A black rectangular device with a black screen

Description automatically generated

**ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**---o0o---**

**HỆ THỐNG PHÂN TÍCH DỮ LIỆU LỚN**

**AMAZON ELASTIC MAPREDUCE**

**(AWS EMR)**

**Giáo viên hướng dẫn: Nguyễn Thị Anh Thư**

**Lớp: LT.K2022.3**

**Sinh viên thực hiện:**

**22410103 - Trần Đức Trung**

**22410069 - Hoàng Xuân Đạt**

**22410089 - Nguyễn Hoàng Phương**

**22410073 - Nguyễn Huỳnh Thúy Hằng**

**22410107 - Trương Văn Tùng**

**22410087 - Vũ Hoàng Phúc**

**22410076 - Nguyễn Tín Hữu**

**22410077 - Lữ Hoàng Khang**

**22410084 - Nguyễn Hoàng Phi**

**22410096 - Nguyễn Quốc Thắng**

**22410104 - Đỗ Anh Tú**

LỜI CẢM ƠN

Trong thời gian nghiên cứu, học tập tại khoa Công nghệ thông tin trường Đại học Công nghệ thông tin, được sự giúp đỡ quý báu của quý thầy cô giáo, các sinh viên cùng các doanh nghiệp là đối tác của nhà trường đã giúp nhóm em hoàn thiện đồ án tốt nghiệp. Nhóm chúng em xin chân thành cảm ơn Cô Nguyễn Thị Anh Thư - Giảng viên hướng dẫn môn Mạng xã hội đã dành thời gian quý báu để hướng dẫn chúng em trong môn học này. Cám ơn thầy đã tạo điều kiện và cung cấp những kiến thức quan trọng để học tập và thực hiện đề tài này một cách tốt nhất. Đồng thời nhóm em cũng xin chân thành cảm ơn Ban giám hiệu nhà trường, các thầy cô trong khoa Công nghệ thông tin cùng các phòng ban nhà trường đã tạo điều kiện tốt nhất cho nhóm em trong suốt thời gian học tập và làm đề tài môn học này.

Trong quá trình thực hiện làm đề tài còn có rất nhiều sai sót, nhóm em mong nhận được sự giúp đỡ nhiệt tình của thầy cô cũng như các để hoàn thiện một cách tốt nhất.

Nhóm em xin chân thành cảm ơn!

Nhóm sinh viên thực hiện

Nhóm 6

**NHẬN XÉT CỦA GIÁO VIÊN BỘ MÔN**

NHẬN XÉT CỦA NHÓM VỀ CÁC THÀNH VIÊN

[I. Giới thiệu, quá trình hình thành của Amazon (AWS) 7](#_Toc152168962)

[1. Giới thiệu 7](#_Toc152168963)

[2. Quá trình hình thành 7](#_Toc152168964)

[3. Ưu điểm và nhược điểm khi sử dụng AWS 8](#_Toc152168965)

[3.1. Ưu điểm: 8](#_Toc152168966)

[3.2. Nhược điểm: 8](#_Toc152168967)

[II. Giới thiệu Big Data, Data Analytics, Lake House 8](#_Toc152168968)

[1. Giới thiệu Big Data 8](#_Toc152168969)

[2. Data Analytics 9](#_Toc152168970)

[3. Lake House 10](#_Toc152168971)

[3.1. Giới thiệu 10](#_Toc152168972)

[3.2. Data Lakehouse hoạt động như thế nào? 10](#_Toc152168973)

[3.3. Các tính năng của một data lakehouse là gì? 10](#_Toc152168974)

[3.4. Những lợi thế của một hồ dữ liệu là gì 11](#_Toc152168975)

[3.5. Nhược điểm của data lakehouse 11](#_Toc152168976)

[3.6. Data Warehouse so với Data Lake so với Data LakeHouse: Đâu là sự khác biệt? 12](#_Toc152168977)

[3.7. Các thành phần của Kiến trúc Data LakeHouse 12](#_Toc152168978)

[3.7.1. Lưu trữ 12](#_Toc152168979)

[3.7.2. Định dạng tệp 12](#_Toc152168980)

[3.7.3. Định dạng bảng 12](#_Toc152168981)

[3.7.4. Công cụ truy vấn 13](#_Toc152168982)

[3.7.5. Ứng dụng 13](#_Toc152168983)

[III. Kiến trúc dữ liệu lớn để xử lý dữ liệu mạng xã hội của AWS. 13](#_Toc152168984)

[1. Kiến trúc dữ liệu lớn để xử lý dữ liệu mạng xã hội của AWS 13](#_Toc152168985)

[1.1. Mở rộng data lake linh hoạt 13](#_Toc152168986)

[1.2. Chuyên dụng cho hiệu suất và chi phí 13](#_Toc152168987)

[1.3. Phi máy chủ và dễ sử dụng 14](#_Toc152168988)

[1.4. Truy cập, bảo mật và quản trị dữ liệu hợp nhất 14](#_Toc152168989)

[1.5. Tích hợp máy học (Machine learning) 14](#_Toc152168990)

[2. Kiến trúc dữ liệu lớn nhóm 6 triển khai 14](#_Toc152168991)

[3. Các thành phần 14](#_Toc152168992)

[3.1. Ingestion Layer 14](#_Toc152168993)

[3.1.1. AWS DataSync 14](#_Toc152168994)

[3.1.2. AWS AppFlow 15](#_Toc152168995)

[3.2. Storage Layer, Catalog Layer hay còn gọi là Lake House Storage 15](#_Toc152168996)

[3.2.1. Amazon Simple Storage Service (AWS S3) 15](#_Toc152168997)

[A. Giới thiệu 15](#_Toc152168998)

[B. Buckets 16](#_Toc152168999)

[C. Objects 16](#_Toc152169000)

[D. Key 16](#_Toc152169001)

[E. Ưu Điểm Của Amazon S3 16](#_Toc152169002)

[F. Tại sao có thể dùng AWS S3 để thay thế cho hệ thống lưu trữa Apache Hadoop HDFS 17](#_Toc152169003)

[3.2.2. AWS Glue 18](#_Toc152169004)

[A. Giới thiệu 18](#_Toc152169005)

[B. Những lợi ích và hạn chế của việc sử dụng Glue 19](#_Toc152169006)

[C. Những hạn chế của AWS Glue 19](#_Toc152169007)

[D. Các hoạt động của AWS Glue 20](#_Toc152169008)

[E. Tính năng của AWS Glue 20](#_Toc152169009)

[F. Trường hợp sử dụng 20](#_Toc152169010)

[3.2.3. AWS Redshift 21](#_Toc152169011)

[A. Giới thiệu 21](#_Toc152169012)

[B. Kiến trúc hệ thống tổng quát 21](#_Toc152169013)

[a. AWS Redshift Spectrum 21](#_Toc152169014)

[b. Client Application và Connection 21](#_Toc152169015)

[c. Lưu trữ dữ liệu 22](#_Toc152169016)

[C. Perfomance 23](#_Toc152169017)

[a. Massively parallel processing 23](#_Toc152169018)

[b. Columnar data storage 23](#_Toc152169019)

[c. Data Compression 23](#_Toc152169020)

[d. Query Optimizer 24](#_Toc152169021)

[D. Cách thức hoạt động 24](#_Toc152169022)

[3.3. Big Data Processing và Enriching 24](#_Toc152169023)

[A. SQL based ELT (với Redshift Spectrum) 24](#_Toc152169024)

[3.4. Data Consumption Layer (Visualazation) 25](#_Toc152169025)

[3.4.1. AWS QuickSight 25](#_Toc152169026)

[A. Giới thiệu 25](#_Toc152169027)

[B. Kiến trúc 26](#_Toc152169028)

[C. Tính năng nổi bật của QuickSight 26](#_Toc152169029)

[a. Trực quan hóa dữ liệu nhanh chóng 26](#_Toc152169030)

[b. Hỗ trợ nhiều nguồn dữ liệu 27](#_Toc152169031)

[c. Phí tính theo phiên sử dụng 27](#_Toc152169032)

[d. Mở rộng dễ dàng 27](#_Toc152169033)

[e. Tính năng học máy ML 27](#_Toc152169034)

[f. Tích hợp vào ứng dụng 27](#_Toc152169035)

[g. Giải pháp toàn cầu 27](#_Toc152169036)

[D. Bảo mật trong Amazon QuickSight 27](#_Toc152169037)

[IV. Kết luận 28](#_Toc152169038)

# Giới thiệu, quá trình hình thành của Amazon (AWS)

## Giới thiệu

Amazon web services là một nền tảng điện toán đám mây phát triển toàn diện được cung cấp bởi Amazon.com. Dịch vụ Web đôi khi được gọi là dịch vụ đám mây hoặc các dịch vụ điện toán từ xa. Các dịch vụ AWS đầu tiên đã được đưa ra vào năm 2006 để cung cấp các dịch vụ trực tuyến cho các trang web và các ứng dụng phía máy khách. Để giảm thiểu việc bị mất điện đột ngột và đảm bảo tính mạnh mẽ của hệ thống, AWS đa dạng về địa lý theo khu vực. Các khu vực này có trung tâm ở Đông Mỹ, Tây Mỹ (hai địa điểm), Brazil, Ireland, Singapore, Nhật Bản và Úc. Mỗi khu vực bao gồm nhiều khu vực địa lý nhỏ hơn được gọi là vùng sẵn có.



1. *AWS*

Amazon web services hoạt động dựa trên nền tảng cloud computing (điện toán đám mây) vậy điện toán đám mây là gì? Điện toán đám mây thường được gọi là "đám mây", là việc phân phối các tài nguyên máy tính theo yêu cầu - tất cả mọi thứ từ các ứng dụng đến các trung tâm dữ liệu-qua Internet trên cơ sở trả cho sử dụng.

## Quá trình hình thành

* **Năm 2003:** Năm 2003, Chris Pinkham và Benjamin Black đã trình bày một bài báo về cơ sở hạ tầng nội bộ của Amazon trông như thế nào. Họ đề nghị bán nó như một dịch vụ và chuẩn bị một tình huống kinh doanh trên đó. Họ đã chuẩn bị một tài liệu dài sáu trang và xem qua nó để xem xét tiến hành nó hay không. Họ quyết định tiến hành tài liệu.
* **Năm 2004:** SQS là viết tắt của “Simple Queue Service” được chính thức ra mắt vào năm 2004. Một nhóm đã đưa ra dịch vụ này tại Cape Town, Nam Phi.
* **Năm 2006:** AWS (Amazon Web Services) chính thức ra mắt.
* **Năm 2007:** Năm 2007, hơn 180.000 developer đã đăng ký AWS.
* **Năm 2010:** Năm 2010, các dịch vụ web bán lẻ amazon.com được chuyển sang AWS, tức là amazon.com hiện đang chạy trên AWS.
* **Năm 2011:** AWS gặp phải một số vấn đề lớn. Một số phần của khối lượng EBS (Elastic Block Store) đã bị kẹt và không thể đọc và ghi các yêu cầu. Phải mất hai ngày để vấn đề được giải quyết.
* **Năm 2012:** AWS tổ chức sự kiện khách hàng đầu tiên được gọi là hội nghị re: Invent. Hội nghị tái phát minh đầu tiên diễn ra trong đó các sản phẩm mới được đưa ra. Trong AWS, một vấn đề lớn khác đã xảy ra ảnh hưởng đến nhiều trang web phổ biến như Pinterest, Reddit và Foursquare.
* **Năm 2013:** Năm 2013, các chứng chỉ đã được đưa ra. AWS đã bắt đầu chương trình cấp chứng chỉ dành cho các kỹ sư phần mềm có chuyên môn về điện toán đám mây.
* **Năm 2014:** AWS cam kết đạt được 100% việc sử dụng năng lượng tái tạo cho dấu ấn toàn cầu của mình.
* **Năm 2015:** AWS phá vỡ doanh thu và đạt tới $ 6 Tỷ USD mỗi năm. Doanh thu tăng trưởng 90% mỗi năm.
* **Năm 2016:** Đến năm 2016, doanh thu tăng gấp đôi và đạt $ 13 Tỷ USD mỗi năm.
* **Năm 2017:** Năm 2017, AWS re: creative phát hành một loạt các Dịch vụ Trí tuệ Nhân tạo, nhờ đó doanh thu của AWS tăng gấp đôi và đạt 27 tỷ USD mỗi năm.
* **Năm 2018:** Vào năm 2018, AWS đã đưa ra Chứng chỉ Chuyên môn về Máy học. Nó tập trung nhiều vào việc tự động hóa Trí tuệ nhân tạo và Máy học.

