PROJECT DATA ENGINEERING – PROCESSING AND ANALYZING TWITTER DATASET

**Name: Doan Nguyen Tuan Anh**

**StudentID: 103848025**

**PhoneNumber: 0373818936**

**Email:** [**anhwin01@gmail.com**](mailto:anhwin01@gmail.com)

**Contact for work:** [**dntanh2001@fpt.vn**](mailto:dntanh2001@fpt.vn)

**📄 Executive Summary**

**Problem Statement**

Trong thời đại số hóa hiện nay, mạng xã hội như Twitter không chỉ là nơi tương tác cá nhân mà còn là nguồn dữ liệu phong phú phản ánh xu hướng, cảm xúc và các chủ đề nóng theo thời gian thực. Tuy nhiên, việc thu thập, lưu trữ và phân tích dữ liệu Twitter đòi hỏi một hạ tầng kỹ thuật mạnh mẽ, có khả năng mở rộng và tự động hóa để chuyển hóa dòng dữ liệu phi cấu trúc này thành tri thức có giá trị.

**Solution Overview**

Dự án “AWS-based Twitter Data Engineering Pipeline” là một mô hình cá nhân được xây dựng như một nền tảng chuyên nghiệp để xử lý dữ liệu xã hội. Dự án tích hợp bộ công cụ chuyên sâu của AWS bao gồm:

* **Apache Spark on EMR**: Xử lý song song khối lượng lớn dữ liệu Twitter từ raw/ đến cleansed/ và processed/.
* **Amazon S3**: Phân tầng dữ liệu theo kiến trúc ELT hiện đại (raw, cleansed, processed).
* **AWS Athena**: Truy vấn dữ liệu bằng SQL trực tiếp trên S3.
* **Amazon Redshift**: Lưu trữ dữ liệu đã xử lý để phục vụ các báo cáo phân tích.
* **Apache Airflow**: Orchestration pipeline (đã tích hợp thành công trên môi trường EC2)
* **PowerBI/QuickSight:** Tích hợp dashboard bằng QuickSight để trực quan hoá dữ liệu.

Hệ thống được thiết kế mô-đun hóa, đảm bảo dễ dàng mở rộng cho doanh nghiệp trong tương lai. Mọi bước trong pipeline đều được tự động hóa, có thể tái sử dụng cho các nguồn dữ liệu xã hội khác.

**Business Benefits & ROI Summary**

* **Tăng hiệu suất phân tích dữ liệu**: giảm đáng kể thời gian xử lý từ hàng giờ xuống còn vài phút nhờ Spark.
* **Tiết kiệm chi phí lưu trữ & tính toán**: nhờ kiến trúc serverless (Athena) và tính đàn hồi của EMR.
* **Tự động hóa quy trình**: giảm sự phụ thuộc vào thao tác thủ công, tăng độ tin cậy và khả năng giám sát.
* **Mở rộng được cho doanh nghiệp**: pipeline đủ linh hoạt để tích hợp sentiment analysis, real-time Kafka streaming và dashboarding BI.

**Investment & Timeline**

Dự án được triển khai trong vòng 4 tuần với chi phí chủ yếu là chi phí thử nghiệm AWS (với free-tier) và sức lao động cá nhân. Trong mô hình doanh nghiệp, chi phí vận hành hàng tháng ước tính từ $100–$300 tùy theo lưu lượng và tần suất xử lý.

**Success Metrics & Expected Outcomes**

* ✅ Thu thập và xử lý >10,000 tweet JSON từ nguồn crawl
* ✅ Tạo ra pipeline tự động hóa 100% từ ingestion đến phân tích
* ✅ Truy vấn dữ liệu từ Athena và Redshift thành công
* ✅ Chuẩn bị nền tảng sẵn sàng tích hợp ML/AI sentiment model
* ✅ Thiết lập được cơ sở hạ tầng chuẩn DE chuyên nghiệp, có khả năng nhân rộng

**2. 🎯 Problem Statement**

**Phân tích tình huống hiện tại**

Trong bối cảnh hiện nay, mạng xã hội – đặc biệt là Twitter – trở thành một nguồn dữ liệu phong phú phản ánh xu hướng, cảm xúc, và hành vi người dùng theo thời gian thực. Tuy nhiên, phần lớn các tổ chức, doanh nghiệp vừa và nhỏ, hoặc cá nhân nghiên cứu chưa có hệ thống chuyên nghiệp để thu thập, lưu trữ và xử lý dữ liệu từ Twitter một cách có hệ thống, mở rộng và hiệu quả trên nền tảng đám mây.

