Ingestão de Dados na Arquitetura Lakehouse na Oracle Cloud

Thais Henrique

Andrea Rigoni

Trilha Inovação com dados em nuvem

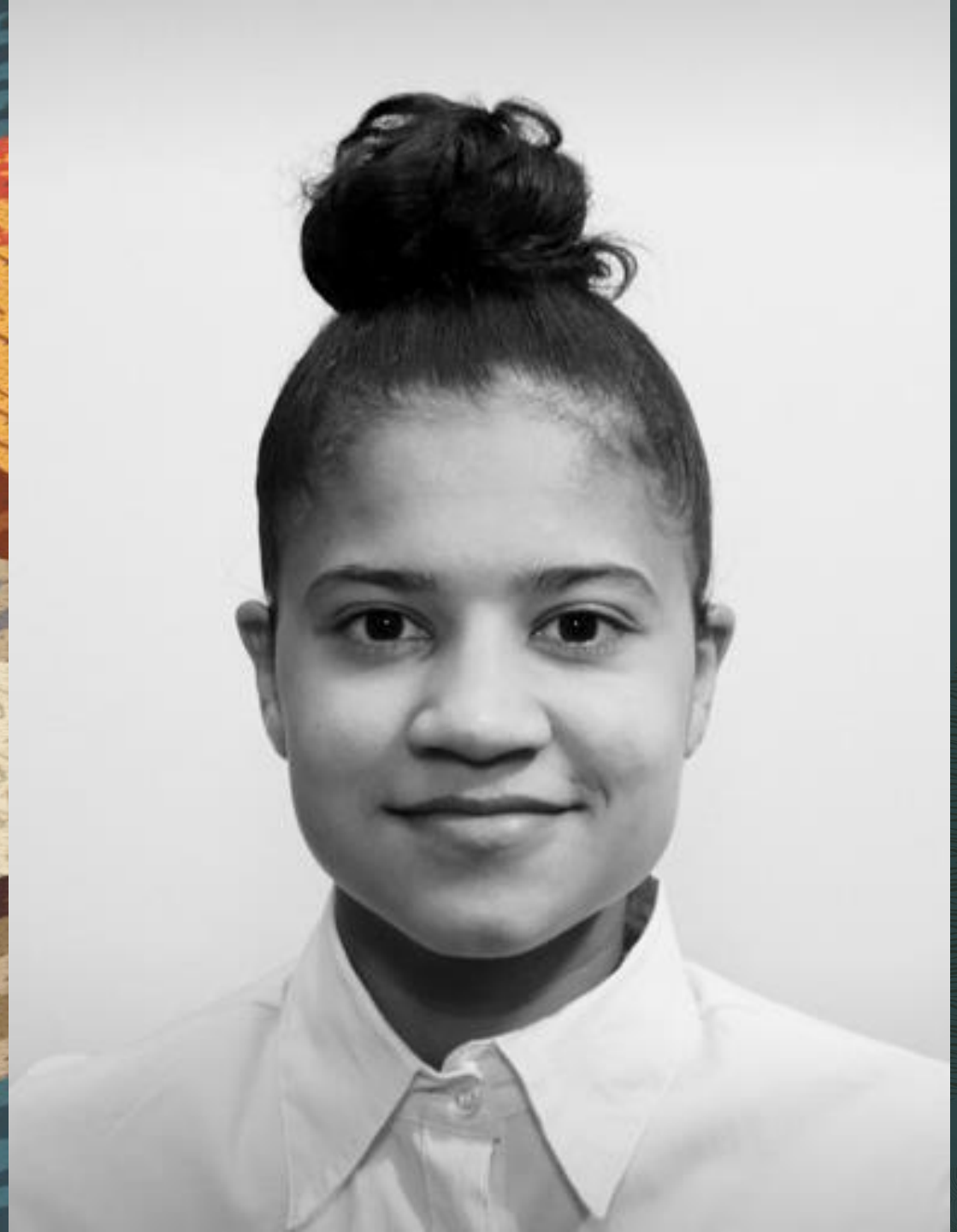
Ingestão de Dados na Arquitetura Lakehouse na Oracle Cloud

Thais Henrique **Andrea Rigoni**

Trilha Inovação com dados em nuvem

Agenda

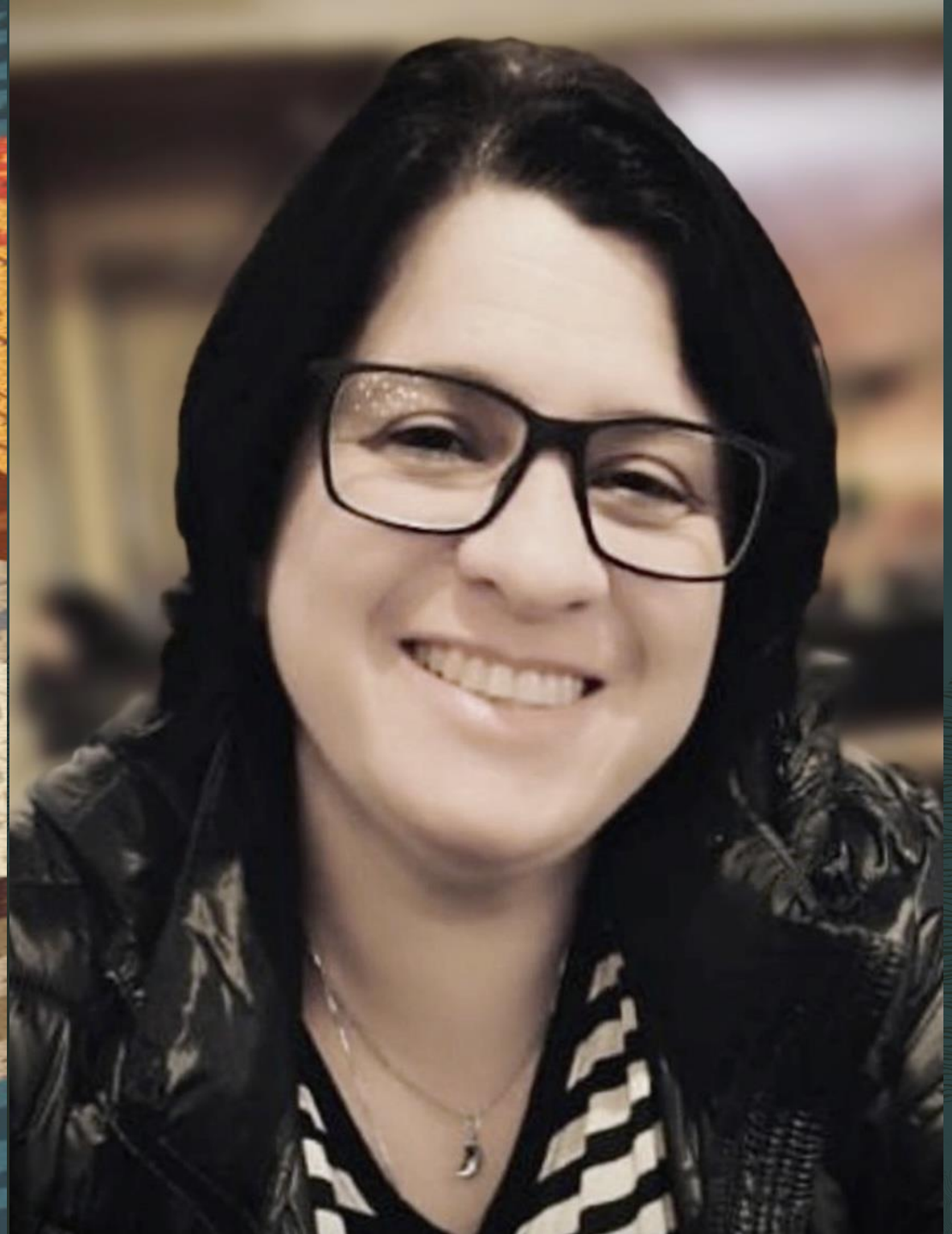
1. Lakehouse
2. Data Lake
3. Ingestão de Dados
4. Governança de Dados
5. Data Warehouse
6. Hands-on Lab