## Ưu điểm và nhược điểm khi sử dụng AWS

### Ưu điểm

* **Dễ sử dụng:** AWS cung cấp một giao diện người dùng trực quan giúp dễ dàng tạo và quản lý. Không cần phải có kiến thức chuyên sâu về về triển khai để sử dụng các dịch vụ AWS.
* **Quản lý tập trung:** AWS giúp quản lý tất cả service của mình từ một vị trí duy nhất. Có thể dễ dàng theo dõi hiệu suất của các service và triển khai các bản cập nhật.
* **Chi phí thấp:** AWS cung cấp một mô hình định giá linh hoạt cho phép chỉ trả tiền cho những gì sử dụng. Có thể tiết kiệm chi phí bằng cách sử dụng các service theo yêu cầu.
* **Khả năng mở rộng:** AWS giúp dễ dàng thêm hoặc xóa các tài nguyên khi cần thiết. Điều này cho phép đáp ứng nhu cầu tăng trưởng của dữ liệu và khối lượng công việc.
* **Tính bảo mật:** AWS cung cấp các tính năng bảo mật tích hợp giúp bảo vệ dữ liệu của bạn
* **Tính tích hợp:** AWS tích hợp với các dịch vụ AWS khác, chẳng hạn như Amazon S3 và Amazon Redshift. Điều này cho phép dễ dàng xử lý và phân tích dữ liệu của mình.

### Nhược điểm

* **Hạn chế dữ liệu:** Điểm yếu đầu tiên của AWS là hạn chế 1 số dữ liệu liên quan đến EC2. Các công ty đang sử dụng AWS sẽ có tài nguyên mặc định để sử dụng. Nhưng vấn đề xảy ra khi giới hạn tài nguyên mặc định khác nhau giữa các khu vực. Mặc dù họ có thể gửi yêu cầu về nhiều tài nguyên hơn. Nhưng các công ty nhận thấy đó là một nhược điểm của AWS
* **Thiếu chuyên gia:** AWS là phát minh mới nhất với cơ sở hạ tầng phức tạp. Các công ty sử dụng AWS làm nền tảng điện toán đám mây phải thuê các chuyên gia về làm việc. Tuy nhiên, chỉ một số chuyên gia có kỹ năng về AWS. Các công ty đã sẵn sàng đầu tư vào AWS. Nhưng họ đang gặp khó khăn trong việc tìm kiếm các chuyên gia phù hợp. Để có thể dẫn dắt các tổ chức này đạt được hiệu suất tốt hơn. Đây có thể là một bất lợi của AWS về thời gian và tiền bạc.
* **Sự thay đổi về giá:** Giá của các dịch vụ AWS khác nhau dựa trên các yếu tố như chi phí đất đai, cáp quang, điện và thuế giữa các khu vực. 3 gói có sẵn cho lựa chọn đó là Developer, Business và Enterprise. Giá của chúng cũng thay đổi tương ứng. Điều này sẽ ảnh hưởng đến chi phí của bạn.

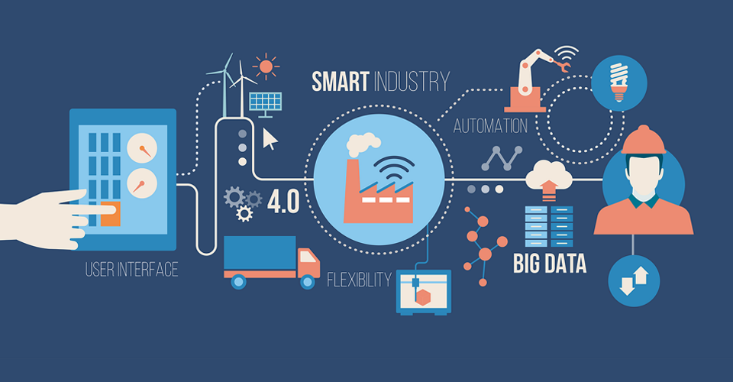
# Giới thiệu Big Data, Data Analytics, Lake House

## Giới thiệu Big Data

Big Data là các tập dữ liệu có khối lượng lớn và phức tạp. Độ lớn đến mức các phần mềm xử lý dữ liệu truyền thống không có khả năng thu thập, quản lý và xử lý dữ liệu trong một khoảng thời gian hợp lý.

Những tập dữ liệu lớn này có thể bao gồm các dữ liệu có cấu trúc, không có cấu trúc và bán cấu trúc, mỗi tập có thể được khai thác để tìm hiểu insights.

Các giải pháp Big Data cung cấp các công cụ, phương pháp và công nghệ được sử dụng để nắm bắt, lưu trữ, tìm kiếm và phân tích dữ liệu trong vài giây để tìm mối quan hệ và hiểu biết về cải tiến và lợi ích cạnh tranh mà trước đây không có.



1. Big Data

Ngày nay, 80% dữ liệu là dữ liệu không có cấu trúc và không thể được xử lý bởi các công nghệ truyền thống. Trước đó, một lượng dữ liệu được tạo ra không cao. Chúng ta tiếp tục lưu trữ dữ liệu vì chỉ cần phân tích lịch sử dữ liệu. Nhưng ngày nay việc tạo dữ liệu tính bằng petabyte và không thể lưu trữ dữ liệu nhiều lần và lấy lại khi cần.

Ứng dụng của Big Data:

* Netflix sử dụng Big Data để cải thiện trải nghiệm của khách hàng
* Phân tích chiến dịch và kế hoạch xúc tiến của Sears Holding
* Phân tích cảm xúc
* Phân tích khách hàng
* Phân tích dự đoán
* Kết hợp và quảng cáo theo thời gian thực

## Data Analytics

Data science là lĩnh vực áp dụng toán học, thống kê ứng dụng kết hợp với kiến thức của nhiều lĩnh vực khác nhau để diễn giải dữ liệu thô, phức tạp thành những thông tin giá trị cho việc ra quyết định.

Khoa học dữ liệu bắt đầu bằng khai thác dữ liệu (data mining). Khai thác dữ liệu sẽ tập trung vào việc thu thập liệu là chính, vì thế những dữ liệu lúc này chỉ là những dữ liệu thô và chưa có giá trị thực.

Data analytics chính là mắt xích chuyển giao quan trọng biến những dữ liệu thô thành những thông tin có giá trị để lý giải hoặc giải quyết một vấn đề cụ thể.

Cụ thể thì data analytics là một quá trình gồm những bước chính như: tổng hợp dữ liệu thô, chọn lọc, phân nhóm, sử dụng những phương pháp khoa học đặc thù để phân tích và đưa ra kết luận. Có thể hiểu theo một cách đơn giản thì data analytics là hoạt động phân tích các dữ liệu thô thành các thông tin dễ hiểu và trực quan nhằm hỗ trợ cho quyết định của doanh nghiệp được giảm thiểu rủi ro, tối ưu nguồn lực và dự đoán tương lai.

Những dữ liệu thô hay “ma trận số” sẽ được giải nghĩa bởi data analytics khiến cho doanh nghiệp có cơ sở chắc chắn trong từng bước đi thậm chí sớm nắm bắt được xu hướng mới.

A diagram of data mining

Description automatically generated

1. Data Analytics

Hiện nay, data analytics được chia thành 4 loại cơ bản như sau:

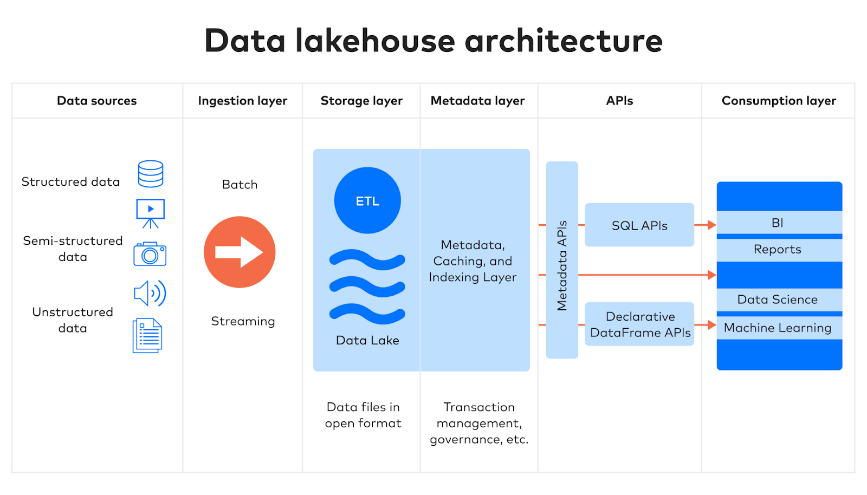
* **Phân tích mô tả (Descriptive analytics):** Phân tích dữ liệu dạng này thường nhắm vào việc mô tả những dữ liệu đã xảy ra trong một khoảng thời gian nhất định. Những dữ liệu trong lịch sử sẽ được tổng hợp, so sánh và đưa ra những kết luận – Ví dụ như tỷ lệ tăng trưởng doanh thu của tháng này so với tháng trước; xu hướng người dùng lựa chọn những thương hiệu có thành phần “organic”, “xanh”, “sản xuất bền vững” …tăng lên trong những năm gần đây.
* **Phân tích chẩn đoán (diagnostic analytics):** Cách tiếp cận này tập trung phân tích LÝ DO của những diễn biến đã xảy ra. Trong quá trình vận hành, doanh nghiệp sẽ đối mặt với không ít những vấn đề không như dự tính. diagnostic analytics này mang nhiệm vụ chẩn đoán nguyên nhân của vấn đề để doanh nghiệp rà soát, điều chỉnh nhanh chóng. Ví dụ như thái độ của nhân viên ca sáng tại một cửa hàng bánh ngọt chưa tốt là nguyên nhân chính khiến doanh thu ca sáng giảm và chênh lệch với ca tối.
* **Phân tích dự đoán (predictive analytics):** Cách tiếp cận này của data analytics nhằm đưa ra những dự đoán trong tương lai gần dựa trên cơ sở dữ liệu trong lịch sử. Ví dụ thời tiết nắng nóng sẽ ảnh hưởng tới tình hình kinh doanh ở năm ngoái. Vậy nếu như tình hình thời tiết năm nay cũng như vậy thì khả năng cao thì doanh thu sẽ bị ảnh hưởng như cùng kỳ năm trước.
* **Phân tích đề xuất (prescriptive analytics):** Phân tích dữ liệu loại này là sự tổng hợp của tất cả phân loại trên để hiểu sâu sắc dữ liệu để từ đó đề xuất các hoạt động phù hợp. Ví dụ như nếu thời tiết nắng nóng năm nay sẽ khắc nghiệt hơn năm ngoái thì cửa hàng nước uống nên chuẩn bị thêm nước đá để bảo quản nguyên vật liệu, áp dụng các chương trình khuyến mãi như mua “mua 1 tặng 1” để thúc đẩy sự mua hàng của người dùng.

## Lake House

### Giới thiệu

Data lakehouse nghĩa là hồ dữ liệu có tính linh hoạt và tiết kiệm chi phí lưu trữ dữ liệu với khả năng truy vấn dữ liệu theo ngữ cảnh và tốc độ cao.

Data lakehouse cung cấp một kho lưu trữ duy nhất cho dữ liệu có cấu trúc và cung cấp nguồn thông tin xác thực cho các tổ chức. Tuy nhiên, các tổ chức phải cấu trúc và lưu trữ dữ liệu đầu vào ở một định dạng cụ thể để cho phép các quy trình trích xuất, chuyển đổi và tải cũng như truy vấn dữ liệu này một cách hiệu quả.