Đa phần dữ liệu Twitter có định dạng phi cấu trúc (unstructured JSON), phân tán và không dễ dàng tích hợp vào quy trình phân tích hay kho dữ liệu truyền thống như Redshift. Việc thiếu một pipeline ELT chuyên nghiệp khiến quá trình khai thác insight từ dữ liệu bị chậm trễ, không đảm bảo tính đúng đắn, và khó mở rộng.

**Xác định pain points và tác động định lượng**

* ❌ Dữ liệu JSON từ Twitter không thể được khai thác trực tiếp bởi các công cụ phân tích truyền thống nếu không qua bước tiền xử lý (cleansing & transformation).
* ❌ Không có pipeline ETL/ELT tự động → mất nhiều thời gian thủ công để làm sạch dữ liệu.
* ❌ Tính mở rộng kém và thiếu khả năng xử lý song song khi khối lượng tweet tăng nhanh theo thời gian.
* ❌ Thiếu khả năng tích hợp dễ dàng với hệ sinh thái AWS để phục vụ phân tích dữ liệu nâng cao (BI, ML, AI).
* ❌ Không thể theo dõi được pipeline hoặc dễ dàng trigger lại nếu có lỗi xảy ra.

| **Stakeholder** | **Mối quan tâm chính** |
| --- | --- |
| Data Engineer | Tự động hóa pipeline, dễ quản lý & mở rộng |
| Nhà phân tích dữ liệu | Dữ liệu sạch, đầy đủ và dễ truy vấn |
| Doanh nghiệp/Startup | Insight nhanh chóng để phục vụ ra quyết định |
| Người học/ứng viên | Mô hình học tập thực tiễn với AWS chuẩn doanh nghiệp |

**Hệ quả nếu không giải quyết**

* Mất cơ hội khai thác insight từ Twitter – nơi chứa dữ liệu phản ánh thị trường theo thời gian thực.
* Hệ thống phân tích dữ liệu thiếu độ tin cậy và không thể mở rộng khi tăng khối lượng dữ liệu.
* Tăng chi phí vận hành vì phải xử lý thủ công và không có pipeline tự động.

**Cơ hội thị trường**

Với việc doanh nghiệp ngày càng chú trọng vào Social Listening và phân tích cảm xúc, nhu cầu xây dựng pipeline dữ liệu xã hội (social data pipeline) theo mô hình chuẩn doanh nghiệp đang tăng mạnh. Việc phát triển mô hình ELT chuyên nghiệp trên AWS sẽ không chỉ giải quyết bài toán của cá nhân mà còn có tiềm năng thương mại hóa cho startup hoặc nhóm phân tích dữ liệu trong tổ chức.

**🏗️ 3. Solution Architecture – Kiến trúc Giải pháp**

Dự án Data Engineering xử lý dữ liệu Twitter được thiết kế theo mô hình hiện đại, chia theo tầng dữ liệu (data lake zone-based architecture), tận dụng toàn bộ sức mạnh của các dịch vụ AWS để đảm bảo **tính linh hoạt, mở rộng, bảo mật, và hiệu quả chi phí**. Dù ở quy mô cá nhân, kiến trúc này vẫn mô phỏng đầy đủ một hệ thống doanh nghiệp có khả năng mở rộng.

**🖼️ High-Level Architecture Diagram**

Twitter API → S3 (raw) → Spark on EMR → S3 (cleansed, processed)

                  ↓

             Glue Crawler (Data Catalog)

                  ↓

         Athena & Redshift → BI tools (Power BI/QuickSight)

A diagram of a computer hardware process

AI-generated content may be incorrect.

