

Guia de Planejamento

Saiba mais sobre Big Data

Medidas que Gerentes de TI Podem Tomar para Avançar com o Software Apache Hadoop*

Por Que Você Deve Ler Este Documento

O guia de planejamento oferece informações valiosas e passos práticos para gerentes de TI que querem implementar iniciativas de Big Data e começar com o software Apache Hadoop, incluindo:

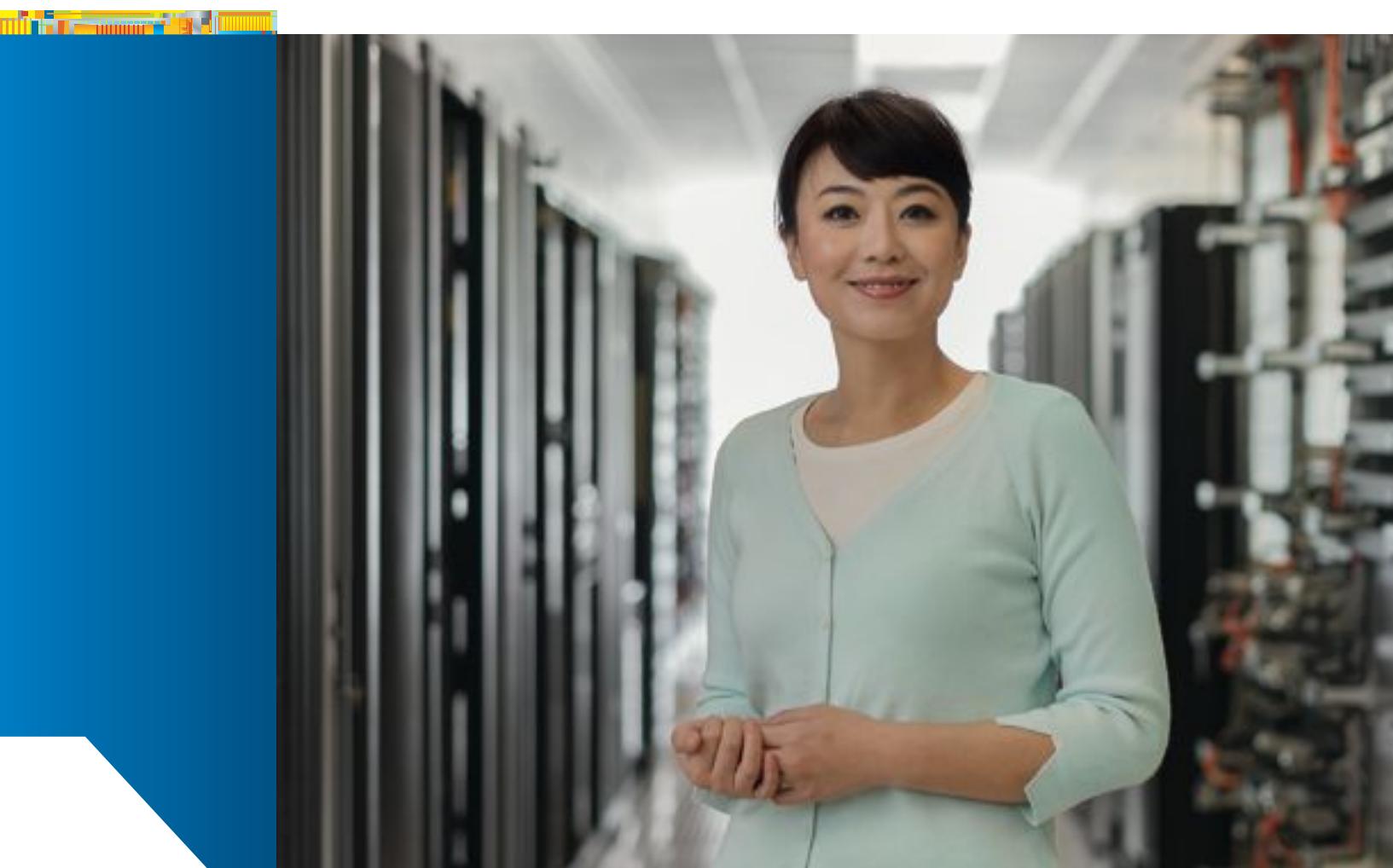
- O cenário de TI para Big Data e os desafios e oportunidades associados com esta força revolucionária
- Introdução ao software Hadoop*, o padrão emergente para ganhar insights a partir de Big Data , incluindo processamento e ferramentas analíticas (software Apache Hadoop MapReduce, Apache HBase*)
- Um guia sobre como aproveitar ao máximo o software Hadoop com foco nas áreas em que a Intel pode ajudar, como tecnologia de infraestrutura, otimização e ajustes
- Cinco “próximos passos” básicos e um checklist para ajudar gerentes de TI a prosseguir com planejamento e implementação de seu próprio projeto Hadoop



Guia de Planejamento

Saiba mais sobre Big Data

Medidas que Gerentes de TI Podem Tomar para Avançar com o Software Apache Hadoop*



Índice

- 3 O Cenário de TI para Big Data
- 4 O Que É (e Não É) Análise de Big Data
- 6 Tecnologias Emergentes para Gestão de Big Data
- 13 Instalando o Hadoop em Seu Data Center
- 18 Cinco Passos e Checklist: Começando seu Projeto de Análise de Big Data
- 20 Recursos de Aprendizagem Intel

O Cenário de TI para Big Data

Cresce a agitação sobre análise de Big Data.

Hoje, toda organização ao redor do mundo encara um aumento sem precedentes no volume de dados. Imagine isso: estima-se que o universo digital de dados tenha alcançado 2,7 zettabytes (ZB) ao final de 2012. Depois disso, estima-se que ele vá dobrar a cada dois anos, alcançando 8 ZB ao final de 2015.¹ É difícil compreender este volume de informação, mas aqui vai um exemplo: se a Biblioteca do Congresso dos Estados Unidos armazena 462 terabytes (TB) de dados digitais, então 8 ZB equivale a quase 18 milhões de Bibliotecas do Congresso.² Isso realmente é Big Data.

O Valor do Big Data

O que exatamente é "Big Data" e de onde ele vem?

Big Data se refere ao imenso volume de conjuntos de dados que alcançam elevadas ordens de magnitude (volume); mais diversos, incluindo dados estruturados, semiestruturados e não estruturados (variedade); e que chegam mais rápido (velocidade) do que você ou sua organização já teve de lidar. Este fluxo de dados é geralmente gerado por equipamentos conectados – PCs e smartphones, sensores como leitores RFID e câmeras de trânsito. Além disso, é heterogêneo e vem em muitos formatos, incluindo texto, documento, imagem, vídeo e outros.

E os 8 ZB de dados projetados para 2015? Quase 15 bilhões de dispositivos conectados – incluindo 3 bilhões de usuários de Internet além de conexões máquina-a-máquina – vão contribuir para este verdadeiro oceano de dados.³

O valor real do Big Data está no insight que ele produz quando analisado – buscando padrões, derivando significado, tomando decisões e, por fim, respondendo ao mundo com inteligência.

Usando Análise de Big Data para Vencer

Big Data é uma força revolucionária, apresentando oportunidades e também desafios a organizações de TI. Um estudo da McKinsey Global Institute indica que dados são tão importantes para organizações quanto a força de trabalho e o capital.⁴ O estudo conclui que, se

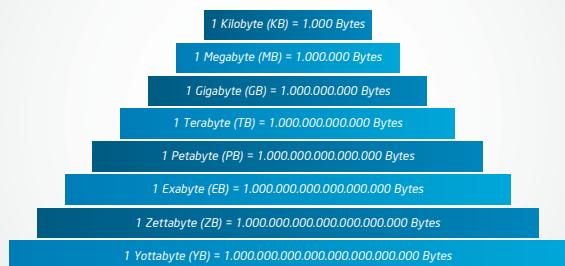
organizações podem realmente capturar, analisar, visualizar e aplicar Big Data às suas metas empresariais, podem diferenciar-se da concorrência e superá-la em termos de eficiência operacional e resultados finais.

