

Big Data e Machine Learning na Nuvem

Fundamentos do GCP: infraestrutura principal



Primeiros passos com o BigQuery



Google Cloud

Última modificação 2018-08-24

© 2017 Google Inc. Todos os direitos reservados. Google e o logotipo da Google são marcas registradas da Google Inc. Todos os outros nomes de empresas e produtos podem ser marcas registradas das respectivas empresas de seus respectivos países.

© 2017 Google Inc. Todos os direitos reservados. Google e o logotipo da Google são marcas registradas da Google Inc. Todos os outros nomes de empresas e produtos podem ser marcas registradas das respectivas empresas de seus respectivos países.

Agenda

Plataforma de Big Data do Google Cloud

Plataforma de aprendizado de máquina do Google Cloud

Teste e laboratório

Os serviços de big data do Google Cloud são totalmente gerenciados e escaláveis



Nuvem
Dataproc

Gerenciou
Hadoop
MapReduce,
Faísca, Porco e
Serviço de colmeia



Nuvem
Fluxo de dados

Transmitir e
lote
em processamento;
unificado e
simplificado
oleodutos



BigQuery

Análise
base de dados;
transmitir dados em
100.000
linhas por segundo



Nuvem
Pub/Sub

Escalável e
flexível
empreendimento
Mensagens



Nuvem
Datalab

Dados interativos
exploração

Google Cloud

As soluções de Big Data do Google Cloud foram projetadas para ajudar você a transformar suas experiências de negócios e de usuários com insights de dados significativos. É uma plataforma integrada e sem servidor. "Sem servidor" significa que você não precisa provisionar instâncias de computação para executar seus trabalhos. Os serviços são totalmente gerenciados e você paga apenas pelos recursos que consumir. A plataforma é "integrada" para que os serviços de dados do GCP trabalhem juntos para ajudar você a criar soluções personalizadas.

O Cloud Dataproc é gerenciado pelo Hadoop

- Maneira rápida, fácil e gerenciada de executar Hadoop e Spark/Hive/Pig no GCP
- Crie clusters em 90 segundos ou menos, em média.
- Dimensione os clusters para cima e para baixo mesmo quando os trabalhos estiverem em execução.



 Google Cloud

Apache Hadoop é uma estrutura de código aberto para big data. É baseado no modelo de programação MapReduce, que o Google inventou e publicou. O modelo MapReduce, em sua forma mais simples, significa que uma função – tradicionalmente chamada de função “map” – é executada em paralelo em um conjunto de dados maciço para produzir resultados intermediários; e outra função – tradicionalmente chamada de função “reduce” – cria um conjunto de resultados final com base em todos esses resultados intermediários. O termo “Hadoop” é frequentemente usado informalmente para abranger o próprio Apache Hadoop e projetos relacionados, como Apache Spark, Apache Pig e Apache Hive.

O Cloud Dataproc é uma maneira rápida, fácil e gerenciada de executar Hadoop, Spark, Hive e Pig no Google Cloud Platform. Tudo o que você precisa fazer é solicitar um cluster Hadoop. Ele será criado para você em 90 segundos ou menos, com base nas máquinas virtuais do Compute Engine cujo número e tipo você pode controlar. Se você precisar de mais ou menos poder de processamento durante a execução do cluster, poderá escalá-lo para cima ou para baixo. Você pode usar a configuração padrão do software Hadoop em seu cluster ou personalizá-la. E você pode monitorar seu cluster usando o Stackdriver.

Por que usar o Cloud Dataproc?

- Migre facilmente trabalhos do Hadoop no local para a nuvem.
- Analisar rapidamente os dados (como dados de log) armazenados no Cloud Storage; crie um cluster em 90 segundos ou menos, em média, e exclua-o imediatamente.
- Use Spark/Spark SQL para realizar mineração e análise de dados rapidamente.
- Use as bibliotecas de aprendizado de máquina Spark (MLlib) para executar algoritmos de classificação.



 Google Cloud

A execução de trabalhos do Hadoop no local requer um investimento em hardware. Por outro lado, a execução desses jobs no Cloud Dataproc permite que você pague apenas pelos recursos de hardware durante a vida útil do cliente efêmero que você cria. Você pode economizar ainda mais usando [instâncias preemptivas para processamento em lote](#).