**🔧 Các thành phần chính và tương tác**

| **Thành phần** | **Vai trò chính** |
| --- | --- |
| **Amazon S3**  **PowerBI** | Là nơi lưu trữ theo tầng (raw/, cleansed/, processed/) dữ liệu Twitter crawl  Visualization cho cả Redshift/Athena |
| **Amazon EMR (Spark)** | Dùng để xử lý dữ liệu theo batch: từ raw → cleansed và từ cleansed → processed |
| **AWS Glue Data Catalog** | Tạo schema cho các file Parquet để truy vấn dễ dàng bằng Athena hoặc load vào Redshift |
| **Amazon Athena** | Cho phép query trực tiếp lên S3 dùng ngôn ngữ SQL (serverless) |
| **Amazon Redshift** | Là DWH để lưu trữ dữ liệu đã xử lý, phục vụ BI, dashboard hoặc AI downstream |
| **Apache Airflow** *(local)* | Orchestrate pipeline từ crawl đến ETL → DWH, được cài riêng cho dễ kiểm soát và demo |

**✅ Lý do chọn các dịch vụ AWS**

| **Dịch vụ** | **Justification** |
| --- | --- |
| **S3** | Tối ưu lưu trữ, phân tầng dữ liệu, dễ tích hợp EMR và Redshift |
| **EMR (Spark)** | Phù hợp xử lý dữ liệu lớn, hỗ trợ cluster-based, autoscaling, cost-saving |
| **Glue Catalog** | Tạo metadata cho Athena và Redshift dễ dàng và tự động hóa schema discovery |
| **Athena** | Truy vấn nhanh, không cần hạ tầng, thích hợp để kiểm tra kết quả xử lý dữ liệu |
| **Redshift** | DWH chuyên nghiệp, tích hợp tốt với BI tools và hỗ trợ query performance cao |
| **Airflow**  **QuickSight** | Quản lý toàn bộ pipeline: ETL, kiểm tra lỗi, tự động hóa định kỳ  Trực quan hoá dữ liệu đã được làm sạch hoặc raw để trực quan bên dashboard |

**🔐 Bảo mật & Compliance**

* Tất cả dữ liệu truyền tải và lưu trữ sử dụng giao thức mã hóa (HTTPS, SSE-S3)
* IAM Roles và Policy được thiết lập đúng theo principle of least privilege.
* Các kết nối giữa Redshift và S3 sử dụng IAM Role với COPY permission chuẩn.

**⚙️ Scalability & Performance**

* EMR hỗ trợ autoscaling cluster giúp tiết kiệm chi phí khi xử lý batch lớn.
* S3 không giới hạn dung lượng, hỗ trợ scale theo vùng dữ liệu (processed/twitter/YYYY/MM/DD)
* Redshift cho phép scale compute và query đồng thời (concurrency scaling).

**🔄 Khả năng tích hợp mở rộng**

* Có thể mở rộng thêm Kafka để hỗ trợ **streaming ingestion** từ Twitter hoặc các nguồn social khác.
* Dễ dàng tích hợp thêm Amazon SageMaker để triển khai **mô hình ML** phân tích sentiment, trend, topic clustering.
* Hỗ trợ tích hợp BI tool như QuickSight hoặc Metabase thông qua Redshift.

**4. 🔧 Technical Implementation**

**4.1. Implementation Phases & Deliverables**

Dự án được triển khai qua các giai đoạn sau:

| **Giai đoạn** | **Mô tả** | **Deliverables** |
| --- | --- | --- |
| **Giai đoạn 1: Data Ingestion** | Crawl dữ liệu từ Twitter bằng snscrape, lưu về S3 bucket theo cấu trúc raw/ | Dữ liệu .json trong s3://<bucket-name>/raw/ |
| **Giai đoạn 2: Data Processing with Spark on EMR** | Xử lý dữ liệu dạng thô thành dữ liệu sạch (cleansed/) và dữ liệu phân tích (processed/) với Apache Spark chạy trên EMR | Hai script Spark (spark\_elt\_twitter.py, processed.py), dữ liệu được lưu trong s3://<bucket>/cleansed/ và processed/ |
| **Giai đoạn 3: Querying with Athena** | Dùng AWS Glue Crawlers tạo bảng trong Data Catalog và query dữ liệu bằng SQL với Athena | Athena query output trong s3://<bucket>/athena-query-results/ |
| **Giai đoạn 4: Data Warehousing** | Load dữ liệu phân tích (processed/) vào Amazon Redshift bằng lệnh COPY | Bảng tweets\_analysis trong Redshift, phục vụ phân tích nâng cao |
| **Giai đoạn 5: Orchestration** | Sử dụng Apache Airflow (cài trên EC2) để tự động hóa pipeline theo DAG | DAG twitter\_full\_etl\_pipeline.py chạy tự động từ crawl → EMR → Redshift |