Thais Henrique

https://www.linkedin.com/in/thaishenrique/

thais.henrique@oracle.com



Andrea Rigoni

https://www.linkedin.com/in/rigoni/

andrea.carmo@oracle.com

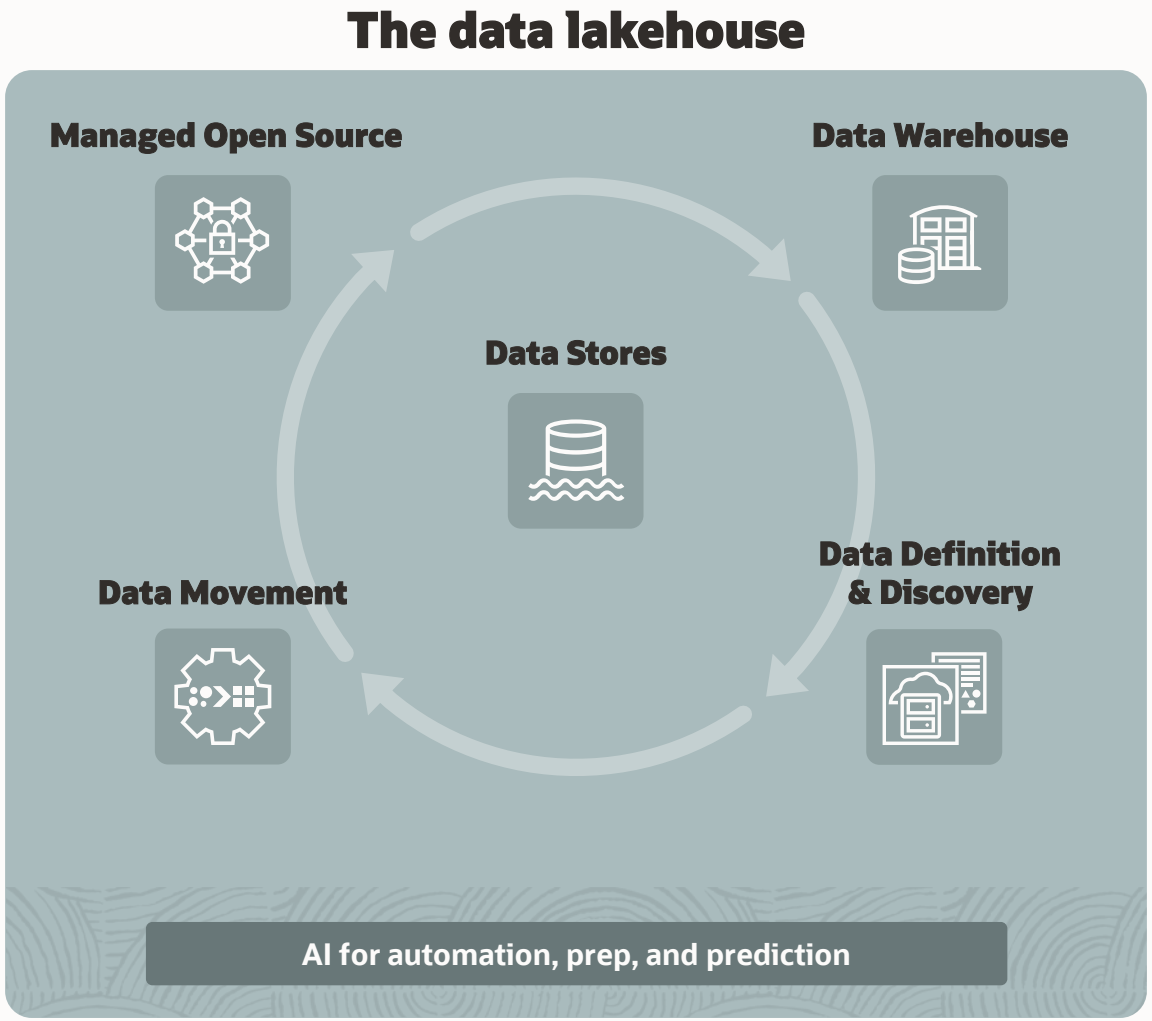
A evolução do gerenciamento de análise de dados

Apresentando o data lakehouse

Data warehouse e data lake integrados
Permitindo análises integradas entre todos os dados

Elimina data silos
Data irão se movimentar entre warehouse, lake e analysis tools conforme necessário

Serviços nativos ou Open
source suportando ambos cenários e todo tipo de dado



Data Warehouse: dados curados e de valor conhecido

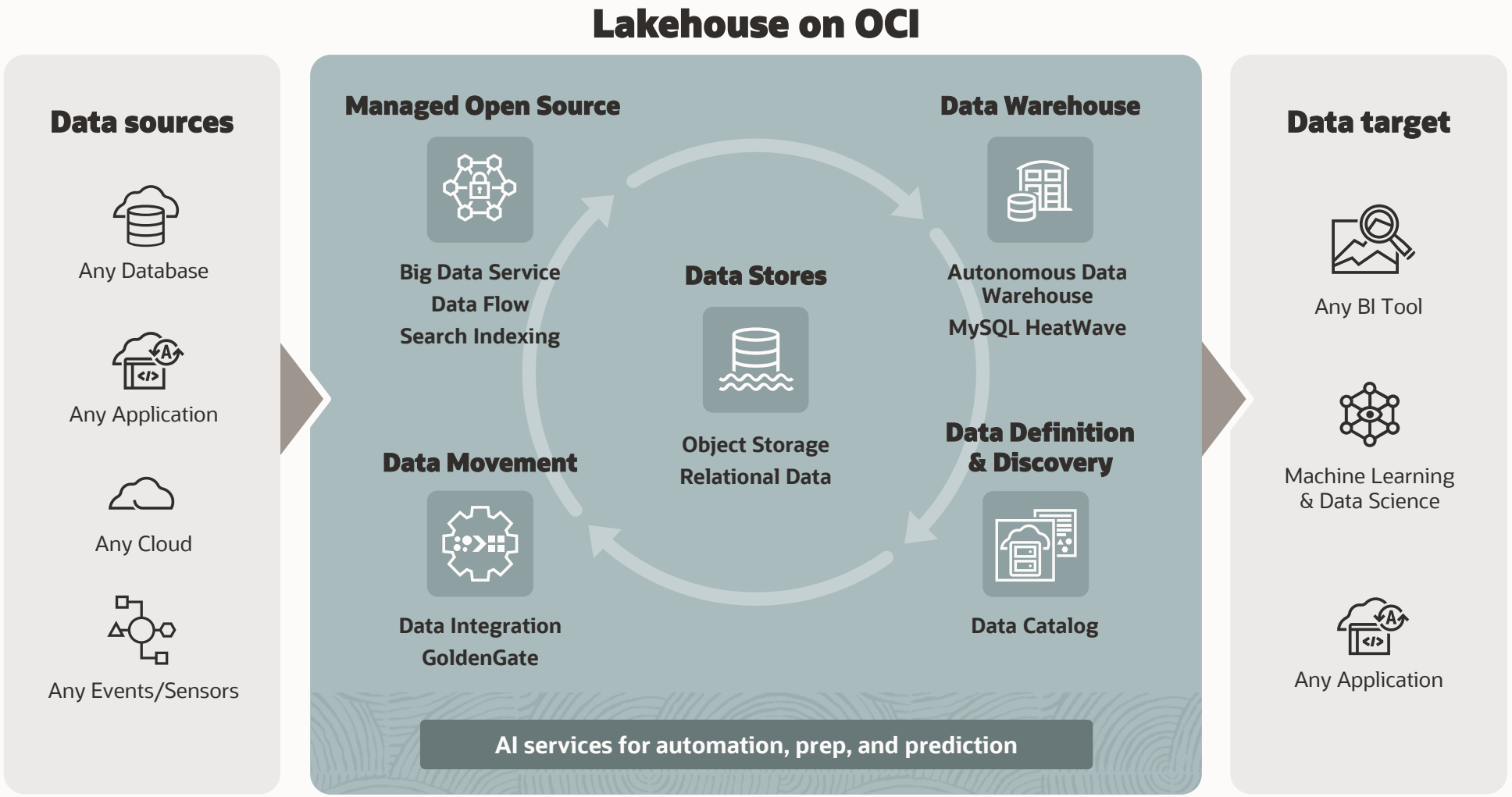
Data Lake: dados brutos, ainda sendo explorados , antigos, ou de menor valor. Também utilizado como staging antes da carga para o data warehouse, archive, e repositório para treinamento de modelos de machine learning

Managed Open-source Services: suporte as principais open-source tools para armazenamento e análise. Apache Hadoop, Apache Spark e Elasticsearch

Data Integration: dado no lakehouse poderá se mover entre lake, warehouse e open-source analytics, dependendo do caso de uso

Data Catalog: mantem uma visão complete de todo o dado disponível, possibilitando descobertas e governança.

