

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
VIỆN TRÍ TUỆ NHÂN TẠO
KỸ THUẬT VÀ CÔNG NGHỆ DỮ LIỆU LỚN



BÁO CÁO MÔN HỌC
KỸ THUẬT VÀ CÔNG NGHỆ DỮ LIỆU LỚN

**Phân tích thị trường việc làm theo dòng dữ liệu
thời gian thực trên nền tảng Big Data**

Real-time Job Market Analytics with Big Data Streaming using Apache
Kafka

Đặng Đức Duy – 23020347
Trịnh Hoàng Đức – 23020359

Giảng viên hướng dẫn:

TS. Trần Hồng Việt
ThS. Ngô Minh Hương

TP. Hồ Chí Minh, 2025

Mục lục

TÓM TẮT	2
1 GIỚI THIỆU	3
1.1 Động cơ nghiên cứu	3
1.2 Mục tiêu	3
1.3 Phạm vi và đóng góp	3
2 DỮ LIỆU VÀ PHƯƠNG PHÁP PHÂN TÍCH	4
2.1 Dữ liệu và đặc điểm dữ liệu lớn	4
2.2 Các biến phân tích chính	4
2.3 Khung phương pháp phân tích	4
3 HẠ TẦNG BIG DATA VÀ DÒNG DỮ LIỆU KAFKA	6
3.1 Vai trò của Kafka trong bài toán phân tích dữ liệu	6
3.2 Kiến trúc dữ liệu: Producer–Topic–Consumer	6
3.3 Tính đúng đắn phân tích trong môi trường streaming	6
4 KẾT QUẢ THỐNG KÊ VÀ TRỰC QUAN HOÁ	7
4.1 Chỉ số tổng quan của lần chạy	7
4.2 Phân bố theo quốc gia: mức độ tập trung dữ liệu	7
4.3 Phân bố job type: cấu trúc hình thức việc làm	8
4.4 Top skills: nhu cầu kỹ năng và vấn đề chuẩn hoá văn bản	10
5 THẢO LUẬN: TỪ KẾT QUẢ ĐẾN HÀM Ý	13
5.1 Tính đại diện và rủi ro diễn giải	13
5.2 Hàm ý cho người học và nhà tuyển dụng	13
5.3 Vai trò của Kafka trong mở rộng phân tích dữ liệu	13
6 HƯỚNG PHÁT TRIỂN	15
6.1 Mở rộng phân tích dữ liệu	15
6.2 Mở rộng hạ tầng Big Data	15
KẾT LUẬN	16

TÓM TẮT

Báo cáo trình bày một hệ thống phân tích thị trường việc làm theo hướng **dữ liệu lớn (Big Data)** và **xử lý theo dòng (streaming)**. Dữ liệu tin tuyển dụng được chuẩn hoá và phát liên tục lên Apache Kafka theo định dạng JSON; consumer đọc theo batch để thực hiện tổng hợp thống kê phục vụ phân tích: (i) phân bố tin theo quốc gia, (ii) phân bố loại hình công việc (job type), và (iii) nhóm kỹ năng nổi bật (top skills). Hệ thống tạo ra các đầu ra ở dạng **CSV/JSON** và hình ảnh trực quan hoá, hỗ trợ diễn giải và báo cáo.

Trọng tâm của báo cáo không chỉ là chạy được pipeline, mà là **phân tích dữ liệu** một cách có phương pháp: nhận diện thiên lệch phân bố (distribution bias), đánh giá chất lượng biến kỹ năng (text normalization), và thảo luận mức độ tin cậy của các kết luận rút ra từ dữ liệu. Kết quả thực nghiệm ghi nhận xử lý **100,000** bản ghi trong **157.6** giây, tương ứng thông lượng trung bình khoảng **634.5 message/giây**. Từ các bảng tổng hợp, báo cáo phân tích xu hướng nổi bật: dữ liệu tập trung ở một số quốc gia nhất định; job type lệch mạnh về một nhóm; kỹ năng mềm xuất hiện với tần suất cao, đồng thời tồn tại hiện tượng trùng nghĩa do không đồng nhất cách viết.