* **Pipeline ELT:**
  + Code Spark xử lý dữ liệu từ raw → cleansed → processed.
  + Airflow DAG điều phối toàn bộ pipeline.
  + Các file .py lưu trên S3 và add step EMR bằng Airflow operator.
* **Glue:**
  + Tạo Glue Database và Crawler → scan S3 processed.
  + Tự động sinh bảng và metadata trong Data Catalog.
* **Athena:**
  + Truy vấn SQL trực tiếp trên processed thông qua Glue Catalog.
  + Kết quả lưu vào athena-query-results (S3).
* **Redshift:**
  + Dữ liệu processed copy sang Redshift.
  + Tạo bảng processed\_tweets dạng PARQUET.
* **Power BI:**
  + Kết nối với Athena và Redshift để tạo dashboard insight (top hashtags, user engagement, sentiment).

**4.2. Technical Requirements**

| **Hạng mục** | **Cấu hình** |
| --- | --- |
| **Compute** | EC2 t2.medium cho Airflow, EMR cluster (1 master + 2 core nodes) |
| **Storage** | S3 cho lưu trữ toàn bộ pipeline, Redshift 160GB SSD |
| **Network** | VPC với Public/Private Subnet, NAT Gateway để EMR truy cập Internet |
| **Security** | IAM Role phân quyền rõ ràng cho EMR, S3, Glue và Redshift. SSH key cho EC2 |

**4.3. Development Methodology**

* **Môi trường phát triển**: Sử dụng Python 3.10 và virtualenv cho từng module riêng biệt.
* **Quản lý mã nguồn**: Toàn bộ mã được quản lý trên GitHub theo mô hình branches (main, dev, feature/\*)
* **Logging & Monitoring**: Spark log lưu trong emr-logs/, Airflow UI giám sát toàn bộ pipeline, CloudWatch hỗ trợ thêm khi cần mở rộng.

**4.4. Testing Strategy**

| **Loại test** | **Mô tả** |
| --- | --- |
| **Unit Test** | Kiểm thử từng phần xử lý JSON trong Spark (select, filter, explode) |
| **Integration Test** | Kiểm thử sự tương thích EMR ↔ S3 ↔ Athena, Redshift |
| **Performance Test** | Chạy batch data 10K–50K tweet để test Spark và Redshift |
| **Manual Smoke Test** | Chạy thử full pipeline bằng DAG Airflow theo từng step |

**4.5. Deployment Plan**

* **Deploy từng phần**:
  + EMR được khởi tạo từ AWS Console và thêm Spark Steps tự động.
  + Airflow được cấu hình local trên EC2 → test DAG trước khi lên sản phẩm.
  + Redshift sử dụng lệnh COPY để import dữ liệu, verify schema.
* **Rollback Procedures**:
  + Spark script có versioning trên GitHub.
  + Nếu pipeline lỗi, Airflow dừng task và log lỗi rõ ràng để debug.
  + Sử dụng timestamp folder trong S3 để đảm bảo dữ liệu không bị ghi đè.

**4.6. Configuration Management**

* **ENV variables**: Quản lý thông tin AWS (bucket, region, Redshift config) qua .env hoặc Airflow Variable.
* **Secrets**: Lưu trong AWS Secrets Manager hoặc .bashrc (khi làm local).
* **EMR bootstrap**: Cấu hình bootstrap.sh tự động cài đặt dependencies như PySpark và boto3.

**💡 Kết luận phần này:**

Việc triển khai kỹ thuật cho dự án được xây dựng chặt chẽ, theo tiêu chuẩn cloud-native, đảm bảo khả năng **mở rộng**, **tự động hóa**, và **chuyển giao dễ dàng**. Dù dự án cá nhân, mô hình đã phản ánh được quy trình chuyên nghiệp, phù hợp với môi trường doanh nghiệp hiện đại.