Construa um Lakehouse em OCI



Construa um Lakehouse em OCI

Armazenamento de Dados

OCI Object Storage

Object Storage

Agnóstico
Repositório capaz de armazenar qualquer tipo de dado

Rápido
Estrutura baseada discos SSD NVMe

Big Data
Construído para atuar como Data Lake

Elástico
Cresça até escala de Petabytes

Object Storage

Controle
Versionamento de objetos

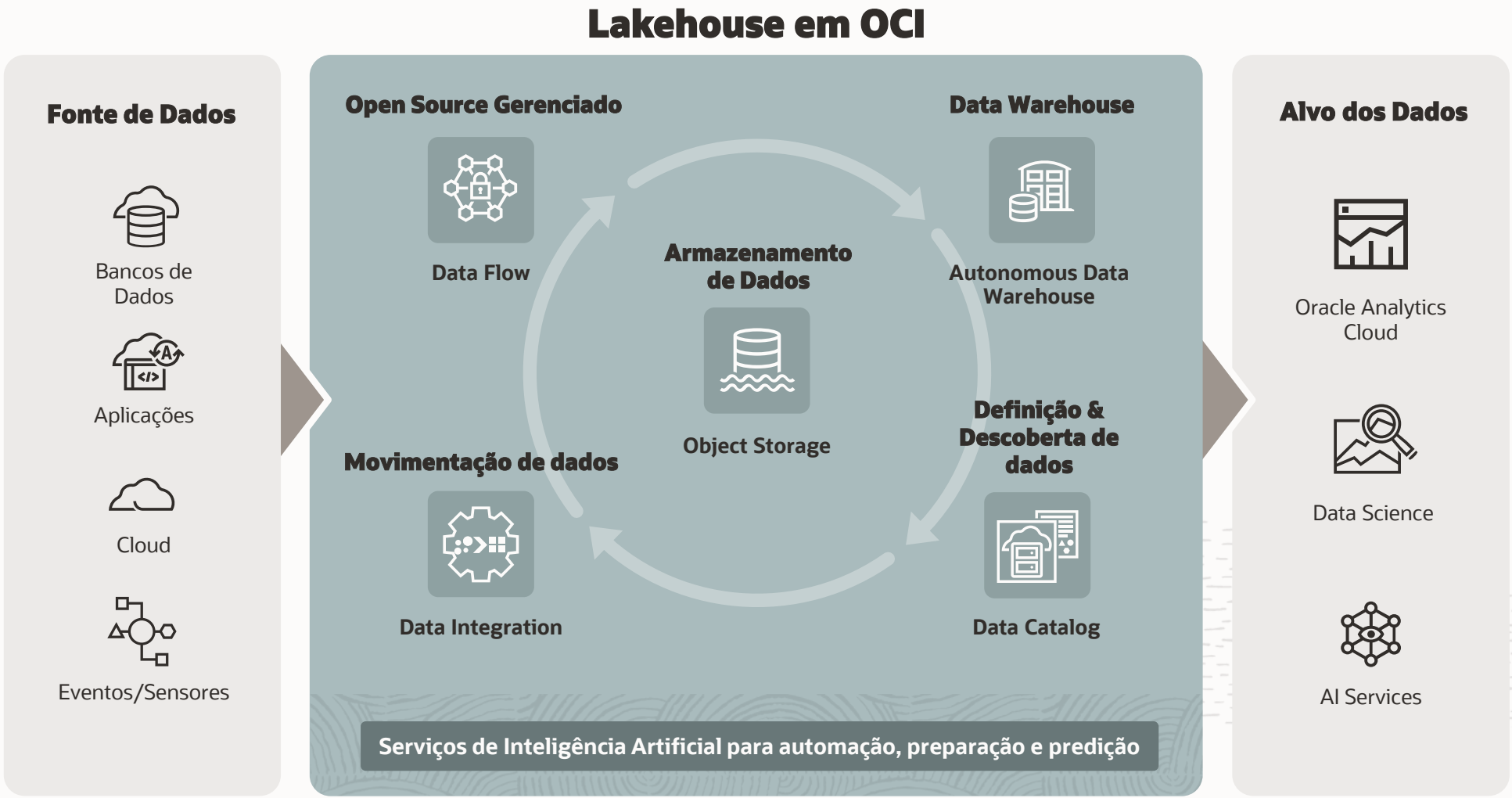
Flexível
Defina a temperatura dos objetos

