Soluções de Dados, Big Data e Machine Learning

Capítulo 1. Introdução à Plataforma de Dados do Azure

PROF. GUSTAVO AGUILAR

Soluções de Dados, Big Data e Machine Learning

AULA 1.0. INTRODUÇÃO

PROF. GUSTAVO AGUILAR

Nesta Aula

- Apresentação do Professor
- Apresentação da Disciplina

Apresentação do Professor

Gustavo Aguilar

Atuação

- Administração de Bancos de Dados
- Ambientes de missão crítica em diversas plataformas
- Modelagem, Arquitetura e Engenharia de Dados
- Persistência e Pesquisa de Dados
- Sistemas Distribuídos e Cloud Computing
- Metodologias Ágeis e DevOps
- Professor e Instrutor Certificado Microsoft (MCT)

Formação

- Bacharelado em Ciência da Computação (PUC-Minas)
- Pós-Graduação em Administração de Banco de Dados
- Especialização em Docência do Ensino Superior
- MBA em Ciência de Dados (IGTI)

Apresentação da Disciplina

- Módulo de apresentação das soluções disponíveis no Azure para:
  - Armazenamento de dados;
  - Armazenamento de dados relacionais / não relacionais em bancos de dados;
  - Distribuição de dados;
  - Big Data;
  - Análise de dados em tempo real;
  - Monitoração e backup das estruturas de dados;
  - Disaster recovery na camada de dados.

Apresentação da Disciplina

Espera-se que o aluno consiga, ao final deste módulo:

- ✓ Compreender os papéis dos tipos de profissionais que trabalham com a plataforma de soluções de dados do Azure;
- ✓ Projetar soluções de dados usando infraestrutura ou plataforma como serviço;
- ✓ Criar soluções usando serviços de banco de dados relacional ou não relacional do Azure;
- ✓ Planejar soluções com distribuição de dados usando o CosmosDB;
- ✓ Projetar soluções de preparação de dados usando o Azure Databricks;
- ✓ Planejar soluções de ingestão de dados usando o Azure Event Hubs;

Apresentação da Disciplina

Espera-se que o aluno consiga, ao final deste módulo:

- ✓ Criar soluções de extração, ingestão e transformação de dados com Azure Data Factory;
- ✓ Projetar soluções para análise de dados com o Azure Synapse Analytics;
- ✓ Projetar soluções para o armazenamento e processamento de dados em larga escala usando o Azure HDInsight;
- ✓ Planejar soluções de aprendizagem de máquina usando o Azure Machine Learning;
- ✓ Planejar o monitoramento das soluções de dados implementadas no Azure;
- ✓ Planejar o backup dos dados no Azure.

Apresentação da Disciplina

- **Material Didático**
  - ✓ Apostila
  - ✓ Aulas gravadas
  - ✓ Slides das aulas gravadas
  - ✓ Aulas interativas ao vivo (*gravação e slides são disponibilizados)
- **Atividades dos Alunos**
  - ✓ Fórum de dúvidas e fórum de debates
  - ✓ Trabalho prático
  - ✓ Desafio final
  - ✓ Questões de reposição da aula interativa (*para quem não participou)

Apresentação da Disciplina

- **Distribuição de pontos**

✓ **10 pontos** de participação na 1ª Aula Interativa (*presença computada através de resposta à enquete)

✓ **25 pontos** do Trabalho Prático

✓ **10 pontos** de participação no Fórum de Debates do módulo 3

✓ **10 pontos** de participação na 2ª Aula Interativa (*presença computada através de resposta à enquete)

✓ **40 pontos** do Desafio do módulo 3

✓ **5 pontos** do Feedback do aluno em relação ao conteúdo e professor do módulo.