**5. 📅 Timeline & Milestones (10%)**

**5.1. Tổng Quan Kế Hoạch Dự Án**

Dự án được thực hiện trong vòng **5 tuần** theo mô hình Agile, chia thành các giai đoạn ngắn (sprint) với các mục tiêu cụ thể. Mỗi giai đoạn đều có kiểm tra đầu ra (deliverables) để đảm bảo tiến độ và chất lượng.

| **Tuần** | **Giai đoạn** | **Hoạt động chính** | **Deliverables** |
| --- | --- | --- | --- |
| **Tuần 1** | Khởi động & Thu thập dữ liệu | - Crawl dữ liệu Twitter- Thiết lập bucket S3 raw/- Tổ chức project structure | File JSON mẫu, S3 bucket hoạt động |
| **Tuần 2** | Xử lý dữ liệu Spark trên EMR | - Tạo EMR cluster- Viết script Spark raw → cleansed và cleansed → processed- Lưu dữ liệu vào S3 | Dữ liệu đã xử lý, EMR logs |
| **Tuần 3** | Query & Phân tích với Athena | - Tạo Glue Crawlers- Định nghĩa schema- Query SQL trên Athena | Athena bảng dữ liệu & kết quả SQL |
| **Tuần 4** | Redshift & Data Warehouse | - Tạo bảng Redshift- Load dữ liệu bằng COPY- Chuẩn hóa schema | Dữ liệu trong Redshift |
| **Tuần 5**  **Tuần 6** | Orchestration với Airflow & Viết báo cáo  Trực quan dữ liệu | - Viết DAG Airflow twitter\_full\_etl\_pipeline- Tự động hoá toàn bộ pipeline- Tổng hợp & viết proposal  - Sau khi đã xử lí dữ liệu, dữ liệu được lưu ở raw, cleansed, và processed sẽ được athena truy vấn làm dashboard trong quicksight | DAG chạy thành công  Chạy được dashboard trên powerBI + Project Proposal hoàn chỉnh |

**5.2. Milestones Quan Trọng**

| **Milestone** | **Thời gian** | **Success Criteria** |
| --- | --- | --- |
| ✅ Dữ liệu Twitter được crawl và lưu vào S3 | Cuối tuần 1 | Có >1000 tweet lưu trong raw/ |
| ✅ Spark xử lý dữ liệu và sinh output cleansed/, processed/ | Cuối tuần 2 | File .parquet tạo ra trong đúng cấu trúc |
| ✅ Athena truy vấn thành công dữ liệu | Cuối tuần 3 | Có ít nhất 3 câu SQL trả về đúng |
| ✅ Dữ liệu vào Redshift, kiểm thử phân tích | Cuối tuần 4 | Có bảng Redshift với dữ liệu đúng schema |
| ✅ Airflow DAG chạy tự động toàn pipeline | Cuối tuần 5 | DAG twitter\_full\_etl\_pipeline success 100% |

**5.3. Dependencies & Resource Planning**

| **Thành phần phụ thuộc** | **Chi tiết** |
| --- | --- |
| 🧱 EMR cluster | Phải sẵn sàng trước tuần 2 |
| 🔐 IAM roles | Được cấu hình đầy đủ để Spark có quyền truy cập S3, Redshift |
| 📊 AWS Glue Crawlers | Phải chạy thành công để tạo schema cho Athena |
| 🌐 Airflow environment | Phải được setup ổn định trước khi tạo DAG |

**5.4. Buffer & Risk Planning**

* **Buffer**: Dự trù 2 ngày cuối cùng cho debugging và cải tiến DAG.
* **Risk dự phòng**: Nếu Airflow gặp lỗi (đã từng gặp), có thể fallback bằng cách gọi script qua cron hoặc shell script.

**6. 💰 Budget Estimation**

**6.1. Tổng Quan Chi Phí Dự Án**

Dự án sử dụng một số dịch vụ cốt lõi của AWS (S3, EMR, Glue, Athena, Redshift), kết hợp với môi trường cục bộ (Airflow). Dự án hiện đang ở quy mô cá nhân nhưng được thiết kế theo mô hình có thể mở rộng cho doanh nghiệp. Do đó, bảng chi phí bao gồm cả **chi phí thực tế** và **ước tính chi phí doanh nghiệp** nếu mở rộng.

