IBM DataStage

Forneça dados prontos para uso nos negócios em tempo real para a IA com o IBM Cloud Pak for Data DataStage

Fornecendo dados prontos para uso nos negócios através de integração de dados

Os empreendimentos digitais de hoje criam e consomem dados como nunca. Isto inclui dados sobre os clientes, transações e funcionários armazenados em diversos sistemas e repositórios. Estes repositórios de dados estão espalhados em diversos ambientes multinuvem e de nuvem híbrida e , e por isso, as organizações estão buscando formas de juntar estas fontes e ambientes separados para gerar informações mais rapidamente **utilizando IA** para auxiliar no fornecimento de experiências diferenciadas e personalizadas aos clientes. De acordo com um estudo da Forrester, cientistas de dados gastam cerca de 80% do seu tempo preparando e gerenciando dados para **iniciativas de IA**. Estes resultados, combinados a uma pesquisa da IBM, registraram que 91% das organizações não usam seus dados de forma eficiente; isto significa que as empresas estão tendo dificuldades para entregar valor através dos silos de dados. As técnicas de arquitetura, práticas e ferramentas usadas para alcançar acessibilidade de dados em tempo real para as vastas quantidades de dados e fornecer dados prontos ao uso nos negócios são chamadas de integração de dados. Com uma tecnologia de integração de dados flexível e escalonável, as empresas são capazes de realizar análise para a segunda melhor alternativa, detecção e análise de churn (rotatividade), previsão de cadeia de fornecimento e execução de detecção de fraude instantânea através de dados de Extração, Transformação e Carregamento (ETL) em diversas fontes de dados.

Para CXOs, arquitetos de empreendimento ou líderes de operações que enfrentam dificuldades no gerenciamento de dados em ambientes de multinuvem ou data lakes e desejam encurtar o tempo necessário para construir e atualizar modelos e aplicações de dados de IA, o IBM® InfoSphere™ DataStage, uma solução líder de mercado em integração de dados que proporciona capacidade de dados confiáveis prontos para uso comercial e além do ETL, fornece uma solução de integração e fornecimento escalonável de dados em um ambiente multinuvem para garantir que informações confiáveis e prontas para uso comercial estejam sendo usadas em tempo real. As capacidades principais do DataStage incluem suporte ao tempo de execução em um ambiente multinuvem que utilizam o modelo uma única vez e são executados em qualquer ambiente de nuvem, enquanto são capazes de escalonar cargas de trabalho com o equilíbrio de carga de trabalho automática e mecanismo paralelo de baixa latência. Além disso, também apresenta fornecimento de dados em tempo real com tecnologia integrada de replicação de dados, tempo e custo reduzido para DevOps com suporte a Integração e Entrega Contínua (CI/CD), rapidez para construir modelos de IA com o Modelo de Integração Autônoma e regras de validação para detectar e resolver problemas de dados automaticamente ao utilizar qualidade de dados em linha.

O DataStage é parte das capacidades do IBM DataOps para operacionalizar dados contínuos de alta qualidade para possibilitar o uso de IA e fornecer um canal de dados automatizado e com função de autoatendimento para as pessoas certas, na hora certa, a partir de qualquer fonte de dados. O IBM InfoSphere DataStage está disponível localmente, na IBM Cloud e em plataformas hiperconvergentes tais como o IBM® Cloud Pak™ for Data que pode ser instalado em qualquer lugar. A IBM® Cloud Pak™ for Data é uma plataforma de IA e dados totalmente integrada construída sobre o Red Hat® OpenShift® que oferece uma arquitetura totalmente nativa da nuvem do DataStage, que pode ser escalonada para o seu negócio. Também fornece uma plataforma para apoio a diversos estilos de entrega de dados, incluindo integração de dados, replicação de dados e virtualização de dados, enquanto a CDC capta mudanças com base em registros conforme elas ocorrem e fornece as informações aos bancos de dados de destino na nuvem e nos data lakes usando fileiras de mensagens com base em Kafka.