Integrável
Conecte outros serviços presentes na OCI

Seguro
Defina políticas de acesso e consumo aos dados

Oracle Lakehouse

Arquitetura de dados moderna

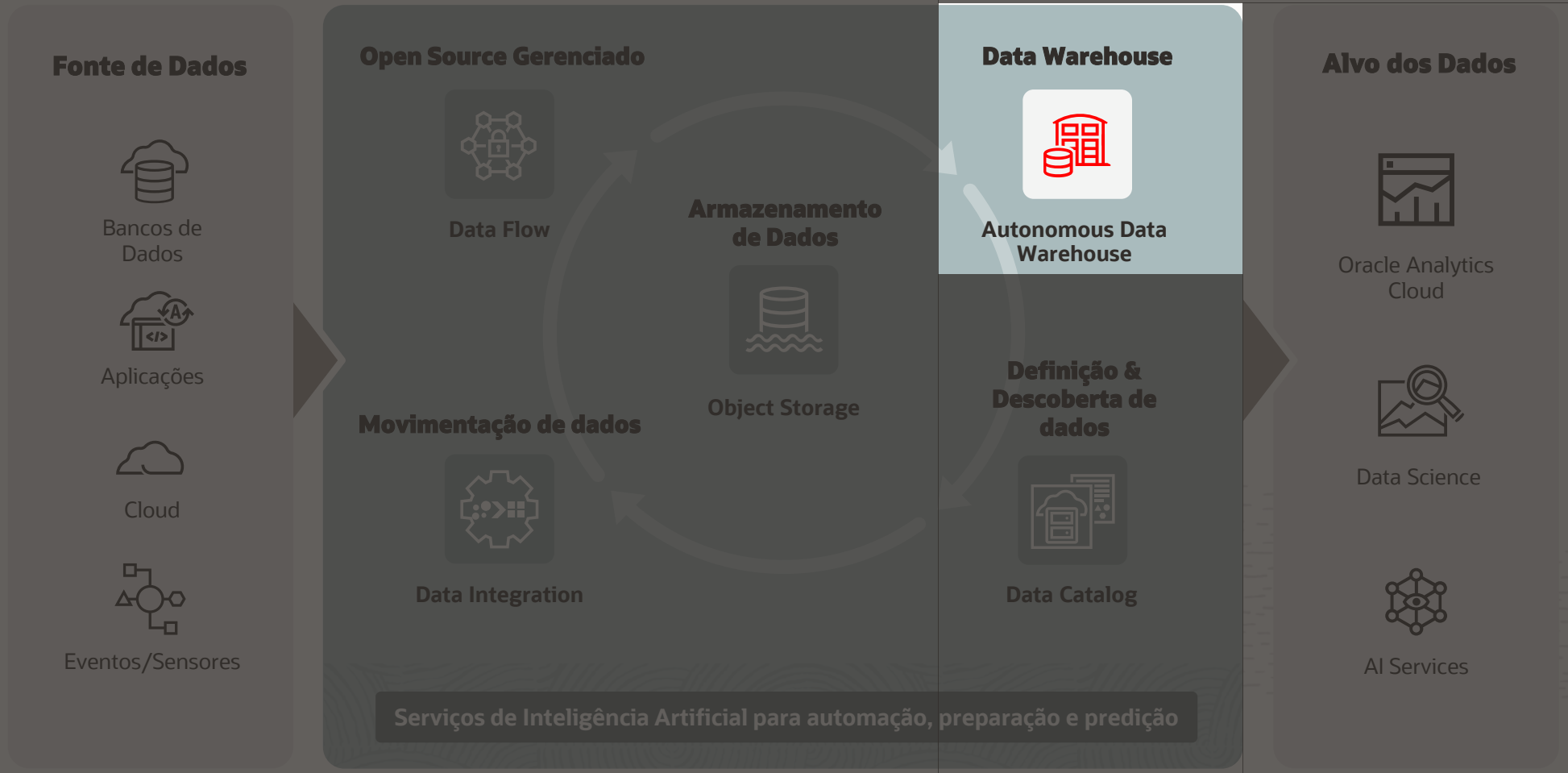


Oracle Lakehouse

Arquitetura de dados moderna

Data Warehouse

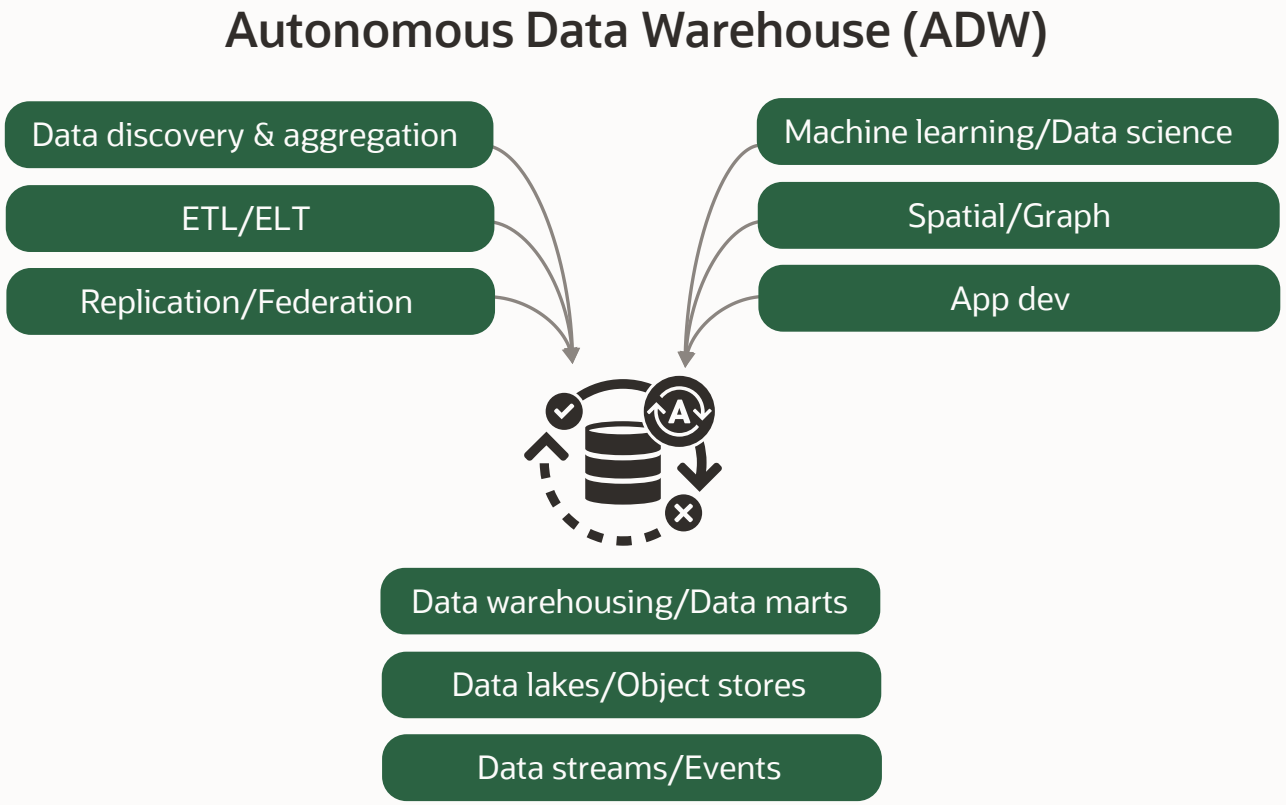
Lakehouse em OCI



OCI Autonomous Data Warehouse

Autonomous Data Warehouse

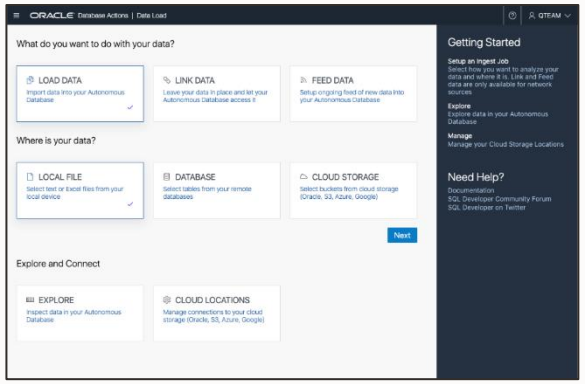
- **Gerenciamento automatizado de data warehouse –** sem administração manual
- **Uma solução completa com análises integradas –** mas também suporte para análises de terceiros
- **Implementação fácil e rápida –** carregue todos os dados e obtenha uma variedade de insights
- **Resposta rápida –** independentemente do tamanho dos dados, tipo de análise ou número de usuários simultâneos
- **Proteção abrangente de dados e privacidade –** sem brechas de segurança ou necessidade de serviços adicionais
- **Escolha** provisionar em nuvem pública ou em seu DC



OCI Autonomous Data Warehouse

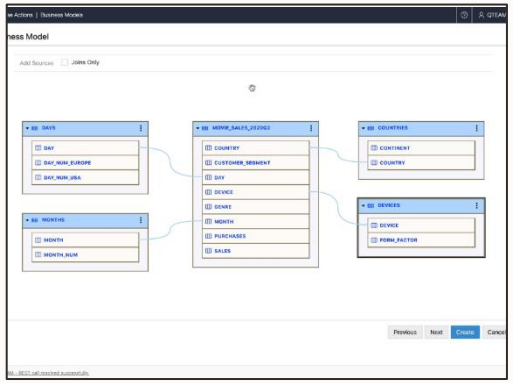
Autonomous Data Warehouse

Load



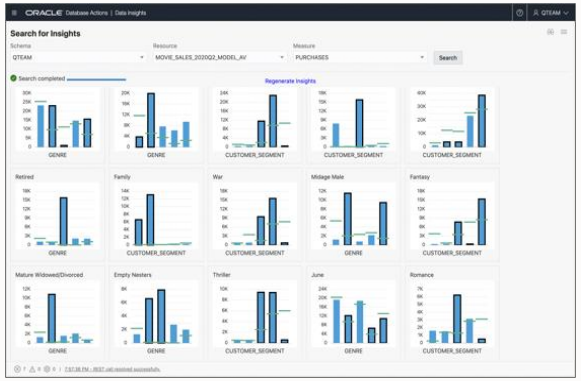
Carga Drag & drop simples

Business Model



Crie Modelos de Negócios Automaticamento

Data Insights



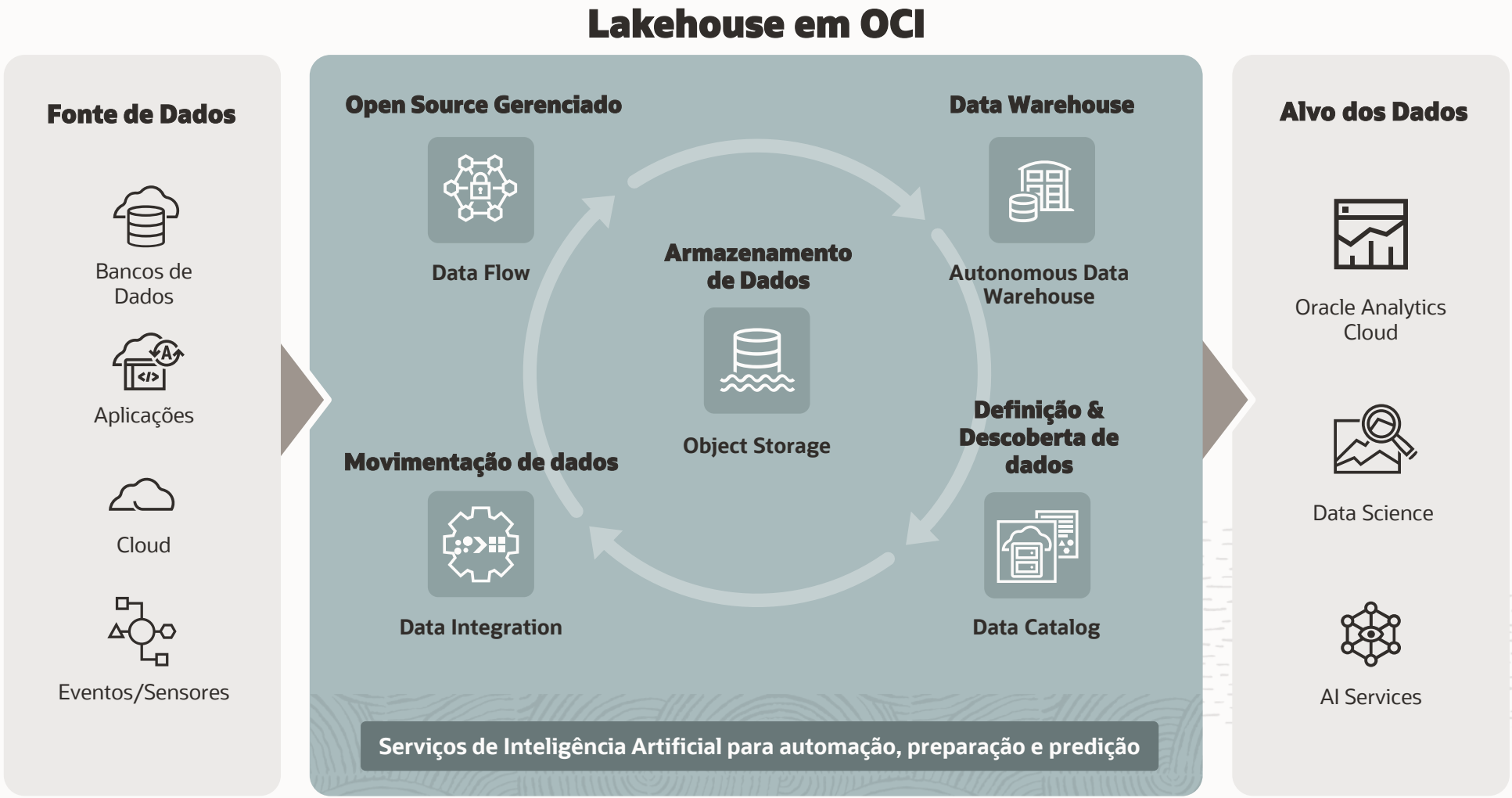
Descubra padrões e Anomalias escondidos Automaticamente