Você também pode economizar dinheiro dizendo ao Cloud Dataproc para usar instâncias preemptivas do Compute Engine para o processamento em lote. Você precisa garantir que seus trabalhos possam ser reiniciados de forma limpa se forem encerrados e você obtiver uma redução significativa no custo das instâncias. Na época em que este vídeo foi feito, as instâncias preemptivas eram cerca de 80% mais baratas. Esteja ciente de que o custo das instâncias do Compute Engine não é o único componente do custo de um cluster do Dataproc, mas é significativo.

Quando seus dados estiverem em um cluster, você poderá usar Spark e Spark SQL para fazer mineração de dados e usar MLlib, que são as bibliotecas de aprendizado de máquina do Apache Spark, para descobrir padrões por meio do aprendizado de máquina.

O Cloud Dataflow oferece pipelines de dados gerenciados

- Processa dados usando instâncias do Compute Engine.

- ☐ Os clusters são dimensionados para você
- ☐ Escalonamento automatizado, sem necessidade de provisionamento de instância

- Escreva o código uma vez e obtenha *lote e transmissão*.

- ☐ Modelo de programação baseado em transformação



 Google Cloud

O Cloud Dataproc é ótimo quando você tem um conjunto de dados de tamanho conhecido ou quando você mesmo deseja gerenciar o tamanho do cluster. Mas e se seus dados aparecerem em tempo real? Ou é de tamanho ou taxa imprevisível? É aí que o Cloud Dataflow é uma escolha particularmente boa. É um modelo de programação unificado e um serviço gerenciado, e permite desenvolver e executar uma grande variedade de padrões de processamento de dados: extrair-transformar-e-carregar, computação em lote e computação contínua. Você usa o Dataflow para criar pipelines de dados, e os mesmos pipelines funcionam para dados em lote e streaming.

O Dataflow é um modelo de programação unificado e um serviço gerenciado para desenvolver e executar uma ampla variedade de padrões de processamento de dados, incluindo ETL, computação em lote e computação contínua. O Cloud Dataflow libera você de tarefas operacionais, como gerenciamento de recursos e otimização de desempenho.

Recursos do Cloud Dataflow:

Gestão de recursos

O Cloud Dataflow automatiza totalmente o gerenciamento dos recursos de processamento necessários. Chega de criar instâncias manualmente.

Sob demanda

Todos os recursos são fornecidos sob demanda, permitindo que você escale para atender às suas necessidades de negócios. Não há necessidade de comprar instâncias de computação reservadas.

Agendamento de trabalho inteligente

Particionamento de trabalho automatizado e otimizado que pode reequilibrar dinamicamente o trabalho atrasado. Chega de perseguir “teclas de atalho” ou pré-processar seus dados de entrada.

Escalonamento automático

O dimensionamento automático horizontal dos recursos do trabalhador para atender aos requisitos de produtividade ideais resulta em melhor relação preço/desempenho geral.

Modelo de Programação Unificado

A API do Dataflow permite que você expresse operações como MapReduce, janelas de dados poderosas e controle de precisão refinado, independentemente da fonte de dados.

Código aberto

Os desenvolvedores que desejam estender o modelo de programação do Dataflow podem bifurcar e/ou enviar solicitações pull no SDK do Cloud Dataflow baseado em Java. Os pipelines do Dataflow também podem ser executados em tempos de execução alternativos, como Spark e Flink.

Monitoramento

Integrado ao Console do Google Cloud Platform, o Cloud Dataflow fornece estatísticas como taxa de transferência e atraso do pipeline, além de inspeção consolidada de registros de trabalhadores, tudo quase em tempo real.

Integrado

Integra-se ao Cloud Storage, Cloud Pub/Sub, Cloud Datastore, Cloud Bigtable e BigQuery para processamento de dados perfeito. E pode ser estendido para interagir com outras fontes e coletores como Apache Kafka e HDFS.

Processamento confiável e consistente

O Cloud Dataflow oferece suporte integrado para execução tolerante a falhas consistente e correta, independentemente do tamanho dos dados, tamanho do cluster, padrão de processamento ou complexidade do pipeline.

