Utilizando Spark como serviço para transformar Dados em Informação

Erik Gama

Trilha Data Cloud Modernization

Erik Gama

THE DEVELOPER'S CONFERENCE

ORACLE

Oracle Cloud
Modo Gratuito

https://bit.ly/TDCConnections2022

Uso Livre

(Always Free)

Serviços que você pode usar por tempo ilimitado

+

Avaliação Gratuita de 30 dias

US$ 500 em créditos gratuitos

O que está incluído no Oracle Cloud – Modo Gratuito (Free Tier)?

Uso Livre (Always Free)

Serviços em nuvem de Uso Livre:

- Dois Oracle Autoinomous Databases com ferramentas avançadas como Oracle APEX e Oracle SQL Developer
- Duas VMs de Computação AMD
- Até 4 instâncias em ARM Ampere A1 Compute
- Armazenamento de Bloco, Objetos e Arquivos; Balanceador de Carga e Saída de Dados; Monitoramento e Notificações

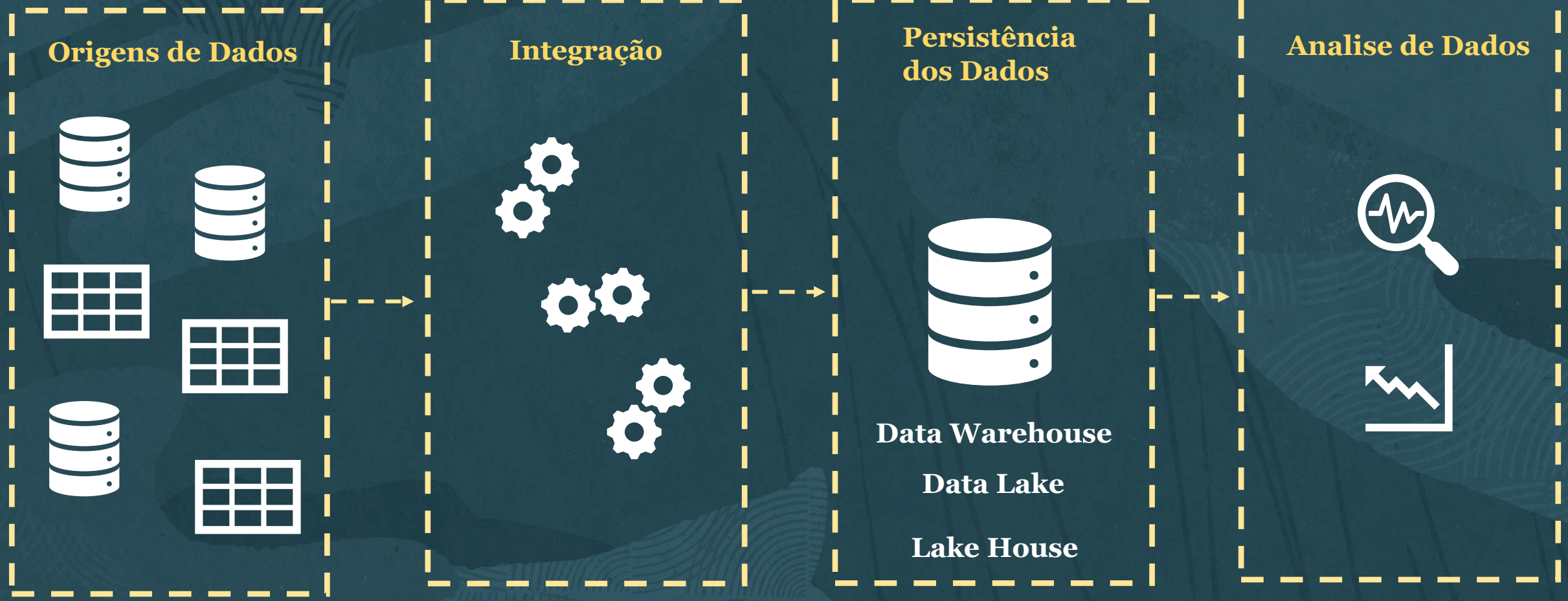
Avaliação Gratuita de 30 dias

US$ 500 em créditos gratuitos (exclusivo para os participantes do TDC Connections 2022)