CHƯƠNG 1. GIỚI THIỆU

1.1 Động cơ nghiên cứu

Phân tích thị trường việc làm là một bài toán dữ liệu có ý nghĩa thực tiễn: doanh nghiệp cần nắm xu hướng kỹ năng để tuyển dụng, người học cần hiểu kỹ năng phổ biến để định hướng năng lực, và nhà quản lý cần quan sát biến động lao động theo khu vực. Điểm khó của dữ liệu tin tuyển dụng nằm ở hai đặc trưng: **(1) tính liên tục** (dữ liệu xuất hiện theo thời gian) và **(2) tính không đồng nhất** (nhiều nguồn, nhiều chuẩn ghi khác nhau). Vì vậy, phương pháp tiếp cận dựa trên Big Data và streaming giúp mô hình hoá bài toán theo hướng hiện đại: ingest dữ liệu liên tục và cập nhật phân tích theo dòng.

1.2 Mục tiêu

Báo cáo đặt ra ba nhóm mục tiêu:

1. **Mục tiêu hệ thống:** xây dựng pipeline streaming dựa trên Apache Kafka để ingest và xử lý dữ liệu tin tuyển dụng.
2. **Mục tiêu dữ liệu:** tạo các bảng tổng hợp phục vụ phân tích (country, job type, skills) và trực quan hoá kết quả.
3. **Mục tiêu phân tích:** diễn giải dữ liệu theo phương pháp: mô tả phân bố, nhận diện thiên lệch, đánh giá chất lượng biến văn bản, và đề xuất cải thiện.

1.3 Phạm vi và đóng góp

Trong phạm vi báo cáo, hệ thống tập trung vào thống kê mô tả và insight nền tảng. Các đóng góp chính:

- Một pipeline Big Data streaming (Producer–Kafka–Consumer) có thể chạy tái lập và xuất kết quả.
- Một khung phân tích dữ liệu định hướng “từ dữ liệu đến kết luận”: phân bố, thiên lệch, chất lượng dữ liệu, và implications.
- Bộ biểu đồ minh hoạ kết quả phân tích giúp tăng tính diễn giải và thuyết phục.

CHƯƠNG 2. DỮ LIỆU VÀ PHƯƠNG PHÁP PHÂN TÍCH

2.1 Dữ liệu và đặc điểm dữ liệu lớn

Dữ liệu tin tuyển dụng có thể xem là dữ liệu lớn theo nghĩa rộng:

- **Volume:** số lượng bản ghi lớn và có thể tăng nhanh theo thời gian.
- **Velocity:** dữ liệu phát sinh liên tục (mang tính dòng).
- **Variety:** nhiều trường dữ liệu, đặc biệt có trường văn bản (job title, skills) không chuẩn hoá.

Trong báo cáo, Apache Kafka đóng vai trò “xương sống” cho đặc trưng velocity: dữ liệu được stream lên topic như một commit log, cho phép consumer cập nhật thống kê theo dòng.

2.2 Các biến phân tích chính

Báo cáo tập trung vào ba nhóm biến:

1. **Biến địa lý:** quốc gia (country) phản ánh phạm vi thị trường và độ tập trung dữ liệu.
2. **Biến loại hình việc làm:** job_type (Onsite/Hybrid/Remote) phản ánh cấu trúc hình thức lao động.
3. **Biến kỹ năng:** skills (tách từ job_skills) phản ánh yêu cầu năng lực của thị trường.

2.3 Khung phương pháp phân tích

Phân tích được tổ chức theo 4 lớp:

- **Mô tả phân bố (descriptive distribution):** thống kê tần suất và trực quan hoá.
- **Nhận diện thiên lệch (bias & representativeness):** đánh giá mức độ lệch phân bố theo quốc gia/job type.