1. LakeHouse

### Data Lakehouse hoạt động như thế nào?

Data lakehouses thường cung cấp hỗ trợ cho việc nhập dữ liệu thông qua nhiều phương pháp. Chúng bao gồm các giao diện lập trình ứng dụng, phát trực tuyến, v.v. Dữ liệu này xuất hiện ở dạng thô, nguyên bản mà không yêu cầu giải mã lược đồ. Kho dữ liệu cung cấp lớp lưu trữ tiết kiệm chi phí cho cả dữ liệu có cấu trúc và phi cấu trúc. Do đó, nó chứa tất cả dữ liệu của một tổ chức.

Nói chung, công nghệ lưu trữ phân loại dữ liệu thành các vùng hạ cánh, vùng thô và được quản lý tùy thuộc vào mức độ sẵn sàng sử dụng của nó. Kết quả là nhận lại là nguồn thông tin chính xác duy nhất và cho phép các công ty tận dụng tối đa khả năng phân tích dữ liệu nâng cao cùng một lúc.

Trong mô hình kho dữ liệu, trước tiên, các tổ chức di chuyển dữ liệu từ các nguồn vào trong kho dữ liệu. Sau đó, một tập hợp con của dữ liệu này sẽ lọc liền mạch để trở thành các tập dữ liệu được quản lý và trở nên đáng tin cậy hơn, trong đó các tổ chức đặt các quy tắc quản trị, sử dụng và truy cập dữ liệu bắt buộc.

### Các tính năng của một data lakehouse là gì?

Mặc dù các chi tiết cụ thể của kho dữ liệu khác nhau dựa trên mục tiêu kinh doanh và trường hợp sử dụng, năm tính năng sau đây là cơ bản:

* **Nhập dữ liệu:** Data lakehouses nhập khối lượng dữ liệu có cấu trúc và phi cấu trúc lớn với tốc độ rất cao ở dạng nguyên bản, thô của chúng. Các công nghệ mở giúp dễ dàng nhập khối lượng lớn dữ liệu kinh doanh, tăng cường khả năng bảo mật và khả năng quan sát.
* **Quản lý dữ liệu**: Kho lưu trữ dữ liệu tận dụng các kho lưu trữ đối tượng chi phí thấp như AWS S3 hoặc Microsoft Azure Blob Storage để lưu trữ và quản lý dữ liệu một cách hiệu quả về chi phí.
* **Ngôn ngữ truy vấn**: Data lakehouses cung cấp một cách để thẩm vấn dữ liệu và gửi hướng dẫn xử lý dưới dạng truy vấn. Hiệu suất của các truy vấn này cần phải ở mức mà chúng có thể hỗ trợ các trường hợp sử dụng phân tích đặc biệt.
* **Xử lý song song hàng loạt**: Data lakehouses cung cấp phản hồi truy vấn với độ trễ tối thiểu. Điều này đơn giản là không thể với các kiến trúc thông thường. Xử lý song song ồ ạt cho phép thực hiện truy vấn phân tán bởi nhiều bộ xử lý để tăng tốc độ phản hồi.

Hỗ trợ khối lượng công việc phân tích đa dạng. Kho dữ liệu hỗ trợ nhiều trường hợp sử dụng xung quanh phân tích và học máy, chẳng hạn như phân tích khám phá để trả lời các câu hỏi đặc biệt.

### Những lợi thế của một hồ dữ liệu là gì

**Giảm dự phòng:** Vì data lakehouse kết hợp các chức năng của hồ và kho dữ liệu nên chúng cung cấp một nền tảng lưu trữ đa năng có thể xử lý mọi loại dữ liệu. Do đó, các tổ chức có thể tránh xa các mô hình kho-hồ kết hợp xem dữ liệu được sao chép để đảm bảo khả năng truy cập.

**Kiểm soát chi phí:** Bằng cách sử dụng các giải pháp lưu trữ đối tượng được tiêu chuẩn hóa, kho dữ liệu mang lại hiệu quả chi phí và khả năng kiểm soát giống như kho dữ liệu. Thêm vào mô hình kho lưu trữ nguồn đơn của kho dữ liệu và doanh nghiệp có thể hợp nhất dữ liệu để kiểm soát chi phí.

**Cải thiện quản trị:** Các mô hình Lakehouse ưu tiên bảo mật dữ liệu và giúp dễ dàng áp dụng các cơ chế quản trị cho phép mọi người làm việc với dữ liệu họ cần, nhưng cũng tuân thủ các tiêu chuẩn tuân thủ quy định ở quy mô toàn cầu.

### Nhược điểm của data lakehouse

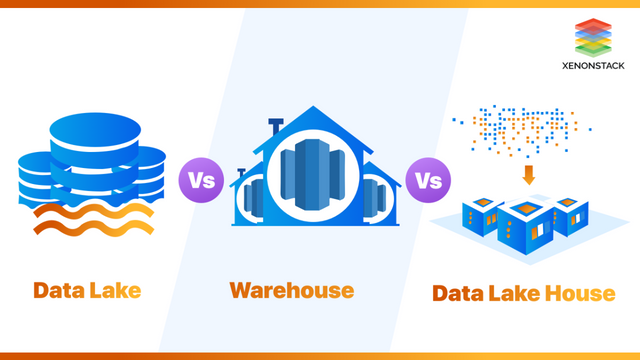
Mặc dù các data lakehouse cho phép các tổ chức lưu trữ bất kỳ khối lượng hoặc loại dữ liệu nào mà không cần xem xét cấu trúc hoặc chi phí lưu trữ, nhưng chất lượng và quản trị của chúng vẫn là những vấn đề quan trọng. Nếu dữ liệu không được quản lý đúng cách, nó có thể dẫn đến việc hồ dữ liệu trở thành đầm lầy.

Cách tiếp cận này dẫn đến các bản sao dữ liệu bổ sung vì dữ liệu trước tiên được trích xuất và tải vào data lakehouse, sau đó cần được trích xuất và tải vào kho dữ liệu cho các ứng dụng xuôi dòng (chẳng hạn như BI). Điều này cũng có thể dẫn đến nhiều lỗi công việc hơn và cuối cùng ảnh hưởng đến các ứng dụng hạ lưu.

Vì dữ liệu được lưu trữ ở định dạng thô và được viết bởi nhiều công cụ và phương pháp khác nhau nên các tệp có thể không được tối ưu hóa cho các công cụ truy vấn và ứng dụng phân tích sẽ có độ trễ thấp.

### Data Warehouse so với Data Lake so với Data LakeHouse: Đâu là sự khác biệt?

Mặc dù data lakehouse kết hợp tính linh hoạt và hiệu quả chi phí của data lake với khả năng truy vấn của data warehouse, nhưng điều quan trọng là phải hiểu các môi trường lưu trữ này khác nhau như thế nào.



1. So sánh Data Lake, Warehouse, Data Lake House

**Data Warehouse**: Kho dữ liệu là tùy chọn lưu trữ dữ liệu lớn ban đầu. Được thiết kế để cung cấp một nguồn thông tin chính xác duy nhất cho dữ liệu có cấu trúc, chúng cung cấp một cách để các tổ chức đơn giản hóa việc quản lý dữ liệu bằng cách tập trung đầu vào. Mặc dù họ cực kỳ thành thạo chức năng này, nhưng họ lại thiếu tính linh hoạt của các hậu duệ hồ dữ liệu của mình.

**Data Lake:** Data Lake đã phát triển như một cách để quản lý đồng thời dữ liệu có cấu trúc và phi cấu trúc tốt hơn. Các hồ này rất linh hoạt và cho phép các công ty lưu trữ dữ liệu ở dạng gốc. Đổi lại, điều này trao quyền cho khoa học dữ liệu chuyên sâu hơn và giảm tổng chi phí quản lý dữ liệu.

**Data LakeHouse:** Data LakeHouse kết hợp các khía cạnh chính của hồ dữ liệu và kho dữ liệu. Điều này cho phép các doanh nghiệp sử dụng mô hình kho dữ liệu như một kho lưu trữ để lưu trữ hợp nhất dữ liệu mà không phải hy sinh tính linh hoạt phân tích của hồ dữ liệu, cho phép kho dữ liệu hoạt động vượt trội ở cả khối lượng công việc phân tích và máy học.

### Các thành phần của Kiến trúc Data LakeHouse

#### Lưu trữ

Lưu trữ là thành phần đầu tiên của một hồ dữ liệu. Đây là nơi dữ liệu hạ cánh sau khi nhập từ các hệ thống hoạt động. Các kho có sẵn từ ba nhà cung cấp dịch vụ đám mây - Amazon S3, Azure Blob Storage và Google Cloud Storage - hỗ trợ lưu trữ bất kỳ loại dữ liệu nào và tạo điều kiện thuận lợi cho hiệu suất và bảo mật cần thiết. Các hệ thống này cũng rất có thể mở rộng và không tốn kém, giúp hợp lý hóa chi phí.

#### Định dạng tệp

Thành phần tiếp theo là nơi lưu trữ dữ liệu thực tế. Thông thường, chúng là các định dạng cột mang lại lợi thế đáng kể trong việc đọc dữ liệu hoặc chia sẻ dữ liệu giữa nhiều hệ thống. Các định dạng tệp phổ biến bao gồm Apache Parquet, ORC, Apache Arrow, v.v. Các tệp này được lưu trữ trong bộ lưu trữ đối tượng.

#### Định dạng bảng

Định dạng bảng data lake là thành phần quan trọng nhất của kiến trúc lakehouse. Phải có một số cách để sắp xếp và quản lý tất cả các tệp dữ liệu thô trong kho lưu trữ hồ dữ liệu. Các định dạng bảng giúp trừu tượng hóa độ phức tạp của cấu trúc dữ liệu vật lý và cho phép các công cụ khác nhau hoạt động đồng thời trên cùng một dữ liệu. Định dạng bảng trong kiến trúc lakehouse tạo điều kiện thuận lợi cho khả năng thực hiện các giao dịch cấp kho dữ liệu (DML) cùng với các đảm bảo ACID. Một số tính năng quan trọng khác của định dạng bảng là tiến hóa lược đồ, SQL biểu cảm, nén dữ liệu, v.v. Apache Iceberg, Hudi và Delta Lake là ba định dạng bảng phổ biến nhất và đang đạt được đà phát triển rộng rãi.

#### Công cụ truy vấn

Các định dạng bảng cung cấp các thông số kỹ thuật và API cần thiết để tương tác với dữ liệu bảng. Tuy nhiên, trách nhiệm xử lý dữ liệu và cung cấp hiệu suất đọc hiệu quả thuộc về công cụ truy vấn. Một số công cụ truy vấn cũng cho phép kết nối riêng với các công cụ BI như Tableau, giúp dễ dàng thực hiện báo cáo trực tiếp trên dữ liệu được lưu trữ trong bộ lưu trữ. Các công cụ truy vấn như Dremio Sonar và Apache Spark hoạt động trơn tru với các định dạng bảng như Apache Iceberg để kích hoạt kiến trúc Data lakehouse mạnh mẽ bằng cách sử dụng các ngôn ngữ thường được sử dụng như SQL.

#### Ứng dụng

Thành phần cuối cùng của data lakehouse là các ứng dụng hạ nguồn tương tác với dữ liệu. Chúng bao gồm các công cụ BI như Tableau và Power BI cũng như các khung máy học như TensorFlow, PyTorch, v.v., giúp các nhà phân tích dữ liệu, nhà khoa học dữ liệu và kỹ sư ML dễ dàng truy cập trực tiếp vào dữ liệu. Thông thường, việc này mất hàng tuần, nếu không muốn nói là hàng tháng, trong các kiến trúc dữ liệu khác.