Análise de Big Data representa um grande desafio para organizações de TI – mesmo assim, um estudo da Intel feito com 200 gerentes de TI mostra que 84 por cento já analisam dados não estruturados e que 44 por cento daqueles que ainda não o fazem esperam poder fazê-lo até 2014.⁵ O potencial para Big Data é irresistível.

Os três Vs mostram o que é Big Data, além de ajudar a definir os principais assuntos com os quais a TI precisa lidar:

- **Volume.** A imensa escala e expansão de dados não estruturados excede soluções tradicionais de armazenagem e analíticas.
- **Variedade.** Sistemas de gestão de dados tradicionais não conseguem lidar com a heterogeneidade do Big Data – também conhecida como “shadow” ou “dark data,” incluindo traços de acessos e históricos de buscas na web.
- **Velocidade.** Dados são gerados em tempo real, o que requer a oferta imediata de informações úteis.

Uma montanha de dados



Big Data é medido em terabytes, petabytes e até mesmo exabytes. Coloque tudo em perspectiva com esta prática tabela de conversão.

O Que É (e Não É) Análise de Big Data

Análise de Big Data tem claramente a capacidade de mudar o jogo, permitindo com que organizações ganhem insights a partir de novas fontes de dados que não foram pesquisadas no passado. Segue um pouco mais sobre o que é, ou não, a análise de Big Data.

Análise de Big Data é ...

- Uma estratégia baseada em tecnologia que permite a coleta de insights mais profundos e relevantes dos clientes, parceiros e sobre o negócio – ganhando assim uma vantagem competitiva.
- Trabalhar com conjuntos de dados cujo o porte e variedade estão além da habilidade de captura, armazenamento e análise de softwares de banco de dados típicos.
- Processamento de um fluxo contínuo de dados em tempo real, possibilitando tomada de decisões sensíveis ao tempo mais rápido do que em qualquer outra época.
- Distribuído na natureza. O processamento de análise vai aonde estão os dados para maior velocidade e eficiência.
- Um novo paradigma no qual a TI colabora com usuários empresariais e “cientistas de dados” para identificar e implementar análises que ampliam a eficiência operacional e resolvem novos problemas empresariais.
- Transferir a tomada de decisão dentro da empresa e permitir com que as pessoas tomem decisões melhores, mais rápidas e em tempo real.

Análise de Big Data não é ...

- Só tecnologia. No nível empresarial, refere-se a explorar as amplamente melhoradas fontes de dados para ganhar insights.
- Somente volume. Também refere-se à variedade e velocidade. Mas, talvez mais importante, refere-se ao valor derivado dos dados.
- Mais gerada ou utilizada somente por grandes empresas online como Google ou Amazon. Embora as empresas de internet possam ter sido pioneiras no Big Data na escala web, aplicativos chegam a todas as indústrias.
- Uso de bancos de dados relacionais tradicionais “tamanho único” criados com base em disco compartilhado e arquitetura de memória. Análise de Big Data usa uma rede de recursos de computação para processamento maciçamente paralelo (PMP).
- Um substituto de bancos de dados relacionais ou centros de processamento de dados. Dados estruturados continuam a ser de importância crítica para as empresas. No entanto, sistemas tradicionais podem não ter capacidade de manipular as novas fontes e contextos do Big Data.

O Propósito deste Guia

O restante deste guia vai descrever as tecnologias emergentes para gestão e análise de Big Data, com foco em começar a utilizar a estrutura do software de código aberto Apache Hadoop*, que fornece a estrutura para processamento distribuído de grandes conjuntos de dados em clusters de computadores. Também fornecemos cinco passos práticos que você pode utilizar para dar início ao planejamento de seu projeto de análise de Big Data utilizando esta tecnologia.

Seu Novo Melhor Amigo: O Cientista de Dados

Um novo tipo de profissional está ajudando as organizações a compreender os imensos fluxos de informação digital: o cientista de dados.

Cientistas de dados são responsáveis por modelar complexos problemas de negócios, descobrindo insights empresariais e identificando oportunidades. Eles trazem ao trabalho:

- Habilidade para integrar e preparar grandes e variados conjuntos de dados
- Habilidade avançada de análise e modelagem para revelar e compreender relacionamentos obscuros
- Conhecimento empresarial para aplicar um contexto
- Habilidades de comunicação para apresentar resultados

A ciência dos dados é um campo emergente. A demanda é elevada, e encontrar pessoal qualificado é um dos principais desafios associados à análise de Big Data. Um cientista de dados pode estar sediado em TI ou no negócio – mas, onde quer que esteja, ele ou ela será seu novo melhor amigo e colaborador no planejamento e implementação de projetos de análise de Big Data.

Tecnologias Emergentes para Gestão de Big Data

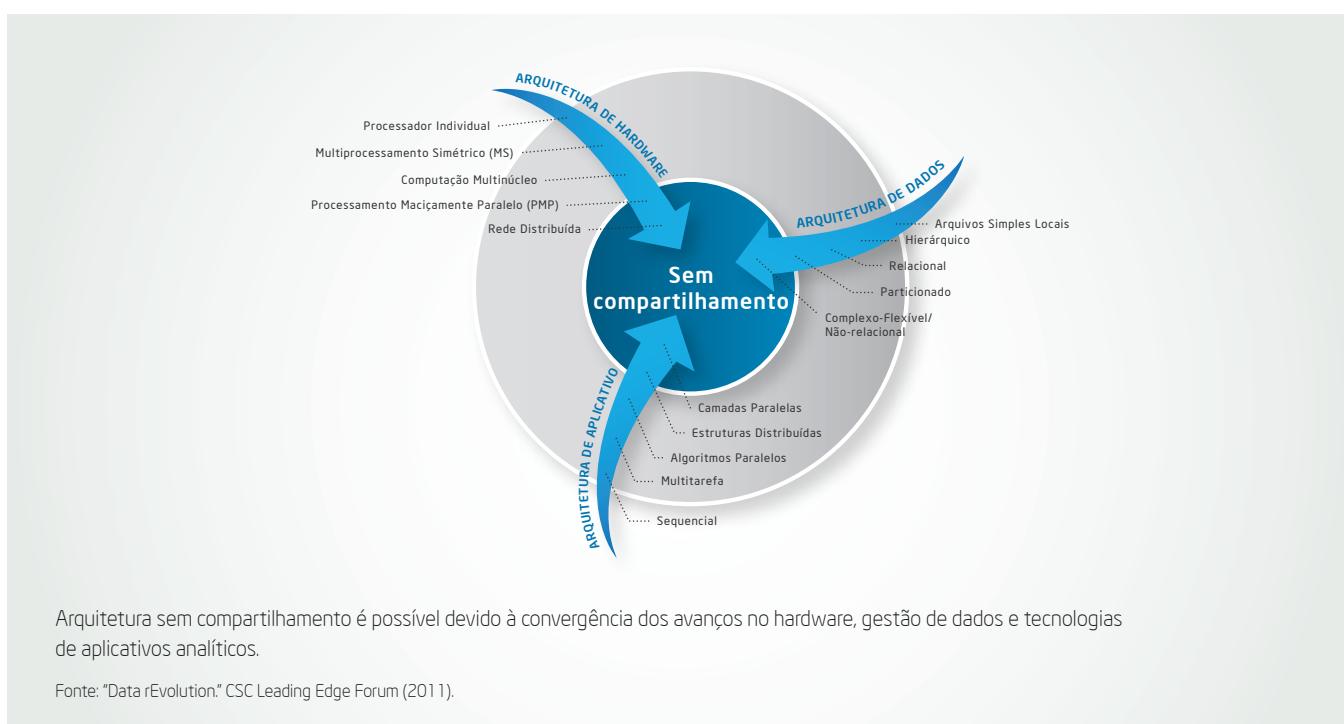
Para organizações compreenderem o potencial total do Big Data, elas devem encontrar uma nova abordagem à captura, armazenamento e análise de dados. Ferramentas e infraestruturas tradicionais não são tão eficientes quando se trabalha com conjuntos de dados maiores e mais variados e que chegam em alta velocidade.

