accenture

Strategy | Digital | Technology | Operations

High performance. Delivered.

Aceleração de dados:
Arquitetura para a cadeia de suprimento de dados em Big Data

Visão Geral

As tecnologias de cadeias de suprimentos estão evoluindo rapidamente, mas as organizações têm adotado a maioria dessas tecnologias de forma fragmentada. Como resultado, a utilização dos dados corporativos, independente de se relacionarem a interações com clientes, desempenho comercial, notificações por computador ou eventos externos no ambiente corporativo, está imensamente abaixo do seu potencial. Além disso, os ecossistemas de dados das empresas tornaram-se complexos e repletos de silos. Isso dificulta o acesso aos dados, o que por sua vez limita o valor que as organizações poderiam obter deles. De fato, de acordo com um recente relatório da Gartner, Inc., 85% das organizações da Fortune 500 serão incapazes de explorar esses grandes volumes de dados (ou big data) para obterem vantagem competitiva até 2015.[1] Além disso, um estudo recente da Accenture descobriu que metade de todas as empresas se preocupa com a precisão dos seus dados, e a maioria dos executivos não tem certeza sobre os resultados comerciais obtidos dos seus programas de análise de dados.[2]

Para poderem aproveitar o valor oculto em seus dados, as empresas devem começar a tratá-los como uma cadeia de suprimento, permitindo que fluam de forma fácil e útil em toda a organização e, finalmente, em cada ecossistema de parceiros da empresa, incluindo fornecedores e clientes. É o momento certo para essa abordagem. Por um lado, novas fontes externas de dados se tornam disponíveis, proporcionando novas oportunidades para insights. Além disso, as ferramentas e a tecnologia necessárias para construir uma melhor plataforma de dados estão disponíveis e já estão sendo usadas. Estas fornecem uma base sobre a qual as empresas podem construir uma cadeia de suprimento de dados integrada e de ponta a ponta.

1. "Big Data Business Benefits Are Hampered by 'Culture Clash'," ("Benefícios comerciais com dados grandes são impedidos pelo 'Choque Cultural'"), Gartner, 12 de setembro de 2013.

2. "Journey to Analytics ROI", ("Jornada para o ROI em Analítica"), 27 de fevereiro de 2013.

Uma cadeia de suprimento de dados moderna começa quando dados são criados, importados ou combinados entre si. Os dados se movem através dos elos da cadeia, adquirindo valor de modo incremental. A cadeia de suprimento termina com insights comerciais acionáveis e valiosos — tais como ideias para um novo produto, serviço ou inovações no processo, campanhas de marketing ou estratégias de globalização. Configurada e gerenciada de forma eficaz, uma cadeia de suprimento de dados permite que as empresas descubram seus dados, aproveitem mais fontes de dados e os acelerem. Essas capacidades, por sua vez, posicionam uma organização para extrair mais valor dos seus dados através de técnicas de computação avançadas, como aprendizado de máquina.

A aceleração de dados desempenha um papel importante em uma cadeia robusta de suprimento de dados. Em sua forma mais simples, a aceleração de dados deriva-se de ferramentas e técnicas que permitem que quantidades maciças de dados sejam ingeridas (transportadas de sua fonte para um sistema desenhado para dados), armazenadas e acessadas em velocidades estonteantes. Especificamente, com a aceleração de dados, as organizações conquistam rápido acesso a dados valiosos, que lhes permitem executar a análise dos dados, obter insights e tomar medidas cabíveis na janela de oportunidade às vezes muito pequena disponível para as empresas. A aceleração de dados, portanto, ajuda as organizações a superar três desafios relacionados a dados: movimentação, processamento e interatividade.

Sob essa perspectiva, a área de Big Data da Accenture, em colaboração com a Accenture Technology Labs, examina atentamente esses desafios e avalia a paisagem de componentes arquitetônicos disponível para abordá-los. Depois, exploramos opções para combinar esses componentes para criar soluções de plataforma de dados.

Três desafios que a aceleração de dados pode abordar

A aceleração de dados ajuda as organizações a enfrentar três desafios: como mover os dados rapidamente da sua origem até os pontos da organização onde são necessários, como processá-los para obter insights acionáveis com a maior rapidez possível e como promover respostas mais rápidas às consultas enviadas por usuários ou aplicativos — o que é conhecido como interatividade.

Movimentação

Tradicionalmente, trazer dados para uma empresa era um processo lento, mas razoavelmente simples: os dados eram reunidos em uma área de preparação e então transformados para o formato apropriado. Depois eram carregados para residirem em uma fonte, tal como um mainframe ou armazém de dados corporativo. Desse ponto, eram transferidos diretamente, de um modo ponto a ponto, para um data mart para o acesso de usuários e aplicativos. No entanto, com o gigantesco aumento nos volumes e variedades de dados, tal processo tradicional já não funciona de forma eficaz.

A Internet das Coisas (IoT) desempenha um papel importante para a geração de novos avanços na movimentação de dados. Em seu sentido mais simples, a IoT é composta de dispositivos conectados — que cobrem desde refrigeradores, medidores inteligentes e câmeras de vídeo a telefones celulares e brinquedos — que podem estar localizados em qualquer lugar do mundo. De acordo com a Gartner, Inc., haverá até 26 bilhões de dispositivos na IoT em 2020.[3] Cada dispositivo conectado gera dados, cada um com seu próprio formato e idiossincrasias.

Independente de uma empresa colocar em produção milhares de sistemas individuais ou simplesmente tentar acompanhar o seu próprio crescimento, o fato de ter uma infraestrutura de dados moderna em operação que possa coletar dados relevantes pode levar à diferenciação, permitindo insights baseados nos dados. Contudo, para extraírem insights úteis dos dados neste novo mundo, as empresas precisam colhê-los de várias fontes sem perdas e entregá-los para processamento e armazenamento. Alguns dados existem na forma de arquivos de log em sistemas externos que precisam ser transportados até a infraestrutura de dados de uma organização para uso futuro. Outras fontes fornecem dados por streaming, que são canalizados para o sistema em tempo real, isto é, à medida que os dados são gerados. Os exemplos incluem informações de consumo de energia provenientes de medidores de energia que estão em constante atualização.

Qualquer que seja a fonte ou o formato, mover os dados de sua origem até onde sejam necessários na organização pode parecer-se com beber água em uma mangueira de incêndio e tentar não perder uma única gota. A aceleração de dados ajuda as organizações a gerenciar essa façanha, permitindo várias formas de levá-los até a infraestrutura de dados de uma empresa e garantir que possam ser consultados rapidamente.

Processamento

Há tempos as empresas processam dados, em um esforço para extrair insights relevantes. No entanto, o volume e a variedade de dados que exigem processamento expandiram-se absurdamente. Para acomodar o crescimento nessas duas frentes e gerar resultados mais rápidos, mas também exatos, as empresas precisam intensificar suas capacidades de processamento. Em particular, elas devem realizar três atividades, com rapidez nunca vista: executar cálculos nos dados, criar e executar modelos de simulação e comparar estatísticas para obterem novos insights a partir dos dados.

A ascensão das tecnologias analíticas em tempo real tem apresentado novas oportunidades nessa frente. Uma boa tecnologia analítica pré-processa os dados recebidos. Por exemplo, ao monitorar a localização de um cliente, uma organização pode enviar uma promoção ou desconto para o dispositivo móvel desse cliente quando ele estiver próximo a um local provável de compra. Contudo, uma tecnologia melhor combina dados em streaming com dados históricos (modelados), para permitir a tomada de decisão mais inteligente. Por exemplo, correlacionando a localização de um cliente com seu histórico de compras anteriores, a empresa pode oferecer uma promoção adaptada para aquele cliente, aumentando a probabilidade de conversão.

3. "Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020." ("Gartner afirma que a base instalada de Internet das Coisas chegará a 26 bilhões de unidades até 2020"), Gartner, 12 de dezembro de 2013.

Para colherem os benefícios integrais do processamento de dados com maior rapidez, as empresas devem fazer melhor uso de clusters de computadores — conjuntos organizados de centenas ou milhares de computadores trabalhando juntos para filtrarem grandes quantidades de dados. Com o custo da memória de acesso aleatório (RAM) em seu ponto mais baixo já visto, novas soluções para a extração de dados de depósitos mais rapidamente têm bombardeado o mercado, cada uma delas prometendo velocidade, durabilidade e precisão.

A aceleração de dados presta apoio ao processamento, alavancando avanços no hardware e software para clusters de computadores, permitindo-lhes operar com eficiência nunca vista.

Interatividade

A interatividade tem a ver com a capacidade de uso da infraestrutura de dados. Fundamentalmente, usuários ou aplicativos enviam consultas para a infraestrutura e esperam receber respostas às consultas em um período aceitável de tempo. As soluções tradicionais facilitaram para as pessoas o envio de consultas para a obtenção dos resultados necessários, com o objetivo de se chegar a insights relevantes. No entanto, o surgimento do Big Data levou a novas linguagens de programação que desencorajam a adoção dos sistemas pelos usuários atuais. Além disso, devido ao grande volume de dados, os usuários podem ter de aguardar vários minutos ou mesmo horas para obterem resultados em uma consulta.

Quanto maior o tempo de espera, mais tempo os usuários levam para obter os insights necessários para a tomada de decisões comerciais e para o atendimento às expectativas dos seus clientes. Isso ocorre, independente de os clientes serem internos (como o diretor de marketing que deseja saber quais são os clientes mais leais e rentáveis da empresa) ou externos (por exemplo, uma empresa cliente de Business Process Outsourcing (BPO) que precisa saber como o desempenho de um processo terceirizado alterou-se, durante a duração do contrato de BPO). Clientes que oferecem serviços críticos para os seus próprios clientes — tais como processamento de transações de varejo — talvez exijam tempos de resposta na faixa de milésimos de segundo. Com tarefas comerciais menos críticas, os tempos de resposta aceitáveis podem ser mais longos.

A aceleração de dados presta apoio à interatividade mais ágil, permitindo que usuários e aplicativos se conectem à infraestrutura de dados de modos aceitos universalmente e garantindo que os resultados das consultas sejam entregues com a rapidez necessária.

Entendimento sobre a paisagem da arquitetura

As organizações podem escolher entre muitos componentes diferentes da tecnologia de dados para a construção da arquitetura necessária para apoio à aceleração de dados. Incluem-se aí as plataformas de Big Data, processamento de eventos complexos, ingestão, bancos de dados na memória, clusters de cache e equipamentos. Cada componente pode abordar a movimentação de dados, processamento e/ou interatividade, e cada um tem características de tecnologia que o diferenciam. Nas seções a seguir, analisaremos com maior atenção esses componentes.