# Kiến trúc dữ liệu lớn để xử lý dữ liệu mạng xã hội của AWS.

## Kiến trúc dữ liệu lớn để xử lý dữ liệu mạng xã hội của AWS

AWS cung cấp nhiều lựa chọn dịch vụ phân tích đa dạng nhất, phù hợp với mọi nhu cầu phân tích dữ liệu đồng thời cho phép các tổ chức thuộc mọi quy mô và lĩnh vực làm mới lại công việc kinh doanh của họ bằng dữ liệu. Từ di chuyển dữ liệu, lưu trữ dữ liệu, hồ dữ liệu, phân tích dữ liệu lớn, phân tích bản ghi, phân tích phát trực tuyến, nghiệp vụ thông minh, máy học, v.v., AWS cung cấp các dịch vụ chuyên dụng có tỷ lệ hiệu năng-mức giá tốt nhất, khả năng điều chỉnh quy mô và có chi phí thấp nhất.

A screenshot of a computer

Description automatically generated

1. Mô hình kiến trúc lake house của AWS

### Mở rộng data lake linh hoạt

Có được sự hỗ trợ từ Amazon S3, dịch vụ có độ sẵn sàng bậc nhất, các hồ dữ liệu do AWS cung cấp có thể mang đến quy mô, tốc độ và tính linh hoạt cần có để kết hợp các phương pháp phân tích và dữ liệu khác nhau. Xây dựng và lưu trữ các hồ dữ liệu của trên AWS để có thông tin chuyên sâu hơn so với khi sử dụng các kho và silo dữ liệu truyền thống.

### Chuyên dụng cho hiệu suất và chi phí

Các dịch vụ phân tích của AWS là những công cụ chuyên dụng giúp nhanh chóng trích xuất thông tin chuyên sâu từ dữ liệu bằng cách sử dụng công cụ phù hợp nhất cho tác vụ, cũng như được tối ưu hóa để đem đến cho hiệu suất, quy mô tốt nhất cũng như chi phí phù hợp với nhu cầu của bạn.

### Phi máy chủ và dễ sử dụng

AWS là dịch vụ có nhiều lựa chọn phi máy chủ nhất cho việc phân tích dữ liệu trên đám mây bao gồm cả các lựa chọn về lưu kho dữ liệu, phân tích dữ liệu lớn, dữ liệu theo thời gian thực, tích hợp dữ liệu, v.v. Chúng tôi quản lý cơ sở hạ tầng cơ bản để có thể chỉ cần tập trung vào ứng dụng của mình.

### Truy cập, bảo mật và quản trị dữ liệu hợp nhất

AWS cho phép xác định và quản lý một cách tập trung các chính sách bảo mật, quản trị và kiểm tra nhằm đáp ứng những quy định cụ thể của ngành và khu vực địa lý. Với AWS, có thể truy cập dữ liệu của mình bất kể dữ liệu nằm ở đâu và chúng tôi sẽ bảo mật dữ liệu đó, bất kể lưu trữ nó ở đâu.

### Tích hợp máy học (Machine learning)

AWS cung cấp tiện ích Machine learning tích hợp sẵn trong các dịch vụ phân tích chuyên dụng. có thể xây dựng, đào tạo và triển khai mô hình máy học bằng các lệnh SQL quen thuộc mà không cần kinh nghiệm về máy học.

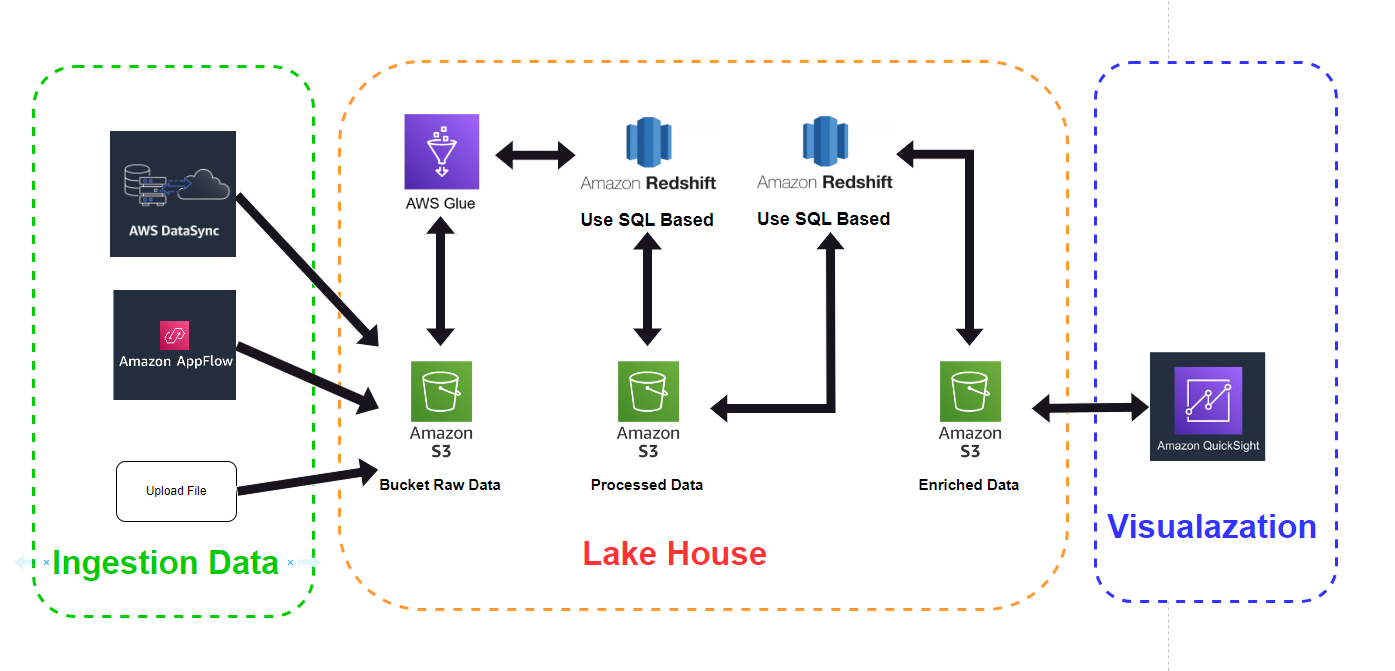
## Kiến trúc dữ liệu lớn nhóm 6 triển khai

Nhóm 6 dùng kiến trúc như sau để triển khai BigData:

**Ingestion Data:** DataSync, AppFlow

**LakeHouse:** S3 + Glue + Redshift

**Visualazation:** QuickSight



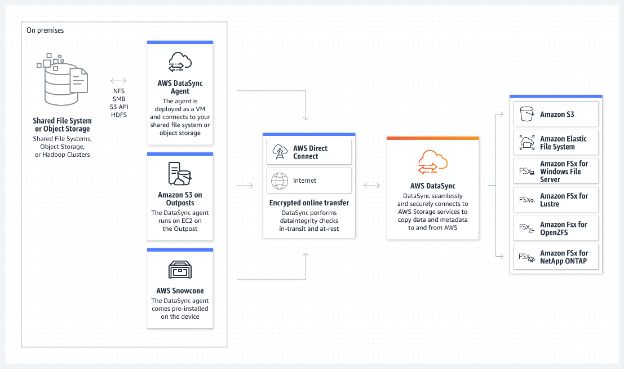
1. AWS BigData nhóm 6 Architecture

## Các thành phần

### Ingestion Layer

#### AWS DataSync

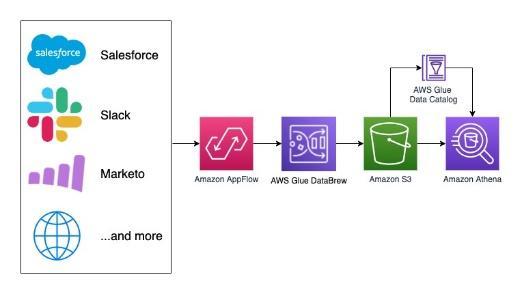
AWS DataSync là dịch vụ bảo mật trực tuyến giúp tự động hóa và đẩy nhanh quá trình di chuyển dữ liệu giữa không gian lưu trữ tại chỗ và dịch vụ Lưu trữ của AWS. DataSync có thể sao chép dữ liệu giữa giao thức chia sẻ Hệ thống tệp mạng (NFS), giao thức chia sẻ Khối thông điệp máy chủ (SMB), Hệ thống tệp phân tán Hadoop (HDFS), vị trí lưu trữ đối tượng tự quản lý, AWS Snowcone, vùng lưu trữ trên Dịch vụ lưu trữ đơn giản của Amazon (Amazon S3), hệ thống tệp trên Hệ thống tệp linh hoạt của Amazon (Amazon EFS), hệ thống tệp Amazon FSx dành cho Windows File Server, hệ thống tệp Amazon FSx dành cho Lustre, hệ thống tệp Amazon FSz dành cho OpenZFS và hệ thống tệp Amazon FSx dành cho NetApp ONTAP.



1. Data Sync

#### AWS AppFlow

AWS AppFlow là một dịch vụ tích hợp được quản lý hoàn toàn, giúp bạn truyền dữ liệu một cách bảo mật giữa các ứng dụng phần mềm dưới dạng dịch vụ (SaaS) và các dịch vụ AWS. Với AppFlow, bạn có thể dễ dàng chuyển dữ liệu từ các ứng dụng SaaS phổ biến như Salesforce, SAP, Google Analytics, Facebook Ads và ServiceNow sang các dịch vụ AWS như Amazon Simple Storage Service (S3), Amazon Redshift, Amazon Relational Database Service (RDS) và Amazon Glue.



1. AWS AppFlow

Amazon AppFlow là một dịch vụ tích hợp mạnh mẽ và dễ sử dụng giúp bạn truyền dữ liệu một cách bảo mật giữa các ứng dụng SaaS và các dịch vụ AWS. AppFlow có thể được sử dụng cho nhiều trường hợp sử dụng khác nhau, giúp bạn dễ dàng kết nối dữ liệu của mình và tận dụng sức mạnh của AWS.

### Storage Layer, Catalog Layer hay còn gọi là Lake House Storage

#### Amazon Simple Storage Service (AWS S3)

##### Giới thiệu

S3 là dịch vụ lưu trữ đối tượng được xây dựng để lưu trữ và truy xuất bất kỳ lượng dữ liệu nào từ bất cứ nơi nào. Đây là dịch vụ lưu trữ đơn giản có độ bền, độ sẵn có, hiệu suất, tính bảo mật dẫn đầu ngành và khả năng thay đổi quy mô gần như không giới hạn với chi phí cực kỳ thấp.

AWS S3 có thể lưu trữ và bảo vệ dữ liệu cho nhiều trường hợp như kho dữ liệu, trang web, ứng dụng di động, sao lưu và khôi phục, lưu trữ, ứng dụng doanh nghiệp, thiết bị IoT và dữ liệu lớn phân tích.

Amazon S3 cung cấp các tính năng quản lý để có thể tối ưu hóa, sắp xếp và thiết lập cấu hình quyền truy cập vào dữ liệu của mình. Nhằm đáp ứng các yêu cầu tuân thủ, tổ chức và kinh doanh cụ thể của bạn

A screenshot of a computer

Description automatically generated

1. AWS S3

Amazon S3 – Có thể sử dụng để:

* Data Lakes & Big data analytics – Lưu trữ và sử lý dữ liệu lớn
* Backup and Storage – Lưu trữ dự phòng dữ liệu và dịch vụ lưu trữ khác
* Application Hosting – Dịch vụ triển khai, cài đặt và quản lý ứng dụng web
* Media Hosting – Xây dựng hạ tầng dự phòng, khả năng mở rộng, tính sẵn sàng cao để lưu video, photo, hoặc music cho phép upload/download.
* Software Delivery – Lưu phần mềm ứng dụng để khách hàng có thể tải về.
* Static webstite – Trang web tĩnh

Amazon S3 lưu trữ dữ liệu dưới dạng objects trong buckets. Một object là một tệp hay bất kỳ siêu dữ liệu (metadata) nào trong tệp đó. Một bucket là một vùng chứa cho các đối tượng. Để lưu trữ dữ liệu của trong Amazon S3, phải tạo một bucket và chỉ định tên vùng chứa cũng như AWS Region. Sau đó, tải dữ liệu lên bucket đó dưới dạng object trong Amazon S3. Mỗi object có một key, là mã định danh duy nhất cho đối tượng trong nhóm.