- **Chất lượng dữ liệu (data quality):** phát hiện trùng biến thể kỹ năng, tác động đến top skills.
- **Diễn giải và hàm ý (insights & implications):** rút ra kết luận có điều kiện và đề xuất cải thiện.

CHƯƠNG 3. HẠ TẦNG BIG DATA VÀ DÒNG DỮ LIỆU KAFKA

3.1 Vai trò của Kafka trong bài toán phân tích dữ liệu

Trong hệ thống, Kafka được dùng như một **hạ tầng ingest và streaming** hơn là mục tiêu cuối cùng. Điều này có ý nghĩa dữ liệu học: thay vì đọc toàn bộ CSV rồi xử lý một lần, dữ liệu được đưa vào một luồng (stream) và consumer có thể:

- cập nhật thống kê theo thời gian,
- replay dữ liệu để tái tạo kết quả khi thay đổi logic,
- mở rộng thêm consumer khác cho các hướng phân tích mới.

Nói cách khác, Kafka tạo “đường ống dữ liệu” (data pipeline) để dữ liệu đi qua nhiều bước phân tích mà không cần thay đổi nguồn.

3.2 Kiến trúc dữ liệu: Producer–Topic–Consumer

Luồng dữ liệu của bài toán:

CSV → Producer → Kafka Topic `jobs_raw` → Consumer Group → CSV/JSON/PNG

3.3 Tính đúng đắn phân tích trong môi trường streaming

Khi xử lý streaming, “tính đúng” không chỉ là thuật toán thống kê mà còn là cách consumer đọc dữ liệu:

- **Offset & commit**: consumer commit theo batch giúp kiểm soát tiến độ xử lý và giảm nguy cơ bỏ sót.
- **Replay**: chạy với group-id mới cho phép đọc lại toàn bộ dữ liệu trong retention, giúp tái tạo kết quả nếu thay đổi pipeline phân tích.

Trong phạm vi báo cáo, các cơ chế này đảm bảo rằng phân tích có thể tái lập và kết quả có thể kiểm chứng bằng file `run_metrics.json`.

CHƯƠNG 4. KẾT QUẢ THỐNG KÊ VÀ TRỰC QUAN HOÁ

4.1 Chỉ số tổng quan của lần chạy

Từ `run_metrics.json`, `hthngghinhn` :

Processed: 100,000 bản ghi.

Runtime: 157.6 giây.

Throughput:

$$\frac{100000}{157.6} \approx 634.5 \text{ message/giây.}$$

Countries_seen: 4 quốc gia.

Bảng 4.1: Chỉ số tổng quan (trích từ `run_metrics.json`).