Novas tecnologias estão emergindo para deixar a análise de Big Data escalável e garantir maior custo-benefício. Uma nova abordagem utiliza o poder de uma grade distribuída de recursos de computação e “arquitetura sem compartilhamento”, estruturas de processamento distribuídas, e bancos de dados não relacionais para redefinir como dados são administrados e analisados.

Arquitetura Sem Compartilhamento para Sistemas Altamente Escaláveis

A nova arquitetura sem compartilhamento pode adequar o imenso volume, variedade e velocidade necessários ao Big Data ao distribuir o trabalho em dezenas, centenas ou milhares de servidores de commodity que processam os dados em paralelo. Inicialmente implementado por grandes projetos de pesquisa comunitária como

SETI@home e também serviços online como Google* e Amazon*, cada nó é independente e sem estado, para que a arquitetura sem compartilhamento escala facilmente – simplesmente agregue outro nó – permitindo com que os sistemas administrem crescentes cargas de processamento.



O processamento é empurrado para os nós onde os dados residem. Isso é completamente diferente da abordagem tradicional, que busca dados para processamento em um ponto central.

Por fim, os dados devem ser reintegrados para entregar resultados significantes. Estruturas de software de processamento distribuídas fazem com que a grade de computação funcione, administrando e empurrando os dados entre máquinas, enviando instruções para os servidores de rede em paralelo, coletando resultados individuais e remontando-os para colher o resultado final.

Estruturas de Processamento Distribuídas e o Surgimento do Apache Hadoop

O Hadoop* está evoluindo como a melhor abordagem para a análise de Big Data. Desenvolvido a partir do projeto de busca de web de código aberto Apache Nutch*⁶ o Hadoop é uma estrutura de software que fornece um modelo de programação simples para permitir processamento distribuído de grandes conjuntos de dados em clusters de computadores. A estrutura é de fácil escalabilidade em hardware como os servidores baseados nos processadores Intel® Xeon®.

O software Hadoop é uma estrutura completa de código aberto para análise de Big Data. Ele inclui um sistema de arquivos distribuídos, uma estrutura de processamento paralelo chamada Apache Hadoop MapReduce e vários componentes de apoio à ingestão de dados, coordenação de fluxos de trabalho, gestão de trabalho e monitoramento dos clusters. O Hadoop tem melhor custo-benefício no manejo de grandes conjuntos de dados não estruturados na comparação com abordagens tradicionais.

O Hadoop oferece várias vantagens chave para análise de Big Data, entre elas:

- **Armazenamento de qualquer dado em formato nativo.** Como os dados não requerem uma proposta de Tag de acordo com um esquema específico, nenhuma informação é perdida.
- **Escala para Big Data.** A capacidade de operação em escala do Hadoop já foi comprovada por empresas como Facebook e Yahoo!, que executam enormes implementações.

- **Entrega de novos insights.** A análise de Big Data está descobrindo relações escondidas que eram difíceis, demoradas e caras – ou até mesmo impossíveis – de identificar utilizando as abordagens de data mining tradicionais.
- **Redução de custos.** O software de código aberto Hadoop opera em servidores padrão e tem custo mais baixo por terabyte para armazenagem e processamento. O armazenamento pode ser agregado em incrementos conforme necessário, e o hardware pode ser acrescentado ou trocado dentro ou fora do cluster.
- **Maior disponibilidade.** O Hadoop se recupera após falhas de hardware, software ou sistema fornecendo tolerância a falhas através da replicação de dados e failover através de nós de computação.
- **Menor risco.** A comunidade Hadoop é ativa e diversa, com desenvolvedores e usuários de muitos setores ao redor do mundo. A tecnologia Hadoop vai continuar avançando.

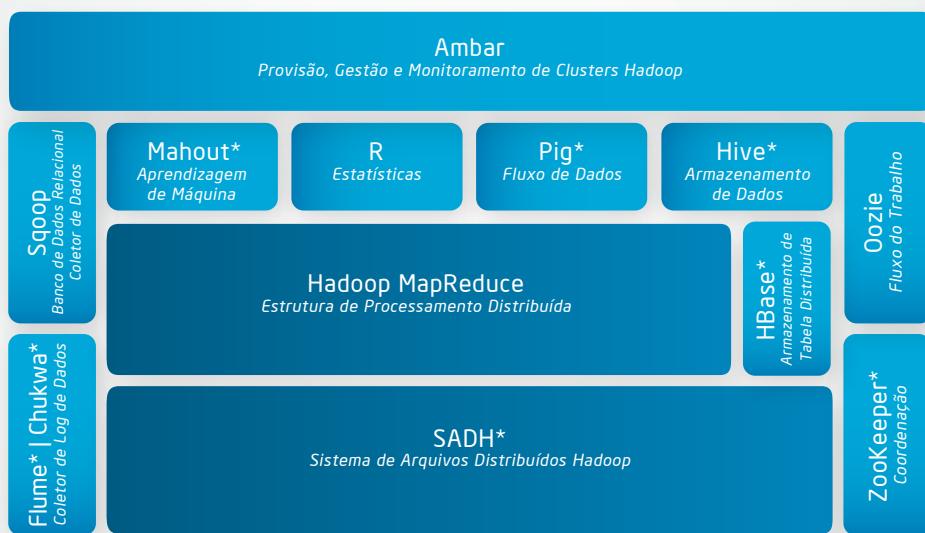
Big Data e a Nuvem

Big Data requer clusters de servidores de apoio às ferramentas que processam grandes volumes, alta velocidade e formatos variados de Big Data. Nuvens já estão em uso em pools de servidores e podem ser ampliadas ou reduzidas de acordo com o Big Data.

O resultado é que a computação em nuvem (cloud computing) oferece maior custo-benefício no suporte de tecnologias Big Data e aplicativos de análise avançados que podem ampliar o valor do negócio.

Descubra mais sobre como Big Data pode funcionar na nuvem em “Big Data in the Cloud: Converging Technologies”, que pode ser acessado no link intel.com/content/www/us/en/big-data/big-data-cloud-technologies-brief.html

Pacote Apache Hadoop*



O pacote de software Hadoop* inclui uma série de componentes.

O Software Hadoop* Substitui meus Sistemas de Bancos de Dados Existentes?

O software Hadoop* é um sistema de armazenamento e processamento de dados massivamente escalável – não é um banco de dados. Na verdade, ele complementa seu sistema existente ao administrar dados que geralmente são problemáticos para eles.

O Hadoop pode simultaneamente absorver e armazenar qualquer tipo de dados de uma série de fontes, agregar e processar de forma arbitrária e entregar para qualquer que seja seu uso – que pode ser a oferta de dados de transações em tempo real ou o fornecimento de inteligência de negócio interativa através de seu sistema existente.

E o Apache Hadoop MapReduce?

O MapReduce é uma estrutura de programação de software do pacote Hadoop que simplifica o processamento de conjuntos Big Data e fornece a programadores um método em comum para entrega e organização de tarefas de processamento complexas através de clusters de computadores. Os aplicativos MapReduce trabalham da seguinte forma: a tarefa de mapeamento divide um conjunto de dados em blocos independentes que serão processados em paralelo. Os resultados de mapa são ordenados e depois oferecidos para a tarefa de redução. Tanto o input quanto o output é armazenado no Apache*

Sistema de Arquivos Distribuídos Hadoop (SADH*) ou outra forma de armazenamento como o Amazon S3, parte da Amazon Web Services. Geralmente, os dados são processados e armazenados no mesmo nó, dando mais eficiência ao agendamento de tarefas onde os dados já residem e resultando em elevada banda agregada ao longo do nó.