Os pipelines do Dataflow fazem o fluxo de dados de uma origem por meio de

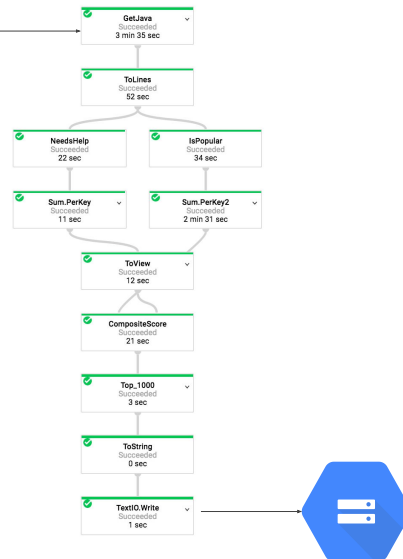
transformações

Fonte



Transformações

Afundar



Armazenamento na nuvem



Este exemplo de pipeline do Dataflow lê dados de uma tabela do BigQuery (a "origem"), processa-os de várias maneiras (as "transformações") e grava sua saída no Cloud Storage (o "coletor"). Algumas dessas transformações que você vê aqui são operações de mapa e algumas são operações de redução. Você pode construir pipelines realmente expressivos.

Cada etapa do pipeline é dimensionada elasticamente. Não há necessidade de iniciar e gerenciar um cluster. Em vez disso, o serviço fornece todos os recursos sob demanda. Ele possui particionamento de trabalho automatizado e otimizado integrado, que pode reequilibrar dinamicamente o trabalho atrasado. Isso reduz a necessidade de se preocupar com "teclas de atalho" – ou seja, situações em que pedaços desproporcionalmente grandes de sua entrada são mapeados para o mesmo usuário.

Por que usar o Cloud Dataflow?

- *ETL*(extrair/transformar/carregar) pipelines para mover, filtrar, enriquecer, moldar dados
- *Análise de dados*: computação em lote ou computação contínua usando streaming
- *Orquestração*: criar pipelines que coordenam serviços, incluindo serviços externos
- Integra-se a serviços do GCP, como Cloud Storage, Cloud Pub/Sub, BigQuery e Bigtable
 - SDKs Java e Python de código aberto



 Google Cloud

As pessoas usam o Dataflow em vários casos de uso. Por um lado, serve bem como uma ferramenta ETL de uso geral.

E seu caso de uso como mecanismo de análise de dados é útil em coisas como: detecção de fraudes em serviços financeiros; Análise de IoT em manufatura, saúde e logística; e clickstream, Ponto de Venda e análise de segmentação no varejo.

E, como esses pipelines que vimos podem orquestrar vários serviços, até mesmo serviços externos, eles podem ser usados em aplicativos em tempo real, como personalizar experiências de usuário de jogos.

O BigQuery é um armazenamento de dados totalmente gerenciado

- Fornece análise interativa quase em tempo real de grandes conjuntos de dados (centenas de TBs)
- Consulta usando a sintaxe SQL (SQL 2011)
- Nenhuma manutenção de cluster é necessária.



 Google Cloud

Se, em vez de um pipeline dinâmico, você quiser fazer consultas SQL ad-hoc em um grande conjunto de dados, é para isso que serve o BigQuery. O BigQuery é o armazenamento de dados analítico de baixo custo e escala de petabytes totalmente gerenciado do Google.

O BigQuery é o armazenamento de dados analítico de baixo custo e escala de petabytes totalmente gerenciado do Google. BigQuery é NoOps: não há infraestrutura para gerenciar e você não precisa de um administrador de banco de dados, então você pode se concentrar na análise de dados para encontrar insights significativos, usar SQL familiar e aproveitar nosso modelo de pagamento conforme o uso. O BigQuery é uma poderosa plataforma de análise de big data usada por todos os tipos de organizações, de startups a empresas da Fortune 500.

Recursos do BigQuery:

Ingestão de dados flexível

Carregue seus dados do Cloud Storage ou do Cloud Datastore ou transmita-os para o BigQuery a 100.000 linhas por segundo para permitir a análise em tempo real dos seus dados.

Disponibilidade Global

Você tem a opção de armazenar seus dados do BigQuery em locais europeus enquanto continua se beneficiando de um serviço totalmente gerenciado, agora com a opção de controle de dados geográficos, sem manutenção de cluster de baixo nível.

Segurança e permissões

Você tem controle total sobre quem tem acesso aos dados armazenados no BigQuery. Se você compartilhar conjuntos de dados, isso não afetará seu custo ou desempenho; aqueles com quem você compartilha pagam por suas próprias consultas.

Controles de custo

O BigQuery oferece mecanismos de controle de custos que permitem limitar seus custos diários no valor que você escolher. Para mais informações, veja [Controles de custo](#).

