Imersão
Azure Databricks

Treinamentos

Fábio Santos

- + de 10 anos na área de TI.
- Hard Skills: Modelagem de dados, BI, Big Data, Data Engineer, SQL Server, Azure SQL Database, Azure Data Factory, Azure Data Lake, Azure Databricks, Azure Synapse Analytics, T-SQL.
- Graduado em Ciências da Computação, Análise e Desenvolvimento de Sistemas | Pós-Graduado em Banco de Dados com Ênfase em Business Intelligence e Ciência de Dados e Big Data.
- Senior Data Engineer | Senior Data Architecture

LS Treinamentos

Eduardo Bispo

- 3 anos como Data Engineer.
- Hard Skills:
  - Modelagem de dados, BI, Big Data, Data Engineer, SQL Server, Azure SQL Database, Azure Data Factory, Azure Data Lake, Azure Databricks, Azure Synapse Analytics, T-SQL.
- Azure Data Engineer
- Data Architecture na Kumulus
- Microsoft Certified Professional

Treinamentos

O que é o Azure Databricks?

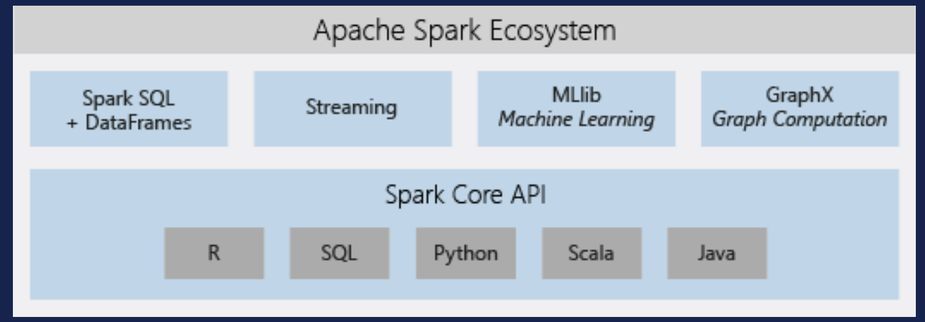
- O Azure Databricks é uma plataforma de análise dados baseada em **Apache Spark** e otimizada para os serviços da Azure.
- O Azure Databricks oferece três ambientes de desenvolvimento:
  - Data Science & Engineering
  - Machine Learning
  - SQL
- Integração com os serviços do Azure: Pode se integrar com uma variedade de serviços de plataforma de dados do Azure e do Power BI

O que é o Apache Spark?

- O Apache Spark é um framework de Big Data que tem o objetivo de processar grandes volumes de dados de forma paralela e distribuída.
- Uma vantagem do Spark é que ele possui componentes que funcionam dentro da própria ferramenta, como o Spark Streamming, Spark SQL e o Graphx.
- Outro ponto importante é que ele permite a programação em: Java, Scala, R, SQL e Python.

Quais os componentes do Apache Spark?

- O Spark tem diversos componentes para diferentes tipos de processamentos, todos construídos sobre o Spark Core, que é o componente que disponibiliza as funções básicas para o processamento como as funções map, reduce, filter e collect.