O MapReduce simplifica o trabalho do programador de aplicativo ao cuidar dos trabalhos de estipulação, monitoramento e re-execução de tarefas que falharam.

Apache Hadoop* em Resumo

O Apache Hadoop* é um esforço comunitário que inclui três principais subprojetos de desenvolvimento e outras iniciativas relacionadas.

Desenvolvimento Chave de Subprojetos

Apache* Sistema de Arquivos Distribuídos Hadoop* (SADH*)	O sistema de armazenamento primário, que utiliza múltiplas réplicas de blocos de dados, realiza a distribuição através dos nós de um cluster e fornece elevado acesso aos dados do aplicativo
Apache Hadoop MapReduce	Um modelo de programação e estrutura de software para aplicações que executa processamento distribuído de grandes conjuntos de dados em clusters de computação
Apache Hadoop Common	Utilitários que suportam a estrutura Hadoop, incluindo Sistemas de Arquivos (uma classe de base abstrata para um sistema genérico de arquivos), chamadas de procedimento remoto (CPR), e bibliotecas de serialização

Outros Projetos Hadoop Relacionados

Apache Avro*	Sistema de serialização de dados
Apache Cassandra*	Banco de dados multimestres escalável sem ponto único de falha
Apache Chukwa*	Sistema de coleta de dados para monitoramento de sistemas distribuídos construídos com base no SADH e MapReduce; inclui uma ferramenta para exibição, monitoramento e análise de resultados
Apache HBase*	Banco de dados escalável e distribuído que suporta armazenamento de dados estruturados para grandes tabelas; utilizado para acesso randômico de leitura e gravação de Big Data em tempo real
Apache Hive*	Infraestrutura de sistema de armazenamento de dados que oferece resumo, perguntas ad hoc e análise de grandes conjuntos de dados em sistemas de arquivos compatíveis com arquivos Hadoop
Apache Mahout*	Uma biblioteca escalável de aprendizagem de máquina e data mining com implementação de uma grande série de algoritmos, incluindo formação de clusters, classificação, filtragem colaborativa e padrão de mineração frequente
Apache Pig*	Uma estrutura de linguagem e execução de fluxo de dados de alto nível para expressão de análise de dados em paralelo
Apache ZooKeeper*	Um serviço de coordenação centralizada de alta performance que mantém informações de configuração e nomenclatura e fornece sincronização distribuída e serviços em grupo para aplicações distribuídas

Fonte: Apache Hadoop, hadoop.apache.org.

Fale com Experts Apache Hadoop*

Uma forma de aprender sobre o software Apache Hadoop* e seus componentes é falar diretamente com os experts profundamente engajados na comunidade de código aberto e seu desenvolvimento. Escute [podcasts](#) de entrevistas de líderes comunitários do Apache Hadoop MapReduce, Apache* SADH*, Apache Hive*, Apache Pig* e HCatalog descrevendo como cada um trabalha, onde se encaixa no pacote Hadoop* e planos para desenvolvimento contínuo. Arquivos PDFs acompanham cada podcast.

Adoção do Hadoop*

Enquanto cada vez mais empresas reconhecem o valor e vantagens associadas a insights de Big Data, a adoção do software Hadoop cresce. O pacote de software de tecnologia de código aberto Hadoop inclui implementação dos bancos de dados distribuídos com MapReduce, SADH e Apache HBase* que suportam grandes tabelas de dados estruturadas.

Após seis anos de refinamento, o Apache lançou a primeira versão completa de produtos do software Apache Hadoop 1.0 em janeiro de 2012. Entre as características certificadas e suportadas por esta versão estão HBase*, melhorias de segurança Kerberos, e uma transferência representacional de estado (RESTful) API para acesso a SADH.⁷

O software Hadoop pode ser baixado a partir de um dos sites de download [Apache](#). Como o software [Hadoop](#) é um projeto voluntário de código aberto, o Hadoop wiki oferece informações sobre como pedir ajuda à comunidade, além de links a tutoriais e documentação de usuário para implementação, solução de problemas e criação de cluster.

Acelerando Padrões de Análise de Big Data

A Open Data Center Alliance (ODCA), um consórcio de TI independente de líderes de TI de mais de 300 empresas, anunciou recentemente a criação do Grupo de Trabalho de Serviços de Dados para documentar as mais recentes necessidades encaradas pela TI na gestão de dados. O grupo de trabalho vai focar primeiro na criação de um modelo de necessidade de uso que cubra segurança, maneabilidade e interoperabilidade de estruturas Big Data com gestão tradicional de dados e soluções de centro de processamento de dados.

Com base nos modelos de uso, os membros do grupo de trabalho vão desenvolver arquitetura de referência e provas de conceito para distribuição comercial com fornecedores independentes de software e parceiros Fabricantes de Equipamento Original (FEQ) para testar a utilização e criar soluções para o mercado empresarial. A aliança também vai colaborar com a comunidade de código aberto para criar unidades de benchmarking. Como conselheira técnica da ODCA, a Intel terá participação fundamental neste desenvolvimento.

O Ecossistema Hadoop

O ecossistema Hadoop é uma paisagem complexa de fornecedores e soluções que inclui players estabelecidos e muitos novos players. Vários fornecedores oferecem sua própria distribuição Hadoop, unindo o pacote básico com outros projetos de software Hadoop como o Apache Hive*, Apache Pig* e Apache Chukwa*. Alguns destes distribuidores podem trabalhar com centros de processamento de dados, bancos de dados e outros produtos de administração de dados para que os dados possam ser transferidos entre clusters Hadoop e outros ambientes para expandir o pool de dados para processos ou pedidos.

Outros fornecedores oferecem software de gestão Hadoop que simplifica a administração e solução de problemas. Um terceiro grupo entrega produtos que ajudam desenvolvedores a escrever aplicativos Hadoop, oferecendo capacidade de busca, ou analisam dados sem o uso do MapReduce. Estes produtos têm como base uma plataforma de software e incluem camadas de abstração que casam um centro de processamento de dados Structured Query Language (SQL) com um cluster Hadoop além de processamento e análise em tempo real. Por fim, há crescente interesse na oferta de serviços de assinatura via nuvem.

Apache Hadoop* Lançado Comercialmente

Apache Hadoop*- ofertas relacionadas disponíveis em várias categorias. Os seguintes fornecedores são uma amostra do crescente ecossistema Hadoop*.

Categoria	Fornecedor/ Oferta
Sistemas integrados Hadoop	<ul style="list-style-type: none">▪ EMC* Greenplum*▪ HP* Big Data Solutions▪ IBM* InfoSphere*▪ Microsoft* Big Data Solution▪ Oracle* Big Data Appliance
Aplicativos Hadoop e bancos de dados analíticos com conectividade Hadoop	<ul style="list-style-type: none">▪ Datameer* Analytics Solution▪ Hadapt* Adaptive Analytic Platform*▪ HP Vertica* Analytics Platform▪ Karmasphere* Analyst▪ ParAccel* Analytic Platform▪ Pentaho* Data Integration▪ Splunk* Enterprise*▪ Teradata* Aster* Solution
Distribuições Hadoop	<ul style="list-style-type: none">▪ Distribuição via Cloudera incluindo Apache Hadoop (DCH)▪ EMC Greenplum HD▪ Hortonworks▪ IBM InfoSphere BigInsights▪ Intel® Distribution for Apache Hadoop Software▪ MapR* M5 Edition▪ Microsoft Big Data Solution▪ Platform Computing* MapReduce
Soluções baseadas em armazenamento em nuvem	<ul style="list-style-type: none">▪ Amazon* Web Services▪ Google* BigQuery
Aplicações Hadoop e bancos de dados analíticos com conectividade Hadoop	<ul style="list-style-type: none">▪ Teradata* Aster* solution▪ ParAccel* Analytic Platform▪ HP Vertica* Analytics Platform▪ Datameer* analytics solution▪ Hadapt* Adaptive Analytic Platform*▪ Karmasphere* Analyst▪ Pentaho* Data Integration▪ Splunk* Enterprise*

Veja [Big Data Vendor Spotlights](#) para encontrar alguns dos parceiros da Intel que oferecem soluções Big Data.