Altamente disponível

A replicação de dados transparente em várias geografias significa que seus dados estão disponíveis e duráveis, mesmo no caso de modos de falha extremos.

Desempenho super rápido

Execute consultas SQL super-rápidas em vários terabytes de dados em segundos, usando o poder de processamento da infraestrutura do Google.

Totalmente integrado

Além das consultas SQL, você pode ler e gravar dados facilmente no BigQuery por meio do Cloud Dataflow, Spark e Hadoop.

Conecte-se com os produtos do Google

Você pode exportar automaticamente seus dados do Google Analytics Premium para o BigQuery e analisar conjuntos de dados armazenados no Google Cloud Storage, Google Drive e Planilhas Google.

O BigQuery pode fazer alterações de criação, substituição, atualização e exclusão em bancos de dados, sujeitas a [algumas limitações](#) e com certo [problemas conhecidos](#).

O BigQuery é executado na infraestrutura de alto desempenho do Google

- A computação e o armazenamento são separados por uma rede terabit entre eles
- Você paga apenas pelo armazenamento e processamento usados
- Desconto automático para armazenamento de dados de longo prazo



 Google Cloud

É fácil inserir dados no BigQuery. Você pode carregar do Cloud Storage ou do Cloud Datastore ou transmiti-lo para o BigQuery em até 100.000 linhas por segundo.

O BigQuery é usado por todos os tipos de organizações, de startups a empresas da Fortune 500. Organizações menores, como as cotas mensais gratuitas do BigQuery. Organizações maiores gostam de sua escala perfeita e seu acordo de nível de serviço disponível de 99,9%.

[Preços de armazenamento de longo prazo](#) é um desconto automático para dados que residem no BigQuery por longos períodos. Quando a idade dos seus dados atingir 90 dias no BigQuery, o Google reduzirá automaticamente o preço do armazenamento de US\$ 0,02 por GB por mês para US\$ 0,01 por GB por mês.

Para mais informações sobre a arquitetura do BigQuery, consulte: <https://cloud.google.com/blog/big-data/2016/01/bigquery-under-the-hood>

O Cloud Pub/Sub é um sistema de mensagens escalonável e confiável

- Suporta mensagens assíncronas de muitos para muitos
 - Os componentes do aplicativo fazem assinaturas push/pull para tópicos
- Inclui suporte para consumidores offline
- Com base em tecnologias comprovadas do Google
- Integra-se ao Cloud Dataflow para pipelines de processamento de dados



 Google Cloud

O Cloud Pub/Sub é um serviço de mensagens em tempo real totalmente gerenciado que permite enviar e receber mensagens entre aplicativos independentes. Você pode aproveitar a flexibilidade do Cloud Pub/Sub para desacoplar sistemas e componentes hospedados no Google Cloud Platform ou em outro lugar na Internet. Com base na mesma tecnologia que o Google usa, o Cloud Pub/Sub foi projetado para fornecer "pelo menos uma vez" entrega em baixa latência com escalabilidade sob demanda para 1 milhão de mensagens por segundo (e além).

Recursos do Cloud Pub/Sub:

Altamente escalável

Qualquer cliente pode enviar até 10.000 mensagens por segundo, por padrão – e milhões por segundo e mais, mediante solicitação.

Entrega de empurrar e puxar

Os assinantes têm opções de entrega flexíveis, sejam acessíveis pela Internet ou por trás de um firewall.

Criptografia

A criptografia de todos os dados de mensagens em trânsito e em repouso fornece segurança e proteção aos dados.

Armazenamento replicado

Projetado para fornecer entrega de mensagens “pelo menos uma vez”, armazenando cada mensagem em vários servidores em várias zonas.

Fila de mensagens

Construa uma fila de mensagens altamente escalável usando um único tópico e assinatura para suportar um padrão de comunicação um para um.

Reconhecimento de ponta a ponta

Construir aplicativos confiáveis é mais fácil com reconhecimentos explícitos no nível do aplicativo.

Espalham

Publique mensagens em um tópico uma vez e vários assinantes receberão cópias para oferecer suporte a padrões de comunicação um para muitos ou muitos para muitos.

API REST

Interface simples e sem estado usando mensagens JSON com bibliotecas de API em muitas linguagens de programação.

Por que usar o Cloud Pub/Sub?