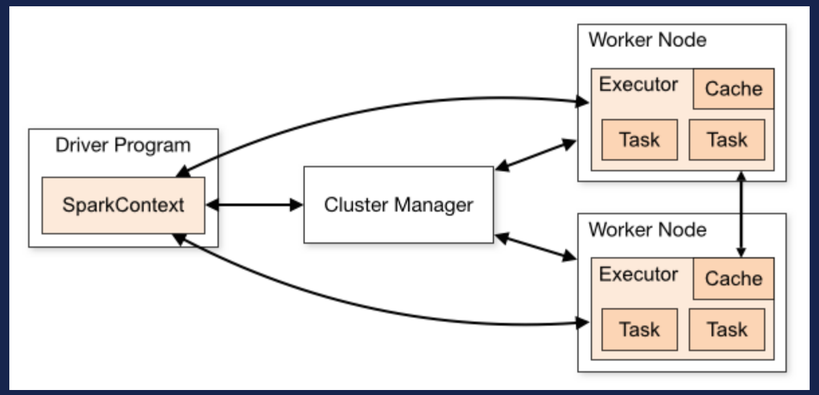


- DataFrame: é um conjunto de dados organizado em colunas. É o equivalente a uma tabela em um banco de dados relacional.
- Spark Streamming: Possibilita o processamento de fluxos em tempo real.
- GraphX: Realiza o processamento sobre grafos.
- SparkSQL: Permite a utilização de SQL na realização de consultas e processamento sobre os dados no Spark.
- Mllib: Biblioteca de aprendizado de máquina, com deferentes algoritmos para as mais diversas atividades, como clustering.

Arquitetura do Apache Spark

A arquitetura de uma aplicação Spark é constituída por três partes principais:

- O Driver Program, que é a aplicação principal que gerencia a criação e é quem executará o processamento definido pelo programador;
- O Cluster Manager é um componente opcional que só é necessário se o Spark for executado de forma distribuída. Ele é responsável por administrar as máquinas que serão utilizadas como workers;
- Os Workers, que são as máquinas que realmente executarão as tarefas que são enviadas pelo Driver Program. Se o Spark for executado de forma local, a máquina desempenhará tanto o papel de Driver Program como de Worker.



APACHE Spark™

Treinamentos

O que é o Azure Databricks?

- O Azure Databricks é uma plataforma de análise dados baseada em **Apache Spark** e otimizada para os serviços da Azure.
- O Azure Databricks oferece três ambientes de desenvolvimento:
  - Data Science & Engineering
  - Machine Learning
  - SQL
- Integração com os serviços do Azure: Pode se integrar com uma variedade de serviços de plataforma de dados do Azure e do Power BI

Data Science & Engineering Workspace

Treinamentos

Data Science & Engineering

Data Science & Engineering: Um ambiente que pemite a colaboração entre engenheiros de dados, analistas de dados e engenheiros de machine learning.

Nele vamos ter acesso a:

- Workspace
- Repos
- Data
- Compute
- Jobs

Data Science & Engineering - Workspace

Um *workspace* do Databricks é um ambiente para acessar todos os ativos do Databricks. O *wokspace* organiza objetos (notebooks, bibliotecas e experimentos) em pastas e fornece acesso a dados e recursos computacionais como *clusters* e *workers*.

Treinamentos

Data Science & Engineering - Repos

Para dar suporte às práticas recomendadas de desenvolvimento de código, o Databricks Repos fornece integração em nível de repositório com provedores Git. Você pode desenvolver código em um notebook do Databricks e sincronizá-lo com um repositório git remoto.

O Databricks suporte os seguintos provedores Git:

- Github
- Bitbucket
- GitLab
- Azure DevOps
- AWS CodeCommit

Data Science & Engineering - Data

O Databricks Database consiste em uma coleção de tabelas. Uma tabela do Databricks consiste em uma coleção de dados estruturads. Você pode armazenar em cache, filtrar e executar operações com suporte ao Apache Spark DataFrames.

Há dois tipos de tabelas: **global** e **local**.

- Tabela global: uma tabela global está disponível em todos os clusters.
- Tabela local: uma tabela local não é acessível de outros clusters. Isso também é conhecido como **TempView.**

Data Science & Engineering - Compute

No ambiente do Databricks é possível criar dois tipos de unidades computacionais **Clusters** e **Pools**.

- Clusters: Existem dois tipos de clusters:
  - All-purpose clusters: Criados usando UI, CLI or REST API, Você pode finalizar ou reiniciar. Vários usuários podem usar o cluster esse tipo de cluster.
  - Job clusters: Criados a partir de um Job, são clusters criado quando um Job é iniciado e encerrado quando o Job é concluído.
- Pools: Os pools do Databricks reduzem os tempos de início do cluster e de dimensionamento automático mantendo um conjunto de instâncias ociosas e prontas para uso.