Atenção: O ecossistema Hadoop emerge rapidamente. Esta lista foi adaptada com base em duas fontes: Dumbill, Edd. "Big Data Market Survey: Hadoop Solutions." O'Reilly Radar (19 de janeiro de 2012). <http://radar.oreilly.com/2012/01/big-data-ecosystem.html> e Data rEvolution: CSC Leading Edge Forum. CSC (2011). http://assets1.csc.com/lef/downloads/LEF_2011Data_rEvolution.pdf

Duas Abordagens ao Uso de Software Hadoop para Análise de Big Data

As empresas estão tendo duas abordagens básicas para a implementação do Hadoop.

Utilização exclusiva do Hadoop. O uso do Hadoop está disponível como software de código aberto que pode ser baixado gratuitamente do Apache ou por meio de distribuições de fornecedores que criam pacotes com a estrutura Hadoop e certos componentes e software de gerenciamento para apoiar a administração do sistema.

Utilização exclusiva do Hadoop é ideal para a criação de um sistema de gestão de Big Data para análise de dados não estruturados e insights. Ferramentas de código aberto também possibilitam consultas a dados estruturados que usam aplicativos MapReduce HBase ou Hive*.

Hadoop integrado com bancos de dados tradicionais.

Organizações com centros de processamento de dados tradicionais e sistemas de análise já em operação podem utilizar sua plataforma existente para incluir uma implementação integrada Hadoop. Ligando recursos de gestão de dados existentes ao software Hadoop, a empresa terá oportunidade de acessar tanto dados estruturados quanto não estruturados para ganhar insights. Por exemplo, a análise das complexas transcrições de um call center pode ser casada com dados estruturados sobre comportamento de compra, como SKUs específicos, lojas, pontos geográficos e assim por diante. Neste caso, conexões prioritárias são utilizadas para transferir dados do Hadoop para os ambientes tradicionais.

Distribuição Intel® para o Software Apache Hadoop*

Distribuição Intel® para o Software Apache Hadoop* (Intel Distribution) inclui Apache Hadoop e outros componentes de software otimizados pela Intel com desempenho de software melhorado e capacidade de segurança. Desenvolvido para permitir um grande leque de análise de dados no Apache Hadoop, o Intel Distribution é otimizado para consultas Apache Hive*, fornece conexão para processamento estatístico R* e permite análise de gráficos utilizando o software Intel Graph Builder for Apache Hadoop, uma biblioteca para recrivar grandes conjuntos de dados e ajudar a viabilizar relação entre dados. Incluso no Intel Distribution, o Intel Manager for Apache Hadoop fornece um console de gerenciamento que simplifica a utilização, configuração e monitoração do Hadoop*.

Intel Distribution está atualmente disponível para avaliação ao redor do mundo. Apoio técnico é fornecido atualmente nos Estados Unidos, China e Singapura, com outras regiões geográficas sendo adicionadas ainda este ano.

Saiba mais sobre [Intel Distribution](#).

Instalando o Hadoop em Seu Data Center

A análise de Big Data é uma estratégia baseada em tecnologia que é muito mais do que o hardware e o software que a alimentam. Ainda assim, como gerente de TI, a responsabilidade por implementar as iniciativas Big Data no Data Center será sua. As implementações do Hadoop podem exigir uma grande infraestrutura e as escolhas

de hardware e software feitas durante o planejamento podem ter impactos significantes na performance e no custo total. Os Data Centers podem aproveitar ao máximo a implementação do Hadoop garantindo que a infraestrutura necessária esteja disponível e que o software Hadoop esteja otimizado e ajustado para melhor performance.

Organize a Infraestrutura correta

A estrutura do Hadoop trabalha sob o princípio de aproximar a computação e os dados, e a estrutura roda tipicamente em um grande cluster de servidores desenvolvidos utilizando hardware padrão. É lá que os dados são armazenados e processados. A combinação da infraestrutura Hadoop com plataformas de servidores padrão oferece a fundação para uma plataforma de análise com bom custo-benefício e alta performance para aplicações paralelas.

Configurando o Sistema de Arquitetura Hadoop

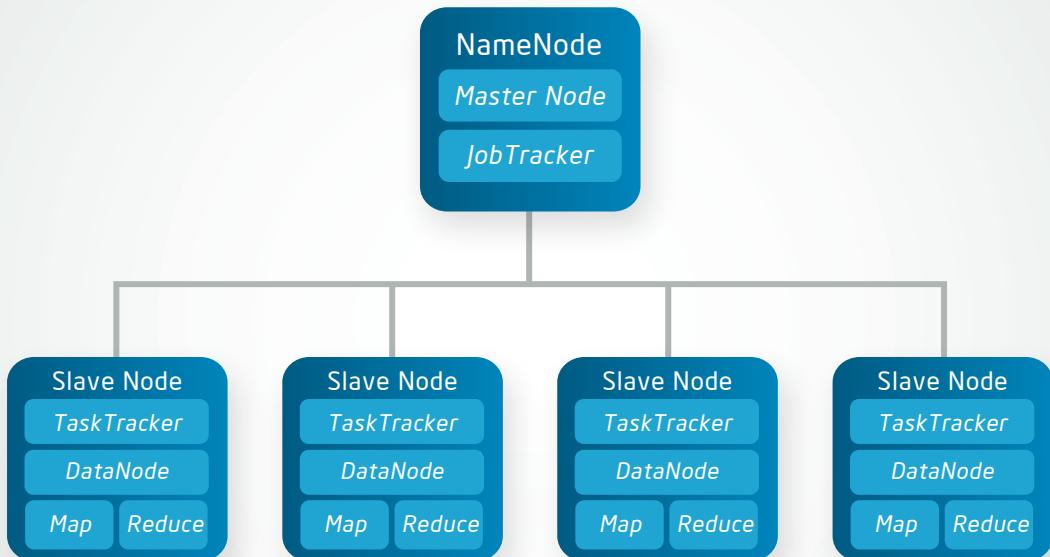
Cada cluster tem um "nó mestre" com múltiplos nós slaves. O nó mestre usa as funções NameNode e JobTracker para coordenar os nós slave e realizar o trabalho. Os slaves utilizam a função TaskTracker para administrar os trabalhos programados pelo JobTracker, SADH para armazenar dados e mapear e reduzir funções para computação de dados. O pacote básico do software inclui Hive e Pig* para linguagem e compiladores, HBase para administração da base de dados NoSQL e Apache Sqoop e Apache Flume* para coletar logs. Apache ZooKeeper* fornece uma coordenação centralizada do pacote.

O Custo da Análise de Big Data

Uma pesquisa de 2012 da InformationWeek aborda a questão econômica de Big Data, descobrindo que restrições de orçamento e outros problemas relacionados com custos são as principais barreiras para gerentes de TI. Desenvolver sua própria implantação de Apache Hadoop* e investir em armazenamento e desenvolvimento de recursos ou implantar uma solução de vendas pode representar custos significativos. Enquanto o armazenamento em nuvem oferece algum alívio potencial, modelos de custo para fornecedores de armazenamento público em nuvem podem não oferecer o suficiente. Com custos de armazenamento e de computação diminuindo constantemente, a implantação e a administração dos seus próprios clusters Hadoop* pode ser a solução mais econômica em relação a nuvens públicas e sistemas comerciais – mesmo adicionando o custo de uma pessoa capacitada para administrar o hardware.

