



企業名
日本経済新聞社



場所
Tokyo Japan



区分
メディア



製品
Elasticスタック



使用事例
ログ管理、可視化

日経電子版の記事検索および ログ解析の両方を1つの仕組みで実現

記事検索は、Elasticsearchと形態素解析のKuromojiおよびNgramを活用して全文検索システムを構築し、ログ解析は、ElasticsearchとKibanaを活用してログ可視化システムを構築。

5

クラスタ

クラスタの数

50

億

ドキュメント総数

2

TB

総データサイズ

1876年の創刊から140年の歴史を持つ日経新聞

日本経済新聞社が発行する日本経済新聞（以下、日経新聞）は、1876年に創刊。現在、販売部数270万部以上の日経新聞と、会員数45万人以上の日経電子版などの媒体を提供しています。2010年に創刊した日経電子版は、日本初の本格的な有料電子版サービスであり、PCやスマートフォン、タブレット端末、携帯電話など、複数のデバイスで利用できます。

日経電子版向けのアプリケーション開発とそれを運営するサーバ管理を担当しているのがデジタル編成局です。Pythonの開発フレームワークであるDjangoやElasticsearchを使って検索APIの開発およびスマートフォンアプリのバックエンドAPIの開発を行っています。また、Elasticsearch、Fluentd、Kibanaをアクセスログ解析ツールとして活用しています。

日経電子版は、2015年8月にデータセンターを「アマゾン ウェブ サービス (AWS)」に移行したのにあわせ、構成管理ツールのAnsibleや自動テストツールのServerspecなどを利用してサーバ環境を構築しています。またウェブサーバのNginx、プロキシサーバのVarnishおよびHAProxyのほか、AWS Elastic Beanstalk上でコンテナ管理ソフトのDockerも利用しています。

課題

日経電子版の記事検索機能および日経電子版を運営するサーバ環境におけるアクセスログ解析の両方を1つの仕組みで実現するためにはどうすればよいのか。

ソリューション

記事検索は、Elasticsearchと形態素解析のKuromojiおよびNgramを活用して全文検索システムを構築し、ログ解析は、ElasticsearchとKibanaを活用してログ可視化システムを構築。

利用範囲の広さでElasticsearchを採用

現在、Elasticsearchを利用しているのは、記事検索とサーバのログ解析の大きく2つの分野です。記事検索は、形態素解析にKuromojiやNgramを使用しており、ひとつひとつのコンテンツに対する検索回数は多いのですが、データ量はそれほど多くないという特徴があります。一方、ログ解析は、FluentdとKibanaを使って可視化しており、検索回数は多くありませんが、データ量が非常に多いという特徴があります。

Elasticsearchを選定したのは、記事検索とログの可視化の両方に使える利用範囲の広さが最大の理由です。日本経済新聞社 デジタル編成局 編成部の梅崎裕利氏は、「特に可視化ツールのKibanaが便利なほか、形態素解析などの豊富なプラグインが用意されていることがElasticsearchを採用した理由でした。またAPIの構造が理解しやすく、ブラウザからAPI経由でスキーマやクラスタの設定ができるのも便利でした」と語ります。

またAWS上でのスケールアウトやスケールアップ、ローリングアップデートなどが容易に運用できることもElasticsearchを採用した理由の1つ。さらに15台のクラスタ構成で、1秒あたり9000リクエスト以上という高い性能を発揮できることも魅力でした。梅崎氏は、「あまりチューニングをしなくても、高いパフォーマンスを発揮できたので、これは使えると感じて本格運用に踏み切りました」と話します。

Elasticsearchとの出会いについて梅崎氏は、「アプリケーション開発の担当者から記事検索ツールがほしいという要望があり、Solrの検証を開始しました。ちょうどそのころ、Elasticsearchのバージョン1.0が登場し、話題になっていたので検証を開始したのが最初でした。当時は、ログ解析への活用事例が多く、記事検索で利用しているという事例はあまり多くありませんでした」と当時を振り返っています。

記事検索とログ解析にElasticsearchを活用

日経電子版の記事は、5年分で約200万件が登録されており、そのデータサイズは5ギガバイトに上ります。記事は執筆されるごとに更新されるので、1日あたり数千回の更新があり、朝刊のデータが更新される午前3時ごろのピーク時には1分間に300件程度の記事が更新されます。現状、利用者からのアクセスは、1秒あたり100回程度ですが、今後は1秒あたり数千回のアクセスも想定しています。

Elasticsearchは、新規開発のアプリケーションから順にバックエンドで利用され始めています。たとえば「もっと日経」と呼ばれるアプリケーションがその1つ。もっと日経は、日経新聞の紙面をスマートフォンのカメラ機能で撮影すると、関連する記事やバックナンバーなどが検索されるサービスです。