##### Buckets

Buckets là nơi chứa các đối tượng được lưu trữ trong Amazon S3. Có thể lưu trữ bất kỳ đối tượng nào trong một nhóm và có thể có tối đa 100 nhóm trong tài khoản của mình. Khi tạo bucket, nhập tên bộ chứa và chọn AWS Region nơi bộ chứa sẽ cư trú. Sau khi tạo một bucket, không thể thay đổi tên hoặc Region (Vùng) của nó. Như thế, một bucket giống như một vùng chứa cho các object lưu trữ trên Amazon S3. Mỗi object sẽ được chứa trong một bucket. Buckets được sử dụng cho những mục đích sau:

* Sắp xếp, quản lý Amazon S3 namespace ở mức cao nhất.
* Xác định tài khoản chịu trách nhiệm về phí lưu trữ và truyền dữ liệu.
* Cung cấp các tùy chọn kiểm soát truy cập, chẳng hạn như chính sách bộ chứa, danh sách kiểm soát truy cập (ACL) và Điểm truy cập S3 mà có thể sử dụng để quản lý quyền truy cập vào tài nguyên Amazon S3 của mình.
* Đóng vai trò là đơn vị tổng hợp cho báo cáo sử dụng.

##### Objects

Objects là các thực thể cơ bản được lưu trữ trong Amazon S3. Nó bao gồm dữ liệu đối tượng (object data) và siêu dữ liệu (metadata). Siêu dữ liệu là một tập hợp các cặp name-value mô tả đối tượng. Các cặp này bao gồm một số siêu dữ liệu mặc định, chẳng hạn như ngày sửa đổi lần cuối và siêu dữ liệu HTTP tiêu chuẩn, ... Một object được xác định duy nhất trong một bucket bằng key (name) và version ID.

##### Key

Key(hay key name) là mã định danh duy nhất cho một object trong bucket. Mỗi object trong một bucket có chính xác một key. Mỗi object trong Amazon S3 có thể được xử lý duy nhất thông qua sự kết hợp của điểm cuối web service, bucket name, key và một version

##### Ưu Điểm Của Amazon S3

**Lưu trữ dữ liệu trong Buckets**: Bucket là khái niệm cơ bản trong Amazon S3 dành cho việc lưu trữ dữ liệu. Có thể chứa vô hạn các loại dữ liệu khác nhau trong một bucket. Mỗi dữ liệu tương tự như một object và mỗi object có thể chứa tới 5TB dữ liệu.

**Tải dữ liệu:** Bên cạnh đó, Amazon S3 còn cho phép tải dữ liệu về bất cứ lúc nào. Thậm chí người khác cũng có thể tải dữ liệu của trong Bucket Amazon S3. Tuy nhiên, nếu không muốn người khác tải về thì có thể phân quyền cho phép hoặc từ chối quyền upload/download dữ liệu của bạn.

**Giao diện tương tác tiêu chuẩn:** Có thể sử dụng REST hoặc SOAP để thiết kế tương tác từ ứng dụng của đến Amazon S3 qua các công cụ lập trình phát triển. Ngoài ra, Amazon S3 cũng được thiết kế các tính năng tối giản nhằm đạt mục tiêu đơn giản và nhanh gọn cho người dùng.

**Tính ổn định cao:** Amazon đảm bảo duy trì tính sẵn sàng cho hệ thống ở mức 99,99%. Chịu được các hỏng hóc và phục hồi hệ thống cực nhanh với thời gian tối thiểu. Dữ liệu của sẽ được bảo vệ khỏi các vấn đề về mạng và nguồn điện cũng như chống lại lỗi phần cứng.

A diagram of a cloud computing network

Description automatically generated

1. AWS S3

**Chi phí rẻ:** Amazon S3 có giá cực rẻ, rất cạnh tranh với các giải pháp của công ty khác trên thị trường. Từ đó các cá nhân hay doanh nghiệp đều có thể sử dụng dịch vụ lưu trữ này với ngân sách tiết kiệm nhất.

**Bảo mật:** Dữ liệu được lưu trữ trong Amazon được bảo mật theo mặc định vì chỉ chủ sở hữu và nhóm mới có quyền truy cập vào tài nguyên Amazon S3. Nó cung cấp mã hóa cho cả lưu trữ an toàn và bảo mật. Trung tâm dữ liệu của Amazon S3 có thể được bảo vệ nhiều hơn chống lại những thứ như hỏa hoạn, bão, trộm cắp…

**Chuyển dữ liệu đơn giản:** Amazon S3 cung cấp cho nhiều tùy chọn cho di chuyển dữ liệu đám mây. Nó làm cho chi phí có hiệu quả và rất đơn giản để chuyển một lượng lớn dữ liệu đến Amazon S3 hoặc ra khỏi Amazon S3.

##### Tại sao có thể dùng AWS S3 để thay thế cho hệ thống lưu trữa Apache Hadoop HDFS

|  |  |  |
| --- | --- | --- |
|  | Amazon S3 | HDFS (Hệ thống tệp phân tán Hadoop) |
| Hỗ trợ kiểu dữ liệu | Lưu trữ đa năng cho nhiều loại dữ liệu khác nhau | Được thiết kế chủ yếu để lưu trữ dữ liệu lớn |
| Khả năng mở rộng | Có thể mở rộng cho nhu cầu lưu trữ. S3 cho phép tách cơ sở hạ tầng điện toán khỏi nhu cầu lưu trữ của mình. Có thể lưu trữ tất cả dữ liệu của mình trong S3 và không phải giám sát việc lưu trữ của mình. Có thể dễ dàng mở rộng quy mô và thêm các nút vào cụm của mình trên đám mây. S3 có khả năng mở rộng, tính linh hoạt và độ đàn hồi cao hơn. | HDFS dựa vào bộ nhớ cục bộ có quy mô theo chiều ngang. Nếu muốn tăng dung lượng lưu trữ, sẽ phải thêm ổ cứng lớn hơn vào các nút hiện có hoặc thêm nhiều máy hơn vào cụm. Điều này khả thi nhưng tốn kém và phức tạp hơn S3. |
| Hội nhập | Tích hợp liền mạch với các dịch vụ AWS | Được thiết kế riêng để sử dụng với hệ sinh thái Hadoop |
| Truy vấn và xử lý | Khả năng truy vấn hạn chế | Được thiết kế để xử lý dữ liệu song song |
| Trường hợp sử dụng | Lưu trữ đối tượng, sao lưu, lưu trữ đám mây, lưu trữ dữ liệu lớn. | Xử lý, phân tích và lưu trữ dữ liệu lớn |
| Giá | Khi nhu cầu lưu trữ và tính toán của công ty tăng lên, S3 sẽ trở nên tiết kiệm chi phí hơn rất nhiều nhờ khả năng tự động mở rộng quy mô. Hơn nữa, việc lựa chọn S3 sẽ tiết kiệm đáng kể chi phí lao động. S3 yêu cầu bảo trì ít hơn nhiều. Với S3, AWS tự động sao lưu dữ liệu và khách hàng chỉ trả tiền cho dữ liệu cần được lưu trữ thay vì bản sao của dữ liệu đó. | Từ quan điểm kiến ​​trúc, HDFS cũng tự động tạo 3 bản sao của mỗi khối dữ liệu. Điều này có nghĩa là chi phí lưu trữ chiếm gấp ba lần dung lượng lưu trữ cần thiết cho dữ liệu cơ sở. |
| Khả dụng | AWS S3 cung cấp cơ sở hạ tầng cực kỳ bền bỉ, việc lưu trữ dữ liệu lớn trong S3 có thời gian ngừng hoạt động ít hơn đáng kể. AWS cung cấp cho khả năng mở rộng kích thước cụm để giải quyết các vấn đề không đủ thông lượng. | HDFS có lợi thế đáng kể về hiệu suất đọc và ghi do vị trí dữ liệu. HDFS có thể cung cấp thông lượng đọc cao hơn nhiều lần so với S3, nhưng vấn đề này được giảm thiểu do S3 cho phép tách riêng dung lượng lưu trữ và khả năng tính toán. |
| Hoạt động | S3 cung cấp khả năng chịu lỗi liền mạch nhờ khả năng sao chép dữ liệu giữa các vùng. S3 cũng cung cấp nhiều cách mã hóa dữ liệu để đáp ứng nhu cầu bảo mật thường được các cơ quan quản lý trong các lĩnh vực cụ thể yêu cầu. | Hadoop cung cấp một số tùy chọn để bảo mật dữ liệu, bao gồm xác thực người dùng kerberized và các lớp cấp phép hệ thống tệp. |
| Sự tồn tại của dữ liệu | dữ liệu luôn ổn định trong S3, nó cung cấp khả năng duy trì dữ liệu. | Với HDFS, dữ liệu không tồn tại khi dừng phiên bản EC2 hoặc EMR. Tuy nhiên, có thể sử dụng khối lượng EBS tốn kém để duy trì dữ liệu trên EC2. |
| Độ bền | S3 cung cấp độ bền 99,999999999 phần trăm vật thể mỗi năm. Điều này có nghĩa là cứ 10.000.000 vật thể thì cứ 10.000 năm lại có một vật thể bị mất đi | Mô hình thống kê về Độ bền của dữ liệu HDFS cho thấy xác suất mất một khối dữ liệu (mặc định là 64 megabyte) trên 4.000 khối dữ liệu lớn cụm nút (tổng dung lượng lưu trữ 16 petabyte, 250.736.598 bản sao khối) là 0,00000057 (5,7 x 10^-7) trong 24 giờ tới và 0,00021 (2,1 x 10^-4) trong 365 ngày tiếp theo. Tuy nhiên, hầu hết các cụm chỉ chứa vài chục phiên bản nên khả năng mất dữ liệu có thể cao hơn nhiều. |

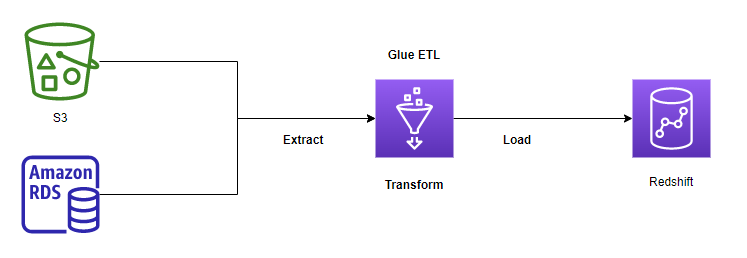
#### AWS Glue

##### Giới thiệu

Amazon Glue là dịch vụ tích hợp dữ liệu không có máy chủ giúp dễ dàng khám phá, chuẩn bị và kết hợp dữ liệu để phân tích, học máy và phát triển ứng dụng. Amazon Glue cung cấp tất cả các khả năng cần thiết cho việc tích hợp dữ liệu để bạn có thể bắt đầu phân tích dữ liệu của mình và đưa dữ liệu vào sử dụng trong vài phút thay vì hàng tháng.

Tích hợp dữ liệu là quá trình chuẩn bị và kết hợp dữ liệu để phân tích, học máy và phát triển ứng dụng. Nó liên quan đến nhiều nhiệm vụ, chẳng hạn như khám phá và trích xuất dữ liệu từ nhiều nguồn khác nhau làm phong phú, làm sạch, chuẩn hóa và kết hợp dữ liệu và tải và sắp xếp dữ liệu trong cơ sở dữ liệu, kho dữ liệu và hồ dữ liệu. Những tác vụ này thường được xử lý bởi nhiều loại người dùng khác nhau, mỗi loại sử dụng các sản phẩm khác nhau.