- Acesso a uma ampla variedade de serviços da Oracle Cloud por 30 dias, incluindo Bancos de Dados, Análise Avançada, Computação e Container Engine for Kubernetes
- Até oito instâncias em todos os serviços disponíveis
- Até 5 TB de armazenamento

Mais detalhes em: https://www.oracle.com/br/cloud/free/

Arquitetura de Dados



Arquitetura de Dados

Persistência dos DADOS

Agnóstico

Rápido

Big data

Elástico

Controle

Flexível

Integrável

Seguro

Persistência dos DADOS

Object Storage


File.parquet


File.parquet


File.parquet


File.parquet

Presentation


File.parquet


File.parquet


File.parquet


File.parquet

Transformation


File.parquet


File.parquet


File.orc


File.avro

Raw Data

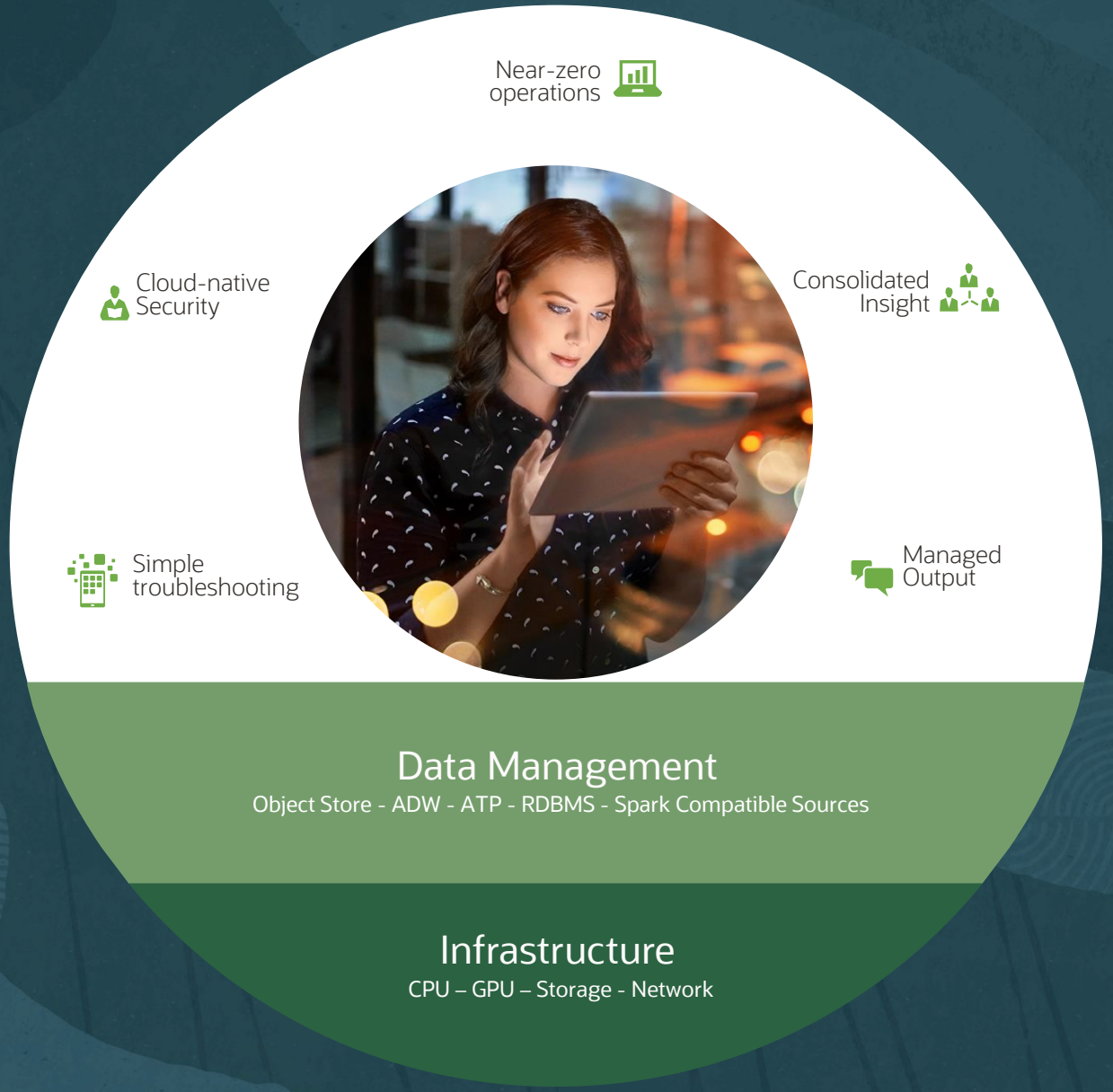
OCI Data Flow

Forneça aplicativos de Big Data e aprendizado de máquina com mais rapidez

Serviço Spark totalmente gerenciado com sobrecarga administrativa quase nula.

Importe / execute aplicativos Spark existentes de EMR, Databricks ou Hadoop.

Serviço elástico de big data.



Oracle Cloud Infrastructure Data Flow

BATCH

Data Flow

Autonomous Database

Database System

Object Storage

Buckets

Data Flow

Sem infraestrutura para gerenciar

O OCI Data Flow provisiona tudo, executa seu trabalho e desliga tudo.

Visão consolidada

Mostramos cada trabalho do Spark, quem o executou, quanto tempo demorou e quanto foi processado.

Diagnostique facilmente trabalhos de longa duração ou com falha