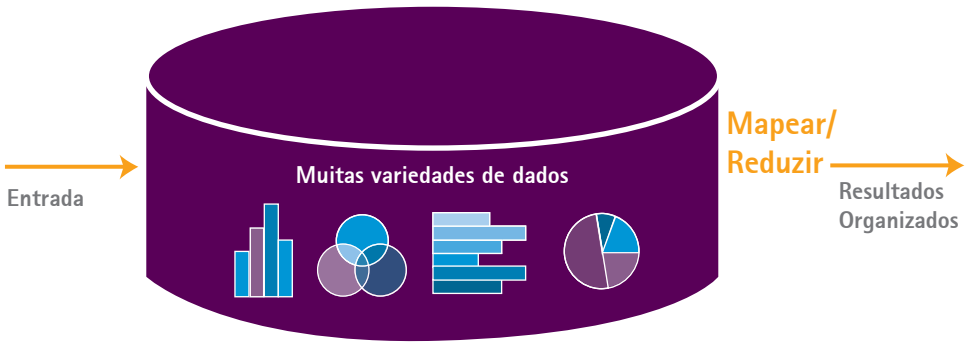
Plataforma de Big Data

Uma plataforma de Big Data (BDP, de Big Data Platform) é um sistema de arquivos distribuídos e mecanismo de computação que pode ser usado para facilitar a movimentação e processamento de dados. BDPs contêm o que chamamos de um núcleos (BDC, de big data core) – um cluster de computadores com armazenamento de dados distribuídos e poder de computação. Os avanços em tecnologias de Big Data permitem que BDCs funcionem como uma plataforma para tipos adicionais de computação, alguns dos quais (como mecanismos de consulta) podem apoiar especificamente a interatividade de dados.

Tradicionalmente, o sistema de arquivos de núcleo de Big Data pode utilizar técnicas como replicação e sharding (partição em um banco de dados que separa bancos de dados muito grandes em partes menores, mais rápidas e de manejo mais fácil) para acelerar e expandir o armazenamento de dados. Além disso, essas técnicas podem ajudar a fortalecer as capacidades de processamento. Adições mais recentes permitem o uso mais poderoso da memória de núcleo como um armazém de dados de alta velocidade, dando suporte à movimentação, processamento e interatividade de dados. Esses aperfeiçoamentos permitem a computação in-memory em um cluster existente de computadores. Além disso, tecnologias de streaming adicionadas ao núcleo podem permitir processamento de eventos complexos em tempo real, e tecnologias de analítica in-memory dão suporte a melhor interatividade de dados.

Aperfeiçoamentos adicionais ao núcleo de Big Data têm seu foco na criação de interfaces rápidas e conhecidas com dados no cluster. Tipicamente, o núcleo armazena dados semiestruturados (como XML e JSON) e dados não estruturados (por exemplo, documentos em Word, PDFs, arquivos de áudio e vídeos) e requer funcionalidade de MapReduce para a leitura. O software de mecanismo de consulta permite a criação de tabelas de dados estruturados no núcleo e funcionalidade de consultas comuns (como SQL).

Figura 1: Plataforma de Big Data



Ingestão

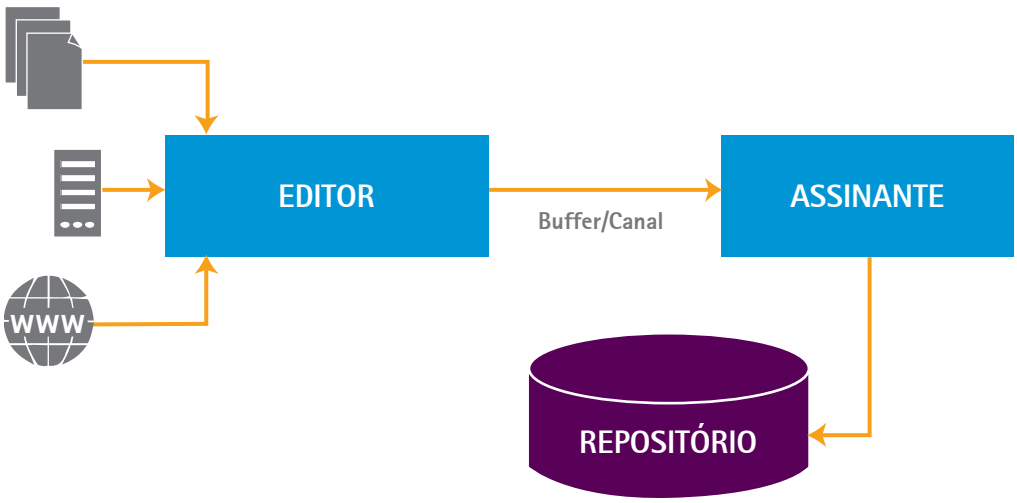
A ingestão diz respeito ao processo de reunir, capturar e movimentar dados de suas origens até repositórios subjacentes, onde os usuários podem processá-los. A ingestão tradicional era executada com um método de extrair-transformar-carregar (ETL, em inglês), que visava garantir dados organizados e completos. A infraestrutura moderna de dados preocupa-se menos com a estrutura dos dados quando entram no sistema e mais com garantir que eles sejam de fato coletados. As técnicas modernas atuam sobre dados de streaming, como cliques contínuos em um site da Web, e envolve filas (processamento dos dados na ordem apropriada).

Como observamos anteriormente, as organizações precisam de um mecanismo para capturar dados de várias fontes externas (cada uma das quais poderia entregar dados em diferentes formatos e ter diferentes requisitos) e transportá-los rapidamente até um local onde os usuários possam acessá-los para o processamento. Os dados podem ser estáticos e residir em um repositório externo à infraestrutura de dados da organização – ou podem ser gerados em tempo real pela origem externa. Soluções de ingestão oferecem mecanismos para acesso e utilização dos dados em ambos os cenários.

Nesse sistema de "editor-assinante", o produtor dos dados publica-os a partir da fonte para uma área de armazenamento temporário (buffer) ou canal (área de retenção dos dados). O assinante (usuário ou consumidor) dos dados pode colhê-los desse local. Um mecanismo de enfileiramento permite que os dados sejam armazenados temporariamente, enquanto o sistema aguarda que os produtores e consumidores adotem as ações relevantes. A velocidade das ações dos produtores e consumidores dos dados determina o tamanho do armazenamento temporário e da fila.

A ingestão robusta auxilia na aceleração dos dados, permitindo que grandes volumes de dados sejam coletados e armazenados rapidamente.

Figura 2: Ingestão



Processamento de eventos complexos

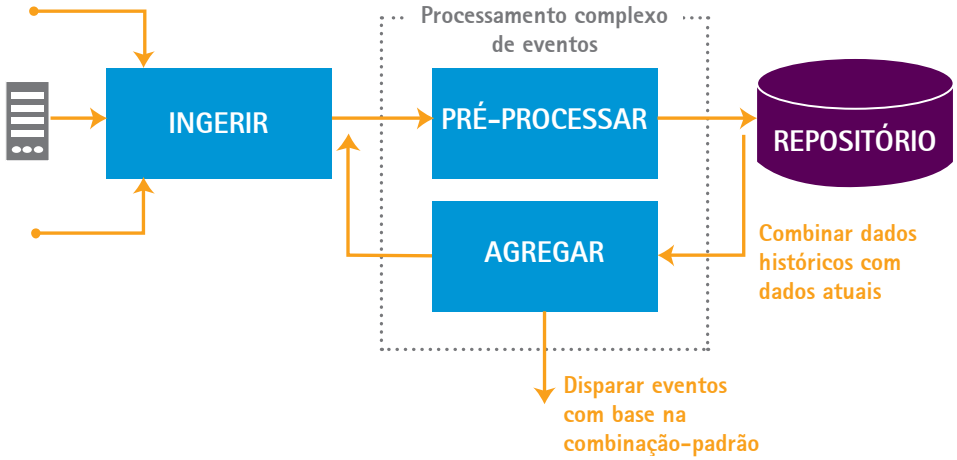
O processamento de eventos complexos (CEP, em inglês) é um método de rastreamento e análise (processamento) de dados sobre eventos (como fluxos de cliques ou feeds de vídeo) e extração de uma conclusão. Um exemplo rápido é o da validação de eventos de segurança contra violações de informações verificadas anteriormente em tempo real para a avaliação de novas ameaças. O processamento de eventos complexos combina dados de várias fontes para inferir eventos ou padrões que sugerem circunstâncias mais complicadas. Ele tem por objetivo identificar eventos significativos (como oportunidades ou ameaças) e permitir que as empresas respondam a eles com a máxima rapidez possível.

O processamento de eventos complexos é particularmente útil para a realização de análises e a extração de insights em tempo real. À medida que os dados são enviados de suas fontes, esses mecanismos realizam o processamento prévio e transformações iniciais para:

- Processar partes da informação e utilizar os totais para agilizar o processamento futuro de lotes maiores, combinando dados históricos com dados novos.
- Combinar os dados com padrões predeterminados, bem como inferir novos padrões nos dados.
- Disparar eventos e ações com base em padrões detectados e entregar insights em tempo real para os tomadores de decisões.

A principal vantagem do CEP é o imediatismo dos insights e ações que proporciona, em comparação com a necessidade de os usuários precisarem aguardar pela conclusão de uma tarefa de processamento de lote durante a noite. A velocidade maior de processamento deriva-se do fato de que a movimentação e processamento dos dados ocorrem em paralelo, apoiados por computações in-memory. Essas soluções diferem de soluções de ingestão, no sentido de terem poder de processamento adicional para a execução de cálculos sobre os dados iniciais, antes de serem consumidos pelo armazém de dados ou sistema de arquivos.

Figura 3: Processamento de eventos complexos

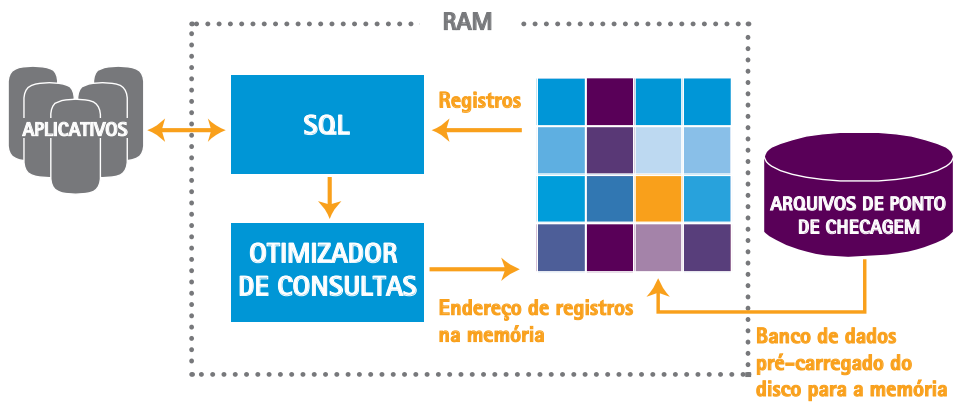


Banco de dados in-memory

Um banco de dados in-memory (IMDB) é um sistema de gestão de banco de dados que utiliza principalmente a memória principal para o armazenamento de dados no computador. Ele difere de sistemas de gestão de banco de dados que utilizam um mecanismo de armazenamento em disco. Bancos de dados in-memory são mais rápidos, porque os algoritmos internos são mais simples e executam menos instruções na unidade de processamento central. Além disso, acessar dados na memória elimina o "tempo de busca" envolvido na consulta de dados no armazenamento em disco, proporcionando assim um desempenho mais veloz e previsível.