Amazon Glue cung cấp cả giao diện trực quan và giao diện dựa trên mã để giúp việc tích hợp dữ liệu trở nên dễ dàng hơn. Người dùng có thể dễ dàng tìm và truy cập dữ liệu bằng Amazon Glue Data Catalog. Các kỹ sư dữ liệu và nhà phát triển ETL (trích xuất, chuyển đổi và tải) có thể tạo và chạy quy trình công việc ETL. Các nhà phân tích dữ liệu và nhà khoa học dữ liệu có thể sử dụng Amazon Glue DataBrew để làm phong phú, làm sạch và chuẩn hóa dữ liệu một cách trực quan mà không cần viết mã.



1. AWS Glue 1

##### Những lợi ích và hạn chế của việc sử dụng Glue

**Tích hợp dữ liệu nhanh hơn:** Các nhóm khác nhau trong tổ chức của bạn có thể sử dụng Amazon Glue để cùng nhau thực hiện các nhiệm vụ tích hợp dữ liệu, bao gồm trích xuất, làm sạch, chuẩn hóa, kết hợp, tải và chạy các quy trình công việc ETL có thể mở rộng. Bằng cách này, bạn giảm thời gian phân tích dữ liệu và đưa dữ liệu vào sử dụng từ vài tháng xuống còn vài phút.

**Khả năng chịu lỗi:** Các công việc bị lỗi trong Glue đều có thể truy xuất được và các nhật ký trong Glue có thể được sửa lỗi

**Filtering:** Bộ lọc dữ liệu xấu.

**Support:** Hỗ trợ một số nguồn dữ liệu Java Database Connectivity (JDBC) không phải nguồn gốc.

**Bảo trì và triển khai:** Bảo trì và triển khai đơn giản vì dịch vụ được quản lý hoàn toàn bởi AWS.

**Tự động hóa việc tích hợp dữ liệu của bạn trên quy mô lớn:** Amazon Glue tự động hóa phần lớn nỗ lực cần thiết để tích hợp dữ liệu. Amazon Glue thu thập thông tin nguồn dữ liệu của bạn, xác định định dạng dữ liệu và đề xuất lược đồ để lưu trữ dữ liệu của bạn. Nó tự động tạo mã để chạy các quá trình chuyển đổi và tải dữ liệu của bạn. Bạn có thể sử dụng Amazon Glue để dễ dàng chạy và quản lý hàng nghìn công việc ETL hoặc kết hợp và sao chép dữ liệu trên nhiều kho dữ liệu bằng SQL.

**Không có máy chủ để quản lý:** Amazon Glue chạy trong môi trường không có máy chủ. Không có cơ sở hạ tầng để quản lý và Amazon Glue cung cấp, đặt cấu hình và thay đổi quy mô các tài nguyên cần thiết để chạy các công việc tích hợp dữ liệu của bạn. Bạn chỉ trả tiền cho những tài nguyên mà công việc của bạn sử dụng khi chạy.

##### Những hạn chế của AWS Glue

**Khả năng tương thích hạn chế:** Mặc dù AWS Glue hoạt động với nhiều nguồn dữ liệu thường được sử dụng nhưng nó chỉ hoạt động với các dịch vụ chạy trên AWS. Các tổ chức có thể cần dịch vụ ETL của bên thứ ba nếu các nguồn không dựa trên AWS.

**Không đồng bộ hóa dữ liệu gia tăng:** Tất cả dữ liệu đều được sắp xếp trên S3 trước tiên, vì vậy Glue không phải là lựa chọn tốt nhất cho các công việc ETL theo thời gian thực.

**Learning curv:** Các nhóm sử dụng Glue phải có hiểu biết sâu sắc về Apache spark.

**Truy vấn cơ sở dữ liệu quan hệ:** Glue hỗ trợ hạn chế cho các truy vấn của cơ sở dữ liệu quan hệ truyền thống, chỉ hỗ trợ các SQL queries.

##### Các hoạt động của AWS Glue

Glue sử dụng các ETL jobs để trích xuất dữ liệu từ sự kết hợp của các dịch vụ đám mây khác do Amazon Web Services (AWS) cung cấp và kết hợp dữ liệu đó vào hồ dữ liệu và kho dữ liệu. Nó sử dụng các giao diện lập trình ứng dụng (API) để chuyển đổi tập dữ liệu được trích xuất để tích hợp và giúp người dùng giám sát công việc.

Người dùng có thể đặt các ETL jobs theo lịch trình hoặc chọn các sự kiện sẽ kích hoạt công việc. Sau khi được kích hoạt, Glue sẽ trích xuất dữ liệu, chuyển đổi dữ liệu dựa trên mã mà Glue tự động tạo và tải dữ liệu đó vào Amazon S3 hoặc Amazon Redshift. Sau đó, Glue ghi siêu dữ liệu từ công việc vào AWS Glue Data Catalog được nhúng.

Dịch vụ có thể tự động tìm thấy dữ liệu có cấu trúc hoặc không cấu trúc của doanh nghiệp khi dữ liệu đó được lưu trữ trong data lake trên S3, kho dữ liệu trong Amazon Redshift và các cơ sở dữ liệu khác là một phần của Amazon Relational Database Service. Glue cũng hỗ trợ cơ sở dữ liệu MySQL, Oracle, Microsoft SQL Server và PostgreSQL chạy trên các phiên bản Amazon Elastic Computing Cloud (EC2) trong Amazon Virtual Private Cloud.

Sau đó, dịch vụ sẽ lập profiles data trong data catalog, đây là kho lưu trữ siêu dữ liệu cho tất cả nội dung dữ liệu chứa các chi tiết như định nghĩa bảng, vị trí và các thuộc tính khác. Một nhóm cũng có thể sử dụng Glue Data Catalog thay thế cho Apache Hive Metastore cho các ứng dụng Amazon Elastic MapReduce.

##### Tính năng của AWS Glue



1. AWS Glue 2

**Automatic schema discovery:** Glue cho phép các nhà phát triển tự động hóa trình thu thập thông tin để lấy thông tin liên quan đến lược đồ và lưu trữ nó trong danh mục dữ liệu, sau đó có thể được sử dụng để quản lý công việc.

**Job schedule:** Glue jobs có thể được đặt và gọi theo lịch trình linh hoạt, bằng trình kích hoạt dựa trên sự kiện hoặc theo yêu cầu. Một số công việc có thể được bắt đầu song song và người dùng có thể chỉ định sự phụ thuộc giữa các công việc.

**Developer endpoints:** Các nhà phát triển có thể sử dụng những thứ này để gỡ lỗi Glue, cũng như tạo các trình đọc, trình ghi và các phép biến đổi tùy chỉnh, sau đó có thể được nhập vào thư viện tùy chỉnh.

**Automatic code generation:** Quá trình ETL tự động tạo mã và đầu vào duy nhất cần thiết là vị trí/đường dẫn để dữ liệu được lưu trữ. Mã này bằng Scala hoặc Python.

**Integrated data catalog:** Hoạt động như một kho lưu trữ siêu dữ liệu duy nhất từ một nguồn khác nhau trong quy trình AWS. Một tài khoản AWS có một danh mục

##### Trường hợp sử dụng

Xây dựng quy trình ETL (trích xuất, chuyển đổi và tải) theo hướng sự kiện

Tạo một danh mục thống nhất để tìm dữ liệu trên nhiều kho dữ liệu

Khám phá dữ liệu với việc chuẩn bị dữ liệu trực quan tự phục vụ

#### AWS Redshift

##### Giới thiệu

Amazon Redshift là kho dữ liệu đám mây được sử dụng rộng rãi nhất.

Nó giúp phân tích tất cả dữ liệu nhanh chóng, đơn giản và tiết kiệm chi phí bằng cách sử dụng SQL tiêu chuẩn và các công cụ Business Intelligence (BI) hiện có.

Nó cho phép chạy các truy vấn phân tích phức tạp từ hàng terabyte (TB) đến hàng petabyte (PB, 1 PB dữ liệu có cấu trúc và bán cấu trúc, sử dụng tối ưu hóa truy vấn phức tạp, lưu trữ dạng cột trên bộ nhớ hiệu suất cao và thực thi truy vấn song song hàng loạt.

##### Kiến trúc hệ thống tổng quát

A diagram of data storage

Description automatically generated

1. Redshift

###### AWS Redshift Spectrum

Amazon Redshift Spectrum là một tính năng của Amazon Redshift cho phép chạy các truy vấn đối với hồ dữ liệu của mình trên Amazon S3 nhưng không cần tải dữ liệu hoặc Extract, transform, and load (ETL). Khi phát hành một truy vấn SQL, truy vấn đó chuyển tới điểm cuối Amazon Redshift để tạo và tối ưu hóa một kế hoạch truy vấn. Amazon Redshift xác định đâu là dữ liệu cục bộ và đâu là dữ liệu trên Amazon S3, tạo một kế hoạch để giảm thiểu lượng dữ liệu trên S3 phải đọc, đồng thời yêu cầu các công nhân Amazon Redshift Spectrum ngoài phạm vi của một nhóm tài nguyên chia sẻ đọc và xử lý dữ liệu từ Amazon S3.

###### Client Application và Connection

Amazon Redshift được xây dựng dựa trên PostgreSQL với chỉ một chút thay đổi. Do đó nếu ứng dụng của đang sử dụng SQL, chỉ cần vài thay đổi nhỏ thì đã có thể chuyển sang Redshift.

Vì được xây dựng dựa trên PostgreSQL nên Redshift hỗ trợ việc kết nối giữa application và hệ thống qua JDBC and ODBC drivers.

Cluster

Bây giờ chúng ta sẽ đi sâu hơn vào cấu trúc của Redshift. Thành phần core trong kiến trúc của Redshift là Cluster. Trong Cluster sẽ có một hoặc nhiều Database. Để truy cập vào Cluster sẽ cần đường link của cluster, tên database và password. Mỗi cluster sẽ được tạo thành bởi 1 hay nhiều Node. Khi Cluster được tạo nên bởi nhiều Node thì Redshift sinh thêm Leader Node.

Leader Node

Nhiệm vụ đầu tiên của Leader Node là kết nối với application để nhận query cũng như trả kết quả. Nhiệm vụ thứ 2 là xử lý, truyền query excution plan tới từng Compute Node. Đồi với những câu lệnh query phức tạp thì Leader Node còn truyền xuống các Compute Node các bước query và chính Leader Node tổng hợp kết quả trả về từ Compute Node để có thể ra được kết quả. Như vậy bước xử lý câu query và tổng hợp kết quả được thực hiện quả Leader Node còn thực hiện nó thì sẽ là ở các Node con

Compute Node

Compute Node chính là nơi thực hiện các câu lệnh query sau đó trả kết quả lại cho Leader Node tổng hợp lại. Điểm đặc biệt ở đây chính là mỗi Compute Node sẽ có CPU, Memory và Storage cho riêng mình. Cấu hình cụ thể sẽ tùy thuộc vào Plan mà chọn khi tạo Cluster. Ví dụ nếu dữ liệu của cực kì khủng khiếp thì có thể chọn cho mình max là 36 vCPU, 244Gb Memory và 16TB Storage.

Node Slices

Mỗi một Compute Node tiếp tục được chia nhỏ ra thành các Node Slice. Mỗi một Node Slice sẽ được phân chia đều CPU, Memory và Storage từ Compute Node đó. Với 1 query mà Compute Node nhận được từ Leader Node, Compute Node tiếp tục truyền xuống cho từng Node Slice để các Slice này thực hiện đồng thời query.

Internal network

Thành phần cuối cùng là Internal network. Redshift sử dụng một hệ thống private network với băng thông rộng, tốc độ cao để có thể kết nối giữa Leader Node và Compute Node. Đây là thành phần bảo đảm truyền query từ Leader Node tới Compute Node và truyền kết quả từ Compute Node tới Leader Node. Đây cũng là một thành phần rất quan trọng ảnh hưởng đến thời gian query mà mình sẽ giới thiệu đến các sau này.

