**Cách TalentNeuron tối ưu hóa hoạt động dữ liệu, cắt giảm chi phí và hiện đại hóa hệ thống với Amazon Aurora I/O-Optimized**

Đây là bài viết khách mời của **Yaya Diawara**, Trưởng nhóm quản trị cơ sở dữ liệu (Team Lead DBA) tại TalentNeuron, phối hợp cùng AWS.

Trong nhiều năm qua, [**TalentNeuron**](https://www.talentneuron.com/)– công ty dẫn đầu trong lĩnh vực **trí tuệ nhân sự** và **lập kế hoạch lực lượng lao động** – đã hỗ trợ các tổ chức bằng cách cung cấp những thông tin sâu sắc dựa trên dữ liệu, thông qua việc thu thập và xử lý một lượng lớn dữ liệu từ các bảng tin tuyển dụng. Tuy nhiên, khi hoạt động dữ liệu của họ ngày càng mở rộng, **độ phức tạp và chi phí của kiến trúc cũ cũng tăng theo**.

Nhận thấy nhu cầu cấp thiết về một giải pháp **hiệu quả hơn và tiết kiệm chi phí hơn**, TalentNeuron đã **hợp tác với AWS** để hiện đại hóa nền tảng dữ liệu của mình.

Trong bài viết này, chúng tôi chia sẻ **ba lợi ích chính** mà TalentNeuron đạt được khi sử dụng [**Amazon Aurora**](https://aws.amazon.com/rds/aurora/) **I/O-Optimized** như một phần trong nền tảng dữ liệu mới của họ:

1. **Giảm chi phí cơ sở dữ liệu hàng tháng xuống 29%**
2. **Cải thiện hiệu suất kiểm tra tính hợp lệ dữ liệu (data validation)**
3. **Thúc đẩy đổi mới nhanh hơn thông qua hiện đại hóa kiến trúc**

**Thách thức trong xử lý dữ liệu**

Kiến trúc dữ liệu của TalentNeuron được xây dựng theo [**medallion pattern**](https://dataengineering.wiki/Concepts/Data+Architecture/Medallion+Architecture). Đây là một cách tiếp cận nhiều lớp trong việc tổ chức dữ liệu trong **data lake**, nơi mà dữ liệu thô đi qua ba tầng:

* **Bronze (Thô)**
* **Silver (Đã làm sạch và xác thực)**
* **Gold (Sẵn sàng cho nghiệp vụ)**

Mỗi tầng lần lượt **nâng cao chất lượng và cấu trúc của dữ liệu**, phục vụ cho các mục đích sử dụng khác nhau.

Hệ thống kiến trúc dữ liệu của TalentNeuron được minh họa trong sơ đồ sau, thể hiện rõ **các tầng riêng biệt cho việc nạp dữ liệu (ingestion), lưu trữ (storage), và tiêu thụ (consumption)**.

A diagram of data quality

AI-generated content may be incorrect.

**Lớp Bronze** lưu trữ dữ liệu thô ban đầu được thu thập từ nhiều nguồn khác nhau. Dữ liệu ở lớp Bronze thể hiện trạng thái **chưa xử lý**, thu thập trực tiếp từ các trang web. Các dữ liệu này được lưu trữ dưới dạng **tài liệu trong** [**Amazon Simple Storage Service**](https://aws.amazon.com/s3/) **(Amazon S3)** và có cơ chế **lưu trữ phiên bản lịch sử** để tra cứu khi cần.

**Lớp Silver**, được lưu trữ trong cơ sở dữ liệu **Amazon Aurora MySQL**, là nơi hệ thống **trích xuất và xác thực dữ liệu**. TalentNeuron có nhiều dịch vụ kiểm tra dữ liệu, giúp xử lý một phần và xác minh tính chính xác của dữ liệu. Họ **trích xuất các thuộc tính từ lớp Bronze** một cách có cấu trúc, đồng thời **chuẩn hóa các thuộc tính chính** để đảm bảo tính nhất quán, sử dụng kết hợp cả dịch vụ bên thứ ba và công cụ nội bộ.