Data Science & Engineering - Job

Um Job é uma maneira de executar código não interativo em um cluster Databricks. Por exemplo, você pode executar uma carga de trabalho de extração, transformação e carregamento (ETL) interativamente ou em um agendamento. Você também pode executar trabalhos interativamente na interface do usuário do notebook.

Treinamentos

Databricks SQL Workspace

Treinamentos

Databricks SQL

O **Databricks SQL** permite que você execute consultas SQL rápidas no seu data lake. As consultas dão suporte a vários tipos de visualização para ajudar você a explorar seus resultados da consulta de diferentes perspectivas.

Nele vamos ter acesso a:

- Data
- SQL Endpoints
- Alerts
- Dashboads

Databricks SQL – Data

O **Databricks SQL Database** consiste em uma coleção de tabelas. Uma tabela do Databricks consiste em uma coleção de dados estruturads. Você pode armazenar em cache, filtrar e executar operações com suporte ao Apache Spark DataFrames.

Há dois tipos de tabelas: **global** e **local**.

- Tabela global: uma tabela global está disponível em todos os clusters.
- Tabela local: uma tabela local não é acessível de outros clusters. Isso também é conhecido como **TempView.**

Treinamentos

Databricks SQL – SQL Endpoints

Um SQL Endpoint é um recurso de computação que permite executar comandos SQL em objetos de dados no Databricks SQL.