- Bloco de construção para ingestão de dados no Dataflow, Internet das Coisas (IoT), Marketing Analytics
- Base para streaming do Dataflow
- Notificações push para aplicativos baseados em nuvem
- Conectar aplicativos no Google Cloud Platform (push/pull entre o Compute Engine e o App Engine)



 Google Cloud

O Cloud Pub/Sub se baseia na mesma tecnologia que o Google usa internamente. É um bloco de construção importante para aplicativos em que os dados chegam em taxas altas e imprevisíveis, como sistemas de Internet das Coisas. Se você estiver analisando dados de streaming, o Cloud Dataflow é um pareamento natural com o Pub/Sub.

O Cloud Datalab oferece exploração interativa de dados

- Ferramenta interativa para exploração, transformação, análise e visualização de dados em larga escala
- Integrado, de código aberto
 - Construído em Jupyter (anteriormente IPython)



 Google Cloud

Para ciência de dados, uma metáfora de notebook de laboratório online é um ambiente útil, porque parece natural intercalar análises de dados com comentários sobre seus resultados. Um sistema de código aberto popular para hospedá-los é o Projeto Jupyter. Ele permite criar e manter notebooks baseados na Web contendo código Python, e você pode executar esse código interativamente e visualizar os resultados.

O Cloud Datalab permite usar notebooks Jupyter para explorar, analisar e visualizar dados no Google Cloud Platform. Ele é executado em uma máquina virtual do Compute Engine. Para começar, você especifica o tipo de máquina virtual que deseja e em qual região do GCP ela deve ser executada. Ao iniciar, ela apresenta um ambiente Python interativo que está pronto para uso. E orquestra vários serviços do GCP automaticamente para que você possa se concentrar em explorar seus dados. Você paga apenas pelos recursos que usa; não há cobrança adicional para o próprio Datalab.

Recursos do Cloud Datalab:

Integrado

O Cloud Datalab lida com a autenticação e a computação na nuvem imediatamente e é integrado ao BigQuery, Compute Engine e Cloud Storage.

Suporte a vários idiomas

Atualmente, o Cloud Datalab é compatível com Python, SQL e JavaScript (para funções definidas pelo usuário do BigQuery).

Formato do bloco de anotações

O Cloud Datalab combina código, documentação, resultados e visualizações em um formato de notebook intuitivo.

Preços de pagamento por uso

Pague apenas pelos recursos de nuvem que usar: o aplicativo do App Engine, o BigQuery e quaisquer recursos adicionais que você decidir usar, como o Cloud Storage.

Visualização de dados interativa

Use o Google Charts ou matplotlib para facilitar as visualizações.

Colaborativo

Controle de fonte de notebooks baseado em Git com a opção de sincronizar com repositórios de código-fonte que não são do Google, como GitHub e Bitbucket.

Código aberto

Os desenvolvedores que desejam estender o Cloud Datalab podem bifurcar e/ou enviar pull requests no [Projeto hospedado no GitHub](#).

Implantação personalizada

Especifique seus requisitos mínimos de VM, o host de rede e muito mais.

Suporte ao IPython

O Cloud Datalab é baseado no Jupyter (anteriormente IPython) para que você possa usar um grande número de pacotes existentes para estatísticas, aprendizado de máquina etc. Aprenda com os notebooks publicados e troque dicas com uma comunidade vibrante do IPython.

Por que usar o Cloud Datalab?

- Crie e gerencie código, documentação, resultados e visualizações em formato de notebook intuitivo.
 - Use o Google Charts ou matplotlib para facilitar as visualizações.
- Analisar dados no BigQuery, Compute Engine e Cloud Storage usando Python, SQL e JavaScript.
- Implante modelos facilmente no BigQuery.



 Google Cloud

O Cloud Datalab é integrado ao BigQuery, Compute Engine e Cloud Storage, portanto, o acesso aos seus dados não envolve problemas de autenticação.

Quando estiver funcionando, você poderá visualizar seus dados com o Google Charts ou matplotlib. E, como há uma comunidade Python interativa vibrante, você pode aprender com os notebooks publicados. Existem muitos pacotes existentes para estatísticas, aprendizado de máquina e assim por diante.

Você pode [anexar uma GPU a uma instância do Cloud Datalab](#) para processamento mais rápido. No momento da redação deste artigo, esse recurso estava em versão beta, o que significa que nenhum SLA está disponível e que o recurso pode ser alterado de maneiras incompatíveis com versões anteriores.