Uma vez que IMDBs restringem todo o banco de dados e os aplicativos a um único espaço de endereço, eles reduzem a complexidade da gestão dos dados. Todos os dados podem ser acessados em milionésimos de segundos. IMDBs não são novos, mas as reduções nos preços de RAM e os aumentos progressivos na capacidade de RAM dos servidores os transformaram em opções altamente econômicas.

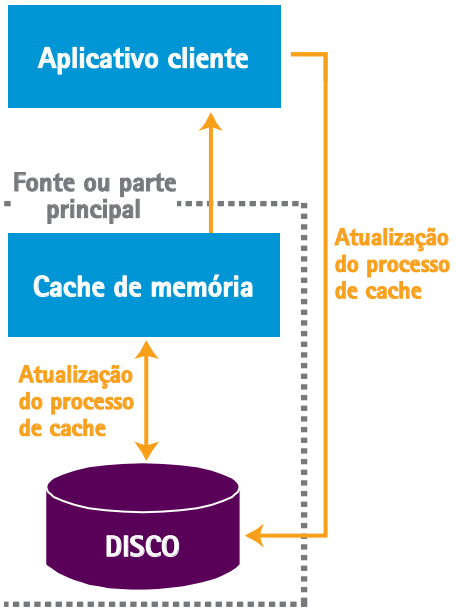
Figura 4: Banco de dados in-memory



Clusters de cache

Clusters de cache são agrupamentos de servidores nos quais a memória é gerenciada por um software central que transfere a carga de fontes de dados a montante, como banco de dados, para aplicativos e usuários. Os Clusters de cache geralmente são mantidos na memória e podem fornecer acesso de alta velocidade a dados acessados com frequência. Eles se situam entre a fonte dos dados e o consumidor dos dados. Os clusters são usados quando o volume de leituras de diferentes fontes de dados que não mudam com frequência é extremamente alto, ou quando um banco de dados é armazenado no disco, onde o tempo de busca pode não ser o ideal.

Figura 5: Cluster de cache

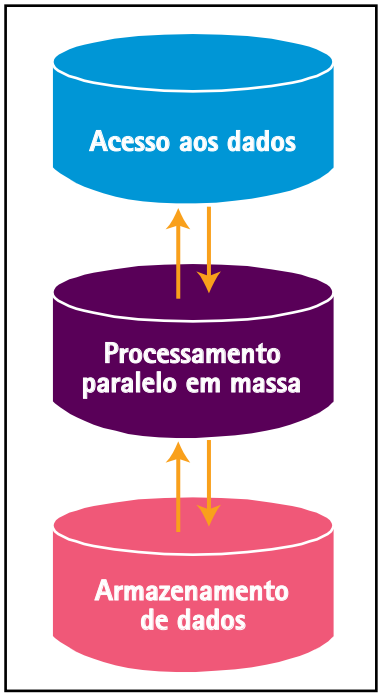


Clusters de cache executam operações de cache em grande escala. Tradicionalmente, eles são adequados para operações simples, tais como valores de leitura e escrita. Muitas vezes, eles são preenchidos quando uma consulta é enviada por um consumidor de dados para uma fonte de dados. Os resultados da fonte de dados são então armazenados no cluster de cache. Assim, se a mesma consulta ocorre novamente, não é preciso seguir todo o trajeto até a fonte dos dados para a recuperação pelo consumidor. "Recibos" de consultas se acumulam com o tempo no cluster. Quando um consumidor solicita dados armazenados no cluster, este responde recorrendo à fonte de dados – a menos que determinados padrões sejam atendidos (geralmente, o tempo desde a última atualização). O preenchimento prévio de um cluster de cache (também conhecido como "aquecimento") com dados que sabidamente são acessados com frequência pode reduzir o estresse nos sistemas subjacentes após uma reinicialização do sistema. Grades de dados levam a operação de cache um passo adiante, adicionando suporte para operações de consulta mais complexas e para certos tipos de computações de processamento paralelo maciço (MPP, em inglês).

Equipamento

Um equipamento é um conjunto de hardware pré-empacotado ou pré-configurado (servidores, memória, armazenamento e canais de entrada/saída), software (sistema operacional), sistema de gestão de banco de dados e software de gestão administrativa) e serviços de suporte. Ele é vendido como uma unidade, normalmente com redundância de hardware integrada, que ajuda a confirmar que o dispositivo continuará disponível em caso de falha de um componente. Um equipamento pode ter um banco de dados comum para o processamento de transações online e o processamento de análises online. Isso reduz atrasos na movimentação de dados, processamento, e interatividade, auxiliando assim na aceleração dos dados.

Figura 6: Equipamentos



Os equipamentos fazem uso de tecnologias semelhantes a núcleos de Big Data que fornecem paralelismo de processamento. Utilizando a arquitetura de MPP, os equipamentos podem dar suporte a bancos de dados mais rápidos e com alto desempenho e ampliá-los quando as cargas aumentam.

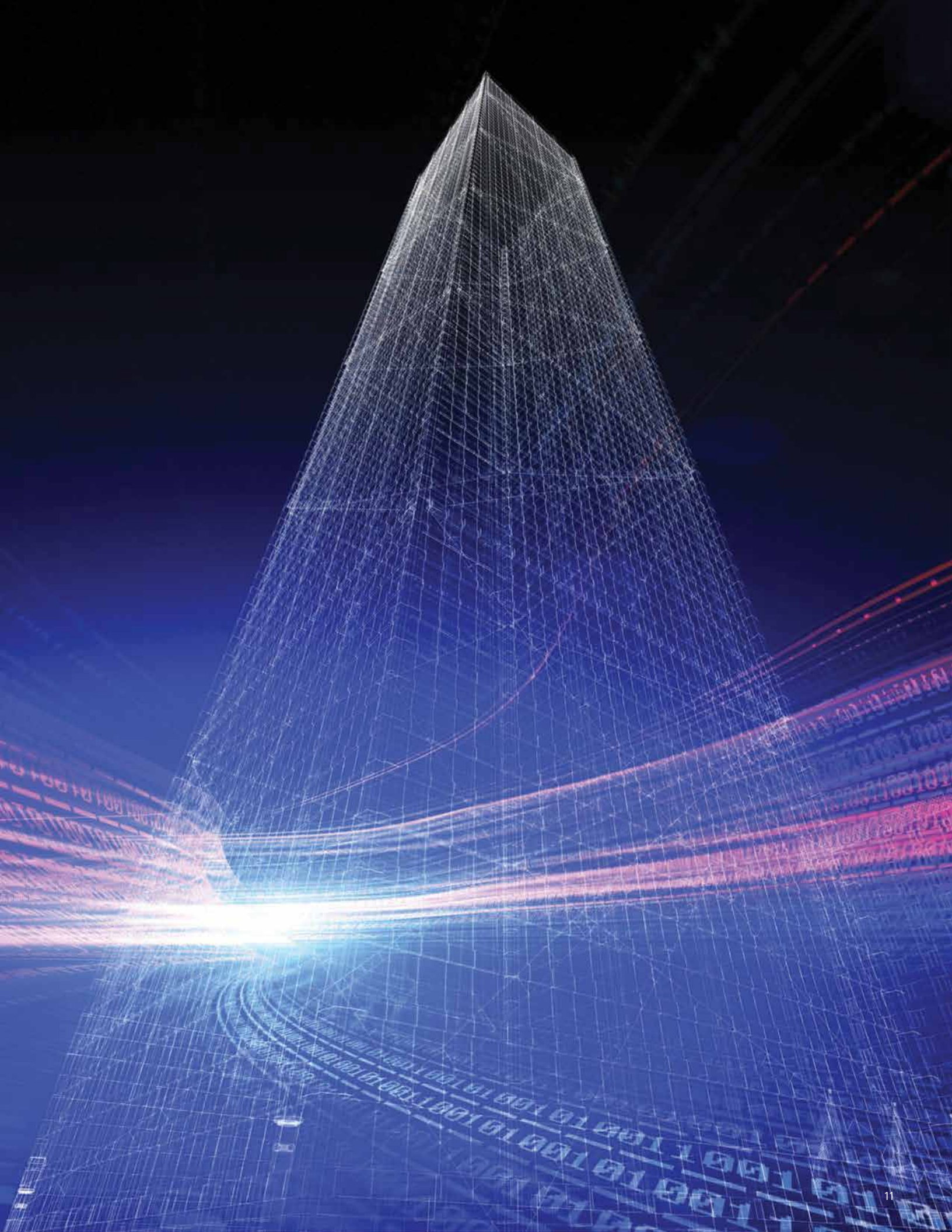
Bancos de dados de alto desempenho em execução em um cluster de servidores são complicados de implementar e requerem conhecimento especializado sobre gestão de sistema, banco de dados e armazenamento. Em organizações que não possuem tal conhecimento, os funcionários administrativos ou de TI podem não adotar tais bancos de dados de boa vontade. A manutenção do sistema e a atualização de software também consomem muito tempo para os administradores de sistema que trabalham nesses bancos de dados. Para essas organizações, os equipamentos proporcionam um modo mais fácil de se obterem os benefícios de banco de dados de alto desempenho, evitando-se os desafios. A maioria dos equipamentos fornece a infraestrutura e as ferramentas necessárias para a construção de aplicações de alto desempenho – incluindo qualquer coisa, desde tecnologia de banco de dados de núcleo e serviços de replicação em tempo real até a gestão do ciclo de vida e provisionamento de dados.