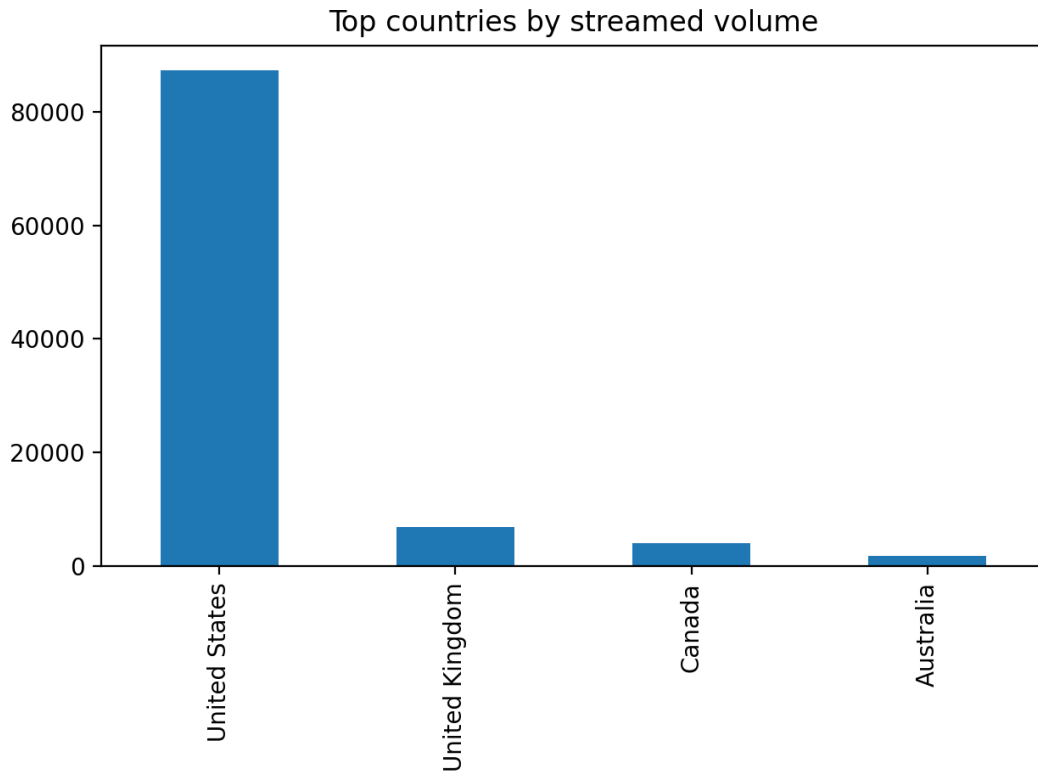
Chỉ số	Giá trị/diễn giải
Topic	jobs_raw
Processed	100,000
Runtime (sec)	157.6
Throughput (msg/s)	≈ 634.5
Countries_seen	4
Commit_every	5,000 (batch commit)

4.2 Phân bố theo quốc gia: mức độ tập trung dữ liệu

Bảng 4.2 cho thấy dữ liệu tập trung mạnh ở một quốc gia, các quốc gia còn lại có tỷ trọng nhỏ hơn đáng kể.

Bảng 4.2: Top quốc gia theo số lượng tin (từ metrics).

Country	Count
United States	87,340
United Kingdom	6,868
Canada	3,964
Australia	1,828



Hình 4.1: Phân bố số lượng tin theo quốc gia.

Phân tích: thiên lệch phân bố và tính đại diện

Một phân bố lệch mạnh (dominant country) tạo ra hai hệ quả phân tích:

- **Insight tổng hợp bị chi phối:** top skills hoặc job type “overall” thường phản ánh quốc gia chiếm đa số hơn là phản ánh toàn thị trường.
- **So sánh liên quốc gia cần chuẩn hoá:** nếu muốn so sánh, cần dùng tỷ lệ (share) hoặc chuẩn hoá theo quy mô mẫu từng quốc gia.