Fonte: Biddick, Michael. "The Big Data Management Challenge." InformationWeek (Abril de 2012). <http://reports.informationweek.com/abstract/81/8766/business-intelligence-and-information-management/research-the-big-data-management-challenge.html>

Implantação do Apache Hadoop* em um Cluster de Nós de Servidor Padrão



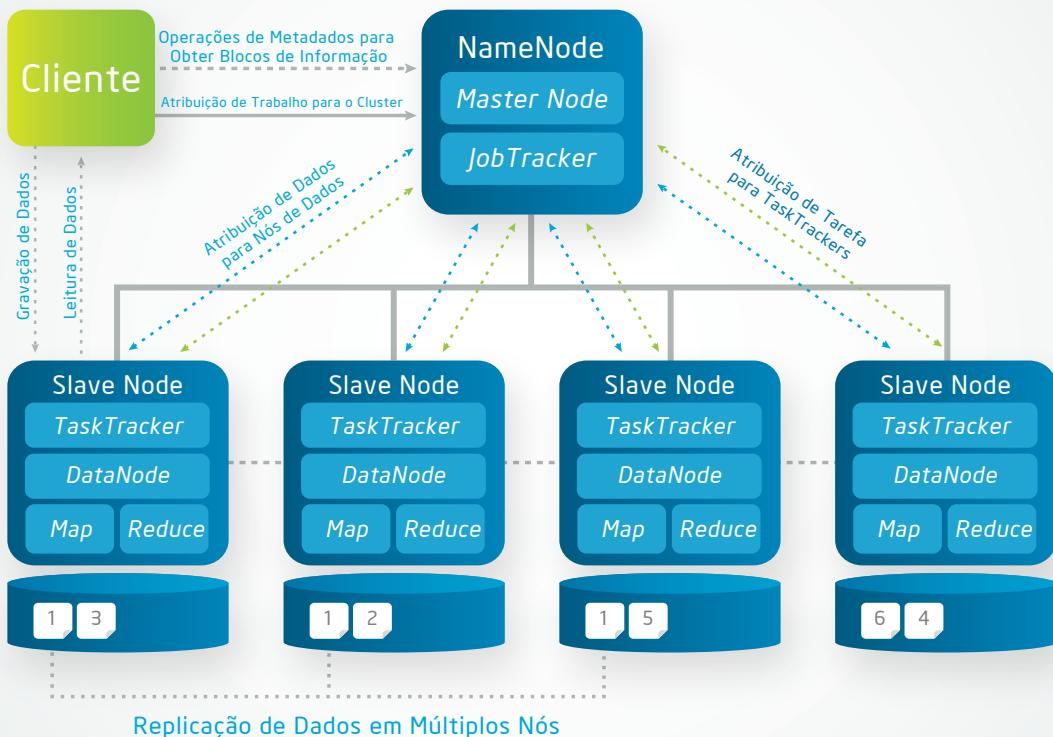
Fonte: Intel® Cloud Builders Guide to Cloud Design and Deployment on Intel® Platforms: Apache Hadoop*. Intel (Fevereiro de 2012).
intelcloudbuilders.com/docs/Intel_Cloud_Builders_Hadoop.pdf

Operando um Cluster de Servidor

Um cliente submete um trabalho ao nó mestre, que o organiza com os slaves no cluster. O JobTracker controla os trabalhos de MapReduce, reportando para o TaskTracker. Caso ocorra alguma falha, o JobTracker reprograma a tarefa para o mesmo nó slave ou para um diferente, qualquer que seja a escolha mais eficiente. O SADH lê dados de localização e de rede, e gerência dados no cluster, replicando dados em vários nós para maior confiabilidade de dados. Se uma das réplicas de dados no SADH for corrompida,

o JobTracker, por saber onde as outras réplicas estão localizadas, pode reprogramar a tarefa onde ela reside, diminuindo a necessidade de mover os dados de um nó para outro. Isso economiza banda larga da rede e mantém alta performance e disponibilidade. Depois de o trabalho ser mapeado, o output é organizado e dividido em vários grupos, que são redistribuídos para redutores. Os redutores podem estar localizados no mesmo nó dos mapeadores ou em outro nó.

Operação de um Cluster Apache Hadoop*



Trabalhos são organizados pelo nó mestre e processados nos nós slave

Infraestrutura Hadoop: Armazenamento e Rede de Big Data

Os clusters Hadoop são aprimorados por consideráveis melhorias nos recursos de computação e armazenamento mainstream e de armazenamento de recursos e são complementados por soluções de Ethernet de 10 Gigabites (10 GbE). A maior banda larga, associada com o 10 GbE, é essencial para importação e replicação de grandes conjuntos de dados entre servidores. Os Adaptadores de Rede Convergentes Intel® Ethernet de 10 Gigabites oferecem conexões de alto rendimento, e as Unidades de Estado Sólido (SSDs) Intel são unidades de disco de alta performance e alto

rendimento para armazenamento bruto. Para aumentar a eficiência, o armazenamento precisa ser apoiado por recursos avançados, como compressão, criptografia, hierarquização e automatização de dados, desduplicação de dados, capacidade de apagar codificação e provisionamento reduzido – tudo isso é atualmente compatível com a família de processadores Intel Xeon E5.

Adquira o guia como ter clusters [Hadoop equilibrados e eficientes com 10 GbE](#).

Arquitetura Intel: Clusters de Alta Performance

Utilizando servidores equipados com a Família de processadores Intel® Xeon® E5 como a plataforma base de servidores para o cluster, uma equipe de experts em Intel Big Data, rede e armazenamento, mediou os resultados de performance do Apache Hadoop* em várias combinações de rede e de componentes de armazenamento. Em geral, foram encontradas as seguintes opções de performance "boa, melhor e Excelente" com a infraestrutura Intel para seu ambiente Big Data. (Observe que certas variáveis podem impactar os resultados para o seu data center.)

Performance	Servidor	Rede	Armazenamento
Boa	Família de processadores Intel Xeon E5	Ethernet de 10 Gigabytes	Discos rígidos (HDs)
Ótima	Família de processadores Intel Xeon E5	10 GbE	Discos rígidos e Unidades de Estado Sólido (SSDs) com capacidade de armazenamento hierarquizada
Excelente	Família de processadores Intel Xeon E5	10 GbE	Unidades de Estado Sólido (SSDs)

Saiba mais sobre a [performance de cada combinação de plataforma](#).

Otimizar e ajustar para melhor performance

A Intel é uma grande apoiadora de iniciativas de código aberto como os softwares Linux*, OpenStack, KVM e Xen*. A Intel tem também dedicado recursos para análise, teste e caracterização de performance do Hadoop, tanto internamente como com parceiros como HP, Super Micro e Cloudera. Por meio desses esforços técnicos, a Intel tem observado que muitas compensações em hardware, software e configurações de sistema têm implicações no data center. Desenvolver um pacote de soluções para maximizar produtividade, limitar consumo de energia e reduzir o custo total pode ajudar a otimizar a utilização de recursos e ao mesmo tempo minimizar os custos operacionais.

A configuração do ambiente Hadoop é um fator essencial para a obtenção de todos os benefícios das soluções de hardware e software. Com base em extensos testes de benchmark no laboratório e em clientes utilizando arquitetura de processadores Intel, [as recomendações da Intel para otimização e ajuste do sistema Hadoop](#) podem ajudar a configurar e administrar seu ambiente Hadoop em termos de performance e custo.

Acertar as configurações exige um tempo inicial significante, pois as exigências para cada sistema empresarial Hadoop irão variar de acordo com o trabalho ou a carga de trabalho. O tempo dedicado à otimização para sua carga de trabalho específica irá valer a pena não apenas com a obtenção de uma melhor performance, mas também na diminuição dos custos totais do ambiente Hadoop.

Benchmarking

Benchmarking é a fundação quantitativa para mensurar a eficiência de qualquer sistema de computação. A Intel desenvolveu a suíte HiBench como uma abrangente série de testes de benchmark para ambientes Hadoop.⁸ As medidas individuais representam cargas de trabalho importantes para o Hadoop com uma série de características de uso de hardware. HiBench inclui microbenchmarks, além de aplicações reais Hadoop que representam uma vasta gama de análise de dados, como indexação de pesquisa e aprendizado da máquina. O HiBench 2.1 está disponível como um software de código aberto sob a Licença Apache 2.0 em <https://github.com/hibench/HiBench-2.1>.