Treinamentos

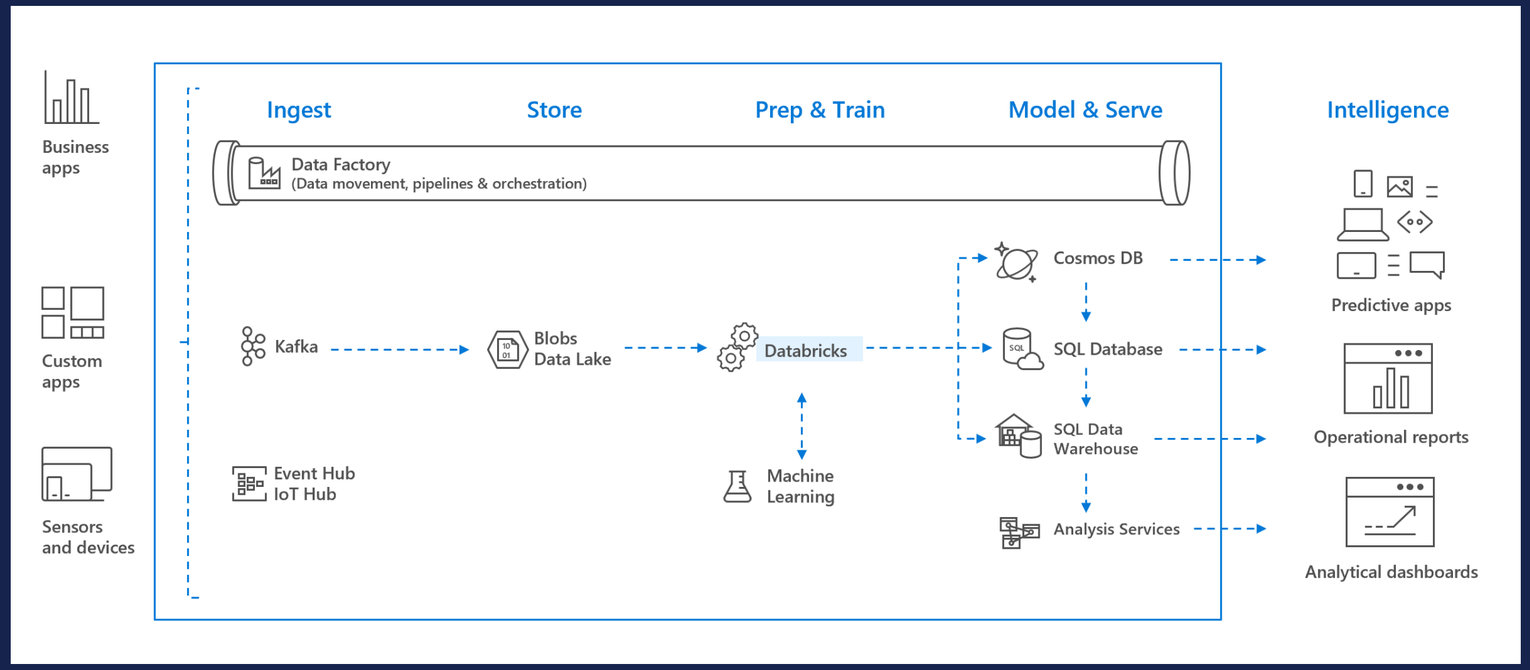
Databricks SQL – Dashboards

Um Dashboard do Databricks SQL permite que você combine visualizações e caixas de texto que fornecem contexto com os dados.

Treinamentos

Arquitetura

Arquitetura – Data Science & Engineering



Treinamentos

Arquitetura – Data Science & Engineering + SQL

Treinamentos

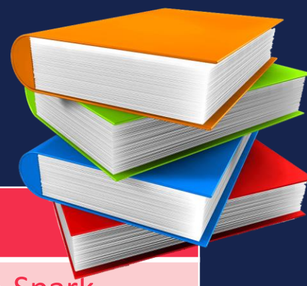
Perguntas?

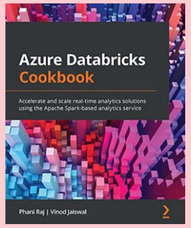
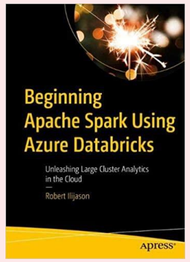
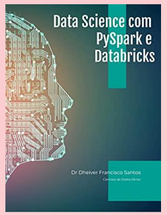
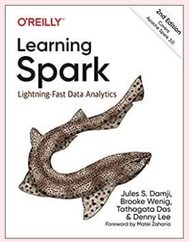
Treinamentos

Links de referência

- https://docs.databricks.com/
- https://docs.microsoft.com/pt-br/azure/databricks/
- https://blog.pragmaticworks.com/topic/azure-databricks
- https://www.advancinganalytics.co.uk/blog/tag/Databricks
- https://www.topcoder.com/thrive/search?tags[]=Databricks
- https://databricks.com/discover/notebook-gallery
- https://www.advancinganalytics.co.uk/blog
- https://blog.dsacademy.com.br/categoria/apache-spark/
- https://delta.io/blog-gallery/
- http://datanrg.blogspot.com/search/label/ADF
- https://www.mssqltips.com/search/?q=azure%20databricks
- https://sparkbyexamples.com/

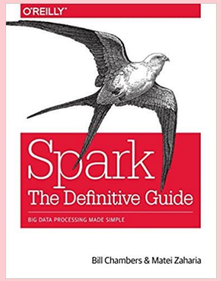
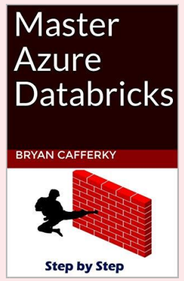
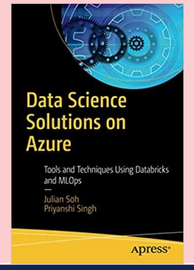
Livros



Livro	Link
	https://www.amazon.com.br/Azure-Databricks-Cookbook-Accelerate-Spark-based-ebook/dp/B0855XX5HS/ref=sr_1_1?__mk_pt_BR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&crid=2TTBGZM659EK4&keywords=databricks&qid=1641221073&sprefix=databricks%2Caps%2C199&sr=8-1&ufe=app_do%3Aamzn1.fos.25548f35-0de7-44b3-b28e-0f56f3f96147
	https://www.amazon.com.br/Beginning-Apache-Spark-Using-Databricks/dp/1484257804/ref=sr_1_3?__mk_pt_BR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&crid=2TTBGZM659EK4&keywords=databricks&qid=1641221073&sprefix=databricks%2Caps%2C199&sr=8-3&ufe=app_do%3Aamzn1.fos.4bb5663b-6f7d-4772-84fa-7c7f565ec65b
	https://www.amazon.com.br/Data-Science-com-PySpark-Databricks-ebook/dp/B09DBVQY1L/ref=sr_1_4?__mk_pt_BR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&crid=2TTBGZM659EK4&keywords=databricks&qid=1641221073&sprefix=databricks%2Caps%2C199&sr=8-4
	https://www.amazon.com.br/Learning-Spark-2e-Jules-Damji/dp/1492050040/ref=sr_1_8?__mk_pt_BR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&crid=2TTBGZM659EK4&keywords=databricks&qid=1641221073&sprefix=databricks%2Caps%2C199&sr=8-8&ufe=app_do%3Aamzn1.fos.4bddec23-2dcf-4403-8597-e1a02442043d