Runs *in* Marketing *Compartment*

Name	Language	State	Owner	Created	Duration	Total oCPU	Data Read	Data Written
PSR_AirlineETL	SQL	Succeeded	user@example.com	Fri, Aug 16, 2019, 21:46:38 UTC	15m	22	151 GB	6 GB
PSR_Scala_test	Scala	Succeeded	user@example.com	Thu, Aug 15, 2019, 18:20:59 UTC	1h 49m 7s	22	62 GB	92 MB
Simple Test App	Java	Succeeded	user@example.com	Mon, Oct 7, 2019, 21:30:33 UTC	1m 40s	2	68 MB	31 MB
PSRTESTING_DJJWKEXEPCESGA	Java	Succeeded	user@example.com	Thu, Oct 17, 2019, 09:30:17 UTC	1m 14s	4	68 MB	31 MB
PSRTESTING_BRNOFFEVEQONTL	Java	Succeeded	user@example.com	Thu, Oct 17, 2019, 09:00:02 UTC	1m 11s	4	68 MB	31 MB
Simple Test App1	Java	Succeeded	user@example.com	Thu, Oct 17, 2019, 07:53:25 UTC	1m 14s	4	68 MB	31 MB

Data Flow

Saída Gerenciada

Capturamos e armazenamos com segurança a saída do job do Spark.

Facilita o envio de análises às pessoas que precisam delas.

Segurança nativa da Nuvem

Integrado com o sistema IAM de nível empresarial da Oracle Cloud

Logs

Name	File Size	Source	Type	Created
spark_application_stdout.log.gz	63 bytes	APPLICATION	STDOUT	Fri, Aug 16, 2019, 22:02:49 UTC
spark_application_stderr.log.gz	320 bytes	APPLICATION	STDERR	Fri, Aug 16, 2019, 22:02:49 UTC
spark_executor_stdout_20190816T215000Z.log.gz	409 KB	EXECUTOR	STDOUT	Fri, Aug 16, 2019, 22:11:02 UTC

Exercício DATA FLOW

Transformar dados da F1 em informação

http://ergast.com/mrd/

Exercício DATA FLOW

F1

Races

Circuits

Drivers

Results


ORACLE®