Design único, execução em qualquer ambiente de nuvem

De acordo com um estudo da IDC, 90% dos clientes de empreendimentos estão utilizando múltiplos ambientes de nuvem. Com a integração de dados em um ambiente multinuvem, os usuários são capazes de separar o design do tempo de execução - você pode projetar suas funções de ETL uma vez e instalar os componentes de tempo de execução através de recipientes em qualquer ambiente de nuvem para reduzir a latência através do processamento de grandes volumes de dados. Você poderá criar e testar uma tarefa no local e executá-la em um ambiente de nuvem, como o Microsoft Azure, fazendo uso do data lake em nuvem do Azure. Os parâmetros de tarefa e seus valores são passados para uma instância remota da DataStage através de uma mensagem no Kafka.

A integração de dados em ambiente multinuvem oferece os seguintes benefícios:

- Capacidade de integração de dados por ambientes locais e de nuvem
- Experiência de concepção automatizada de tarefas para simplificar o processo de tarefa
- Execução remota de tarefas para minimizar custos de saída de remoção de dados
- Cumprimento de requisitos geopolíticos
- Latência reduzida para processamento de grandes conjuntos de dados, já que estes não precisam ser deslocados e ficam onde estão

Equilíbrio automático de carga de trabalho e processamento paralelo

Com uma arquitetura inteiramente nativa da nuvem, você poderá usar recipientes locais ou compartilhados à DataStage para escalonar suas cargas de trabalho de forma dinâmica, além de otimizar para grandes conjuntos de dados com a melhor tecnologia de mecanismo paralelo (PX). Os usuários podem escolher criar uma tarefa paralela, sequência ou Apache Spark na IBM DataStage Flow Designer.

Você pode executar tarefas da DataStage Flow Designer em dois mecanismos de tempo de execução:

- Tarefas do tipo paralela ou sequencial podem ser executadas apenas em mecanismo paralelo. Tarefas normalmente intensas em recursos são executadas no mecanismo paralelo, e como resultado, o tempo médio para conclusão das tarefas utilizando processamento paralelo é de dois minutos.
- Tarefas do tipo Spark podem ser executadas apenas em mecanismo Spark.

Fornecimento de dados em tempo real

A DataStage com tecnologia de Captura de Mudança de Dados (CDC) para captação em tempo real instalada como recipientes pode fornecer o melhor dos mundos da integração de dados e replicação dos mesmos . A DataStage permite uma transformação complexa com grandes conjuntos de dados, enquanto a CDC capta mudanças com base em registros conforme ocorrem, modificando-as usando transformações complexas e entrega aos bancos de dados de destino em nuvem e data lakes, usando fileiras de mensagens baseadas em Kafka. A DataStage também permite envio de tarefas de transformação de dados baseadas em lote e baseada em equivalência aos depósitos de dados.

Tempo e custo reduzidos para DevOps com suporte a CI/CD

Para enfrentar o desafio de gerenciar uma quantidade de aplicações em recipiente por diversos sistemas operacionais, as organizações precisam de uma ferramenta de código aberto como a Red Hat OpenShift, disponível em Cloud Pak for Data. A plataforma Cloud Pak for Data ajuda a escalonar e fornecer recipientes para suportar as iniciativas de TI como microsserviços e estratégias de migração de nuvem. Os recipientes do DataStage permitem a criação e automação dos canais de Integração Contínua / Entrega Contínua (CI/CD) para tarefas do desenvolvimento ao teste, até a produção e suporte ao canal de CI/CD ao sustentar ferramentas de controle de fonte, como GitHub, para publicação frequente de tarefas e liberação para produção.

Modelo de Integração Autônoma para abastecimento de IA

Acelere a coleta e a integração de dados para IA de maneira mais rápida e em escala ao descobrir e classificar ativos automaticamente, gerando fluxos de integração com base em transformações personalizadas integradas e regras de qualidade e detecção e proteção de informações sensíveis.

Valorização Rápida com o modelo de tarefas automatizado

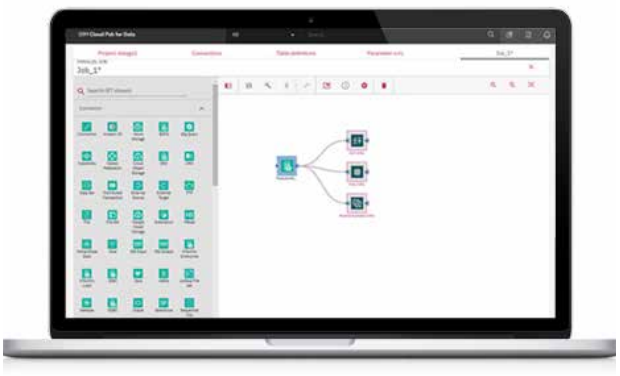

Figura 1. DataStage Flow Designer com recursos de design automatizados.