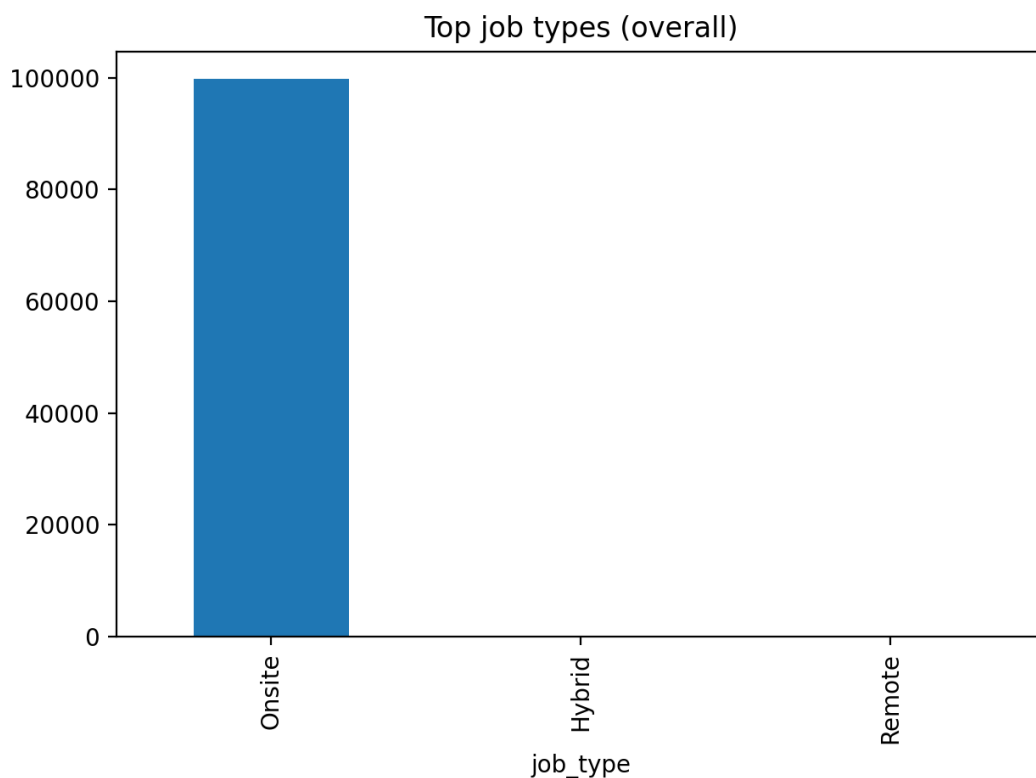
Do đó, khi diễn giải top skills toàn cục, báo cáo sẽ coi đây là một “tổng hợp có điều kiện”, phụ thuộc cấu trúc dữ liệu đầu vào.

4.3 Phân bố job type: cấu trúc hình thức việc làm

Kết quả tổng hợp job type được trình bày trong Bảng 4.3 và Hình tương ứng.

Bảng 4.3: Phân bố job type (tổng hợp).

Job type	Count
Onsite	99,712
Hybrid	186
Remote	102



Hình 4.2: Phân bố job type trong dữ liệu.

Phân tích: hàm ý và giả thuyết dữ liệu

Tỷ trọng Onsite cao có thể xuất phát từ:

- cách gắn nhãn job type trong nguồn tin (ví dụ mặc định Onsite nếu thiếu thông tin),
- quy tắc chuẩn hoá quy về một nhãn chung,
- phạm vi thu thập dữ liệu có thiên hướng về nhóm công việc Onsite.

