**INTRODUCTION TO FLINK**

**1. Vấn đề:**

Các hệ thống truyền thống như Hadoop chỉ hỗ trợ *batch processing*, dẫn đến độ trễ cao và không phù hợp với các tác vụ cần phản hồi tức thì.

Không thể xử lý dữ liệu *streaming* với độ trễ thấp và tính chính xác cao.

**2. Những vấn đề trong stream processing:**

- Khi một hệ thống xử lý một sự kiện, có thể xảy ra lỗi (như mất kết nối, node bị crash), khiến hệ thống xử lý lại cùng một sự kiện. Điều này dẫn đế hệ thống có thể xử lý hai lần cùng một sự kiện hoặc sự kiện hoàn toàn không đuợc xử lý.

=> *Đảm bảo exactly-once processing*: Đảm bảo mỗi sự kiện chỉ được xử lý đúng một lần — dù có lỗi, sự kiện không bị mất hay trùng lặp.

- Trong môi truờng phân tán, các sự kiện xuất hiện đến nơi xử lý không đúng với thứ tự thời gian các sự kiện đuợc sinh ra, do độ trễ và tốc độ mạng khác nhau.

=> *Đảm bảo thứ tự sự kiện*: Đảm bảo hệ thống xử lý sự kiện theo đúng thời gian xảy ra *thực tế (event time)*, không phải thời gian *nhận được (processing time)*.

- *Đảm bảo phục hồi lỗi (fault tolerance)*.

**3. Flink**

**Apache Flink** là một nền tảng mã nguồn mở dùng để xử lý dữ liệu phân tán theo *thời gian thực* và *theo batch*. Flink nổi bật với khả năng xử lý dòng dữ liệu chính xác (exactly-once processing), khôi phục lỗi tự động (fault tolerance) và hỗ trợ tính toán theo thời gian sự kiện (event time processing), giúp xây dựng các ứng dụng phân tích dữ liệu mạnh mẽ, linh hoạt và đáng tin cậy.