OCI Autonomous Data Warehouse

Autonomous Data Warehouse

Oracle Lakehouse

Arquitetura de dados moderna



Oracle Lakehouse

Arquitetura de dados moderna

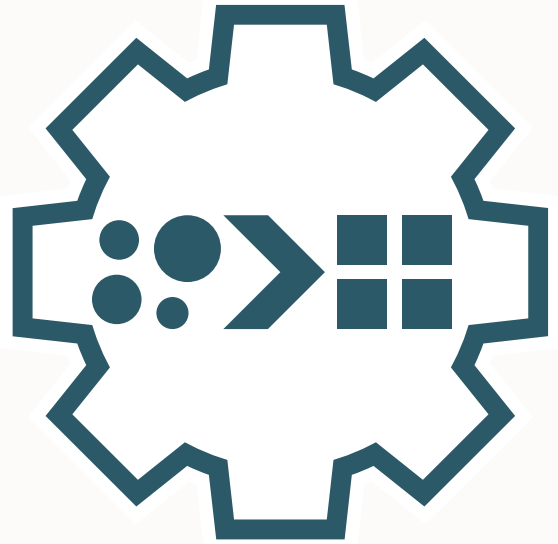
Ingestão de Dados

Oracle Cloud Infrastructure (OCI) Data Integration

Data Integration

O serviço de ETL nativo em nuvem e serveless para integração, transformação e movimentação de dados dentro do ecossistema OCI

- Interface gráfica, code-free
- Preparação e criação de perfis de dados interativos
- Proteção de Schema evolution (Drift)
- Powered by Spark ETL ou E-LT Push-Down



OCI DI

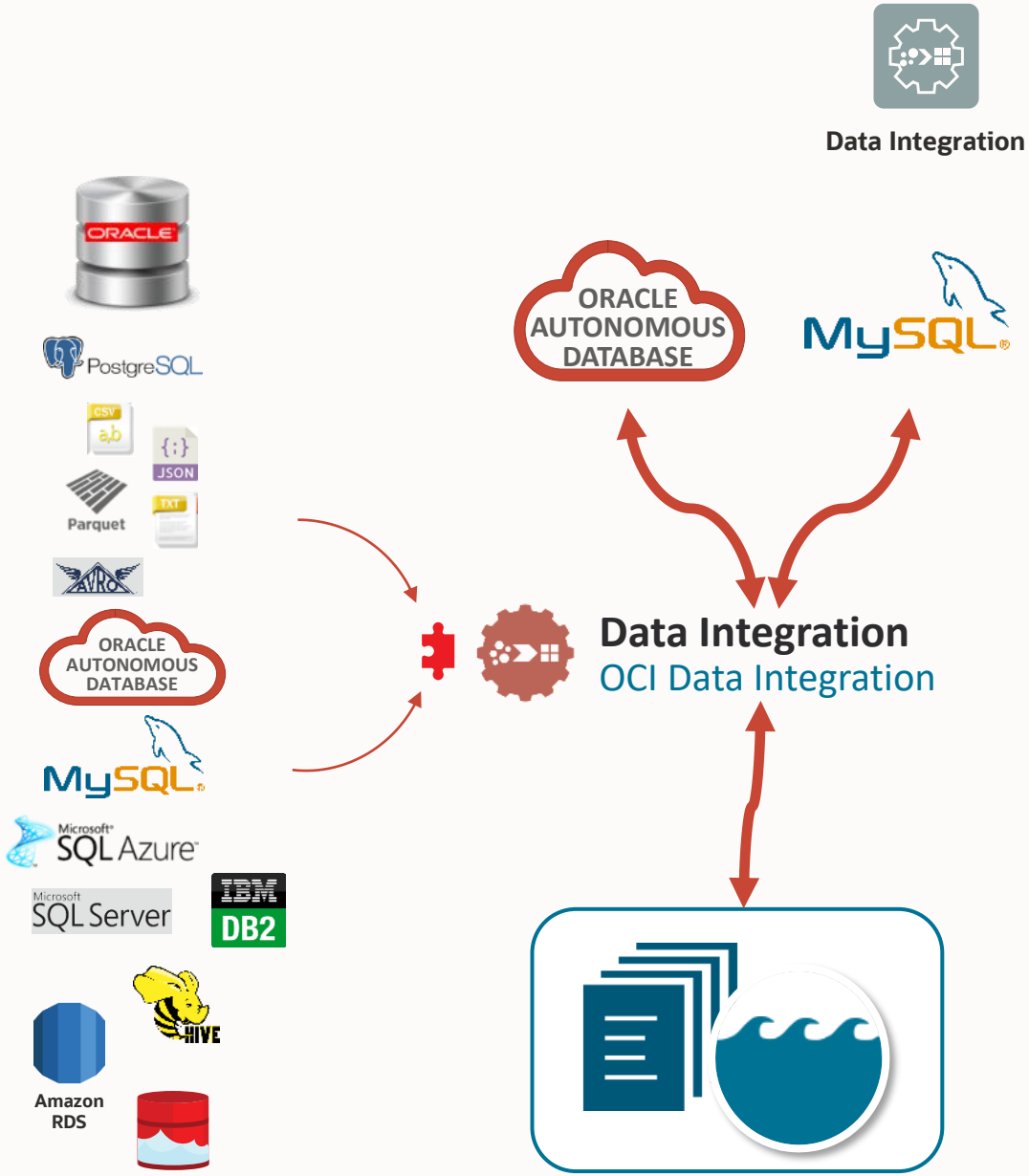
Ampla Conectividade

Optimized for Oracle Cloud

- Acesso seguro – público e privado – inclusive para on-premises
- Conexão para:
  - Oracle Autonomous Database (ADW/ATP)
  - Oracle Database & Exadata DB Systems
  - Oracle Object Storage: CSV, JSON, Parquet, Avro
  - Oracle Fusion
  - MySQL / PostgreSQL / Apache Hive
  - Microsoft SQL Server & Azure SQL Database & Azure Synapse Analytics
  - Amazon RDS (MySQL, Oracle, Microsoft SQL Server)
  - IBM DB2

Benefícios

- A melhor conectividade para Oracle Cloud
- Expansão de um conjunto de adaptadores nativos fáceis de usar



Maximize a produtividade do desenvolvedor

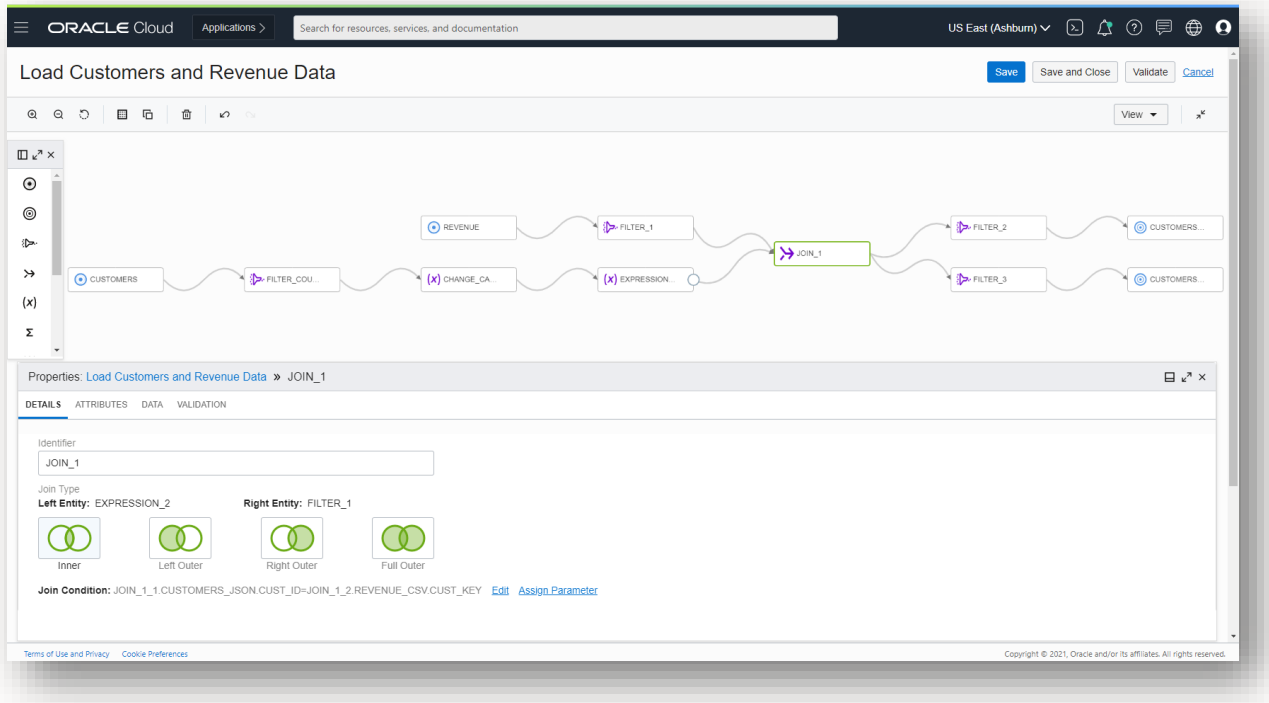
Data Integration

No Code Data Flow Design

- Poderoso editor gráfico para criar fluxos de dados (data flows)
- Preveja dados com Data Xplorer
- Fluxos de dados parametrizáveis para máxima flexibilidade