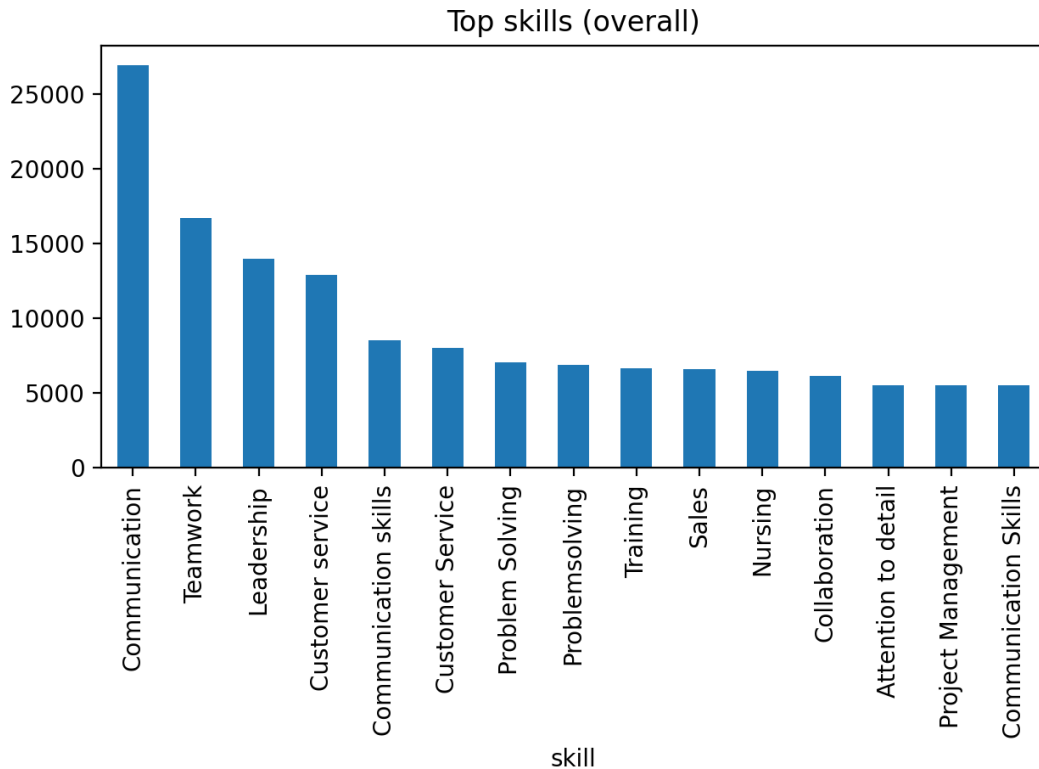
Điểm quan trọng trong Big Data analytics là: thay vì khẳng định nguyên nhân, phân tích cần nêu **giả thuyết hợp lý** và đề xuất kiểm chứng (ví dụ kiểm tra tỷ lệ missing của job type, hoặc phân tích theo quốc gia/nhóm ngành).

4.4 Top skills: nhu cầu kỹ năng và vấn đề chuẩn hoá văn bản

Bảng 4.4 trình bày top kỹ năng. Nhiều kỹ năng thuộc nhóm kỹ năng mềm xuất hiện ở tần suất cao (communication, teamwork, leadership), phản ánh xu hướng phổ quát trong mô tả công việc.

Bảng 4.4: Top 10 kỹ năng (tổng hợp).

Skill	Count
Communication	26,886
Teamwork	16,724
Leadership	13,950
Customer service	12,906
Communication skills	8,538
Customer Service	8,034
Problem Solving	7,040
Problemsolving	6,882
Training	6,654
Sales	6,632



Hình 4.3: Top kỹ năng (tổng hợp).

Phân tích sâu: biến thể kỹ năng và tác động đến kết luận

Kết quả cho thấy hiện tượng **split** do biến thể cách viết:

- `Customer service` và `Customer Service` gần như cùng nghĩa.
- `Problem Solving` và `Problemsolving` khác biệt chủ yếu ở chuẩn hoá khoảng trắng.
- `Communication` và `Communication skills` có thể là hai mức độ diễn đạt cùng một nhóm năng lực.

Về mặt phân tích dữ liệu, điều này làm giảm độ “sạch” của top skills và có thể làm sai lệch thứ hạng nếu gộp đúng synonym. Tuy nhiên, chính phát hiện này là một insight quan trọng về **data quality**: dữ liệu kỹ năng cần chuẩn hoá sâu trước khi đưa vào phân tích ra quyết định (ví dụ thiết kế chương trình đào tạo).

Đề xuất chuẩn hoá kỹ năng (hướng phân tích dữ liệu)

Một quy trình chuẩn hoá có thể gồm:

1. lowercase + strip + chuẩn hoá khoảng trắng,

2. loại bỏ ký tự thừa,
3. mapping synonym (customer service \rightarrow customer service),
4. gộp theo cụm kỹ năng (soft skills vs technical skills) để tăng tính diễn giải.

Các bước này sẽ giúp kết quả top skills phản ánh đúng “nhu cầu năng lực” hơn là phản ánh lỗi biểu diễn văn bản.

