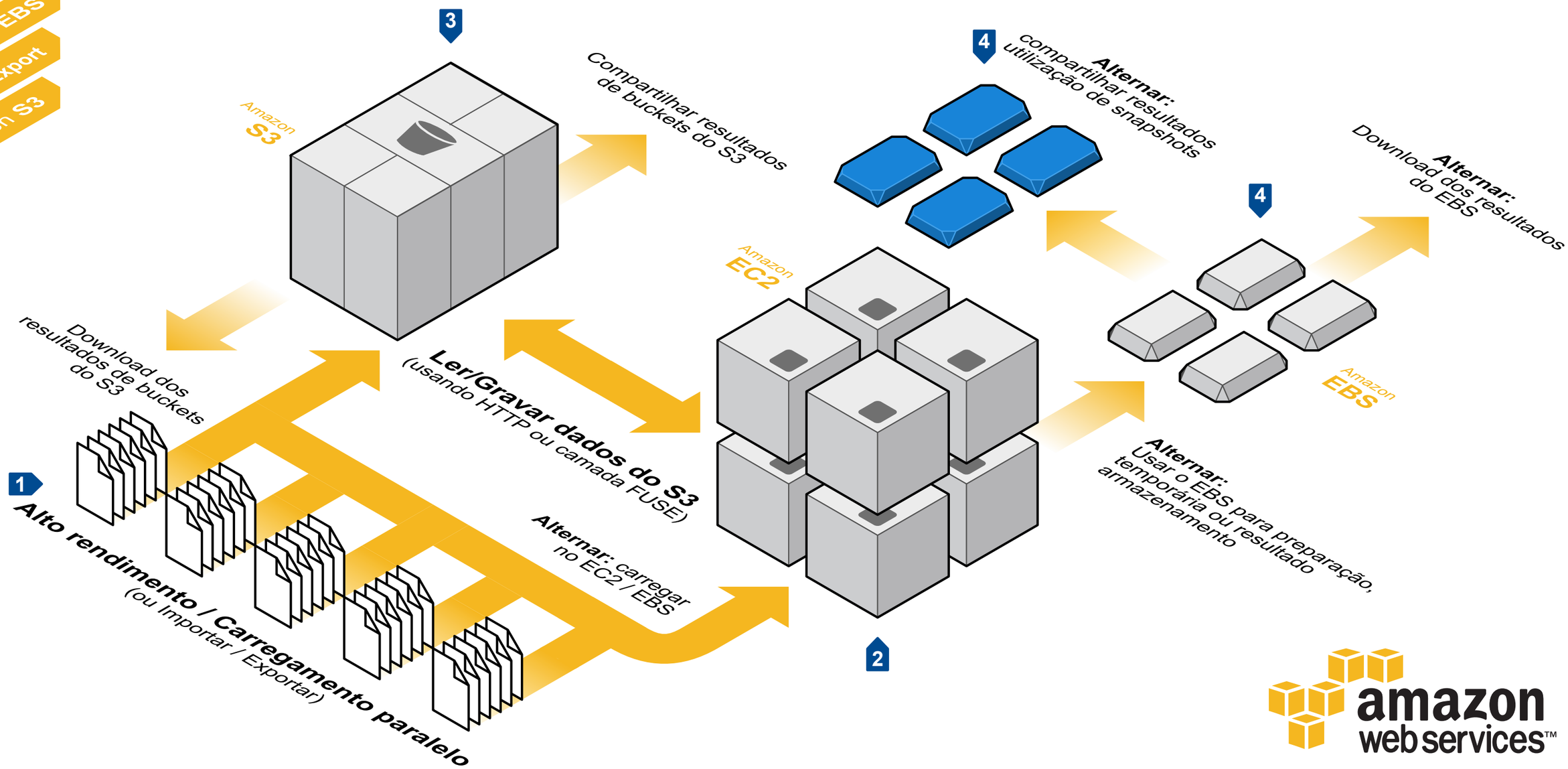
AWS Referência Arquiteturas

- Amazon EC2
- Amazon EBS
- AWS Import/Export
- Amazon S3

COMPUTAÇÃO EM LARGA ESCALA & CONJUNTOS DE DADOS ENORMES

A Amazon Web Services é bastante popular para cenários de computação em larga escala, como computação científica, simulação e projetos de pesquisa. Esses cenários envolvem conjuntos de dados enormes coletados de equipamentos científicos, dispositivos de medição ou outros trabalhos de computação. Após a coleta, esses conjuntos de dados precisam ser analisados por trabalhos de computação em larga escala para gerar conjuntos de dados de resultados. De forma ideal, os resultados serão disponibilizados assim que os dados forem coletados. Com frequência, esses resultados serão disponibilizados para um público maior.



Sistema Visão geral

1 Para carregar conjuntos de dados grandes no AWS, é fundamental usufruir ao máximo a largura de banda disponível. Você pode fazê-lo ao carregar dados no **Amazon Simple Storage Service (S3)** em paralelo a vários clientes, cada um deles usando multithreading para habilitar carregamentos simultâneos ou carregamentos de multipartes para outras paralelizações. As configurações TCP como o escalonamento de janelas e a confirmação seletiva podem ser ajustadas para aprimorar ainda mais o rendimento. Com as otimizações adequadas, é possível carregar vários terabytes por dia. Outra alternativa para conjuntos de dados enormes poderá ser **Amazon Import/Export**, que oferece suporte ao envio de dispositivos de armazenamento à AWS e à inserção de seu conteúdo diretamente em volumes do **Amazon S3** ou **Amazon EBS**.

2 O processamento paralelo de trabalhos em larga escala é essencial e os aplicativos paralelos existentes normalmente podem ser executados em várias instâncias do **Amazon Elastic Compute Cloud (EC2)**. Um aplicativo paralelo às vezes poderá considerar grandes áreas de rascunho que todos os nós podem ler e fazer gravações com eficiência. O S3 pode ser usado como uma área de rascunho, seja diretamente usando HTTP ou uma camada FUSE (por exemplo, s3fs ou SubCloud) se o aplicativo esperar um sistema de arquivos no estilo POSIX.

3 Assim que o trabalho for concluído e os dados do resultado forem armazenados no **Amazon S3**, **Amazon EC2** poderão ser desativadas e o conjunto de dados do resultado poderá ser baixado. Os dados resultantes poderão ser compartilhados com outras pessoas, seja ao conceder permissões de leitura para selecionar usuários ou para todos, ou ao usar URLs limitadas por período.

4 Em vez de usar o **Amazon S3**, você pode usar o **Amazon EBS** para preparar um conjunto de entrada, atuar como uma área de armazenamento temporário e/ou capturar o conjunto resultante. Durante o carregamento, os conceitos de fluxos de carregamento paralelo e o tweak de TCP também serão aplicados. Além disso, carregamentos que usam UDP poderão aumentar ainda mais a velocidade. O conjunto de dados resultante pode ser gravado em volumes de EBS, no qual os snapshots por período dos volumes podem ser obtidos para compartilhamento.