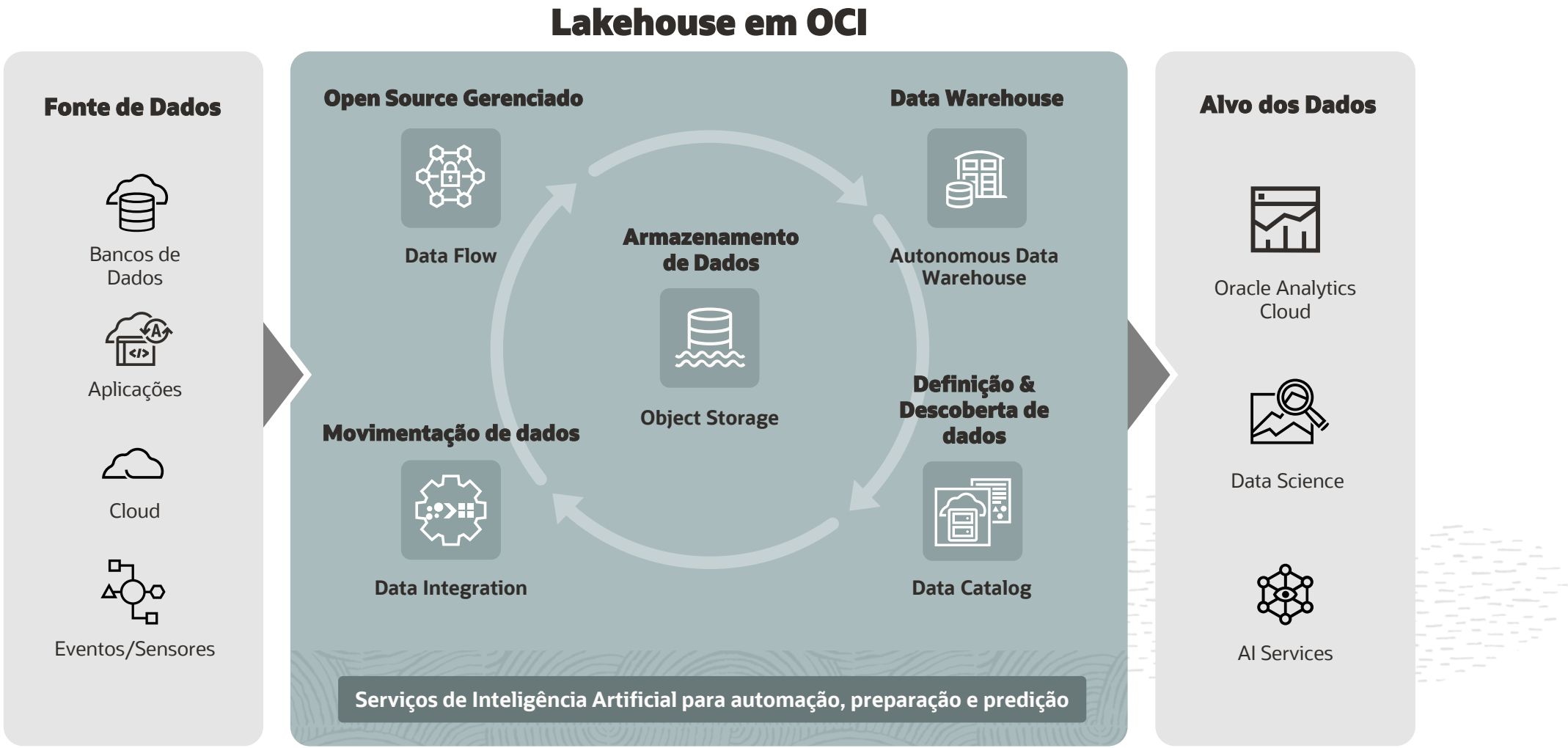
Benefícios

- Permite o foco em inovação
- Simplifica os processos de ETL / Manutenção
- Permite integrações e transformações poderosas



Oracle Lakehouse

Arquitetura de dados moderna



Oracle Lakehouse

Arquitetura de dados moderna

Governança de Dados

Lakehouse em OCI

Fonte de Dados

- Bancos de Dados
- Aplicações
- Cloud
- Eventos/Sensores

Open Source Gerenciado

- Data Flow

Armazenamento de Dados

- Object Storage

Movimentação de dados

- Data Integration

Data Warehouse

- Autonomous Data Warehouse

Definição & Descoberta de dados

- Data Catalog

Alvo dos Dados

- Oracle Analytics Cloud
- Data Science
- AI Services

Serviços de Inteligência Artificial para automação, preparação e predição

Por que organizações precisam de um Catálogo de Dados na Nuvem Oracle?

Data Catalog

Analista de Dados

Cientista de Dados

Engenheiros de Dados

Data Stewards

Dificuldade de encontrar o dado correto para Analytis

- Falta de visão holística sobre os dados
- Difícil de solucionar problemas de dados
- Poucas informações sobre os dados

Dificuldade de entender dados em um Data Lake

- Sem dicionário de dados
- Definição manual de schema
- Incapacidade de compartilhar modelos de dados entre aplicativos

Necessidade de melhora na Governança de Dados

- Falta de visão sobre donos dos dados
- Falta de conceitos de negócios comuns
- Sem colaboração para resolver problemas de dados
- Proliferação de dados sensíveis

Overview de Capacidades

Coleta de Metadados

- Coleta automatizada de metadados técnicos
- Suporta fontes de dados na nuvem e on-premise
  - OCI Object Storage, Autonomous Database
  - Oracle DB, MySQL, Hive, Kafka
  - MS SQL Server, Azure SQL DB, IBM DB2, PostgreSQL
- Entidades lógicas para agrupar arquivos relacionados em data lakes

Curadoria de Metadados

- Glossários de Negócio com Termos e Categorias
- Propriedades de enriquecimento personalizadas definidas pelo usuário
- Tags de formato livre para anotações
- Vincule ativos a termos comerciais, tags

Busque e Filtre

- Ambiente colaborativo para consumidores de dados
- Pesquise com base em nomes técnicos, termos comerciais, tags e propriedades personalizadas
- Busque dados por hierarquias do sistema
- Ver detalhes técnicos e de contexto de negócios