**6.2. Ước Tính Chi Phí (Mô Hình Cá Nhân – Thực Tế Đã Triển Khai)**

| **Dịch vụ AWS** | **Mục đích sử dụng** | **Thời gian dùng** | **Ước tính chi phí** |
| --- | --- | --- | --- |
| Amazon S3 | Lưu dữ liệu raw/, cleansed/, processed/ | 1 tháng | ~$0.05 USD |
| Amazon EMR (on-demand, m5.xlarge) | Chạy Spark jobs xử lý dữ liệu | 5–6 giờ | ~$1.50 USD |
| AWS Glue Crawler | Sinh schema cho Athena | 1 crawler, chạy 2 lần | ~$0.10 USD |
| Amazon Athena | Truy vấn dữ liệu bằng SQL | ~10 queries | ~$0.10 USD |
| Amazon Redshift (Free Tier) | Lưu trữ và phân tích dữ liệu | 1 bảng nhỏ | ~$0.00 USD |
| EC2 (cho Airflow – optional)  QuickSight (Dashboard) | Chạy UI và DAG orchestration  Chạy UI showcase kết quả | ~10 giờ  1 giờ | ~$1.20 USD  Free Trial |
| **Tổng cộng** | **~3.00 USD** |  |  |

**6.3. Ước Tính Chi Phí Khi Mở Rộng Doanh Nghiệp (Monthly)**

| **Thành phần** | **Sử dụng thực tế** | **Chi phí ước tính/tháng** |
| --- | --- | --- |
| Amazon S3 (100GB) | Lưu dữ liệu Twitter và ML output | ~$2.50 USD |
| EMR (job hàng ngày) | Chạy Spark ELT mỗi đêm | ~$50 USD |
| AWS Glue | Catalog schema và job sync | ~$10 USD |
| Athena (1TB query) | Truy vấn dữ liệu phân tích | ~$5 USD |
| Redshift (2-node) | Dữ liệu phân tích nâng cao | ~$180 USD |
| Amazon MSK (Kafka) | Streaming xử lý real-time | ~$250 USD |
| EC2 (Airflow + Bastion)  QuickSight | Orchestration + remote access  Dashboard để visualize | ~$40 USD  ~$50 USD |
| **Tổng cộng** | **~577.50 USD/month** |  |

Có thể tối ưu thêm bằng cách dùng **Spot Instance cho EMR**, hoặc chuyển sang **Redshift Serverless**.

**6.4. ROI & Tối Ưu Hóa Chi Phí**

| **Mục tiêu** | **Chi tiết** |
| --- | --- |
| **ROI (Return on Investment)** | Nếu áp dụng cho doanh nghiệp, giúp giảm thời gian xử lý dữ liệu thủ công xuống 80%, cải thiện tốc độ báo cáo và phân tích social media. |
| **Chiến lược tối ưu** | - Tự động tắt EMR sau job (idle cluster timeout)- Dùng Glue catalog thay vì schema thủ công- Tận dụng S3 Intelligent-Tiering- Dùng Redshift Spectrum nếu phân tích dữ liệu lớn mà không cần import |

**7. ⚠️ Risk Assessment**

**7.1. Tổng Quan Quản Trị Rủi Ro**

Dự án Data Engineering này được triển khai trên môi trường AWS với nhiều dịch vụ như EMR, S3, Redshift, Glue và Athena. Mặc dù hệ thống đã hoạt động ổn định ở quy mô cá nhân, một số rủi ro tiềm ẩn vẫn cần được đánh giá và quản lý khi dự án mở rộng hoặc tích hợp thêm Machine Learning (ML), streaming (Kafka), hoặc CI/CD.

**7.2. Bảng Đánh Giá Rủi Ro (Risk Matrix)**

| **Rủi ro** | **Loại rủi ro** | **Mức độ ảnh hưởng** | **Xác suất** | **Ưu tiên** | **Chiến lược giảm thiểu** |
| --- | --- | --- | --- | --- | --- |
| Cấu hình sai dịch vụ AWS (IAM, S3, EMR) | Kỹ thuật | Cao | Trung bình | Cao | Sử dụng AWS IAM với principle of least privilege, kiểm thử kỹ script |
| Chi phí vượt dự toán khi dùng EMR, Redshift | Tài chính | Trung bình | Trung bình | Trung bình | Giới hạn thời gian cluster, bật auto-terminate, monitoring bằng CloudWatch |
| Airflow DAG fail do lỗi Python/syntax | Vận hành | Thấp | Cao | Trung bình | Kiểm tra DAG cục bộ trước khi upload, cấu hình email alert |
| Dữ liệu không đồng nhất (Twitter JSON thay đổi format) | Dữ liệu | Cao | Thấp | Trung bình | Dùng Spark xử lý với try/except, validate schema thường xuyên |
| Mất key .pem, không truy cập được EC2 | Operational | Cao | Thấp | Trung bình | Backup key trên S3 private + MFA, hạn chế dùng EC2 lâu dài |
| Tích hợp Kafka/ML phức tạp | Phát triển | Trung bình | Trung bình | Trung bình | Tách module, thử nghiệm local trước khi deploy vào AWS |