Pelo lado do hardware dos equipamentos, circuitos impressos "de silício personalizado", que não estão disponíveis para uso fora do equipamento, oferecem benefícios imensos. Um exemplo é o uso de silício personalizado em circuitos integrados específicos à aplicação, que permitem que os desenvolvedores criem soluções exclusivas, adaptadas para as suas necessidades específicas. O silício personalizado também permite o desenvolvimento em dispositivos otimizados para casos de uso específicos, sem o custo de desenvolver individualmente toda a propriedade intelectual subjacente. O silício personalizado para a otimização da rede, por exemplo, fornece uma solução exclusiva, que integra lógica interna, memória, tecnologia de serializador/desserializador, núcleos de rede e núcleos de processador, que podem gerar ganhos adicionais de desempenho, proporcionando vantagens sobre soluções não personalizadas.

Graças a esses recursos avançados, os equipamentos podem dar suporte e executar cálculos complexos em volumes enormes de dados em toda a empresa. Assim, os tomadores de decisões podem analisar imensos volumes de dados em tempos de resposta sem precedentes, com flexibilidade impressionante, sem a necessidade de suporte constante e auxílio dos fornecedores. Para muitas organizações, este aspecto de "plug-and-play" dos equipamentos apresenta um apelo considerável.

Componentes de arquitetura e recursos de tecnologia

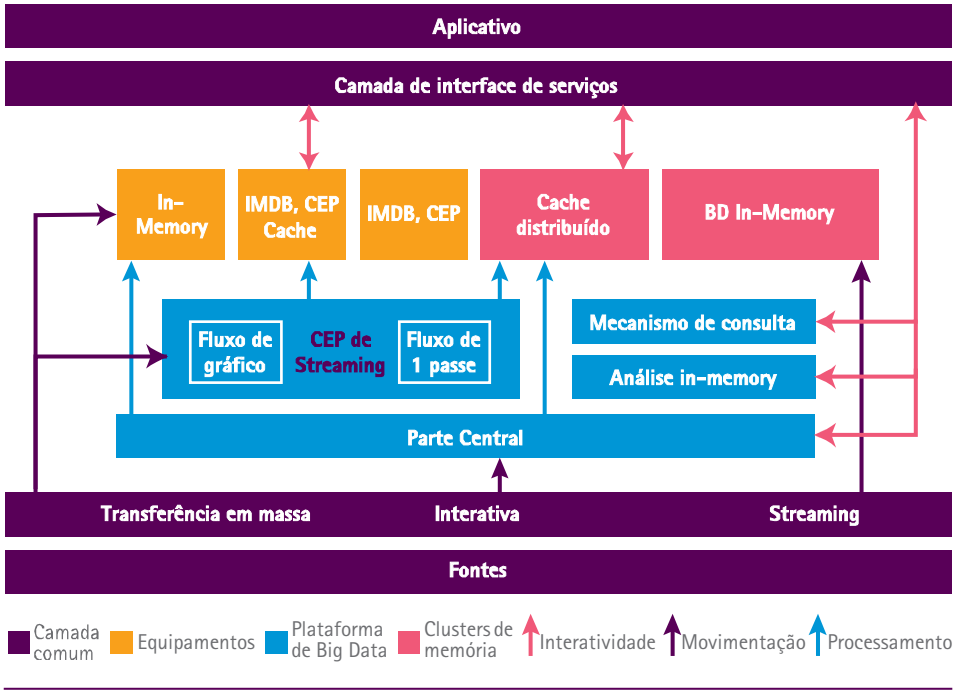
Componente	Recursos de tecnologia
Plataforma de Big Data	• Computação distribuída • In-memory • Streaming • Rede otimizada
Ingestão	• Computação distribuída • In-memory • Streaming
Processamento de eventos complexos	• Computação distribuída • In-memory • Streaming
Banco de dados in-memory	• Computação distribuída • In-memory
Clusters de cache	• In-memory
Equipamentos	• Computação distribuída • In-memory • Rede otimizada • Silício personalizado



Combinar componentes para criar soluções

Os componentes da arquitetura descritos acima não podem funcionar isoladamente para o suporte à aceleração de dados. Em vez disso, eles devem "interagir" com os outros, aproveitando as vantagens recíprocas. Nesta seção, exploraremos quatro pilhas de tecnologia fundamentais que atendem a esses imperativos. Nós utilizamos uma abordagem incremental, de complemento, para mostrar como essas pilhas (todas com camadas comuns) são construídas para permitir a movimentação, processamento e interatividade de dados.

Figura 7: O cenário de soluções



Camadas comuns

Para qualquer uso de dados corporativos, os dados ingressam na infraestrutura de dados da organização, são processados e os usuários interagem com eles, enviando consultas e recebendo respostas que servem para a tomada de decisões. Os pontos onde os dados se originam e onde saem do sistema são o que chamamos de camadas comuns, que podem ser imaginadas como acima e abaixo dos componentes da arquitetura discutidos anteriormente. As fontes de dados alimentam o componente de ingestão, na parte inferior do sistema. Na parte superior do sistema, os usuários e aplicativos interagem com os dados.

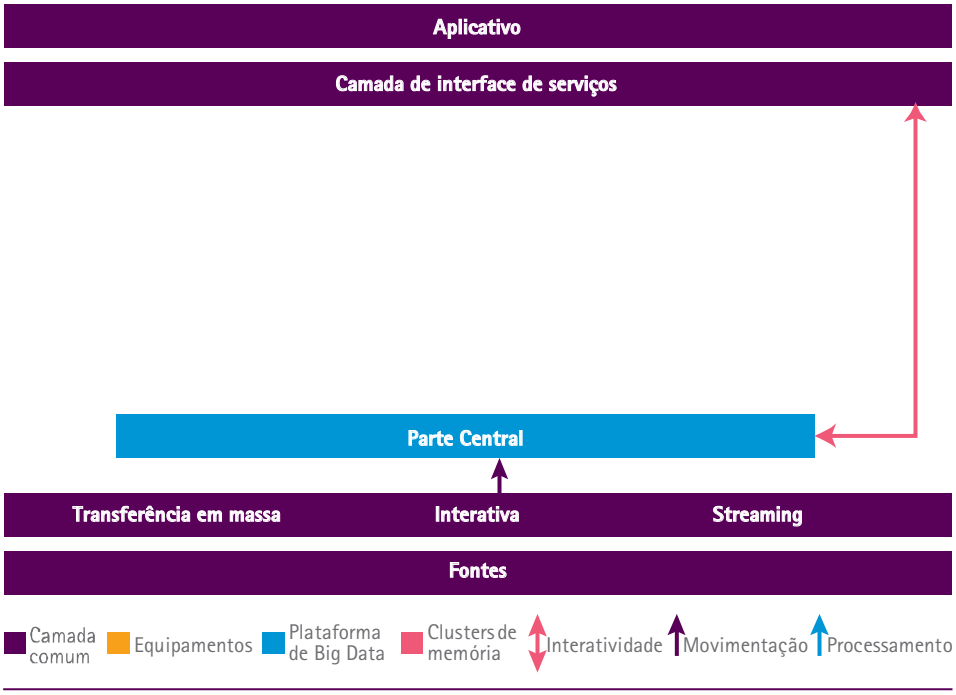
Tipos de Problemas				
	Movimentação "Beber água em uma mangueira de incêndio sem perder uma única gota".		**Processamento "Necessidade de processar grandes volumes de dados mais rápido"**	**Interatividade "Pergunta urgente que deve ser respondida imediatamente"**
	ETL	Streaming		
1. Apenas equipamento	Básico	Aprimorado	Aprimorado	Aprimorado
2. BDP para equipamento	Básico	Aprimorado	Aprimorado	Aprimorado
3. Streaming para equipamento	Básico	Aprimorado+	Aprimorado+	Aprimorado+
4. Apenas BDP	Básico	Aprimorado	Básico	Básico
5. Streaming para BDP	Básico	Aprimorado+	Aprimorado	Básico
6. BDP com análise in-memory	Básico	Aprimorado	Aprimorado	Aprimorado
7. Streaming para BDP com análise in-memory	Básico	Aprimorado+	Aprimorado+	Aprimorado
8. BDP com mecanismo de consulta	Básico	Aprimorado	Básico	Aprimorado
9. Streaming para BDP com mecanismo de consulta	Básico	Aprimorado+	Aprimorado+	Aprimorado
10. Apenas cluster de cache distribuído	Básico	Aprimorado	Básico	Aprimorado
11. BDP para cluster de cache	Básico	Aprimorado	Básico	Aprimorado
12. Apenas cluster de banco de dados in-memory	Básico	Aprimorado	Básico	Aprimorado
13. BDP para cluster de banco de dados in-memory	Básico	Aprimorado	Básico	Aprimorado
14. Streaming para cluster de banco de dados in-memory	Básico	Aprimorado+	Aprimorado+	Aprimorado
	Processamento de eventos complexos pode melhorar a ingestão por streaming		Processamento de eventos complexos pode aumentar a velocidade com o pré-processamento de dados	Caches e bancos de dados in-memory podem permitir interatividade em tempo real

Plataforma de Big Data

Apenas núcleo de Big Data

Neste cenário, os dados normalmente entram no cluster de computadores através de um processo em lote ou streaming. No entanto, os eventos não são processados imediatamente. O núcleo baseia-se na tarefa – os cálculos são programados para execução em determinado intervalo, em vez de em tempo real. Ele aproveita a replicação e o processamento paralelo distribuído em grandes conjuntos de dados, o que permite análises avançadas. Os aplicativos e serviços podem acessar o núcleo diretamente e oferecer melhor desempenho em grandes conjuntos de dados não estruturados. Este está se tornando rapidamente o padrão de fato; portanto, consideramos essa tecnologia a referência para a movimentação, processamento e interatividade excepcionais de dados.