Otimizado para a nuvem Oracle

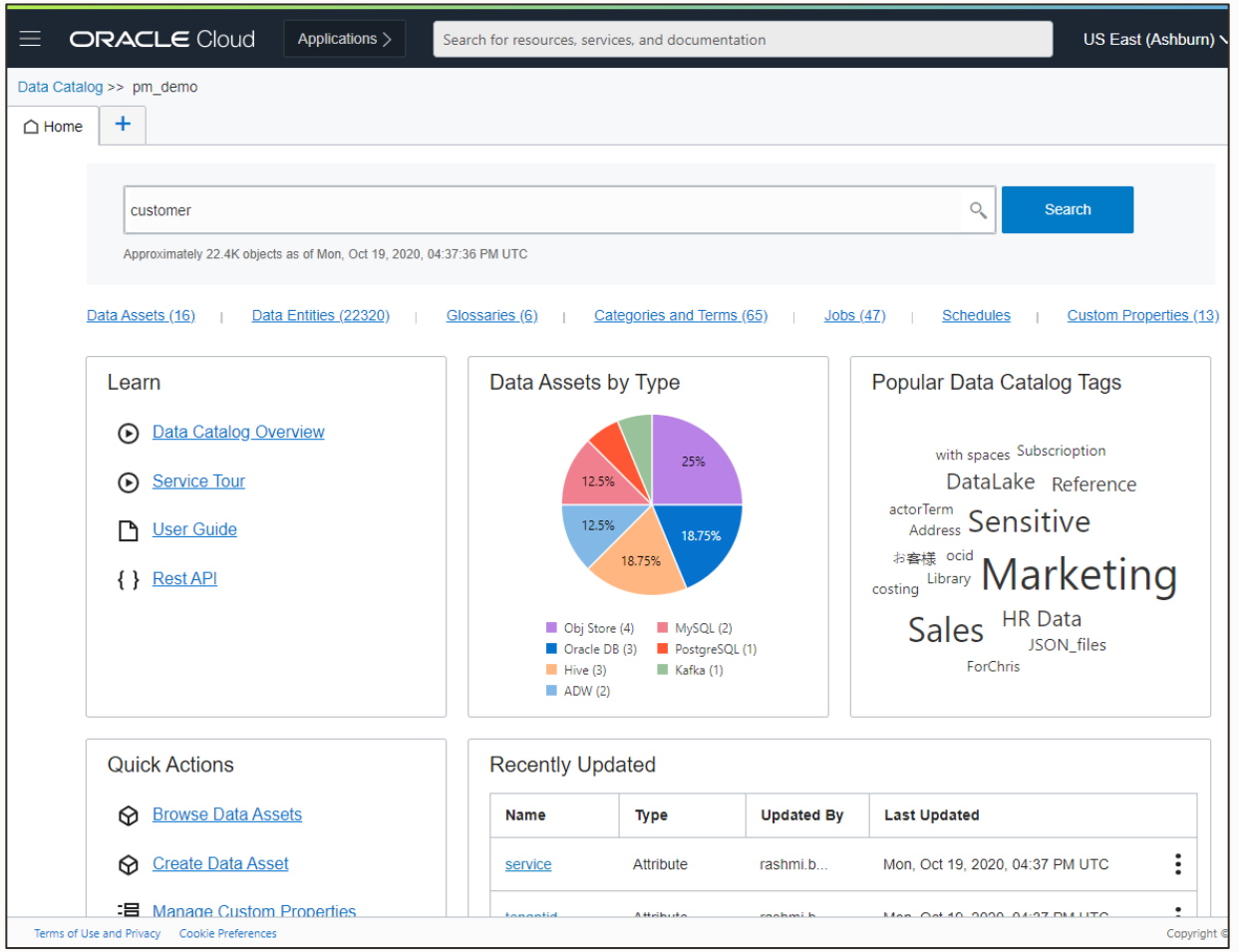
- Seguro, escalável, serverless, nativo da nuvem
- REST APIs e SDKs em Java, Python, Ruby, Go para integração
- Políticas baseadas em IAM para gerenciamento de controle de acesso

* Capacidades atuais disponíveis sem custo

Ambiente Colaborativo

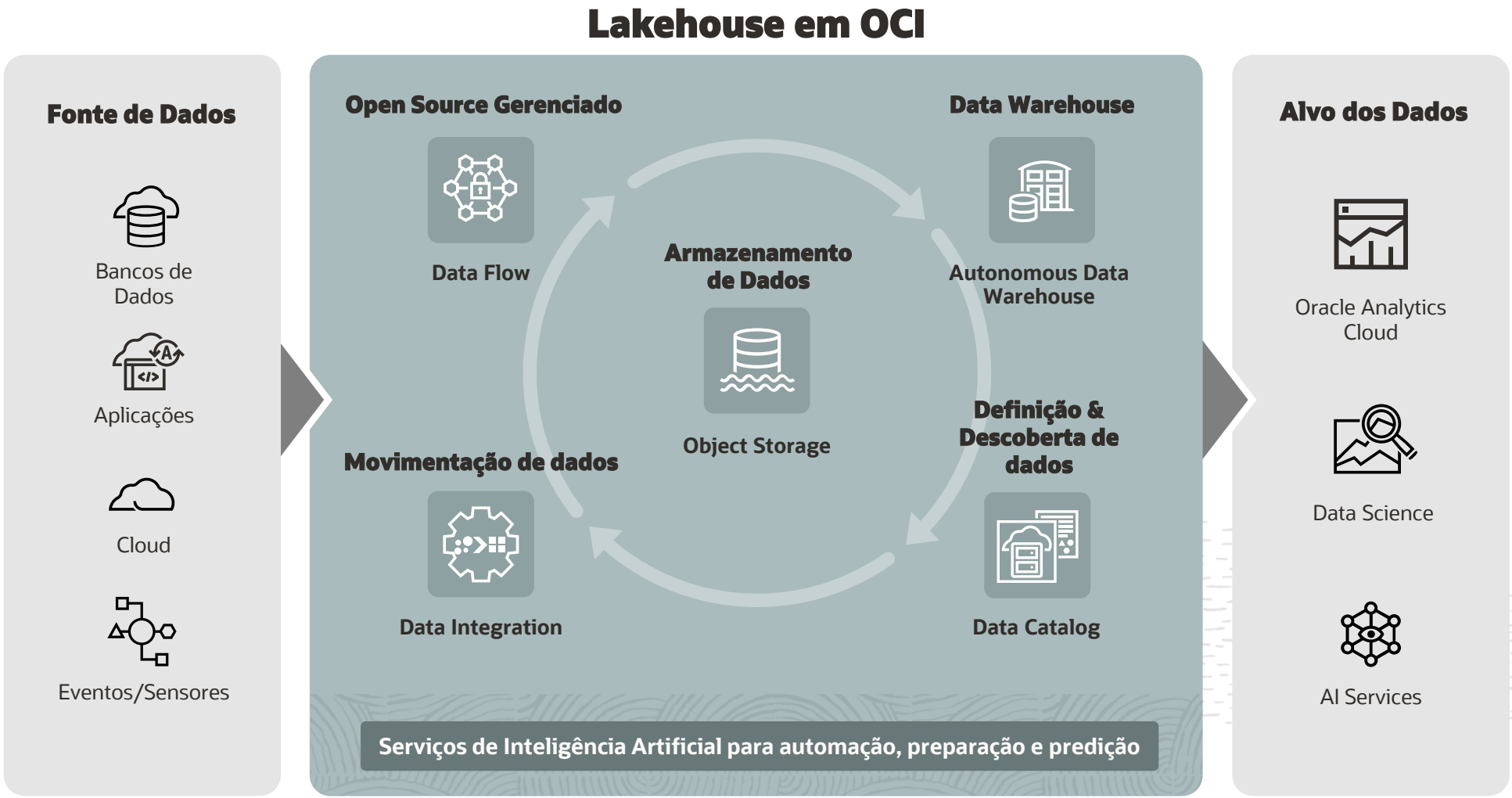
Data Catalog

- Página inicial com atalhos e estatísticas operacionais
- Ações rápidas para gerenciar ativos de dados, trabalhos, propriedades personalizadas, padrões de nome de arquivo, etc.
- Utilize as opções de Tags populares e Objetos atualizados recentemente para acesso rápido