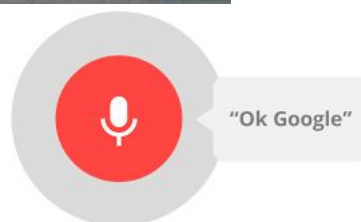
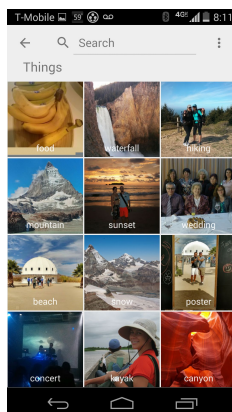
Agenda

Plataforma de Big Data do Google Cloud

Plataforma de aprendizado de máquina do Google Cloud

Teste e laboratório

APIs de aprendizado de máquina habilitar aplicativos que veem, ouvem e entendem



O aprendizado de máquina é um ramo do campo da inteligência artificial. É uma maneira de resolver problemas sem codificar explicitamente a solução. Em vez disso, os codificadores humanos constroem sistemas que se aprimoram ao longo do tempo, por meio da exposição repetida a dados de amostra, que chamamos de “dados de treinamento”.

Os principais aplicativos do Google usam aprendizado de máquina, como YouTube, Fotos, o aplicativo para dispositivos móveis do Google e o Google Tradutor. A plataforma de aprendizado de máquina do Google agora está disponível como um serviço de nuvem, para que você possa adicionar recursos inovadores aos seus próprios aplicativos.

Plataforma de aprendizado de máquina em nuvem



Ferramenta de código aberto para criar e executar modelos de rede neural

- Amplo suporte a plataformas: CPU ou GPU; móvel, servidor ou nuvem



Cloud ML

Serviço de aprendizado de máquina totalmente gerenciado

- Experiência familiar de desenvolvedor baseada em notebook
- Otimizado para infraestrutura do Google; integra-se ao BigQuery e ao Cloud Storage



APIs de aprendizado de máquina

Modelos de aprendizado de máquina pré-treinados criados pelo Google

- Fala: Transmita resultados em tempo real, detecta 80 idiomas
- Visão: identificar objetos, pontos de referência, texto e conteúdo
- Traduzir: tradução de idiomas, incluindo detecção
- Linguagem natural: estrutura, significado do texto



A Cloud Machine Learning Platform oferece serviços modernos de machine learning, com modelos pré-treinados e uma plataforma para gerar seus próprios modelos personalizados. Assim como em outros produtos do GCP, há uma gama de serviços que vai do mais geral ao pré-personalizado.

O TensorFlow é uma biblioteca de software de código aberto excepcionalmente adequada para aplicativos de aprendizado de máquina, como redes neurais. Foi desenvolvido pelo Google Brain para uso interno do Google e, em seguida, de código aberto para que o mundo pudesse se beneficiar. Você pode executar o TensorFlow onde quiser, mas o GCP é o local ideal para isso, porque os modelos de machine learning precisam de muitos recursos de computação sob demanda e muitos dados de treinamento. O TensorFlow também pode aproveitar as Tensor Processing Units, que são dispositivos de hardware projetados para acelerar as cargas de trabalho de machine learning com o TensorFlow. O GCP os disponibiliza na nuvem com as máquinas virtuais do Compute Engine. Cada Cloud TPU oferece até 180 teraflops de desempenho e, como você paga apenas pelo que usa, não há necessidade de investimento de capital inicial.

Suponha que você queira um serviço mais gerenciado. O Google Cloud Machine Learning Engine permite criar facilmente modelos de machine learning que funcionam em qualquer tipo de

dados, de qualquer tamanho. Ele pode usar qualquer modelo do TensorFlow e realizar treinamento em larga escala em um cluster gerenciado.

Por fim, suponha que você queira apenas adicionar vários recursos de aprendizado de máquina aos seus aplicativos, sem precisar se preocupar com os detalhes de como eles são fornecidos. O Google Cloud também oferece uma variedade de APIs de aprendizado de máquina adequadas para finalidades específicas, e as discutirei em breve.

Por que usar a plataforma Cloud Machine Learning?

Para dados estruturados



Classificação e regressão



Recomendação



Detecção de anomalia

Para dados não estruturados



Imagem e vídeo análise



Análise de texto



As pessoas usam a plataforma Cloud Machine Learning para muitos aplicativos. Geralmente, eles se enquadram em duas categorias, dependendo se os dados em que trabalham são estruturados ou não estruturados.