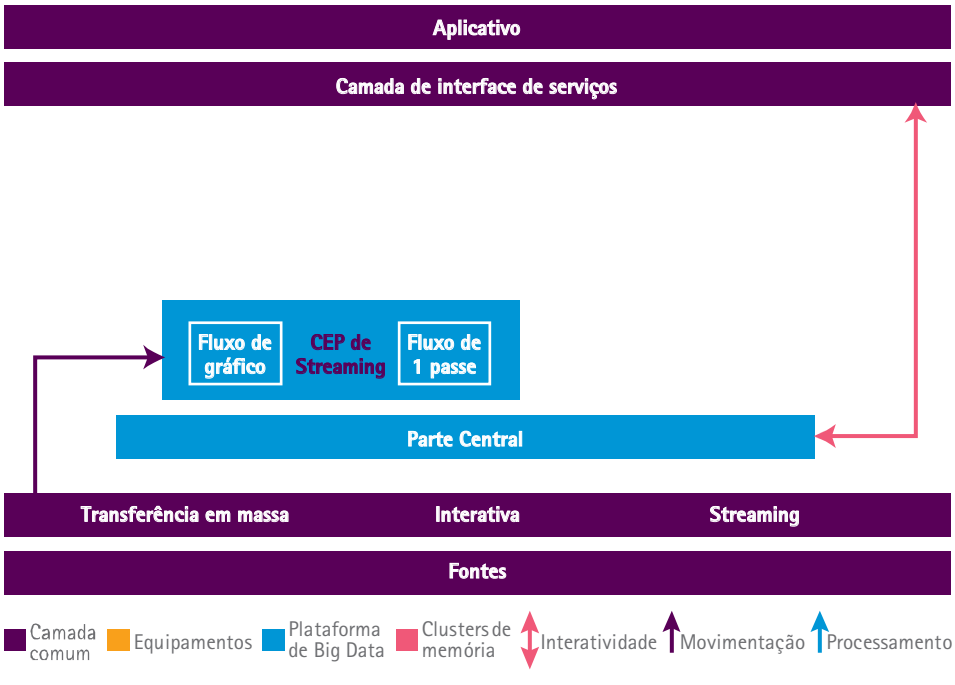
Figura 8: Apenas núcleo de Big Data



Núcleo de Big Data e processamento de eventos complexos

Adicionar CEP aumenta a capacidade de processamento do núcleo de Big Data, já que a ingestão de dados através do CEP permite detecção em tempo real de padrões nos dados e disparo do evento. Essa funcionalidade é útil para correlacionarmos informações em tempo real com um modelo analítico; por exemplo, quando uma organização deseja ser alertada sobre um evento de segurança em tempo real. Ao aproveitarem as capacidades de processamento em um conjunto de dados existente no núcleo, os cientistas de dados podem criar um modelo de aprendizado de máquina e transferi-lo para a unidade de CEP. Sem a necessidade de aguardar a execução das tarefas do núcleo, o CEP pode agir imediatamente, com base em critérios gerados no modelo. Assim, ele aumenta a capacidade de processamento do núcleo e amplia componentes de interatividade, ativando painéis animados em tempo real.

Figura 9: Núcleo de Big Data e processamento de eventos complexos



Núcleo de Big Data e processamento de eventos complexos

As capacidades analíticas tradicionais de Big Data derivam-se da capacidade para aproveitar o poder de computação distribuída do hardware. À medida que tal poder computacional se fortalece com o tempo, o mesmo ocorre com as aplicações que utilizam tal hardware. O software de analítica IMDB, por exemplo, pode ser adicionado ao núcleo de Big Data para melhorar a computação, colocando os principais dados na RAM em nós no cluster, evitando o problema de operações de disco lentas. Além disso, novos softwares prometem ajudar a reduzir o tempo de processamento necessário em várias ordens de magnitude.

Figura 10: Núcleo de Big Data e banco de dados in-memory

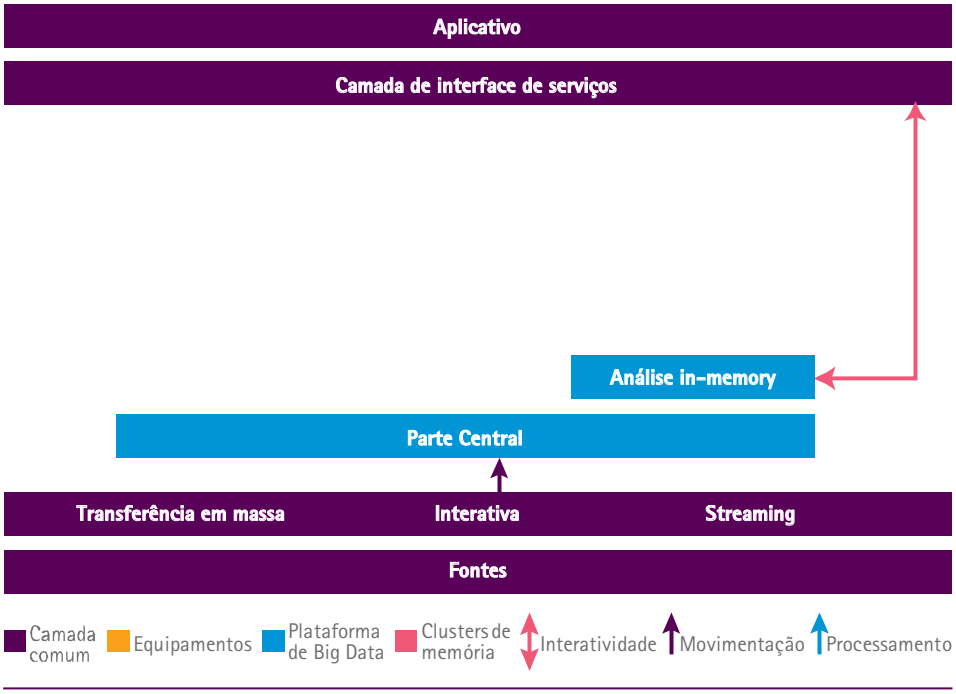
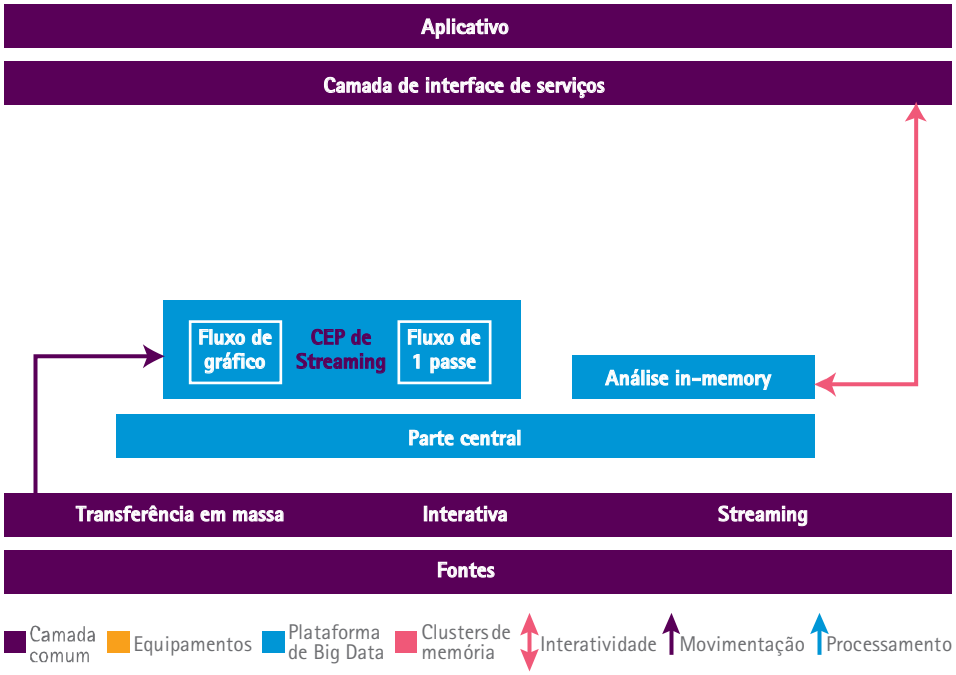


Figura 11: Processamento de eventos complexos e analítica de banco de dados in-memory

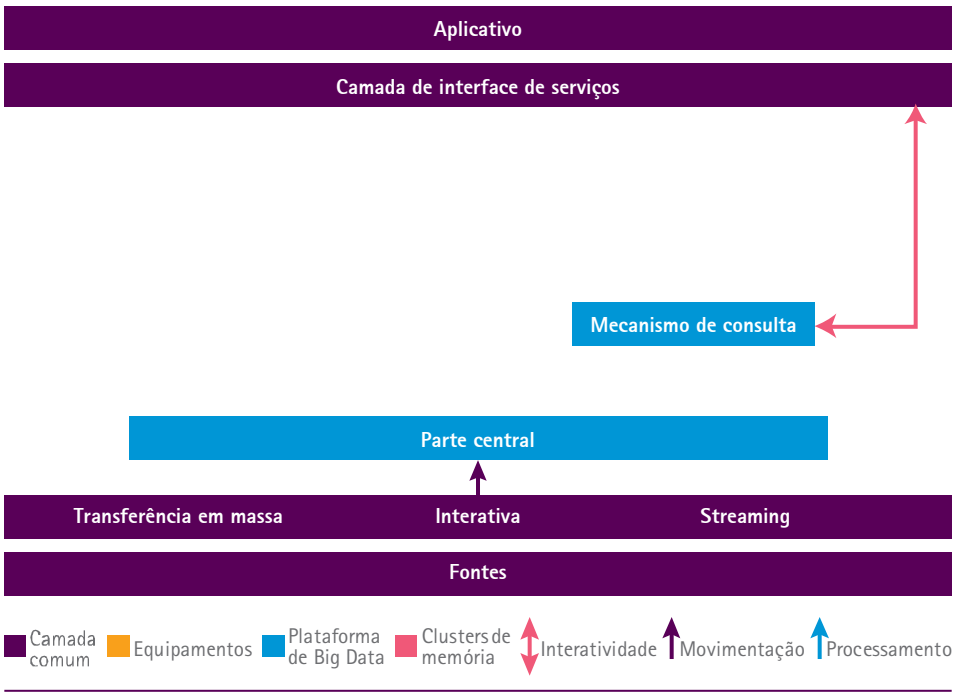
A união desses três componentes permite processamento e interatividade mais rápidos.



Big Data com mecanismo de consulta

Adicionar a tecnologia de mecanismo de consulta a um BDC abre interfaces comuns para o acesso a dados por aplicações com menos demora. Isso torna a platataforma de Big Data mais acessível aos usuários e aplicações.

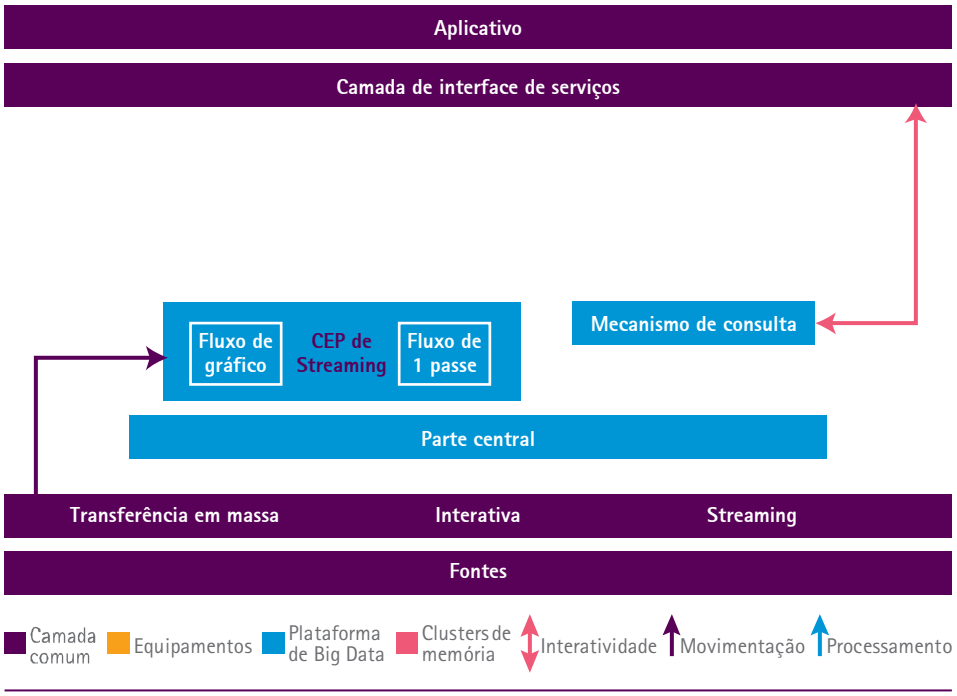
Figura 12: Big Data com mecanismo de consulta



Processamento de eventos complexos e mecanismo de consulta

Com certas tecnologias, os resultados do CEP podem ser acessados diretamente a partir de tecnologias de mecanismos de consulta, fomentando melhor movimentação, processamento e interatividade dos dados.