Oracle Lakehouse

Arquitetura de dados moderna

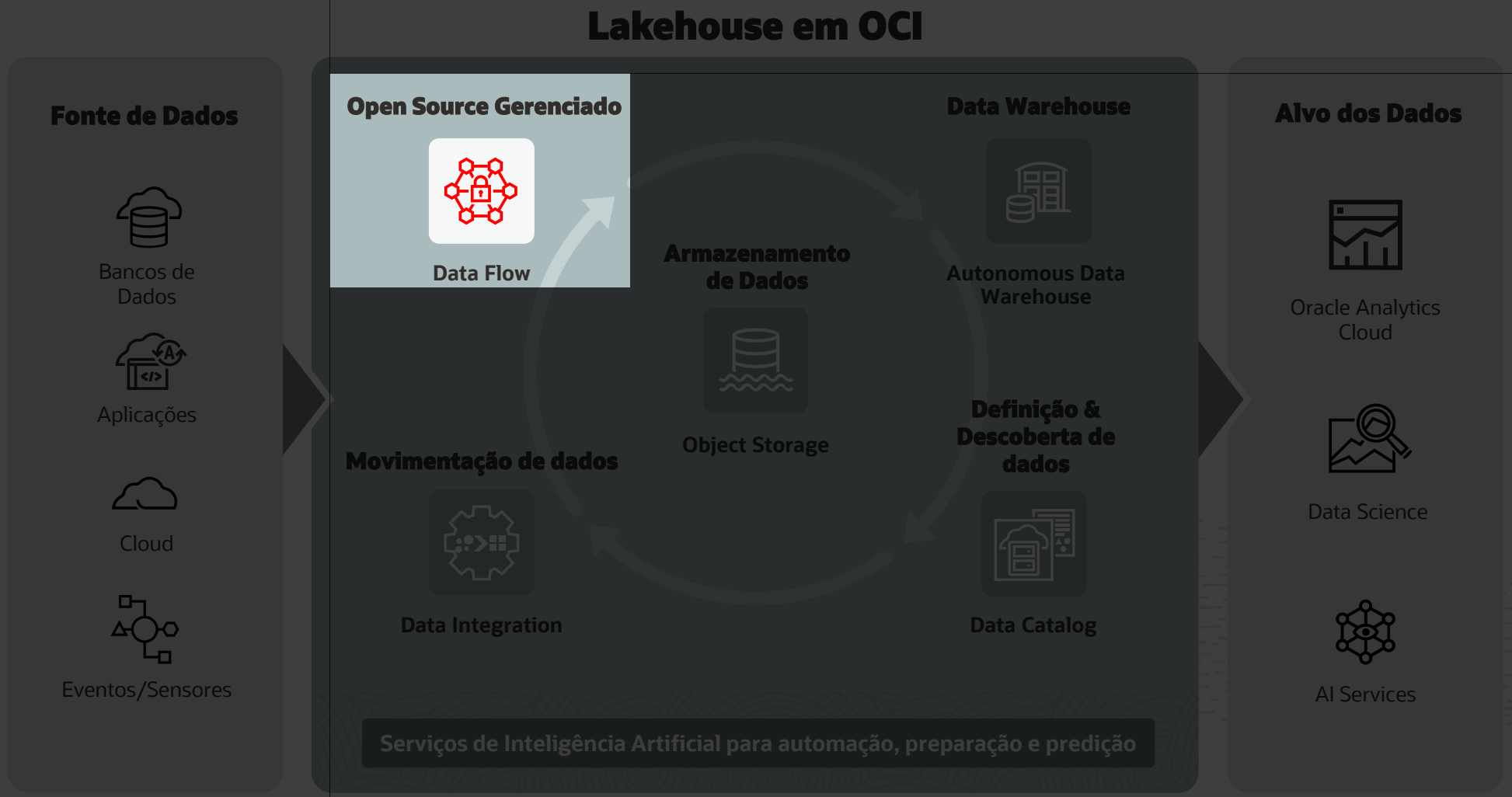


Oracle Lakehouse

Arquitetura de dados moderna

Open Source Gerenciado

- Processamento
- Aplicações efemêras
- Ingestão de Dados



OCI Data Flow

- Data Flow permite entregar Spark-based Big Data ou ML applications de forma rápida.
- Mantenha o foco em suas aplicações e esqueça a Infraestrutura.
- Pague somente pelo que for utilizar e quando utilizar.

OCI Data Flow

Data Flow

Spark on Demand
Implante clusters Spark em minutos sem nada para manter

Operação simplificada
Uis para monitoramento, alertas e diagnósticos.

Enterprise Security
RBAC e Impersonation para controle total.

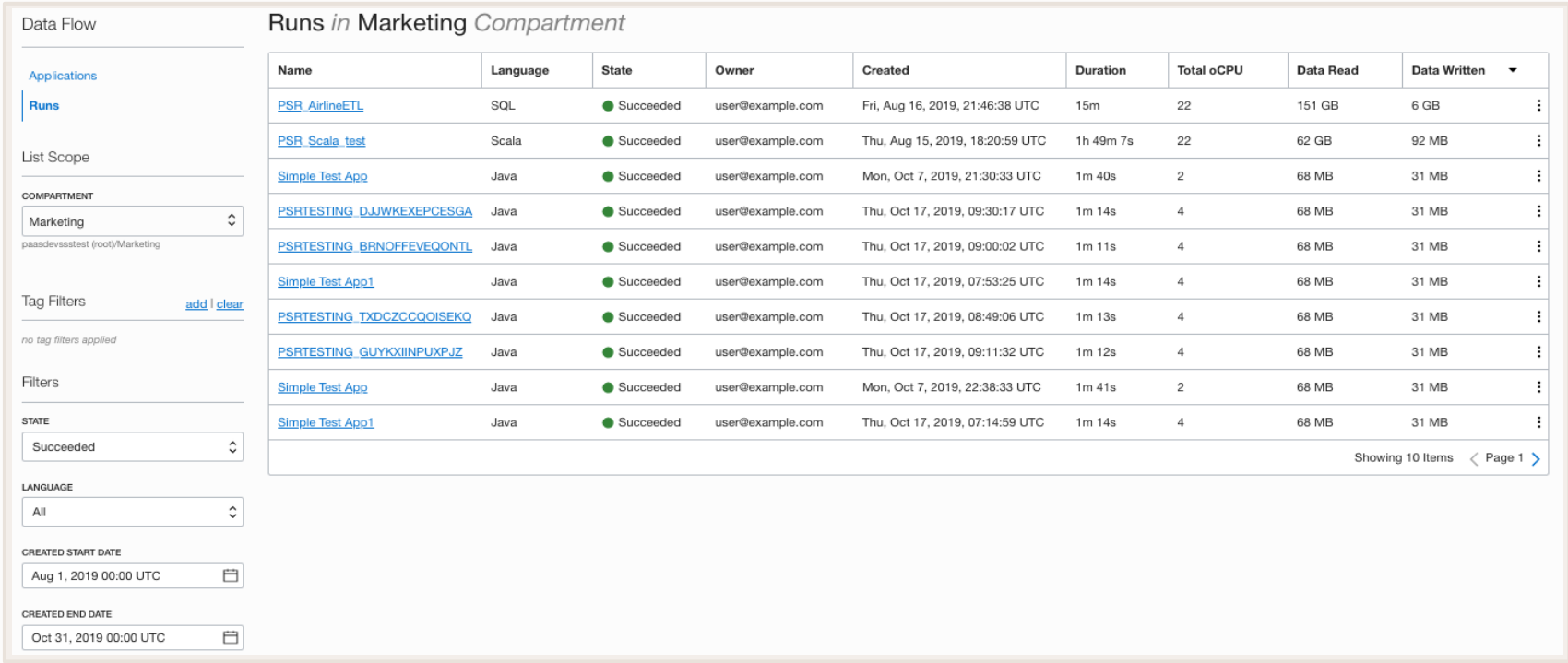
Baixo Custo
Não existe custo pelo serviço,
Pague apenas pela Infraestrutura utilizada.

Data Flow torna a operação mais simples.

Data Flow

Operations

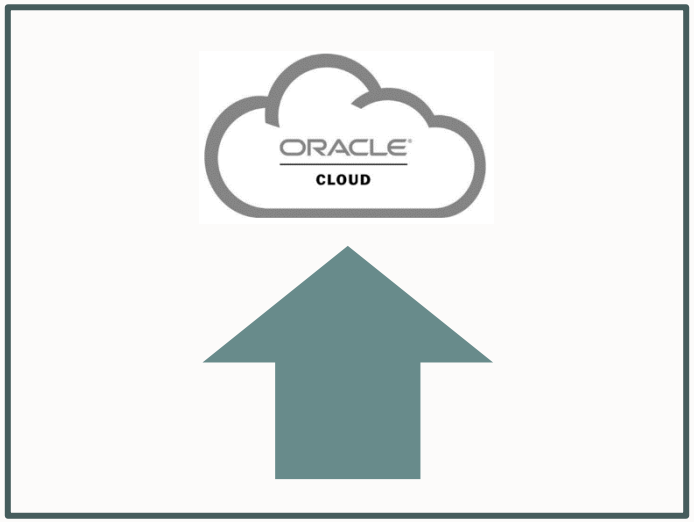
- Sort and filter to quickly find the biggest jobs.
- Find out who is consuming the most.
- Stop jobs that are running too long.



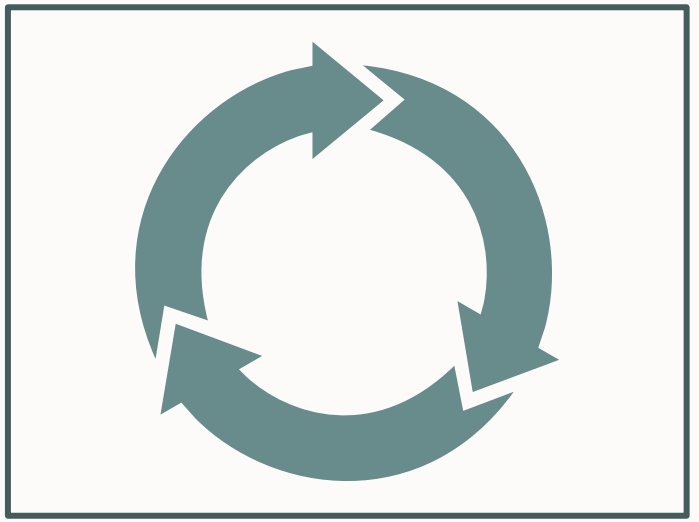
OCI Data Flow

Data Flow

Desenvolva e teste aplicativos Spark localmente usando o IDE de sua escolha.



Carregue aplicativos como estão, sem modificações.



Execute na nuvem em qualquer escala.

OCI Data Flow

Convertendo CSV para Parquet

- Validação de Pré-requisitos
- Edição do script Python
- Criação da Data Flow Application
- Execução da aplicação
- Validação do resultado

Data Flow

Hands-on – Lab Data Flow