HiBench: Os Detalhes

A suite HiBench da Intel analisa 10 cargas de trabalho em quatro categorias.

Categoria	Carga de Trabalho	Descrição
Microbenchmarks	Tipo	<ul style="list-style-type: none">▪ Esta carga de trabalho classifica seus dados de entrada de forma binária, que são gerados usando o exemplo RandomTextWriter do Apache Hadoop*.▪ Representação dos trabalhos do MapReduce, que transforma dados de um formato em outro.
	Word Count	<ul style="list-style-type: none">▪ Esta carga de trabalho conta as ocorrências de cada palavra nos dados de entrada, utilizando o RandomTextWriter do Hadoop*.▪ Representação de trabalhos do MapReduce, que extrai uma pequena quantidade de dados interessantes de um grande conjunto de dados.
	TeraSort	<ul style="list-style-type: none">▪ Um benchmark padrão para classificação de dados de grande porte gerado pelo programa TeraGen.
	Enhanced DFSIO	<ul style="list-style-type: none">▪ Testes do rendimento do sistema SADH* Apache* de um cluster Hadoop.▪ Calcula a banda larga agregada por meio de amostras do número de bytes lidos ou transmitidos em intervalos de tempo fixos em cada mapeamento de tarefa.
Busca na Web	Apache Nutch* Indexing	<ul style="list-style-type: none">▪ Esta carga de trabalho testa o subsistema de indexação em Nutch*, uma famosa ferramenta de busca de código aberto Apache. O subsistema crawler na ferramenta Nutch é usado para crawlear um espelho interno da Wikipedia* e gerar 8,4 GB de dados comprimidos (para cerca de 2,4 milhões de páginas da web) no total como entrada de carga de trabalho.▪ O sistema de indexação de larga escala é um dos usos mais significantes do MapReduce (por exemplo, nas plataformas do Google* e do Facebook*).
	Page Rank	<ul style="list-style-type: none">▪ Esta carga de trabalho é uma implementação de código aberto do algoritmo de ranqueamento de página, um algoritmo de análise de link bastante utilizado por ferramentas de busca da Web.
Aprendizagem da máquina	K-Means Clustering	<ul style="list-style-type: none">▪ Aplicação de área típica do MapReduce para data mining em larga escala e aprendizagem de máquina (por exemplo, nas plataformas Google e Facebook).▪ K-Means é um algoritmo de clustering bastante conhecido.
	Bayesian Classification	<ul style="list-style-type: none">▪ Área de aplicação típica do MapReduce para data mining de larga escala e aprendizado de máquina (por exemplo, nas plataformas Google e Facebook).▪ Esta carga de trabalho testa o treinador Bayesiano (um algoritmo de classificação bastante conhecido para descoberta e data mining) na biblioteca de código aberto Apache Mahout*.
Analytical query	Join Apache Hive*	<ul style="list-style-type: none">▪ Esta carga de trabalho simula complexas consultas de análise de estruturas de tabelas (relacionais) calculando a soma de cada grupo em uma única tabela apenas para leitura
	Hive* Aggregation	<ul style="list-style-type: none">▪ Esta carga de trabalho simula complexas consultas de análise de estruturas de tabelas (relacionais) calculando a média e a soma de cada grupo unindo duas tabelas diferentes.

Cinco Passos e Checklist: Começando seu Projeto de Análise de Big Data

Se você leu tudo até aqui, já adquiriu um bom entendimento da configuração de TI do Big Data, seu valor potencial para organizações e das tecnologias emergentes que podem ajudar a ter insights desses recursos de dados não estruturados. Além disso, você já adquiriu uma boa visão geral do básico para organizar a infraestrutura necessária e ajustar suas iniciativas Hadoop.

Você pode iniciar seu projeto de análise Big Data seguindo esses cinco passos.

Passo 1: Trabalhe com os usuários do seu negócio para articular as grandes oportunidades.

- Identifique e colabore com usuários do negócio (analistas, cientistas de dados, profissionais de marketing, etc.) para encontrar as melhores oportunidades de negócio para análises de Big Data em sua organização. Por exemplo, considere um problema de negócio existente – especialmente um que seja difícil, extenso ou impossível de resolver com suas fontes de dados e sistemas analíticos atuais. Ou considere um problema que nunca foi abordado antes porque as fontes de dados são novas e não estruturadas.
- Priorize sua lista de oportunidades e selecione um projeto com retorno sobre o investimento discernível.
- Determine as habilidades necessárias para você ser bem-sucedido na sua iniciativa.

Passo 2: Faça sua pesquisa para se atualizar em termos de tecnologia.

- converse com seus colegas de TI.
- Aproveite os recursos do Intel IT Center sobre Big Data.
- Entenda as ofertas dos fornecedores.
- Leia tutoriais e examine as documentações de usuários oferecidas pela Apache.

Passo 3: Desenvolva caso(s) de uso para seu projeto.

- Identifique os casos de uso necessários para realizar seu projeto.
- Mapeie fluxos de dados para ajudar a definir qual competências de tecnologia e de Big Data são necessárias para resolver seu problema de negócio.
- Decida quais dados incluir e quais deixar de fora. Identifique apenas os dados estratégicos que levarão a insights significantes.
- Determine como os dados irão interagir e a complexidade das regras de negócio.
- Identifique as consultas analíticas e os algoritmos necessários para gerar os resultados desejados.

Passo 4: Identifique lacunas entre as capacidades atuais e futuras.

- Quais exigências adicionais de qualidade de dados você terá para coletar, filtrar e agregar dados em formatos utilizáveis?
- Quais políticas de governança de dados você precisará estabelecer para classificar dados, definir sua relevância, além de armazenar, analisar e acessar dados?
- Quais capacidades de infraestrutura serão necessárias para garantir escalabilidade, baixa latência e performance?
- Como os dados serão apresentados para usuários? Os resultados precisam ser apresentados em um modo fácil de compreender para uma variedade de usuários do negócio, de executivos seniores e profissionais de informação.

Passo 5: Desenvolva um ambiente de teste para uma versão em produção.

- Adapte arquitetura de referência para o seu projeto. A Intel está trabalhando com parceiros de ponta para desenvolver arquitetura de referência para poder ajudar como parte do programa Intel Cloud Builders com base em casos de usuários Big Data.
- Defina a apresentação, a aplicação analítica, o armazenamento de dados e, se aplicável, a gestão de dados em nuvem privada ou pública.
- Determine quais ferramentas os usuários precisarão para apresentar resultados de modo significativo. A adoção de ferramentas pelo usuário irá influenciar significativamente o sucesso geral do seu projeto.

Recursos da Intel para Saber Mais

Sobre Big Data

Análise de Big Data (Página de Resumo)

Essa página agrega os principais recursos Intel para ajudar a implementar suas próprias iniciativas Big Data. Visite essa página no Intel IT Center para ler guias de planejamento, pesquisas, informações de soluções de vendas e estudos de casos reais.

intel.com/bigdata

Big Data Mining em Projetos para melhor Inteligência de Negócio

Essa página institucional de Intel IT descreve como a Intel está desenvolvendo os sistemas e a capacidade para analisar Big Data para motivar eficiência operacional e vantagem competitiva. A Intel IT, em parceria com outros grupos de negócios Intel, está desenvolvendo várias mostras de conceito para plataforma Big Data, incluindo detecção de malware, design de validação de chip, inteligência de mercado e sistema de recomendação.

intel.com/content/www/us/en/it-management/intel-it-best-practices/mining-big-data-in-the-enterprise-for-better-business-intelligence.html

Por dentro de TI: Big Data

Neste podcast, Moty Fania, que lidera a equipe de estratégia Intel para inteligência de negócios Big Data, fala sobre desenvolver as habilidades necessárias e a plataforma correta para lidar com Big Data.