梅崎氏は、「当初は、スマートフォンのカメラの性能などが原因で、OCRの精度がそれほど高くなく、OCRの結果で直接検索してもヒットしないという課題がありました。そこで、Elasticsearchの転置インデックスを利用してキーワードAPIを作成。記事中の単語の出現回数を抽出し、特徴量に活用して、重みづけを行い、クエリに反映する仕組みを実現することで、検索精度を大幅に向上しています」と話します。

またログ解析では、時刻やユーザーのIP、「GET」なのか「POST」なのか、ホスト名、パス、クエリ、エージェント、ステータスなど、必要なログはすべて収集し、Elasticsearchに保存する仕組みを構築しています。収集されるログのデータ量は、1秒間に1万件を超えており、1日あたりでは約3億件、120ギガバイトになります。このログを常時1週間分保持しています。

ログ解析用のElasticsearchは、AWS上で稼働する6台のサーバで運用されています。梅崎氏は、「ログ解析は、アクセス数がそれほど多くないので、6台で運用できます。構成としては、レプリカを作らず、サーバ1台に1つずつシャードを配置しています。現在、1シャードあたり20ギガバイト程度ですが、ベストプラクティスでは5ギガバイトということなので4分割も検討しています」と話しています。

秘伝の“タレ”なしにログ解析を実現

日経電子版の記事検索にElasticsearchを利用した最大のメリットは、新規アプリケーションの開発速度が格段に向上したことです。Elasticsearchは、ドキュメントが豊富で、ウェブサイトには各バージョンのドキュメントが用意されています。また今回、Elasticsearchを中心とした検索エンジン・APIを内製化したことにより、容易にチューニングや機能追加が可能になりました。技術者が少ない企業でも、手軽に採用できるメリットがあります。

さらにエイリアスを活用することで、本番稼働中でもスキーマを変更できるのも便利です。梅崎氏は、「クラウド上で稼働しているメリットとして、容易にサーバを拡張することができます。通常は3台のサーバで運用していますが、負荷が高くなったときには15台程度まで容易に拡張できます。またアップデートも容易で、シームレスにローリングアップデートができるのも便利です」と話します。

一方、ログ解析でElasticsearchを利用した最大のメリットは、専門的な深い知識がなくてもログ解析が可能になることです。ElasticsearchとKibanaを利用したログ解析により、たとえば、記事検索のレスポンスにどれくらいの時間がかかっているかを容易に解析できます。これにより、アプリケーションのどこにボトルネックがあるのか、改善すべき機能はどのなのかを迅速に把握できます。

またレスポンスのステータス一覧により、エラーやキャッシュミスが増えていないかどうかを解析できます。さらにGEO IPを利用することで、どの地域からのアクセスが多いのかを解析することもできます。そのほか、レスポンスサイズの合計をパスごとに表示することで、どのファイルが多くの帯域幅を使っているのかを解析できます。これにより、データ通信量の削減が期待できます。



ウェブサーバのApacheやNginxなどのアクセスログを、LTSV (Labeled Tab-separated Values) 形式で出力し、ログ収集・転送ツールのFluentdやログ管理ツールのLogstashを使ってElasticsearchに転送し、グラフ化ツールのKibanaで解析できる仕組みを構築しています。これにより潜在的な問題や課題を容易に把握し、トラブル発生時には即時に解決することができるようになりました。

- 梅崎氏

梅崎氏は、「トップページで使われているSVGファイルがgzipされていないことがログ解析により分かりました。SVGファイルをgzip対応することで、1カ月あたり1テラバイトの通信量を削減できました」と語ります。また、デプロイするときにサーバがダウンしていないか、サーバ台数を変更したときに正しくロードバランスしているかなども確認できます。さらに、Kibanaを利用することで、急にアクセス数が増えたときの原因調査も容易です。

そのほか、Kibana 4で複数アグリゲーションが利用できるようになり、複雑な解析も容易にできるようになっています。また以前は、ログの確認に早くも10分、遅いと丸1日かかっていましたが、ElasticsearchとKibanaを利用することで、1分程度に短縮できました。梅崎氏は、「ただしKibanaを見ていると楽しいので、いろいろなことを調べてしまい、結果的に時間短縮はできていません（笑）」と話しています。

今後はSenseやTimelionなどにも期待

今後の取り組みについて梅崎氏は、「Elasticsearchは、日経社内で広く活用されています。記事検索では、新規アプリケーションやプロトタイプ開発で利用され、アクセスログ解析は、3億件以上のログから、検索ログ、行動ログ、スロークエリなどを分析しています。今後は、既存の検索エンジンを置き換え、利用範囲も拡大する計画です。SenseやTimelionなど、新しいツールの活用も検討しています」と話している。



ElasticsearchとKibanaを活用することで、エンジニア秘伝の“タレ (shell芸)”を使って可視化していたログ解析をURLや画像で即座に社内共有できるようになりました。これまではログ解析ができるエンジニアが2人しかいなかったのですが、ElasticsearchとKibanaを利用することで、秘伝のタレが不要になり、ログ解析ができるエンジニアを6人以上に増やすことができました

- 梅崎氏