Figura 13: Processamento de eventos complexos e mecanismo de consulta

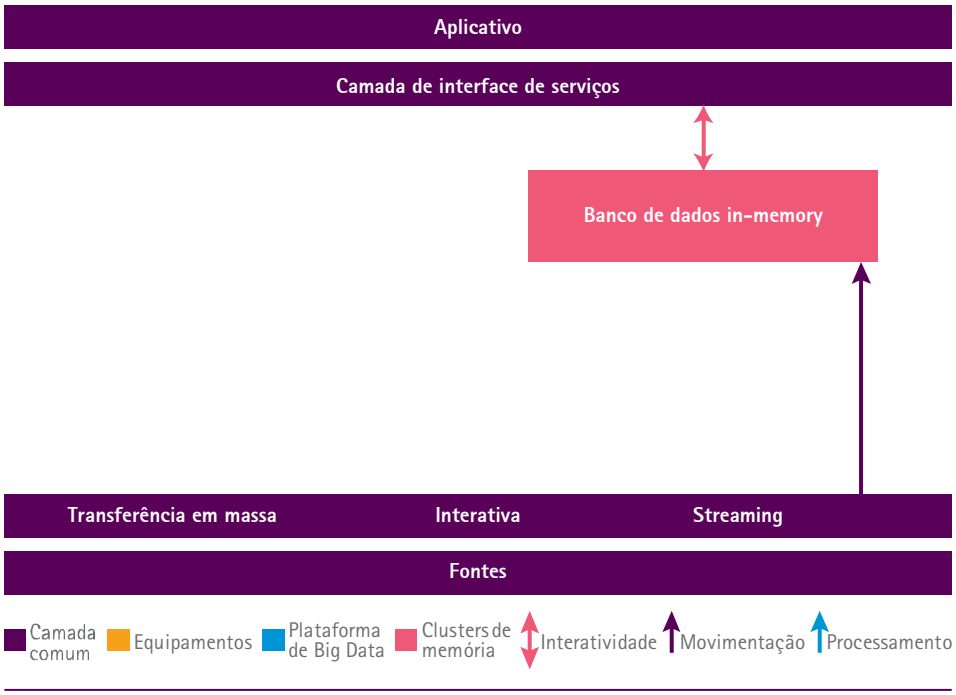


Banco de dados *in-memory*

Apenas cluster de banco de dados in-memory

Para agilizar a movimentação, processamento e interatividade, dados de diferentes fontes externas são transferidos por streaming ou em massa diretamente para o IMDB. O processamento inclui cálculos simples e complexos, execuções de modelos e comparações estatísticas – que ocorrem de forma in-memory dentro do banco de dados. Sem a necessidade de chamar informações da memória continuamente, o IMDB melhora o desempenho da leitura e escrita, acelerando o processamento de dados. Usuários e aplicações podem consultar diretamente o IMDB, como fariam com qualquer outro banco de dados, para obter informações específicas. Essas consultas normalmente usam estruturas do tipo SQL, tornando os dados facilmente acessíveis. Além disso, ocorre otimização das consultas na memória. Por exemplo, ao retornarem dados, os computadores no cluster com mais recursos disponíveis serão selecionados para a resposta. Essa otimização oferece tempos de resposta mais rápidos.

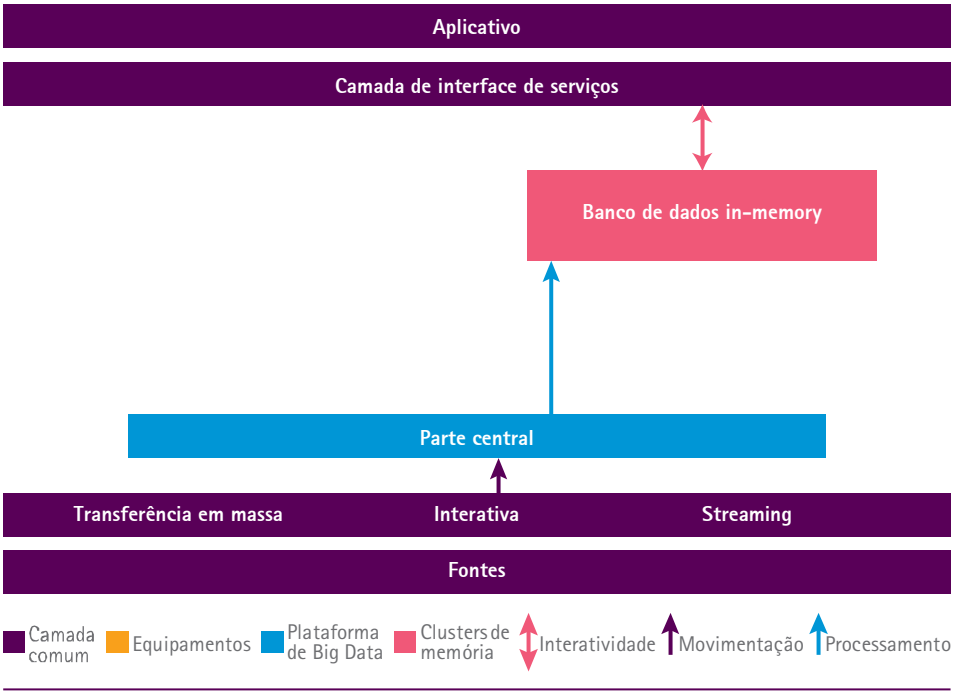
Figura 14: Apenas cluster de banco de dados in-memory v



Cluster de banco de dados in-memory e plataforma de Big Data

Inicialmente, os dados são ingeridos no sistema como transferência em massa ou como processo em streaming através da plataforma. Os dados são armazenados no sistema de arquivos distribuídos da plataforma. Essa abordagem permite a ocorrência de algum reprocessamento na plataforma, antes da transferência dos dados para o IMDB. Esse cálculo prévio acelera o processamento futuro. O banco de dados executa a maior parte do processamento analítico completamente in-memory, proporcionando desempenho mais veloz de leitura e escrita. Tal como acontece com o caso de apenas cluster, as consultas solicitadas por uma aplicação são otimizadas e executadas no banco de dados in-memory, e os resultados são retornados rapidamente para a aplicação.

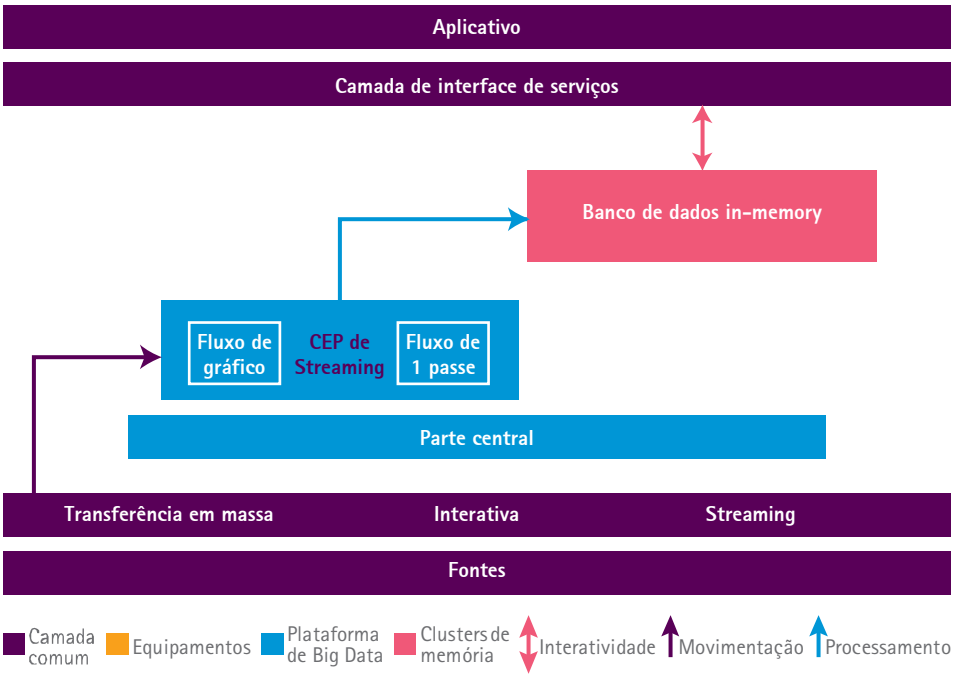
Figura 15: Cluster de banco de dados in-memory e plataforma de Big Data



A parte central da plataforma de Big Data in-memory e processamento de eventos complexos

Os dados de fontes são ingeridos primeiro no sistema, através de um mecanismo de CEP. A maior parte do processamento analítico, incluindo a execução do modelo e comparação estatística, ocorre no IMDB. As consultas solicitadas por uma aplicação são executadas no banco de dados e retornadas para a aplicação, para interatividade mais rápida.