**7.3. Kế Hoạch Ứng Phó (Contingency Plans)**

* **Kỹ thuật**: Sử dụng Terraform hoặc CloudFormation để có thể khôi phục hạ tầng nhanh nếu cần.
* **Tài chính**: Giới hạn dịch vụ AWS bằng Budget Alert, kích hoạt “Billing Alarm” trên AWS.
* **Con người**: Ghi lại chi tiết các bước triển khai vào README.md để dễ bàn giao hoặc mở rộng nhóm.
* **Mở rộng (AI/ML sau này)**: Triển khai thêm môi trường staging để thử nghiệm mô hình AI mà không ảnh hưởng production.

**7.4. Kế Hoạch Giám Sát (Monitoring Plan)**

* Dùng **AWS CloudWatch** để theo dõi job EMR, truy vấn Athena, và tài nguyên Redshift.
* Dùng **Airflow Email Alerts** để phát hiện DAG failure hoặc retry bất thường.
* Thiết lập **SLA Monitoring** trong Airflow nếu dùng cho doanh nghiệp.

**8. 🎯 Expected Outcomes**

**8.1. Success Metrics – Chỉ số thành công**

Dự án được xem là thành công khi đạt được các chỉ số định lượng và định tính sau:

| **Loại chỉ số** | **Mục tiêu cụ thể** |
| --- | --- |
| ✅ **Kỹ thuật** | Pipeline hoạt động ổn định: raw → cleansed → processed → Redshift |
| ✅ **Thời gian xử lý** | Dưới 15 phút từ khi upload đến khi có thể query qua Athena |
| ✅ **Tối ưu hóa chi phí** | EMR cluster chỉ tồn tại đúng thời gian xử lý, tiết kiệm 50–70% so với luôn bật |
| ✅ **Tích hợp** | Spark + S3 + Redshift + Athena hoạt động liền mạch |
| ✅ **Mở rộng** | Sẵn sàng tích hợp Apache Kafka (streaming) hoặc mô hình ML nếu cần |

**8.2. Lợi ích ngắn hạn (0–6 tháng)**

* Hoàn thiện một pipeline xử lý dữ liệu dạng batch chuyên nghiệp, có thể demo và áp dụng vào các dataset khác.
* Sử dụng được đầy đủ các dịch vụ AWS trong mô hình ELT thực tế.
* Có thể xuất dữ liệu sạch từ Twitter để phục vụ trực tiếp cho phân tích, dashboard hoặc machine learning.
* Đạt được kỹ năng thực chiến với các dịch vụ AWS quan trọng (EMR, Athena, Redshift, IAM…).

**8.3. Lợi ích trung hạn (6–18 tháng)**

* Dễ dàng mở rộng tích hợp Apache Kafka (Amazon MSK) để xử lý dữ liệu Twitter theo thời gian thực.
* Có thể tích hợp module phân tích cảm xúc (sentiment analysis) hoặc mô hình AI để phân tích các chủ đề trending.
* Tái sử dụng pipeline cho các nguồn dữ liệu khác (news, YouTube, Reddit, v.v.).
* Gắn kết được với các dashboard như Metabase hoặc QuickSight để phân tích real-time.

**8.4. Lợi ích dài hạn (18+ tháng)**

* Mở rộng mô hình thành giải pháp phân tích mạng xã hội cho doanh nghiệp (B2B).
* Tích hợp CI/CD, kiểm thử tự động, data quality check để phục vụ các tổ chức lớn.
* Đóng gói thành sản phẩm SaaS hoặc white-label analytics platform phục vụ nhiều ngành.

**8.5. Trải nghiệm người dùng và giá trị chiến lược**

* Cung cấp giao diện truy vấn dễ dàng qua Athena hoặc kết nối từ BI tool.
* Giúp nhà phân tích, marketer, hoặc data scientist rút ngắn 80% thời gian thu thập và xử lý dữ liệu.
* Là tiền đề cho các chiến dịch AI/ML quy mô lớn nhờ dữ liệu đã được làm sạch, cấu trúc rõ ràng.