Com base em dados estruturados, você pode usar o ML para vários tipos de tarefas de classificação e regressão, como análise de perda de clientes, diagnóstico de produtos e previsão. Pode ser o coração de um mecanismo de recomendação, para personalização de conteúdo e vendas cruzadas e adicionais. Você pode usar o ML para detectar anomalias, como na detecção de fraudes, diagnósticos de sensores ou métricas de log.

Com base em dados não estruturados, você pode usar o ML para análise de imagens, como identificar remessa danificada, identificar estilos e sinalizar conteúdo. Você também pode fazer análises de texto, como análise de log do call center, identificação de idioma, classificação de tópicos e análise de sentimentos.

Em muitos dos aplicativos mais inovadores para aprendizado de máquina, vários desses tipos de aplicativos são combinados. E se, sempre que um de seus clientes publicasse elogios a um de seus produtos nas mídias sociais, seu aplicativo pudesse alcançá-los automaticamente com um desconto personalizado

em outro produto que eles provavelmente vão gostar? A plataforma Google Cloud Machine Learning torna esse tipo de interatividade ao seu alcance.

API Cloud Vision

- Analise imagens com uma API REST simples

- Detecção de logotipo, detecção de etiqueta, etc.

- Com a API Cloud Vision, você pode:

- Obtenha insights de imagens
 - Detectar conteúdo impróprio
 - Analise o sentimento
 - Extrair texto



A API Cloud Vision permite que os desenvolvedores entendam o conteúdo de uma imagem encapsulando modelos de aprendizado de máquina avançados em uma API REST fácil de usar. Ele rapidamente classifica as imagens em milhares de categorias ("veleiro", "leão", "Torre Eiffel"), detecta objetos individuais nas imagens e encontra e lê palavras impressas contidas nas imagens. Você pode criar metadados em seu catálogo de imagens, moderar conteúdo ofensivo ou habilitar novos cenários de marketing por meio de análise de sentimento de imagem. Analise as imagens enviadas na solicitação ou integre ao seu armazenamento de imagens no Cloud Storage.

API Cloud Speech

- Reconhece mais de 80 idiomas e variantes
- Pode retornar texto em tempo real
- Altamente preciso, mesmo em ambientes ruidosos
- Acesso de qualquer dispositivo
- Com tecnologia de aprendizado de máquina do Google



 Google Cloud

A API Cloud Speech permite que os desenvolvedores convertam áudio em texto. Como você tem uma base de usuários cada vez mais global, a API reconhece mais de 80 idiomas e variantes. Você pode transcrever o texto de usuários ditando para o microfone de um aplicativo, habilitar comando e controle por voz ou transcrever arquivos de áudio.

API Cloud Natural Language

- Usa modelos de aprendizado de máquina para revelar a estrutura e o significado do texto.
- Extraia informações sobre itens mencionados em documentos de texto, artigos de notícias e postagens em blogs.
- Analise o texto carregado na solicitação ou integre-o ao Cloud Storage.



 Google Cloud

A API Cloud Natural Language oferece uma variedade de tecnologias de compreensão de linguagem natural para desenvolvedores.

Ele pode fazer análise de sintaxe, quebrando frases fornecidas por seus usuários em tokens, identificar os substantivos, verbos, adjetivos e outras partes do discurso e descobrir as relações entre as palavras.

Ele pode fazer o reconhecimento de entidades: em outras palavras, pode analisar texto e sinalizar menções de pessoas, organizações, locais, eventos, produtos e mídia.

Ele pode entender o sentimento geral expresso em um bloco de texto.

E possui esses recursos em vários idiomas, incluindo inglês, espanhol e japonês.

Recursos da API Cloud Natural Language

Análise de sintaxe

- Extraia tokens e frases, identifique partes do discurso (PoS) e crie árvores de análise de dependência para cada frase.

Reconhecimento de Entidade

- Identifique entidades e rotule por tipos como pessoa, organização, local, eventos, produtos e mídia.

Análise de sentimentos

- Entenda o sentimento geral expresso em um bloco de texto.

Multi Idiomas

- Permite analisar facilmente texto em vários idiomas, incluindo inglês, espanhol e japonês.

API REST integrada

- Acesso via API REST. O texto pode ser carregado na solicitação ou integrado ao Cloud Storage.

Para obter mais informações sobre a API Natural Language, consulte: <https://cloud.google.com/natural-language/docs/>.