Figura 16: Cluster de Big Data in-memory e processamento de eventos complexos

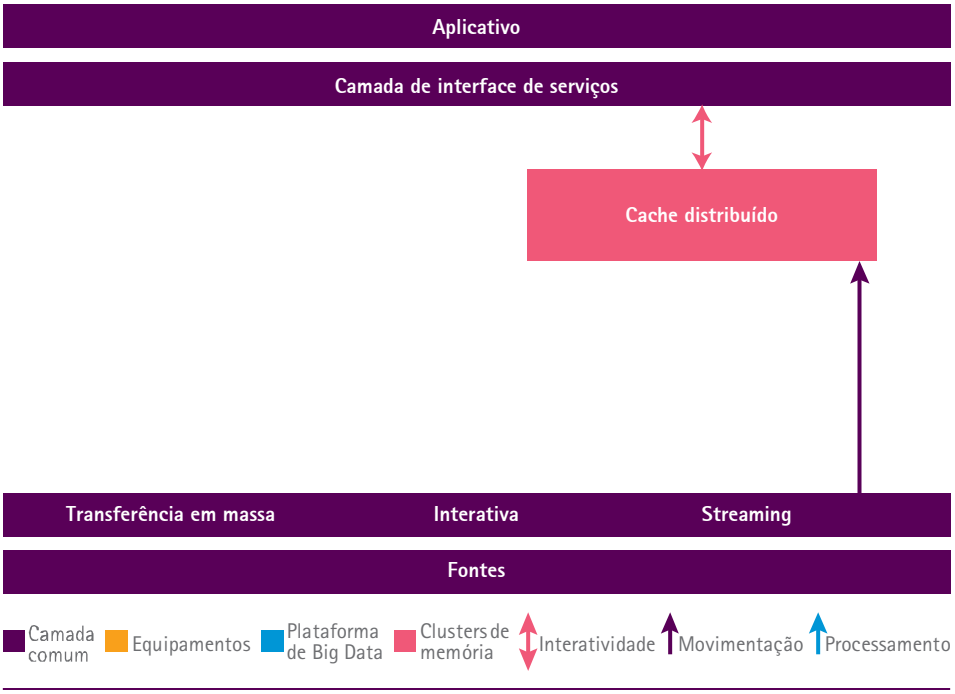


Cache distribuído

Apenas cache

Esta pilha consiste em uma estrutura simples de cache localizada no topo do repositório da fonte de dados e conectada a um aplicativo. O aplicativo recupera os dados. Para otimizar o tempo de consulta, o cache deve ser "ajustado", de modo que os subconjuntos de dados mais relevantes para o aplicativo sejam colocados no cache. Uma vez que o cache simplesmente armazena dados, o processamento de dados recai sobre o aplicativo, o que pode causar velocidades mais lentas e atrasos no processamento.

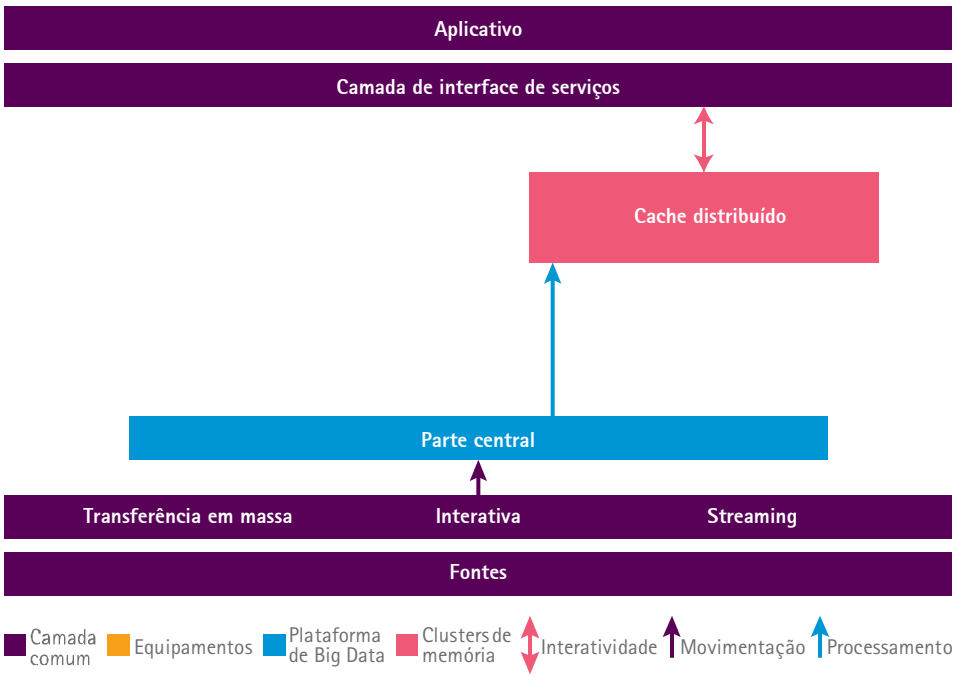
Figura 17: Apenas cache



Cache, aplicativo e plataforma de Big Data

A plataforma ingere dados da fonte e realiza a maior parte do processamento antes de carregar um subconjunto de dados no cache. Isso move a carga de processamento de dados do aplicativo para a plataforma, que pode executar processos analíticos complexos em grandes conjuntos de dados de forma mais eficiente. Um cache se situa no topo da plataforma, que alimenta os resultados da consulta ao aplicativo.

Figura 18: Cache, aplicativo e plataforma de Big Data

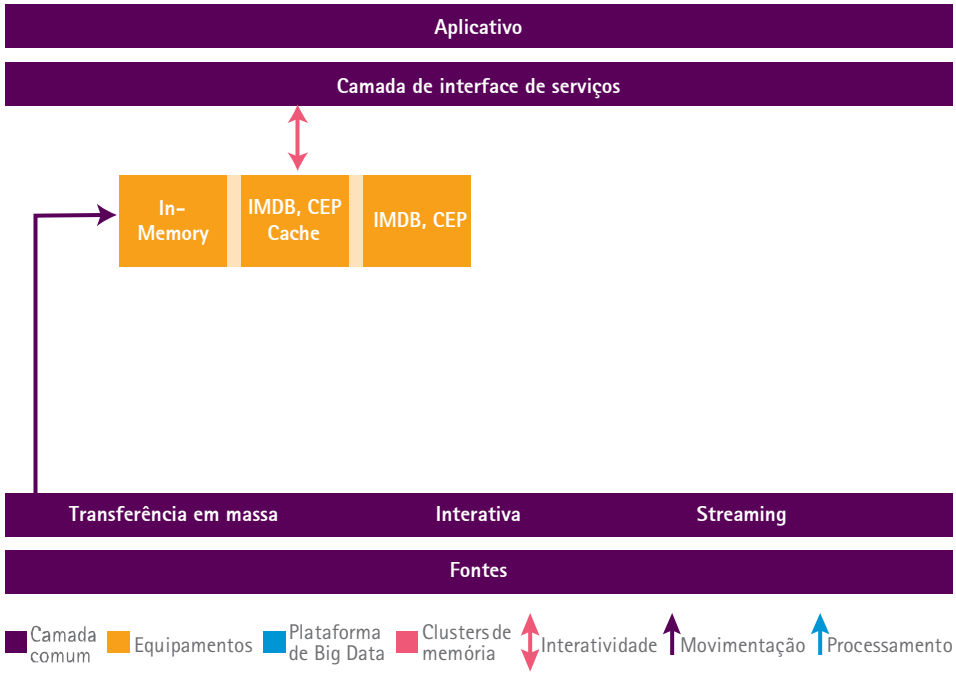


Equipamento

Apenas equipamento

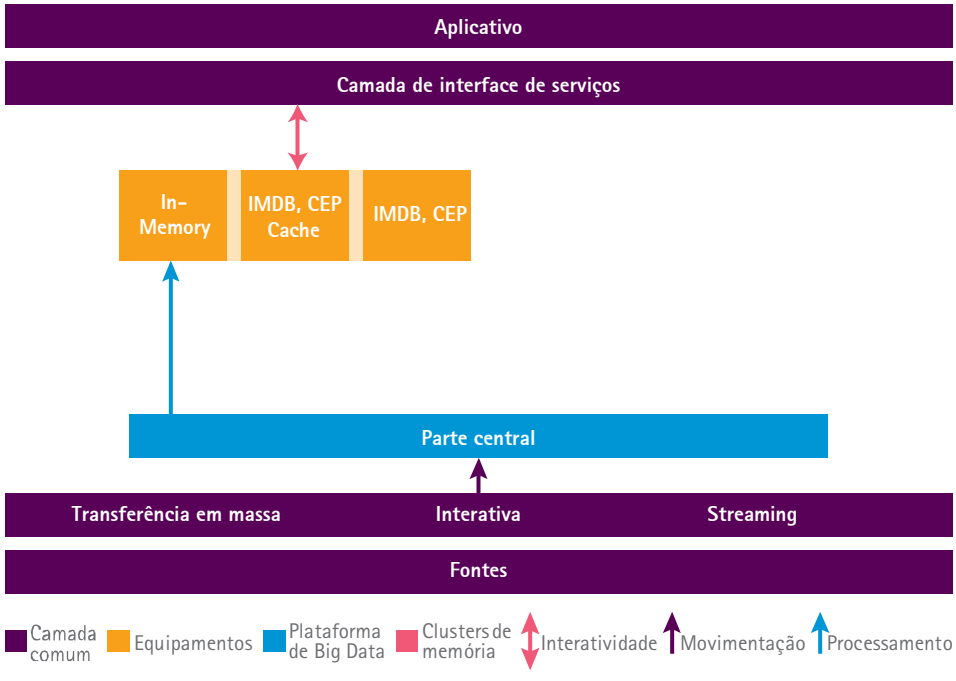
Os dados da fonte são transferidos por streaming diretamente para o equipamento, que completa o processamento, análise e cálculos. O aplicativo "fala" diretamente com o equipamento para solicitações de consultas.