**Lớp Gold** lưu trữ dữ liệu cuối cùng đã được tinh chỉnh, cũng được đặt trong cơ sở dữ liệu **Aurora MySQL**. Các hệ thống phía sau (downstream systems) sử dụng dữ liệu từ lớp này. Đây cũng là nơi hệ thống thực hiện **loại bỏ trùng lặp** dựa trên các quy tắc nghiệp vụ, **chỉnh lý dữ liệu** và **làm giàu thông tin** để phục vụ mục đích phân tích.

**Mặc dù mô hình này hoạt động tốt, nhưng vẫn gặp một số thách thức:**

* **Chi phí cao** – Việc lưu trữ dữ liệu ở lớp Silver trong  [Aurora MySQL-Compatible](https://aws.amazon.com/rds/aurora/features/) rất **tốn I/O**, dẫn đến **chi phí hàng tháng cao** do hoạt động đọc/ghi liên tục trong quá trình chuẩn hóa dữ liệu.
* **Điểm nghẽn trong xác thực dữ liệu** – Sự tích hợp quá chặt giữa các lớp, cùng với các ràng buộc dữ liệu cứng nhắc, tạo nên độ phức tạp trong vận hành và việc xử lý dữ liệu trùng lặp. Một dịch vụ quan trọng để kiểm tra dữ liệu thu thập được **bị phụ thuộc chặt chẽ vào cơ sở dữ liệu**, khiến quá trình hiện đại hóa hoặc thay thế trở nên khó khăn.
* **Cản trở đổi mới** – Người dùng nội bộ phải **phụ thuộc vào đội ngũ DBA** để viết truy vấn SQL, làm chậm việc truy cập dữ liệu và khiến đội DBA bị kéo khỏi những công việc mang tính chiến lược hơn. Họ phải dành nhiều thời gian để **duy trì pipeline dữ liệu**, cản trở khả năng **trao quyền truy cập dữ liệu cho toàn tổ chức** và đổi mới nhằm phục vụ khách hàng.

**TalentNeuron cần một cách tiếp cận mới để:**

* Giảm chi phí vận hành
* Đơn giản hóa kiến trúc dữ liệu
* Khai thác thêm giá trị kinh doanh từ tài sản dữ liệu sẵn có

**Tổng quan giải pháp**

Để bắt đầu hành trình hiện đại hóa, TalentNeuron đã **chuyển sang sử dụng**[**Amazon Aurora I/O-Optimized**](https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/Aurora.Overview.StorageReliability.html)Khác với cấu hình Aurora tiêu chuẩn – nơi **chi phí I/O được tính riêng biệt** – Aurora I/O-Optimized mang lại **mức giá ổn định hơn** bằng cách **gộp chi phí I/O vào chi phí của instance và lưu trữ**.

Điều này rất lý tưởng cho các **tác vụ nặng về I/O**, như quy trình **chuẩn hóa dữ liệu** của TalentNeuron.

Sơ đồ sau minh họa kiến trúc giải pháp này.

A diagram of a diagram

AI-generated content may be incorrect.

**Lợi ích kinh doanh đạt được**

**Giảm chi phí**

TalentNeuron đã chuyển cơ sở dữ liệu Aurora MySQL ở lớp **Silver-tier** sang sử dụng **Aurora I/O-Optimized**, mang lại **mức giảm chi phí cơ sở dữ liệu thuần túy lên đến 29,1%** – mà **không phải đánh đổi hiệu năng**.

Sau khi áp dụng Aurora I/O-Optimized, họ cũng ghi nhận **mức cải thiện lên đến 40%** đối với cả chỉ số VolumeReadIOPs và VolumeWriteIOPs tại lớp Silver.

