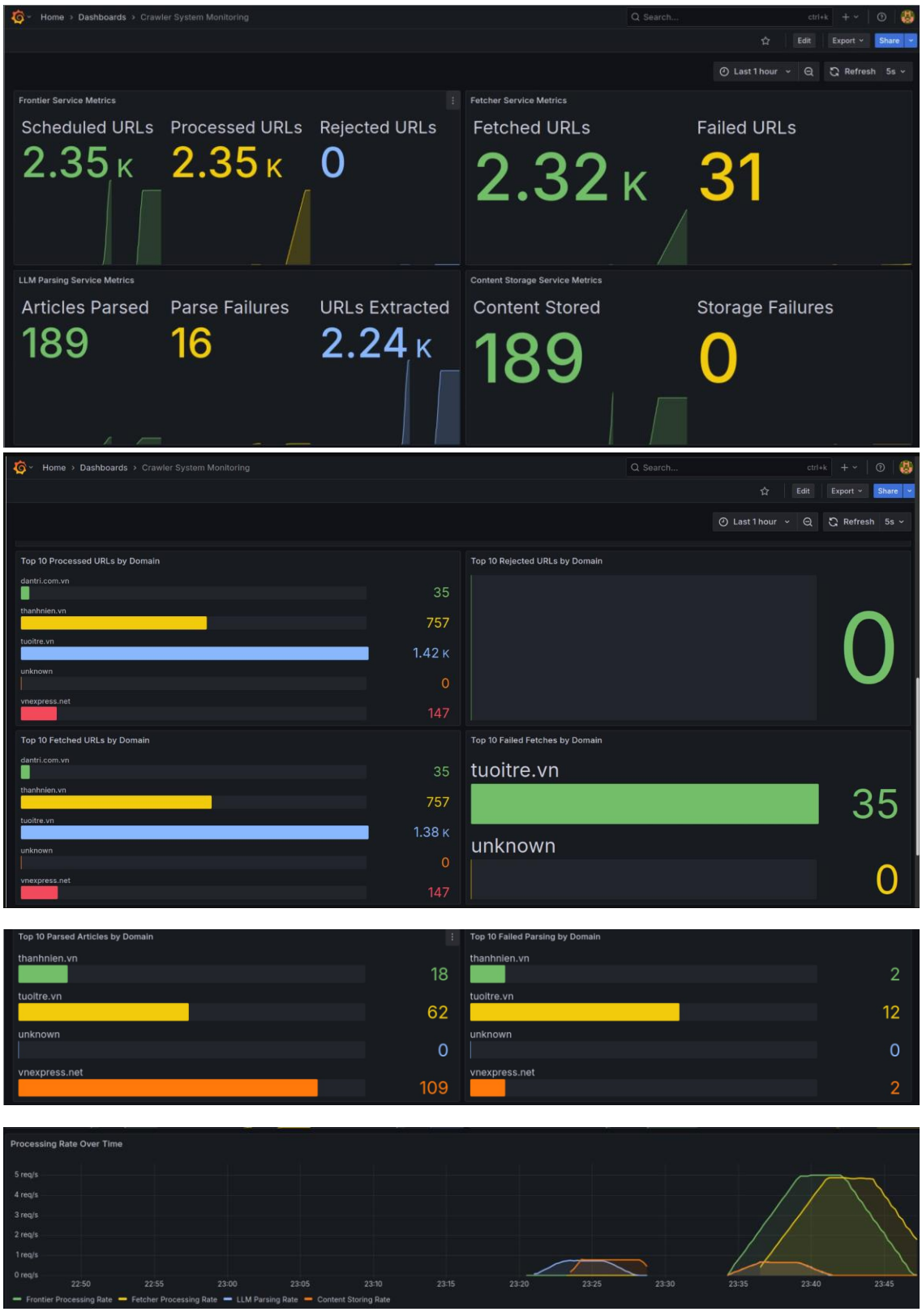
Cập nhật tiến độ tuần 3

Hoàn thiện flow, cập nhật một số thuật toán

1. Frontier:
   - Hiện tại đang bị trường hợp khi đưa vào các queue sẽ set duy nhất 1 key theo tính toán, khiên không linh hoạt, dễ bị tất cả các url, hay quá nhiều url bị vào chung 1 queue, khiến việc tách ra không có ý nghĩa, một số domain sẽ có thể bị chờ rất lâu mới đến lượt.
   - Cải tiến ở lớp queue selector 1: Phân phối lại key vào các queue cho đều hơn (nếu quá nhiều url cùng độ ưu tiên thì cũng sẽ chia thành các queue khác). Thuật toán lấy ra cũng sửa lại thành round robin với cài đặt thêm trọng số, để các queue có độ ưu tiên lớn hơn vẫn sẽ được lấy trước và lấy nhiều lần hơn, vẫn đảm bảo được sẽ lấy cả các queue có dộ ưu tiên thấp mà không phải chờ quá lâu.
2. Fetcher:
3. Parser:
   - Dùng Gemini để thực hiện lấy CSS selector, có thực hiện rate limit, block thread nếu đạt đến limit của gemini API. Nếu hết free hoặc lỗi thì sẽ dùng css selector default.
   - Có thực hiện ném lỗi nếu như phần nội dung parse được bị khuyết (chưa kiểm chứng được độ chính xác nhưng nhìn kết quả thì thấy khá ổn, vẫn focus được nội dung chính)
4. Storing
   - Thực hiện hiện lưu thông tin parse được vào elasticsearch (lưu luôn cả embedding vector)
   - Dùng model Gemini text-embedding-004 để embed nội dung cũng như phần đầu vào của tìm kiếm (so với của open AI thì không tốt bằng nhưng độ chính xác cũng khá tốt. Hơn hết nó free, 1500req/min, 2048 input token là đủ dùng cho bài toán này)
   - Triển khai các API search: có search theo keyword (trong nội dung, nội dung + title, title), semantic search (embedding title + content), hybrid (keyword có trọng số cao hơn). API lấy recent content.
5. Monitoring:
   - Thống kê số thành công, số lỗi, thành công/mỗi domain, lỗi / mỗi domain
   - Kết quả chạy 10 phút:



Có sự chênh lệch giữa số url xử lý của mỗi domain nhiều là do chưa clear message kafka cũ, từ trước lúc tối ưu thuật toán của frontier, nên số lượng url của một domain lúc đó bị nhiều, dồn dập. Với monitoring cũ thì chỉ có 1 domain thực process cho đến hết rồi mới đến các domain khác.

6. Mới chỉ thực hiện viết test với frontier, viết test các service khác, chỉ làm và chạy xem log để debug. Nếu còn thời gian sẽ tối ưu code, refactor, viết test