp tác cùng khách hàng để **thiết kế, di chuyển và tối ưu hóa các giải pháp cơ sở dữ liệu trên nền tảng Amazon Web Services (AWS)**, với trọng tâm là **khả năng mở rộng, bảo mật và hiệu suất tối đa**.

Với niềm đam mê dành cho **các cơ sở dữ liệu phân tán**, Rajesh giúp các tổ chức **chuyển đổi hạ tầng dữ liệu của họ một cách hiệu quả**.

Khi không làm việc trong vai trò kiến trúc giải pháp cơ sở dữ liệu, anh thường dành thời gian tham gia **các hoạt động ngoài trời cùng gia đình và bạn bè**.