*“Amazon Aurora I/O-Optimized đã mang lại cho chúng tôi hiệu quả chi phí và hiệu suất cần thiết để tái cấu trúc chiến lược dữ liệu,”*  
– **Scott Roan**, Phó Chủ tịch Cấp cao Phụ trách Công nghệ tại TalentNeuron.  
*“Đây là một bước quan trọng trong việc tối ưu hóa hoạt động của chúng tôi.”*

**Cải thiện quy trình xác thực dữ liệu**

Một lợi ích trực tiếp khác từ khả năng I/O được nâng cao là **cải thiện hiệu suất của các quy trình xác thực dữ liệu**. Trong một kiến trúc medallion điển hình, dữ liệu đi qua các giai đoạn bronze–silver–gold và trải qua nhiều bước xác thực và chuyển đổi. Điều này thường yêu cầu các hoạt động đọc và ghi cơ sở dữ liệu thường xuyên.

Bên cạnh các bước xử lý dữ liệu trong kiến trúc medallion được đề cập ở trên, khả năng I/O-Optimized của Amazon Aurora đã **tăng tốc các bước xác thực dữ liệu bổ sung** song song với quá trình di chuyển dữ liệu trong pipeline. Cụ thể, TalentNeuron đã thấy **thời gian xác thực dữ liệu giảm 28%**, từ **72 phút xuống còn 51 phút**, đáp ứng **thỏa thuận mức dịch vụ nội bộ (SLA)** để xử lý và xác thực trong **dưới một giờ** tại tầng Silver – nơi các xác thực này diễn ra – điều mà TalentNeuron không thể đạt được trước khi áp dụng Aurora I/O-Optimized.

Điều này đặc biệt quan trọng với TalentNeuron vì khi khách hàng của họ tìm kiếm các vị trí phù hợp trên thị trường việc làm, cơ sở dữ liệu cần thực hiện các truy vấn phức tạp và truy xuất tập dữ liệu một cách nhanh chóng.

**Thúc đẩy đổi mới**

Vượt ra ngoài Aurora, TalentNeuron đang bắt đầu chuyển dịch sang kiến trúc data lake. Các thành viên trong nhóm TalentNeuron đang khám phá việc sử dụng **định dạng bảng mở Apache Iceberg (OTF)** từ các quy trình **ETL thử nghiệm**. Giờ đây, khi đội ngũ DBA dành ít thời gian hơn để quản lý pipeline dữ liệu, họ có thể tập trung nhiều hơn vào việc khám phá các trường hợp sử dụng mới như **ELT với Amazon EMR** được điều phối bằng **AWS Step Functions**.

TalentNeuron vẫn chưa hoàn toàn nhận được tất cả lợi ích từ việc chuyển sang **Amazon EMR** và **Amazon S3**, nhưng đây là bước đầu tiên trong việc **tách rời thiết kế ban đầu**. Các nhóm nội bộ hiện đã sử dụng **notebook của Amazon EMR để truy vấn trực tiếp data lake trên Amazon S3**, giúp **giảm sự phụ thuộc vào DBA** và **tăng tốc độ khai thác thông tin**. Với dữ liệu hiện được lưu trữ dưới định dạng **Iceberg OTF** và siêu dữ liệu được quản lý bằng **AWS Glue Data Catalog**, các nhóm nội bộ tại TalentNeuron hiện nhận thấy **lợi ích từ việc phân loại và truy vấn nâng cao** – trực tiếp thúc đẩy các phân tích nâng cao và các dịch vụ dữ liệu dạng DaaS (Data-as-a-Service).

TalentNeuron đã **rút ngắn thời gian xử lý** cho các tác vụ này từ **nhiều ngày xuống chỉ còn vài giờ**, và đây là một bước quan trọng hướng tới mục tiêu doanh nghiệp: **trao quyền truy cập dữ liệu cho toàn tổ chức**.

**Một lợi ích khác sẽ đạt được là các dòng doanh thu bổ sung**

Khi TalentNeuron tiếp tục tiến xa hơn trong hành trình hiện đại hóa dữ liệu của mình, họ cũng đang **khám phá các cách mới để cung cấp DaaS** cho các bên liên quan bên ngoài.

**Định Hướng Tiếp Theo**

Hành trình hiện đại hóa dữ liệu của TalentNeuron cùng AWS mới chỉ bắt đầu. Với một nền tảng dữ liệu hiện đại đã được thiết lập, họ hiện có đủ sự linh hoạt để tiếp tục khám phá thêm các dịch vụ AWS khác, nhằm hướng tới mục tiêu trở thành **một tổ chức lấy dữ liệu làm trung tâm**.

