

OTIMIZAÇÃO DE PIPELINE PARA EXTRAÇÃO DE BIG DATA CONTENDO FOCOS DE CALOR DO PROGRAMA QUEIMADAS DO INPE

Bruno Denardo¹; Abner Rodrigo da Silva Costa²; Fabrício Galende Marques de
Carvalho³; Fabiano Morelli⁴

¹ Instituto Nacional de Pesquisas Espaciais. (bruno.denardo@inpe.br).

² Instituto Nacional de Pesquisas Espaciais. (abner.costa@inpe.br).

³ Instituto Nacional de Pesquisas Espaciais/Fatec São José dos Campos. (fabricao.galende@inpe.br).

⁴ Instituto Nacional de Pesquisas Espaciais. (fabiano.morelli@inpe.br).

Resumo: O Programa Queimadas, do Instituto Nacional de Pesquisas Espaciais (INPE), recebe, processa e disponibiliza ao público, desde 1998, os focos de calor, que representam queimadas ou incêndios florestais. Esses dados são tratados e armazenados em bancos de dados relacionais que ocupam vários terabytes. Foi nesse contexto que surgiu a demanda por um serviço web capaz de filtrar e exportar esses dados em CSV e em formatos georreferenciados. A solução originalmente construída no instituto foi desenvolvida de forma modular, com microsserviços em containers (serviços de recepção, geração de arquivos, envio de link e download). Porém, devido ao alto volume de dados demandados pelos usuários, passou a apresentar instabilidades recorrentes por comunicação direta e encadeada entre serviços, além de fragilidades no tratamento de requisições e na forma de consultas feitas ao banco de dados, o que prejudicava o desempenho, aumentava o consumo de recursos e dificultava a manutenção. Os objetivos deste projeto foram melhorar os aspectos de estabilidade operacional, desempenho e segurança, além de deixar a solução mais flexível e configurável, para ser reaproveitada em outros cenários de exportação de grandes volumes de dados do programa. Além disso, o projeto se justifica pela necessidade de garantir a continuidade da disponibilização pública desses dados, especialmente em períodos de alta demanda. Quanto à metodologia, foi feito um diagnóstico do que existia, identificando onde estavam os gargalos. Aplicou-se o desenvolvimento incremental, com monitoramento e testes, para garantir que cada mudança fosse funcional e integrada ao restante. Também foi projetada uma arquitetura baseada em configuração, permitindo via arquivos ajustar rotas da API, validações e as extrações no banco de dados, sem precisar alterar código a cada novo contexto de aplicação. Quanto aos resultados, uma mudança estrutural significativa foi intermediar a comunicação entre serviços: criou-se uma fila com Redis, em container próprio e desacoplado, para segurar a carga entre a recepção e a geração, reduzindo sobrecarga e abrindo espaço para escalabilidade horizontal quando necessário. Um cache foi implementado para registrar consultas já processadas e evitar recomputação de arquivos prontos, esvaziando a fila e reduzindo a latência na entrega dos dados. Também foi refatorado o módulo de geração de arquivos, que antes carregava resultados inteiros em memória e não possuía escalabilidade. Na geração dos formatos georreferenciados, foi substituído o GeoPandas pela ferramenta ogr2ogr da GDAL, reduzindo o tempo de criação em mais de cinquenta por cento em cenários típicos e permitindo consultas em

lotes, com melhor controle de RAM. A mesma lógica de lotes foi aplicada aos dados tabulares com pandas. Sob a ótica da segurança, além de validações mais rígidas, foram limitadas as requisições excessivas por janela de tempo e foram aplicadas medidas contra injeção SQL. Com esse trabalho, conclui-se que o projeto envolvendo desacoplamento por fila, aliado à geração em lotes e ao cache, melhorou a estabilidade, o uso de recursos e o tempo percebido pelo usuário. O aspecto configurável aumentou a flexibilidade sem comprometer a segurança, facilitando a evolução e a manutenção da pipeline de extração.

Palavras-chave: Big Data; Microsserviços; Dados geoespaciais; Exportação de dados.