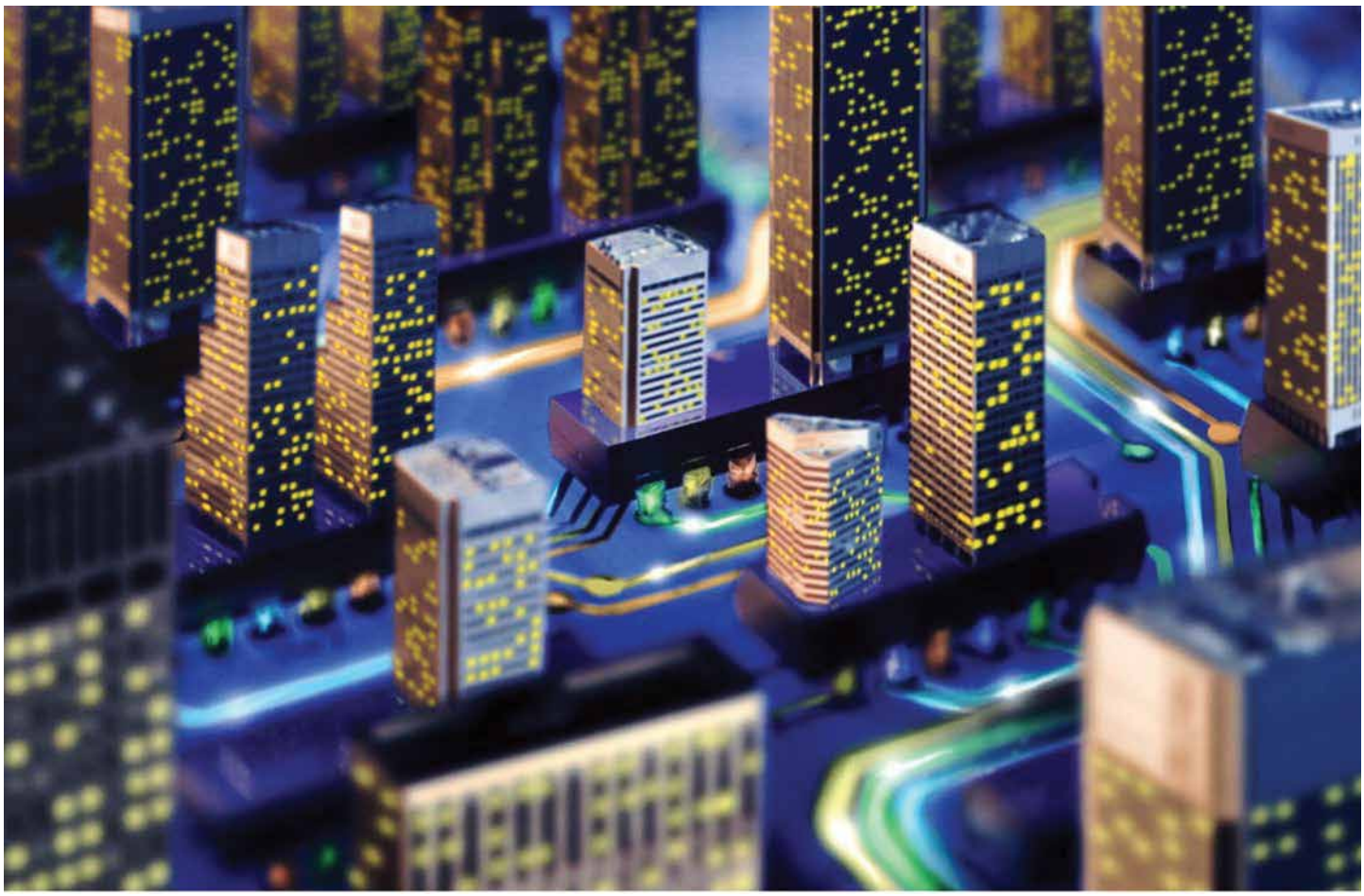
Figura 19: Apenas equipamento



Equipamento e plataforma de Big Data

Os dados da fonte são importados e armazenados dentro da plataforma. A pilha pode tratar os dados dentro da plataforma antes de transferi-los para o equipamento, para a obtenção de velocidade maior de processamento. O aplicativo também pode falar diretamente com o equipamento para solicitações de consultas.

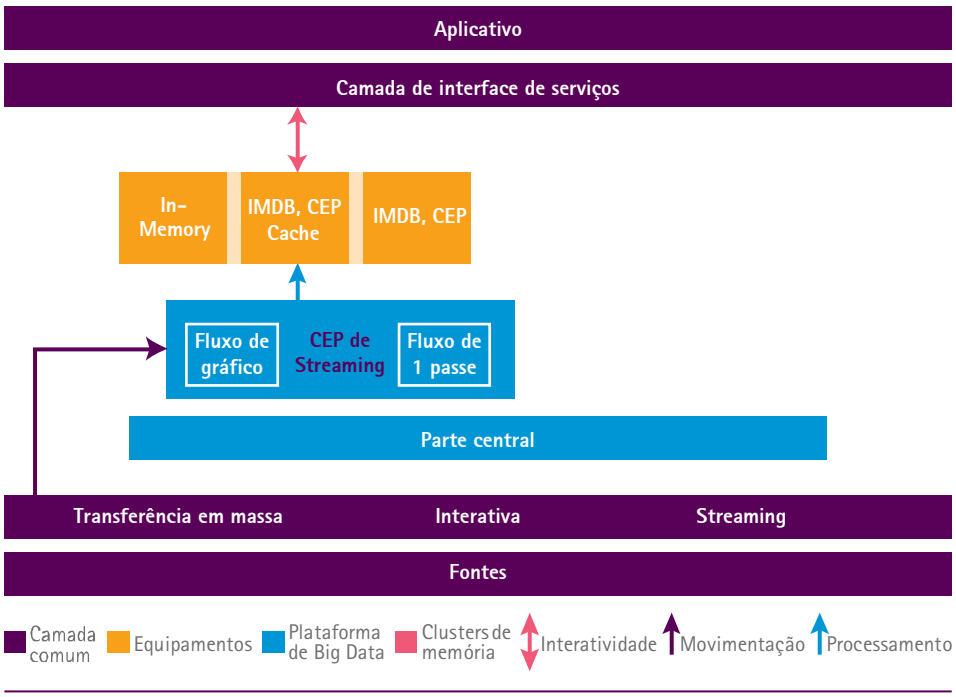




Equipamento e streaming

Primeiro, os dados da fonte são importados e armazenados dentro da plataforma de Big Data por meio de streaming. A pilha pode também processar os dados no interior da plataforma, antes de transferi-los para o equipamento para a obtenção da velocidade ideal de processamento. O aplicativo pode consultar o diretório do equipamento para obter informações.

Figura 21: Equipamento e streaming



Conclusão

Para obter vantagem competitiva com seus dados corporativos, uma organização deve ser capaz de gerar insights comerciais a partir desses dados. A barreira de entrada está mais baixa do que nunca, graças ao advento da plataforma em Big Data. No entanto, existem grandes desafios. Para superá-los, as organizações precisam estabelecer uma cadeia de suprimento de dados que (entre outras realizações) acelere a movimentação, processamento e interatividade dos dados, permitindo que os tomadores de decisões capturem e ajam rapidamente com base nos insights obtidos com seus dados, além de obterem retornos com seus investimentos em analítica.

No entanto, a paisagem de soluções destinadas a promover a aceleração de dados está mais complexa do que nunca. Para construírem a arquitetura correta de aceleração de dados, os executivos devem, primeiro, aprofundar a compreensão sobre os desafios inerentes à movimentação, processamento e interatividade dos dados. Depois, eles precisam familiarizar-se com os componentes arquitetônicos disponíveis atualmente no mercado – cada um dos quais é capaz de dar suporte à aceleração de dados de forma única.

Contudo, mesmo esse entendimento não é suficiente: os componentes da arquitetura entregam valor máximo apenas quando são combinados de maneira a capitalizar suas vantagens complementares. Ao explorarem quatro possíveis configurações de arquitetura, os executivos podem iniciar discussões valiosas acerca das melhores configurações para as necessidades de suas empresas. Igualmente importante, eles podem trazer uma perspectiva mais informada para discussões com os fornecedores sobre soluções potenciais.

Este ponto de vista fornece uma visão geral que os executivos podem usar como ponto de partida, tanto para entenderem essa paisagem em evolução quanto para começarem a se familiarizar com soluções arquitetônicas apropriadas para abordarem suas necessidades comerciais e obterem ROI em analítica.

Próximos passos

Para começar a construir uma estratégia de cadeia de suprimento de dados que auxilie na aceleração de dados em sua organização:

- Inventarie os seus dados. Comece com os seus dados mais acessados e relevantes em relação ao tempo. Estes receberão o primeiro acesso à sua plataforma de dados e serão acelerados na plataforma.
- Identifique os processos ineficientes. Investigue qualquer processo manual e demorado de coleta de dados, como etiquetagem ou limpeza. Tal processo pode ser candidato a substituição por algoritmos de aprendizado de máquina.
- Identifique silos de dados. Juntamente com os silos, identifique necessidades de dados correspondentes que não são atendidas atualmente entre os negócios.
- Simplifique o acesso aos dados. Crie uma estratégia para padronizar o acesso aos dados através da plataforma. As soluções podem ser híbridas, combinando middleware tradicional e gestão de API, ou até mesmo uma oferta de "plataforma como serviço".
- Priorize as cadeias individuais de suprimento de dados. Priorizar ajuda no desenvolvimento de um roteiro para a implementação da cadeia de suprimento de dados em grande escala.
- Considere fontes de dados externas. Busque fora da sua empresa as fontes de dados externas que podem ser incorporadas para complementar os dados atuais e ajudar a gerar insights mais completos.
- Escolha a pilha de tecnologia de aceleração de dados para os seus dados e pesquise os métodos de implantação ideais.

Para obter mais informações, entre em contato com:

Daniel Lázaro
Líder da prática em Big Data da América Latina
daniel.lazaro@accenture.com
Direto: +55 11 5188-3018

Sobre a Accenture Analytics

A Accenture Analytics faz parte da Accenture Digital e entrega resultados em escala baseados em insights para ajudar as organizações a aprimorarem sua performance. Com profunda experiência em diversas indústrias, funções, processos de negócio e tecnologias, a Accenture Analytics desenvolve serviços inovadores de consultoria e terceirização para os clientes, ajudando a garantir que obtenham retorno de seus investimentos em inteligência analítica. Para mais informações, siga-nos em @ISpeakAnalytics e acesse www.accenture.com/analytics.

Sobre Accenture Technology Labs

Accenture Technology Labs, a organização dedicada a pesquisa e desenvolvimento (P&D) de tecnologia dentro da Accenture, transforma inovação tecnológica em resultados de negócio há mais de 20 anos. Nossa equipe de P&D explora tecnologias novas e emergentes para criar uma visão de como a tecnologia moldará o futuro e inventará a nova onda de soluções de negócio de ponta. Trabalhando com a rede global de especialistas da Accenture, Accenture Technology Labs ajuda os clientes a inovar para alcançar a alta performance. Os laboratórios estão localizados no Vale do Silício, Califórnia, EUA; em Sophia Antipolis, França; em Arlington, Virginia, EUA; em Beijing, China e em Bangalore, Índia. Para mais informações, siga-nos em @AccentureLabs e acesse www.accenture.com/accenturetechlabs.

Sobre a Accenture

A Accenture é uma empresa global de consultoria de gestão, serviços de tecnologia e outsourcing, com mais de 293.000 profissionais atendendo a clientes em mais de 120 países. Combinando experiência ímpar, conhecimento profundo sobre todos os setores econômicos e funções de negócio, e extensa pesquisa junto às mais bem-sucedidas organizações no mundo, a Accenture colabora com seus clientes, quer sejam empresas ou governos, para ajudá-los a alcançar altos níveis de performance. A companhia teve receitas líquidas de US$ 28,6 bilhões no ano fiscal encerrado em 31 de agosto de 2013. Sua página na internet é www.accenture.com.br