API Cloud Translation

- Traduza strings arbitrárias entre milhares de pares de idiomas
- Detectar programaticamente o idioma de um documento
- Suporte para dezenas de idiomas



 Google Cloud

A API Cloud Translation fornece uma interface programática simples para traduzir uma string arbitrária para qualquer idioma compatível. A API de tradução é altamente responsiva, portanto, sites e aplicativos podem ser integrados à API de tradução para uma tradução rápida e dinâmica do texto de origem do idioma de origem para um idioma de destino (por exemplo, francês para inglês). A detecção de idioma também está disponível nos casos em que o idioma de origem é desconhecido.

A API Translation é compatível com as bibliotecas de cliente padrão da API do Google em Python, Java, Ruby, Objective-C e outras linguagens.

Você pode tentar no seu navegador: <https://developers.google.com/apis-explorer/#p/translate/v2/>

API de inteligência de vídeo em nuvem

- Anote o conteúdo dos vídeos
- Detectar mudanças de cena
- Sinalizar conteúdo impróprio
- Suporte para vários formatos de vídeo



 Google Cloud

A Google Cloud Video Intelligence API permite que os desenvolvedores usem a tecnologia de análise de vídeo do Google como parte de seus aplicativos. A API REST permite que os usuários anotem vídeos armazenados no Google Cloud Storage com informações contextuais de vídeo e de nível de quadro (1 fps). Ele ajuda você a identificar entidades-chave, ou seja, substantivos, em seu vídeo e quando elas ocorrem. Você pode usá-lo para tornar o conteúdo de vídeo pesquisável e detectável.

A API suporta a anotação de formatos de vídeo comuns, incluindo .MOV, .MPEG4, .MP4 e .AVI.

Agenda

Plataforma de Big Data do Google Cloud

Plataforma de aprendizado de máquina do Google Cloud

Teste e laboratório

Questionário

Quando você usaria o Cloud
Dataproc?

Cite dois casos de uso para o Cloud
Dataflow.

Cite três casos de uso para a plataforma
de aprendizado de máquina do Google.

Respostas do questionário

Quando você usaria o Cloud Dataproc?

Você pode usá-lo para migrar trabalhos locais do Hadoop para a nuvem. Você também pode usá-lo para mineração de dados e análise de dados baseados em nuvem.

Cite dois casos de uso para o Cloud Dataflow.

Cite três casos de uso para a plataforma de aprendizado de máquina do Google.

Respostas do questionário

Quando você usaria o Cloud Dataproc?

Você pode usá-lo para migrar trabalhos locais do Hadoop para a nuvem. Você também pode usá-lo para mineração de dados e análise de dados baseados em nuvem.

Cite dois casos de uso para o Cloud Dataflow.

ETL, orquestração

Cite três casos de uso para a plataforma de aprendizado de máquina do Google.

Respostas do questionário

Quando você usaria o Cloud Dataproc?

Você pode usá-lo para migrar trabalhos locais do Hadoop para a nuvem. Você também pode usá-lo para mineração de dados e análise de dados baseados em nuvem.

Cite dois casos de uso para o Cloud Dataflow.

ETL, orquestração

Cite três casos de uso para a plataforma de aprendizado de máquina do Google.

Deteção de fraudes, análise de sentimentos, personalização de conteúdo

Instruções de laboratório

Neste laboratório, você carregará os dados de registro do servidor no BigQuery e fará uma consulta SQL nele.

- Carregar dados do Cloud Storage no BigQuery.
- Faça uma consulta nos dados no BigQuery.

Neste laboratório, você carregará um arquivo CSV em uma tabela do BigQuery. Depois de carregar os dados, você os consulta usando a interface do usuário da Web do BigQuery, a CLI e o shell do BigQuery.

Mais recursos

Plataforma de Big Data do Google [https://
cloud.google.com/products/big-data/](https://cloud.google.com/products/big-data/)

Plataforma de aprendizado de máquina do Google [https://
cloud.google.com/products/machine-learning/](https://cloud.google.com/products/machine-learning/)



© 2017 Google Inc. Todos os direitos reservados. Google e o logotipo do Google são marcas registradas da Google Inc. Todos os outros nomes de empresas e produtos podem ser marcas registradas das respectivas empresas de seus respectivos países.

© 2017 Google Inc. Todos os direitos reservados. Google e o logotipo do Google são marcas registradas da Google Inc. Todos os outros nomes de empresas e produtos podem ser marcas registradas das respectivas empresas de seus respectivos países.