###### Lưu trữ dữ liệu

Distribution Style

Tiếp theo sẽ nói một chút về việc lưu trữ dữ liệu của Redshift. Như ở trên đã nói đến cấu trúc nhỏ nhất trong Redshift là các Node Slices và data của được lưu trữ trên node slices này. Vậy thì Redshift phân bổ data đến các Node Slices như thế nào.

Khi tạo bảng thì sẽ có trường nên khai báo là Distribution Key. Distribution Key là một column trong bảng của bạn, và Redshift sẽ sử dụng giá trị của column này để phân chia data đến node slices tương ứng. Hiểu nôm na ví dụ chọn distribution key là cột date. Thì những bản ghi có data là ngày 2016-04-24 sẽ ở trên cùng một node slice, hoặc trên cùng một node nếu như số lượng bản ghi có date là ngày 2016-04-24 quá nhiều... Việc phân bố cũng như lưu trữ xem node nào, nodes slice nào lưu trữ giá trị nào được quyết định và lưu trữ bởi Leader Node.

Ví dụ khi query là date BETWEEN 2016-04-24 and 2016-04-25, leader node thấy với những date là 2016-04-24 và 2016-04-25 thì chỉ được lưu ở node 1 và node 2. Khi đó thì query sẽ chỉ được 2 node này thực hiện. Qua đó tránh việc phải xử lý lượng dữ liệu không cần thiết. Do đó việc chọn Distribution Key là cực kì quan trọng khi thiết kế table.

Columnar Data Storage

Columnar Data Storage hiểu đơn giản là lưu trữ dữ liệu theo cột. Đây là một trong những yếu tố cực kì quan trọng giúp cho việc giảm đi một lượng cực kì lớn I/O requirements và lượng dữ liệu phải load trong quá trình query.

Cấu trúc lưu data kiểu truyền thống là lưu theo row (row-wise database storage) như thấy dưới đây.

A screenshot of a computer

Description automatically generated

1. Columnar Data Storage Block

Mỗi một block sẽ lưu data của cả 1 row. Nghĩa là nếu mà row size lớn hơn block size thì sẽ cần nhiều hơn 1 block để lưu data, còn nếu như row size nhở hơn block size thì trên 1 block sẽ lưu data của ít nhât 2 row. Kiểu lưu này phù hợp với những application là online transaction processing (OLTP). Ở những application kiểu này thì việc đọc và ghi liên quan đến toàn bộ dữ liệu hoăc chỉ là 1 lượng nhỏ dữ liệu. Do đó việc phải đọc và ghi qua bao nhiêu block không cần phải tối ưu vì hoặc là xử lý trên toàn bộ block, hoặc chỉ cần xử lý trên 1 lượng block nhỏ.

Redshift sử dụng Columnar Data Storage, lưu theo cột để lưu dữ liệu như ban thấy dưới đây.

A screenshot of a computer

Description automatically generated

1. Columnar Data Storage

Như thấy, thì trên 1 block thay vì lưu cả 1 row thì sẽ lưu theo cột. Tất nhiên trong bài viết này mình sẽ không đi sâu vào so sánh giữa Columnar Storage và Row-Wise Storage (chắc là sẽ có một bài riêng cho phần này (hihi)). Ở đây mình chỉ nói về điểm mạnh của Columnar Data Storage là:

* Khả năng query cực kì mạnh mẽ khi phải query lượng dữ liệu lớn.
* Khả năng nén dữ liệu rất cao, qua đó giảm được lượng dữ liệu phải xử lý.
* Tốc độc query rất tốt nếu như chỉ liên quan đến một hay một vài cột trong bảng.
* Column được lưu trên block theo thứ tự nào thì tùy thuộc vào Sort Key mà khai báo khi tạo bảng.

##### Perfomance

Quay trở lại với vấn đề chính là tại sao Redshift có thể có tốc độ query cực kì tốt với lượng dữ liệu lớn. Thì có lẽ qua phần giới thiệu về system của Redshift các cũng có thể đoán ra được phần nào. Dưới đây là những lý do chính.

###### Massively parallel processing

Với cấu trúc chia nhở data vào các compute node, các compute node lại được chia nhỏ thành các node slices. Và các node slices lại được cấp CPU, Memory, Storage để xử lý dữ liệu song song thì hiển nhiên là tốc độ query sẽ nhanh hơn rất nhiều. Ví dụ như chọn cluster là ds1.xlarge với max node có thể là 32, mỗi node có max là 2 slices, thì với 64 process xử lý query song song hiển nhiên là nhanh hơn so với 1 process. Con số này sẽ còn khủng khiếp hơn khi sử dụng cluster dc1.8xlarge với tối đa 128 nodes và max là 32 slices trên 1 node (tất nhiên là nếu chọn con này thì tốn lượng tiền lớn rồi (hihi)).

###### Columnar data storage

Như đã nói ở trên thì lưu dữ liệu theo cột sẽ tốn ít I/O và lượng dữ liệu phải load hơn. Columnar data storage cực kì hữu dụng nếu data của cực kì lớn. Đây cũng là một trong những lý do chính giúp Redshift có tốc độ query rất nhanh khi so sánh vs MongoDB ở bài trước của mình.

###### Data Compression

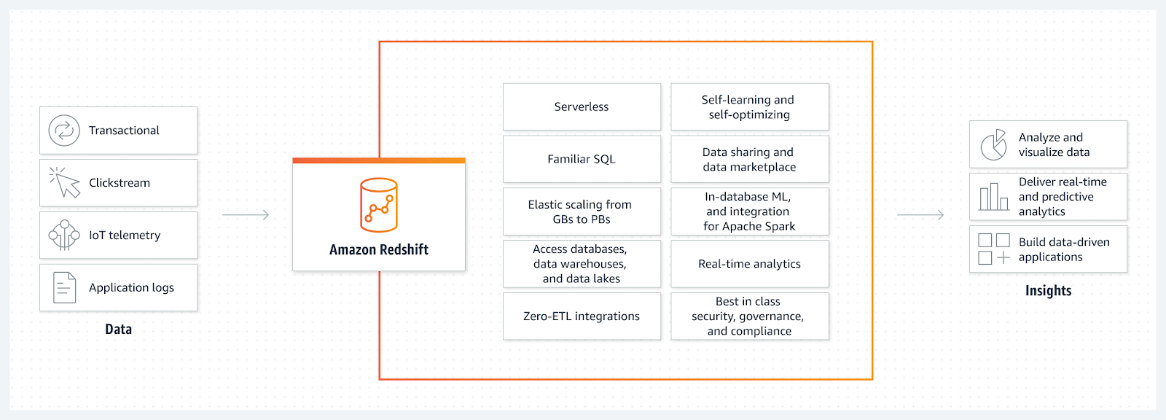
Với kiểu lưu theo Columnar Data Storage thì Redshift cũng gây ấn tượng mạnh với khả năng nén dữ liệu. Với table có 8.5 triệu bản ghi như của mình, nếu như ở MongoDB lượng dữ liệu chiếm là 4.5Gb thì ở Redshift chỉ gần 1Gb, điều đó giúp cho lượng dữ liệu phải xử lý giảm đi rất rất nhiều qua đó tăng cao được performance. Redshift cung cấp rất nhiểu kiểu encode cho dữ liệu của như raw, byte, delta, lzo... để lựa chọn, qua đó nén một cách tối ưu nhất dữ liệu của mình.

###### Query Optimizer

Phân Engine xử lý query của Redshift có khả năng tối ưu hóa query với việc sử dụng (MPP-aware (Massive Parallel Processing)) và những lợi thế có được từ Columnar data storage. Khi nhận được query thì Engine sẽ chuyển query plan thành code và sau đó chuyển code đó đến Compute Node để thực hiện.

Ngoài ra Redshift còn cung cấp cho 1 loạt các bảng có sắn để có thể phân tích flow của query qua đó giúp cải thiện câu query. Ví dụ như bảng SVL\_QUERY\_SUMMARY cung cấp cho thống kê query theo stream, bảng SVL\_QUERY\_REPORT cung cấp cho thông tin query theo Node Slice, hay STL\_ALERT\_EVENT\_LOG lưu trữ những thông báo mà nó ảnh hưởng đến performance của query...

##### Cách thức hoạt động



1. Amazon Redshift

Amazon Redshift sử dụng SQL để phân tích dữ liệu có cấu trúc và bán cấu trúc trên các kho dữ liệu, cơ sở dữ liệu hoạt động và hồ dữ liệu bằng cách sử dụng phần cứng và công nghệ máy học được AWS thiết kế để mang đến hiệu suất với mức giá tốt nhất ở bất kỳ quy mô nào.

### Big Data Processing và Enriching

##### SQL based ELT (với Redshift Spectrum)

Để chuyển đổi dữ liệu có cấu trúc trong storage layer Lakehouse, bạn có thể xây dựng các pipeline ELT mạnh mẽ bằng cách sử dụng ngữ nghĩa SQL quen thuộc. Các pipeline ELT này có thể sử dụng khả năng xử lý song song hàng loạt (MPP) trong Amazon Redshift và khả năng trong Redshift Spectrum để tạo ra hàng nghìn node tạm thời để mở rộng quy mô xử lý đến petabyte dữ liệu. Các pipeline ELT dựa trên quy trình được lưu trữ tương tự trên Amazon Redshift có thể biến đổi những điều sau:

* Flat structured data được AWS DMS hoặc Amazon AppFlow phân phối trực tiếp vào các staging tables của Amazon Redshift
* Dữ liệu được lưu trữ trong data lake sử dụng các định dạng tệp nguồn mở như CSV, JSON, Avro, Parquet và ORC

Đối với các bước data enrichment, các pipeline này có thể bao gồm các câu lệnh SQL nối các internal dimension tables với các large fact tables được lưu trữ trong data lake S3 (sử dụng lớp Redshift Spectrum). Bước cuối cùng, các pipeline xử lý dữ liệu có thể chèn curated, enriched, và modeled data vào internal table của Amazon Redshift hoặc external table được lưu trữ trong Amazon S3.

### Data Consumption Layer (Visualazation)

#### AWS QuickSight

##### Giới thiệu

Amazon QuickSight là dịch vụ thông tin kinh doanh (BI) quy mô đám mây mà bạn có thể sử dụng để cung cấp thông tin chi tiết dễ hiểu cho những người làm việc cùng bạn, bất kể họ ở đâu. Amazon QuickSight kết nối với dữ liệu của bạn trên đám mây và kết hợp dữ liệu từ nhiều nguồn khác nhau. Trong một bảng thông tin dữ liệu duy nhất, QuickSight có thể bao gồm dữ liệu AWS, dữ liệu của bên thứ ba, dữ liệu lớn, dữ liệu bảng tính, dữ liệu SaaS, dữ liệu B2B, v.v. Là một dịch vụ dựa trên đám mây được quản lý hoàn toàn, Amazon QuickSight cung cấp khả năng bảo mật cấp doanh nghiệp, tính khả dụng toàn cầu và tính năng dự phòng tích hợp. Nó cũng cung cấp các công cụ quản lý người dùng mà bạn cần để mở rộng quy mô từ 10 người dùng lên 10.000, tất cả đều không cần cơ sở hạ tầng để triển khai hoặc quản lý.

QuickSight mang lại cho người ra quyết định cơ hội khám phá và giải thích thông tin trong môi trường trực quan tương tác. Họ có quyền truy cập an toàn vào trang tổng quan từ bất kỳ thiết bị nào trên mạng của bạn và từ thiết bị di động.

Để tìm hiểu thêm về các thành phần và quy trình chính của Amazon QuickSight cũng như quy trình làm việc điển hình để tạo trực quan hóa dữ liệu, hãy xem các phần sau. Hãy bắt đầu ngay hôm nay để khai thác tiềm năng dữ liệu của bạn và đưa ra quyết định tốt nhất có thể.

Mỗi ngày, mọi người trong tổ chức của bạn đưa ra quyết định ảnh hưởng đến doanh nghiệp của bạn. Khi họ có thông tin chính xác vào đúng thời điểm, họ có thể đưa ra lựa chọn Đưa công ty của bạn đi đúng hướng.