Ví dụ, **Amazon Athena** có thể giúp thực hiện các truy vấn một lần trực tiếp trên data lake, còn **Amazon DataZone** và **AWS Data Exchange** mở ra cơ hội **đóng gói và chia sẻ dữ liệu dưới dạng sản phẩm DaaS (Data-as-a-Service)**.

*“Bằng cách tối ưu nền tảng với Aurora I/O-Optimized và AWS, chúng tôi không chỉ tiết kiệm chi phí – mà còn đang xây dựng một nền tảng để phát triển lâu dài,”*  
– **Roan** chia sẻ.  
*“Mục tiêu ở đây là trao quyền cho cả đội ngũ nội bộ và khách hàng của chúng tôi, để khai thác sức mạnh của dữ liệu theo cách chưa từng có trước đây.”*

**Kết Luận**

Quá trình tái kiến trúc dữ liệu và tư duy lấy dữ liệu làm trọng tâm của TalentNeuron đã cho thấy **AWS có thể biến những thách thức dữ liệu thành cơ hội thực sự**.

Việc áp dụng **Aurora I/O-Optimized** không chỉ giúp họ **giảm chi phí**, **đơn giản hóa kiến trúc**, mà còn **mở đường cho một tương lai nơi dữ liệu trở nên dễ tiếp cận hơn và có giá trị hơn**.

Khi tiếp tục đổi mới cùng AWS, TalentNeuron đang chứng minh rằng **hiện đại hóa đúng chiến lược không chỉ mang lại lợi ích trước mắt mà còn là nền móng cho thành công bền vững** về lâu dài.

**Thông Tin Về Tác Giả**

**Yaya Diawara**

Yaya là **Trưởng nhóm Kỹ sư Dữ liệu (Data Engineering Lead)** tại TalentNeuron với hơn **20 năm kinh nghiệm**, chuyên sâu trong việc thiết kế và tối ưu các hệ sinh thái dữ liệu hiện đại. Anh có kiến thức vững chắc về **cơ sở dữ liệu, phân tích dữ liệu, di chuyển hệ thống lên đám mây (cloud migrations), nền tảng big data** và **tinh chỉnh hiệu suất hệ thống**. Yaya đặc biệt thành thạo trong việc xây dựng **công cụ tìm kiếm, kho dữ liệu (data warehouses)** và quy trình **ETL**.

Ngoài công việc, anh yêu thích **âm nhạc, làm vườn, bơi lội**, và thường xuyên theo dõi các lĩnh vực **khám phá vũ trụ** cùng những **tiến bộ khoa học mới nhất**.

****

**Anupa Bhattacharyya**

Anupa là **Trưởng nhóm Hỗ trợ Doanh nghiệp (Enterprise Support Lead)** tại Amazon Web Services (AWS), nơi cô **hướng dẫn các khách hàng doanh nghiệp trong suốt hành trình chuyển đổi lên đám mây**. Với hơn **15 năm kinh nghiệm trong lĩnh vực dữ liệu và phân tích**, Anupa nổi bật với khả năng **xây dựng các sáng kiến chiến lược** dành cho khách hàng doanh nghiệp quy mô lớn.

Ngoài công việc, cô yêu thích **vẽ tranh, du lịch, dành thời gian cho gia đình**, và **thưởng thức các món ăn mới lạ**.

****

**Peter Chung**

Peter là **Kiến trúc sư Giải pháp Cấp cao (Senior Solutions Architect)** tại **AWS**, hiện đang làm việc tại **New York**. Anh hỗ trợ các công ty phần mềm và internet trong nhiều lĩnh vực **mở rộng quy mô, hiện đại hóa và tối ưu hóa hệ thống** trên nền tảng AWS. Peter là **tác giả của cuốn sách “AWS FinOps Simplified”** và cũng là **thành viên tích cực trong cộng đồng FinOps**, nơi tập trung vào tối ưu chi phí vận hành trên đám mây.