Treinamentos

Livros

Livro	Link
	https://www.amazon.com.br/Spark-Definitive-Guide-Bill-Chambers/dp/1491912219/ref=sr_1_15?__mk_pt_BR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&crid=2TTBGZM659EK4&keywords=databricks&qid=1641221073&sprefix=databricks%2Caps%2C199&sr=8-15&ufe=app_do%3Aamzn1.fos.25548f35-0de7-44b3-b28e-0f56f3f96147
	https://www.amazon.com.br/Master-Azure-Databricks-Step-English-ebook/dp/B08JFWCNDF/ref=sr_1_2?__mk_pt_BR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&crid=2TTBGZM659EK4&keywords=databricks&qid=1641221073&sprefix=databricks%2Caps%2C199&sr=8-2
	https://www.amazon.com.br/Data-Science-Solutions-Azure-Techniques/dp/1484264045/ref=sr_1_6?__mk_pt_BR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&keywords=databricks&qid=1641226673&sr=8-6&ufe=app_do%3Aamzn1.fos.e05b01e0-91a7-477e-a514-15a32325a6d6

Treinamentos

Canais do YouTube

- https://www.youtube.com/c/Databricks/playlists
- https://www.youtube.com/playlist?list=PLMWaZteqtEaKi4WAePWtCSQCfQpvBT2U1
- https://www.youtube.com/playlist?list=PLppGISR503dWC9HKPMYmi_E29Dr_wzPtq
- https://www.youtube.com/playlist?list=PLfOYknl0ZnIane4rJrtbcdsD5WYetWRVW
- https://www.youtube.com/playlist?list=PL7_h0bRfL52oWNfE0GhwbnNjeJSmf8Q35
- https://www.youtube.com/playlist?list=PL7_h0bRfL52rUU6chVIygk7eEiB3Htj-C
- https://www.youtube.com/playlist?list=PL7_h0bRfL52qWoCcS18nXcT1s-5rSa1yp
- https://www.youtube.com/playlist?list=PLW0Bbnox7aDviJh3bBhnPTmilA9E_mEgs
- https://www.youtube.com/playlist?list=PLW0Bbnox7aDtGJpjevHcU7xJNm7PtLI5f

Treinamentos

Cursos

- https://docs.microsoft.com/pt-br/learn/browse/?terms=databricks
- https://docs.microsoft.com/pt-br/learn/certifications/azure-data-fundamentals/
- https://academy.databricks.com/
- https://www.udemy.com/course/azure-databricks-spark-core-for-data-engineers/
- https://www.udemy.com/course/databricks-fundamentals-apache-spark-core/
- https://www.udemy.com/course/data-engineering-using-databricks-on-aws-and-azure/
- https://www.udemy.com/course/apache-spark-3-databricks-certified-associate-developer/
- https://www.udemy.com/course/databricks-unificando-seus-dados-e-analise/
- https://www.udemy.com/course/databricks-e-pyspark-analisando-dados/
- https://www.udemy.com/course/big-data-apache-spark-com-pyspark-para-iniciantes/

Treinamentos

Contatos

Muito Obrigado

in fabiofsantos

✉ softwaressantoss@icloud.com

in eduardo-bispo-ferreira-963200174

✉ eduardobispof@gmail.com

Treinamentos