A IBM DataStage Flow Designer é uma UI (interface de usuário) baseada em web para o DataStage com habilidades de aprendizado de máquina (ML) para auxiliar os usuários, mesmo usuários não técnicos, para construir canais e estágios dentro de uma tarefa.

A DataStage Flow Designer oferece os seguintes benefícios:

- Compatibilidade retroativa. Não precisa migrar tarefas. Muitas empresas possuem milhares de tarefas em um único projeto, e dependem da execução destas tarefas 24 horas por dia, 7 dias por semana. A migração, com possibilidade alta de erros e paradas, não é uma opção. Estas empresas podem pegar qualquer tarefa existente de DataStage e executá-la no IBM DataStage Flow Designer, para que não haja necessidade de migrar estas tarefas para um novo local.
- Aumento da produtividade do desenvolvedor A IBM DataStage Flow Designer possui recursos como busca integrada, um tour rápido para acelerar o início das empresas, propagação automática de metadados, paleta inteligente, estágios sugeridos e destaque simultâneo de todos os erros de compilação. Os desenvolvedores podem usar esses recursos para aumentar a produtividade ao projetar tarefas, e sua produtividade pode aumentar até nove vezes mais rápido que as tarefas tradicionais codificadas manualmente.
- Grande quantidade de operadores e conectividade. Além das capacidades de design e desenvolvimento, a DataStage oferece centenas de operadores pré-construídos, prontos para usar e em primeira mão. Eles reduzem drasticamente o tempo gasto pelos desenvolvedores na preparação de dados para ações analíticas. Com os novos operadores adicionados com intervalo de algumas semanas, a produtividade do desenvolvedor é melhorada com o tempo.

Qualidade e segurança de bordo dos dados para fornecimento confiável de dados

A DataStage oferece uma experiência única de usuário para integração dos dados usando a DataStage Flow Designer para execução de validação de dados, padronização e correspondência de regras no momento de entrega dos dados aos ambientes de destino, tais como data lakes, para evitar problemas de qualidade e problemas em potencial com a segurança, no que tange acessos de usuários não autorizados aos seus dados sensíveis. Este conceito de qualidade de dados também pode ser estendido para dar apoio à governança abrangente de dados em todo o depósito de dados (DWH).

Resumo

A DataStage fornece:

- Design único, execução em qualquer lugar, com equilíbrio de carga de trabalho automático integrado, paralelismo e escalabilidade.
- Captação de atualizações em tempo real ou com estilos de entrega a base de lotes.
- Resiliência integrada, operação fácil e CI/CD.
- Integração de dados otimizada para IA
- Concepção de tarefas automatizada com capacidades de ML
- Qualidade e segurança de bordo dos dados para fornecimento confiável de dados

A IBM oferece uma amplitude de capacidades de integração de dados em todos os ambientes multinuvem, nas dependências ou em sistemas hiperconvergentes como o IBM Cloud Pak for Data ou qualquer outra plataforma de nuvem à escolha. Estas diferentes capacidades fornecem uma solução de integração de dados flexível e escalonável para acesso rápido a grandes volumes de dados de alta qualidade para IA, no modelo de instalação de sua escolha.

Obtenha uma demonstração guiada e gratuita para saber mais sobre a IBM InfoSphere DataStage

Por que a IBM?
As capacidades do IBM DataOps ajudam a criar uma base de análise pronta para uso comercial ao fornecer tecnologia líder de mercado com automação habilitada por IA, governança infundida e catálogo robusto de conhecimento para operacionalizar dados de alta qualidade, de forma contínua por toda a empresa. Aumento da qualidade dos dados para fornecer um canal de dados eficiente e com capacidade de autoatendimento, para o pessoal certo, no momento certo, a partir de qualquer fonte.

Para saber mais sobre o DataOps visite
ibm.com/dataops
Para saber mais sobre a IBM InfoSphere DataStage visite
ibm.com/products/infosphere-datastage
Visite o hub de Análise e Big Data em
ibmbigdatahub.com

IBM Corporation
New Orchard Road, Armonk, NY 10504
Produzido nos Estados Unidos da América
Abril de 2020