<http://connectedsocialmedia.com/intel/5773/inside-it-big-data/>

Pesquisas: Análise de Big Data

Leia os resultados de uma pesquisa com 200 gerentes de TI que fornece insights sobre como as organizações estão utilizando análises de Big Data atualmente, incluindo o que as organizações precisam para avançar e o que a pesquisa diz sobre a indústria de TI. Os principais destaques estão no [vídeo](#) "Gerentes de TI falam sobre Análises de Big Data".

intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html

Grandes pensadores sobre Big Data

Uma série de entrevistas com líderes do pensamento sobre Big Data, incluindo o CEO da LiveRamp, Auren Hoffman, que fala sobre a revolução do Big Data aumentando a competição do Mercado; o principal analista da Forrester, Mike Gualtieri, sobre os próximos passos; e o CEO da Cognito, Joshua Feast, sobre Big Data, comportamento humano, e resultados de pesquisa

intel.com/content/www/us/en/big-data/big-thinkers-on-big-data.html

Sobre o Software Hadoop

Principais destaques do Apache Hadoop

Visite essa página para ouvir os especialistas da comunidade de código aberto Apache Hadoop explicando como os componentes de software do pacote Hadoop trabalham e para onde o desenvolvimento irá nos levar. Os podcasts trazem entrevistas com Alan Gates (Hortonworks) sobre HCatalog e Pig, Konstantin Shvachko (AltoScale) sobre SADH, Deveraj Das (Hortonworks) sobre MapReduce e Carl Steinbach (Cloudera) sobre Hive.

intel.com/content/www/us/en/big-data/big-data-apache-hadoop-framework-spotlights-landing.html

Guia Intel® Cloud Builders para Cloud Design e Implantação de Plataformas Intel®: Apache Hadoop**

Essa arquitetura de referência é para organizações que querem desenvolver sua própria infraestrutura de computação em nuvem (cloud computing), incluindo clusters Apache Hadoop para administrar Big Data. Inclui passos para desenvolver a implantação de seu ambiente de laboratório de Data Center e contém detalhes sobre topologia, hardware, software, instalação, configuração e teste do Hadoop. A implementação desta arquitetura de referência irá ajudar a começar a desenvolver e operar sua própria infraestrutura Hadoop..

intelcloudbuilders.com/docs/Intel_Cloud_Builders_Hadoop.pdf

*Otimizando Implementações do Hadoop**

Essa página institucional oferece orientação para organizações sobre como planejar implementações do Hadoop. Com base em extensivos testes de laboratório com o software do Hadoop na Intel, descreve as melhores práticas para estabelecer especificações de servidores hardware, discute o ambiente de servidor do software e dá dicas de configuração e ajustes que podem melhorar a performance.

intel.com/content/www/us/en/cloud-computing/cloud-computing-optimizing-hadoop-deployments-paper.html

Recursos Adicionais

Big Data: Como Aproveitar um Ativo que Pode Mudar o Jogo

Este relatório da Economist Intelligence Unit, com patrocínio do SAS, analisa Big Data e seu impacto em empresas. A pesquisa examina as características organizacionais de empresas que já extraem valores de dados e encontrou uma forte relação entre gestão eficiente de dados e desempenho financeiro. Essas empresas estabelecem modelos para como as organizações precisam evoluir para administrar e agregar valor com eficiência usando Big Data.

sas.com/resources/asset/SAS_BigData_final.pdf

A Onda Forrester™: Soluções Empresariais Hadoop, Trimestre 1 2012

Este relatório, elaborado por James Kobielski na Forrester, avalia 13 soluções empresariais Hadoop, aplicando uma avaliação de 15 critérios para cada. Os líderes incluem Amazon Web Services, IBM, EMC Greenplum, MapR, Cloudera e Hortonworks.

forrester.com/The+Forrester+Wave+Enterprise+Hadoop+Solutions+Q1+2012/quickscan/-/E-RES60755

Notas de fim

1. Gens, Frank. IDC Predictions 2012: Competing for 2020. IDC (Dezembro de 2011). <http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf>.
2. "Big Data Infographic and Gartner 2012 Top 10 Strategic TechTrends." Business Analytics 3.0 (blog) (11 de novembro de 2011). <http://practicalanalytics.wordpress.com/2011/11/11/big-data-infographic-and-gartner-2012-top-10-strategic-tech-trends/>
3. "Global Internet Traffic Projected to Quadruple by 2015." The Network (press release) (1º de junho de 2011). <http://newsroom.cisco.com/press-release-content?type=webcontent&articleId=324003>
4. *Big Data : The Next Frontier for Innovation, Competition, and Productivity.* McKinsey Global Institute (Maio de 2011). mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.pdf
5. *Peer Research on Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data .* Intel (Agosto de 2012) intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html
6. O software Nutch* foi inicialmente um projeto independente de código aberto criado por Doug Cutting e Mike Cafarella. Em 2005, Nutch começou a ser administrado pela Apache Software Foundation, primeiro como um subprojeto do software de busca Apache Lucene* e, em 2010, como um projeto de alto nível da Apache Software Foundation. Fonte: "Nutch Joins Apache Incubator" (release para imprensa). Apache Software Foundation (Janeiro de 2005). nutch.apache.org/#January+2005%3A+Nutch+Joins+Apache+Incubator
7. "Hadoop Hits Primetime with Production Release." Datanami (6 de janeiro de 2012). datanami.com/datanami/2012-01-06/hadoop_hits_primetime_with_production_release.html
8. Huang, Shengsheng, Jie Huang, Jinquan Dai, Tao Xie, Bo Huang. The HiBench Benchmark Suite: Characterization of the MapReduce-Based Data Analysis. IEEE (Março de 2010).

Mais sobre o Intel® IT Center

Guia de Planejamento: Saiba Mais sobre Big Data é oferecido pelo [Intel® IT Center](#), o programa da Intel para profissionais de TI. O Intel IT Center foi desenvolvido para oferecer informação objetiva e direta para ajudar profissionais de TI a implementar projetos estratégicos em sua agenda, incluindo virtualização, desenvolvimento de Data Center, nuvem e segurança para clientes e infraestrutura. Visite o Intel IT Center para:

- Guias de Planejamento, pesquisas e soluções de destaque para ajudar na implementação de projetos importantes
- Estudos de casos reais que mostram como seus colegas de mercado abordaram os mesmos desafios que você enfrenta
- Informações sobre como a própria organização de TI da Intel está implantando nuvem, virtualização, segurança e outras iniciativas estratégicas
- Informações sobre eventos nos quais você pode ouvir os especialistas de produtos da Intel e outros profissionais da Intel

Saiba mais em intel.com/ITCenter.

Compartilhe com seus Colegas

Este documento tem fins exclusivamente informativos. ESTE DOCUMENTO É OFERECIDO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIAS DE QUALQUER ESPÉCIE, INCLUINDO NENHUMA GARANTIA DE COMERCIALIZAÇÃO, LEGALIDADE, ADEQUAÇÃO PARA UM DETERMINADO FIM OU QUALQUER OUTRA GARANTIA DE QUALQUER OUTRO PROPÓSITO, ESPECIFICAÇÃO OU AMOSTRA. A Intel se isenta de qualquer responsabilidade, incluindo responsabilidade por violação de qualquer direito de propriedade, em relação ao uso destas informações. Nenhuma licença, expressa ou implícita, por embargo ou de qualquer outra forma, a qualquer direito de propriedade intelectual é concedida aqui.

Copyright 2013, Intel Corporation. Todos os direitos reservados. Intel, o logotipo da Intel, Intel Apaixonados pelo Futuro, o logotipo do Intel Apaixonados pelo Futuro e Xeon são marcas comerciais da Intel Corporation nos Estados Unidos e/ou em outros países.

*Outros nomes e marcas podem ser propriedade de terceiros.

Microsoft é uma marca registrada da Microsoft Corporation nos Estados Unidos e/ou em outros países..