Dưới đây là một số lợi ích của việc sử dụng Amazon QuickSight để phân tích, trực quan hóa dữ liệu và Báo cáo:

* Công cụ trong bộ nhớ, được gọi là SPICE, đáp ứng với tốc độ cực nhanh.
* Không có chi phí trả trước cho giấy phép và tổng chi phí sở hữu (TCO) thấp.
* Phân tích cộng tác mà không cần cài đặt ứng dụng.
* Kết hợp nhiều loại dữ liệu thành một phân tích.
* Xuất bản và chia sẻ phân tích của bạn dưới dạng bảng điều khiển.
* Kiểm soát các tính năng có sẵn trong bảng thông tin.
* Không cần quản lý các quyền cơ sở dữ liệu chi tiết—người xem bảng thông tin có thể Chỉ xem những gì bạn chia sẻ.

Đối với người dùng nâng cao, phiên bản QuickSight Enterprise cung cấp nhiều tính năng hơn nữa:

* Tiết kiệm thời gian và tiền bạc cho bạn với thông tin chi tiết về dữ liệu tự động và có thể tùy chỉnh, được hỗ trợ bởi máy học (ML). Điều này cho phép tổ chức của bạn thực hiện Sau đây, mà không yêu cầu bất kỳ kiến thức nào về học máy:
* Tự động đưa ra dự báo đáng tin cậy.
* Tự động xác định các ngoại lệ.
* Tìm xu hướng ẩn.
* Hành động dựa trên các động lực kinh doanh chính.
* Dịch dữ liệu thành các câu chuyện dễ đọc, như lát xếp dòng tiêu đề cho bảng điều khiển của bạn.

Cung cấp các tính năng bảo mật Enterprise bổ sung, bao gồm:

* Người dùng, nhóm và đăng nhập một lần có liên kết (Trung tâm nhận dạng IAM) với AWS Identity and Access Management (IAM) Liên kết, SAML, OpenID Connect hoặc AWS Directory Service dành cho Microsoft Active Directory.
* Quyền chi tiết để truy cập dữ liệu AWS.
* Bảo mật cấp hàng.
* Mã hóa dữ liệu bảo mật cao trên thiết bị lưu trữ.
* Truy cập vào dữ liệu AWS và dữ liệu tại chỗ trong Amazon Virtual Private Cloud
* Cung cấp giá trả cho mỗi phiên cho người dùng mà bạn đặt trong "trình đọc" Vai trò bảo mật—Người đọc là Bảng điều khiển Người đăng ký, những người xem báo cáo nhưng không tạo báo cáo.
* Cho phép bạn biến QuickSight thành một phần của các trang web và ứng dụng của riêng bạn bằng cách Triển khai các phiên bảng điều khiển và bảng điều khiển nhúng.
* Biến doanh nghiệp của chúng tôi thành doanh nghiệp của bạn với các tính năng đa nhiệm cho giá trị gia tăng đại lý (VAR) của các dịch vụ phân tích.
* Cho phép bạn lập trình các mẫu bảng điều khiển kịch bản có thể được chuyển sang các tài khoản AWS khác.
* Đơn giản hóa việc quản lý và tổ chức truy cập với các thư mục chia sẻ và cá nhân cho tài sản phân tích.
* Cho phép hạn ngạch nhập dữ liệu lớn hơn để nhập dữ liệu SPICE và thường xuyên hơn Làm mới dữ liệu theo lịch trình.

##### Kiến trúc

**Data Sources:** QuickSight có thể kết nối và trực quan hóa dữ liệu từ nhiều nguồn khác nhau như Amazon Redshift, Amazon RDS, Amazon Athena, Amazon S3, và nhiều nguồn dữ liệu khác.

**SPICE Engine (Super-fast, Parallel, In-memory Calculation Engine):** Đây là một thành phần quan trọng của QuickSight, giúp tăng cường hiệu suất bằng cách lưu trữ dữ liệu tạm thời (cache) và thực hiện các phép toán tính toán trên dữ liệu này.

**Data Preparation:** QuickSight cung cấp các công cụ để chuẩn bị và biến đổi dữ liệu trước khi trực quan hóa, giúp người dùng hiểu và làm sạch dữ liệu một cách dễ dàng.

**Analysis Workspace:** Đây là nơi người dùng có thể tạo, tùy chỉnh và chia sẻ bảng điều khiển (dashboards) và báo cáo. Nó cung cấp các công cụ trực quan hóa mạnh mẽ và linh hoạt.

**Access Control and Security:** QuickSight tích hợp với IAM để quản lý quyền truy cập và đảm bảo tính an toàn của dữ liệu.

**Integration with Other AWS Services:** QuickSight có thể tích hợp với nhiều dịch vụ AWS khác như Amazon S3, Athena, Redshift, và nhiều dịch vụ lưu trữ và xử lý dữ liệu khác



1. QuickSight

##### Tính năng nổi bật của QuickSight

###### Trực quan hóa dữ liệu nhanh chóng

Bạn có thể dễ dàng nhanh chóng tạo và truy cập bảng điều khiển (Dashboard) và báo cáo. Truy cập bảng điều khiển từ trình duyệt hoặc thiết bị di động một cách an toàn.

Có thể hiển thị dữ liệu bằng biểu đồ, … và có thế gửi báo cáo phân tích bằng email.

###### Hỗ trợ nhiều nguồn dữ liệu

Kết nối với nhiều nguồn dữ liệu khác nhau, bao gồm cơ sở dữ liệu, dịch vụ đám mây, và các tệp dữ liệu:

* Ứng dụng SaaS: Salesforce, Square, ServiceNow, Twitter, Github, JIRA
* Cơ sở dữ liệu của bên thứ ba: Teradata, MySQL, PostgreSQL, SQL Server
* Dịch vụ AWS: Redshift, Athena, S3, RDS, Aurora
* Subnet VPC cá nhân
* Các loại file khác nhau như Excel, CSV, JSON, Presto

###### Phí tính theo phiên sử dụng

QuickSight có ba loại quyền là Quản trị viên (Admin), Tác giả (Author) và Người đọc (Reader). Phí sẽ được tính theo số phiên mà Reader truy cập vào Dashboard do Author tạo, giúp tiết kiệm chi phí, và không phát sinh phí đầu kì.

Lưu ý: chỉ có Enterprise Edition mới có khả năng sử dụng quyền Người đọc (Reader).

###### Mở rộng dễ dàng

Có thể mở rộng phạm vi sử dụng từ hàng chục người đến hàng chục nghìn người. Dịch vụ không cần máy chủ, không cần quản lý cơ sở hạ tầng, bất kể sử dụng bao nhiêu dữ liệu. Ngay cả khi chưa có kế hoạch cụ thể, bạn vẫn có thể thử dịch vụ một cách dễ dàng. Thêm vào đó, với SPICE của DB in-memory có khả năng truy cập tốc độ cao giúp giúp tăng tốc độ truy vấn dữ liệu và hỗ trợ dữ liệu lớn. SPICE có các tính năng như tốc độ cao, không cần quản lý, tính khả dụng cao (Multi – AZ dự phòng, tự động lưu vào S3), và bằng cách nhập dữ liệu vào SPICE, có thể phân tích tốc độ cao mà không làm nặng nguồn dữ liệu. Hỗ trợ lên đến 250 triệu hàng hoặc 500GB cho một bộ dữ liệu (dataset).

###### Tính năng học máy ML

Bạn có thể tận dụng các tính năng học máy (ML) đã được chứng nhận của AWS để thực hiện phân tích. Nếu sử dụng ML Insight, có thể phát hiện bất thường, tiên đoán, tường thuật tự động, hiển thị liên kết với các mô hình được tạo bằng SageMaker.

###### Tích hợp vào ứng dụng

Có thể nhúng vào trang web công ty nội bộ hoặc SaaS bằng liên kết API.

Không cần viết mã để nhúng, và không phát sinh phí cho các bảng điều khiển nhúng chưa sử dụng.

###### Giải pháp toàn cầu

Thiết kế và phát triển đã dựa trên việc trở thành sản phẩm đối ứng toàn cầu từ giai đoạn phát triển đầu tiên. Đã được dịch sang 10 ngôn ngữ chính bao gồm tiếng Anh, Đức, Tây Ban Nha, Pháp, Ý, Bồ Đào Nha, Nhật Bản, Hàn Quốc, Trung Quốc (phồn thể), Trung Quốc (giản thể).

##### Bảo mật trong Amazon QuickSight

Dịch vụ Amazon QuickSight có sẵn ở nhiều khu vực nhằm đảm bảo tính sẵn sàng và dự phòng. Ngoài ra, QuickSight cung cấp nhiều tính năng bảo mật mạnh mẽ như sau:

* Kiểm soát truy cập dựa trên vai trò: QuickSight cho phép bạn xác định người dùng và vai trò của họ, từ đó kiểm soát quyền truy cập vào tài liệu và dữ liệu.
* Tích hợp với Active Directory: Bạn có thể tích hợp QuickSight với hệ thống Active Directory của tổ chức để quản lý danh tính và quyền truy cập một cách trung tâm.
* Kiểm duyệt thông qua CloudTrail: QuickSight tích hợp với AWS CloudTrail để ghi lại và kiểm tra các hoạt động trong dịch vụ, giúp bạn theo dõi và xác minh tương tác của người dùng.
* Hỗ trợ đăng nhập đơn (IAM, bên thứ ba): QuickSight hỗ trợ đăng nhập thông qua AWS Identity and Access Management (IAM) cũng như các dịch vụ đăng nhập bên thứ ba khác.
* Sử dụng subnet VPC cá nhân: QuickSight cho phép bạn sử dụng các subnet trong môi trường Amazon Virtual Private Cloud (VPC) để tăng cường bảo mật.
* Sao lưu dữ liệu: Dịch vụ này cung cấp các tính năng sao lưu dữ liệu, giúp bảo vệ và khôi phục dữ liệu quan trọng.

Ngoài ra, QuickSight tuân thủ nhiều tiêu chuẩn bảo mật quan trọng như FedRamp, HIPAA, PCI DSS, ISO, SOC, giúp bạn đáp ứng các yêu cầu bảo mật và tuân thủ ngành trong lĩnh vực của bạn.

# Kết luận

Tài liệu tham khảo:

LakeHouse:

<https://meeyland.com/tin-tuc/data-lakehouse-la-gi-nhung-uu-diem-cua-data-lakehouse-378178688>

AWS Architecture:

<https://vticloud.io/aws-big-data-introduction/>

<https://aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws/>

AWS Data Sync:

<https://aws.amazon.com/vi/datasync/>

AWS AppFlow:

<https://aws.amazon.com/appflow/>

AWS S3:  
<https://aws.amazon.com/s3/>

<https://cloudfly.vn/blog/amazon-s3-la-gi-tat-tan-tat-dieu-can-biet-ve-amazon-s3>  
<https://viblo.asia/p/bat-dau-voi-amazon-simple-storage-service-157G5n85vAje>

AWS Glue:

<https://vticloud.io/aws-big-data-introduction/>

<https://viblo.asia/p/aws-redshift-E7bGoxYKv5e2>

[https://aws.amazon.com/vi/redshift/#](https://aws.amazon.com/vi/redshift/)

<https://kungfutech.edu.vn/bai-viet/aws/glue>

<https://www.amazonaws.cn/en/glue/>

<https://www.techtarget.com/searchaws/definition/AWS-Glue>

AWS Redshift

<https://viblo.asia/p/aws-redshift-E7bGoxYKv5e2>

[https://aws.amazon.com/vi/redshift/#](https://aws.amazon.com/vi/redshift/)

AWS QuickSight

<https://docs.aws.amazon.com/pdfs/athena/latest/ug/athena-ug.pdf>

<https://docs.aws.amazon.com/quicksight/latest/user/welcome.html>

<https://docs.aws.amazon.com/pdfs/quicksight/latest/user/amazon-quicksight-user.pdf>

<https://docs.aws.amazon.com/athena/latest/ug/what-is.html>