CHƯƠNG 5. THẢO LUẬN: TỪ KẾT QUẢ ĐẾN HÀM Ý

5.1 Tính đại diện và rủi ro diễn giải

Hai nguồn rủi ro lớn khi diễn giải kết quả:

- **Thiên lệch mẫu:** dữ liệu tập trung ở một quốc gia có thể làm “overall insight” mang tính địa phương.
- **Chất lượng biến văn bản:** kỹ năng bị split làm giảm độ tin cậy của xếp hạng top.

Do đó, báo cáo khuyến nghị khi dùng kết quả cho quyết định, cần đi kèm mô tả điều kiện dữ liệu: phạm vi quốc gia, mức độ chuẩn hoá skills và quy tắc gắn nhãn job type.

5.2 Hàm ý cho người học và nhà tuyển dụng

Từ top skills, có thể rút ra một hàm ý định hướng:

- nhóm kỹ năng mềm (communication, teamwork, leadership) xuất hiện với tần suất cao, do đó nên được xem là nền tảng ở nhiều vai trò.
- các kỹ năng như sales/training/customer service cho thấy nhu cầu lớn ở các vị trí tuyển đầu hoặc vận hành.

Tuy nhiên, để đưa ra khuyến nghị sâu hơn (theo ngành/level), cần mở rộng phân tích kết hợp giữa job title, job level và skills.

5.3 Vai trò của Kafka trong mở rộng phân tích dữ liệu

Kafka hỗ trợ mở rộng phân tích theo hai hướng:

- **Mở rộng theo thời gian:** thêm consumer để tổng hợp theo cửa sổ thời gian (ngày/tuần/tháng), theo dõi xu hướng thay đổi.
- **Mở rộng theo mục tiêu:** song song hoá nhiều consumer khác nhau (ví dụ consumer A tổng hợp skills, consumer B phân tích job title, consumer C lưu vào DB).

Nhờ mô hình commit log, cùng một luồng dữ liệu có thể phục vụ nhiều bài toán phân tích mà không cần thay đổi producer.

CHƯƠNG 6. HƯỚNG PHÁT TRIỂN

6.1 Mở rộng phân tích dữ liệu

Một số hướng phát triển mang tính phân tích:

- **Phân tích theo quốc gia:** thay vì overall, xây top skills per country và so sánh khác biệt.
- **Phân tích theo thời gian:** dùng window để xem xu hướng remote/hybrid tăng giảm theo thời gian.
- **Phân tích theo job level:** tìm kỹ năng đặc trưng cho junior vs senior.
- **Chuẩn hoá skills nâng cao:** mapping synonym + clustering để nhóm kỹ năng.

6.2 Mở rộng hạ tầng Big Data

Các bước nâng cấp hạ tầng (khi cần):

- lưu kết quả vào hệ lưu trữ (PostgreSQL/Elastic) thay vì chỉ CSV,
- giám sát consumer lag/throughput,
- sử dụng stream processing framework (Kafka Streams/Flink) nếu cần exactly-once cho pipeline phức tạp.

KẾT LUẬN

Báo cáo đã xây dựng một hệ thống phân tích thị trường việc làm theo hướng Big Data streaming, trong đó Apache Kafka đóng vai trò nền tảng ingest và phân phối dòng dữ liệu. Trên nền tảng đó, pipeline tạo ra các thống kê có giá trị phân tích: phân bố tin theo quốc gia, cơ cấu job type và top skills; đồng thời chỉ ra hai vấn đề dữ liệu quan trọng ảnh hưởng trực tiếp đến kết luận: thiên lệch phân bố và chất lượng chuẩn hoá kỹ năng. Các phân tích và đề xuất trong báo cáo tạo nền tảng để mở rộng thành hệ thống quan sát xu hướng kỹ năng và thị trường việc làm theo thời gian, với mức độ tin cậy cao hơn khi dữ liệu được chuẩn hoá và phân tích theo nhóm.