- **TOTAL: 100 pontos**

Próxima Aula

- Modalidades de Serviços

Soluções de Dados, Big Data e Machine Learning

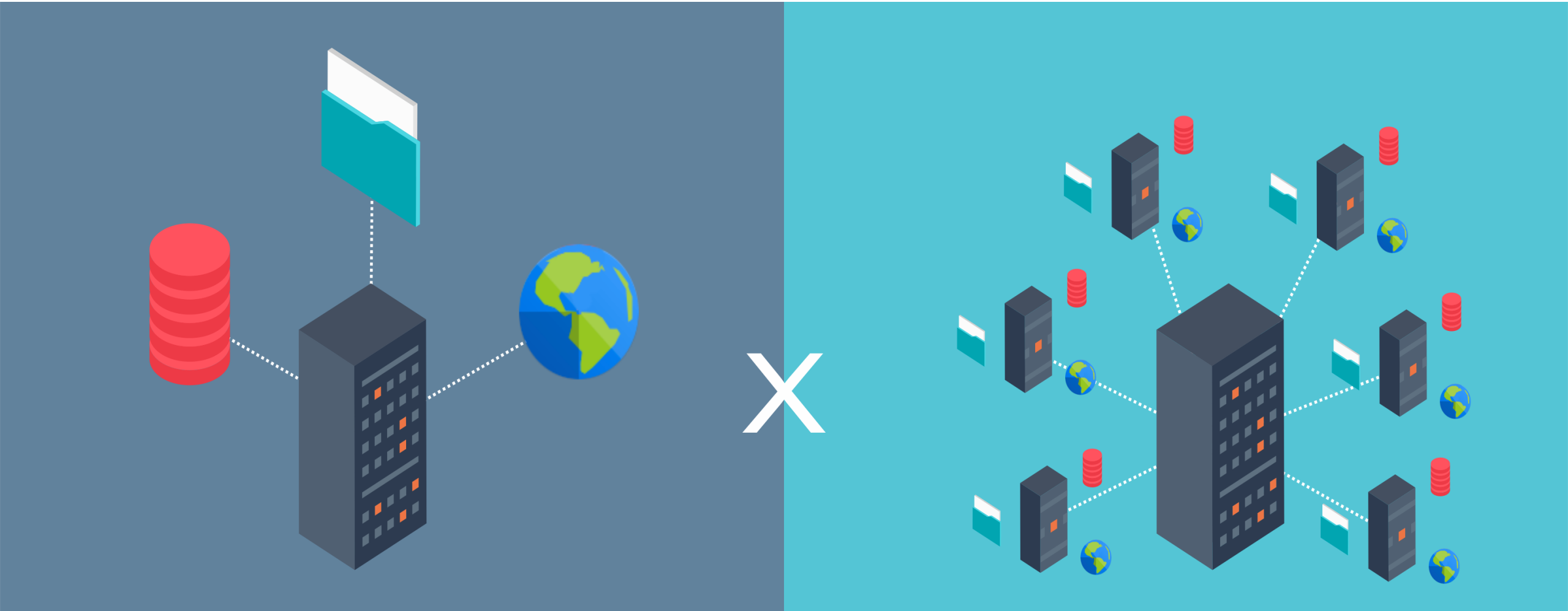
AULA 1.1. MODALIDADES DE SERVIÇOS

PROF. GUSTAVO AGUILAR

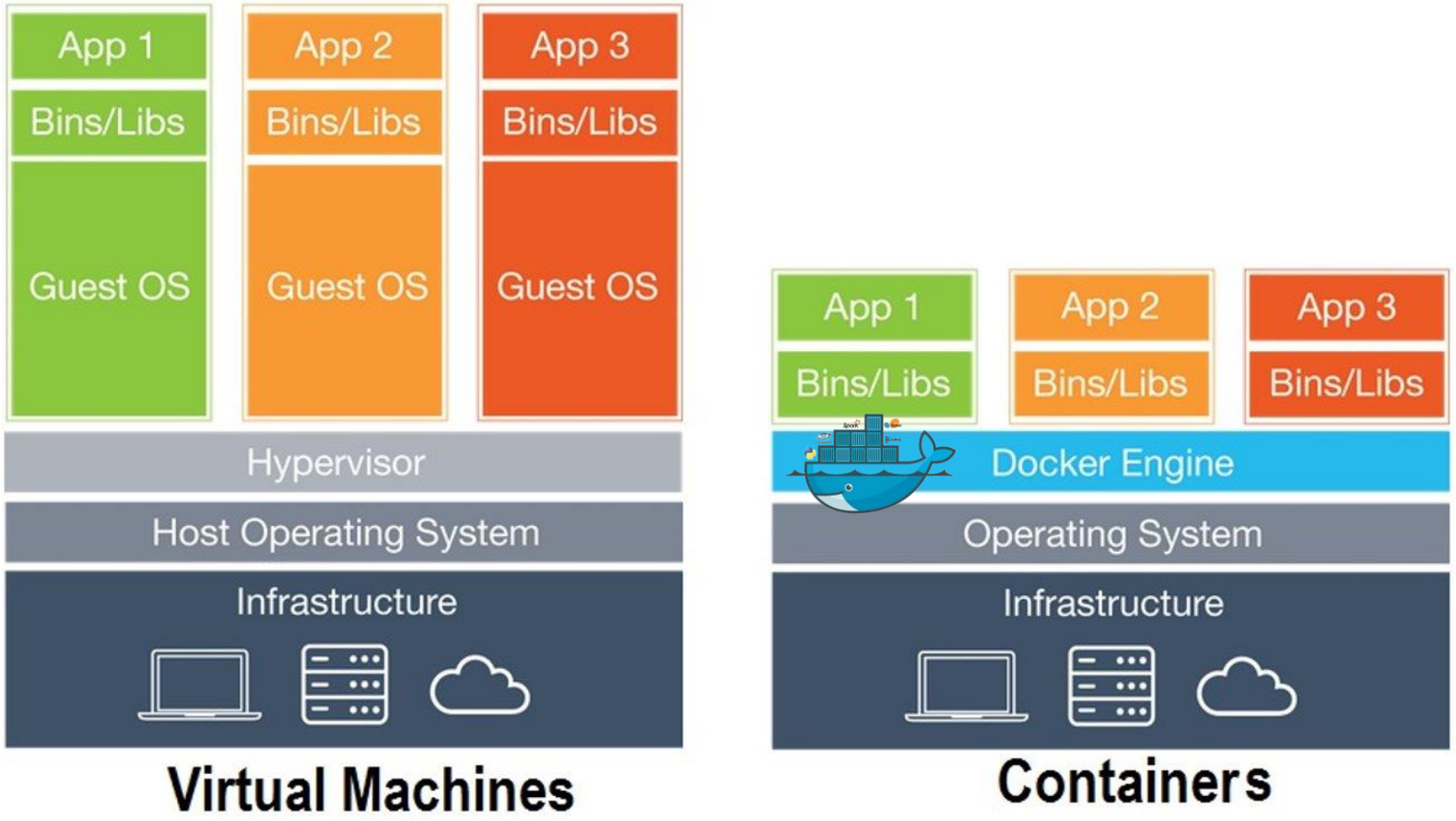
Nesta Aula

- Servidor Físico x Virtual x Container
- Modalidades de Serviços
- IaaS x PaaS x SaaS
- Quando Usar IaaS

Servidor Físico x Virtual



Servidor Virtual x Container



Modalidades de Serviços

IaaS

Infraestrutura como serviço

Servidores, armazenamento, segurança, rede

Windows Azure VM, EC2, VPCs, S3, etc.

PaaS

Plataforma como serviço

Ferramentas Dev., SO, BD

Windows Azure, SQL Azure, etc.

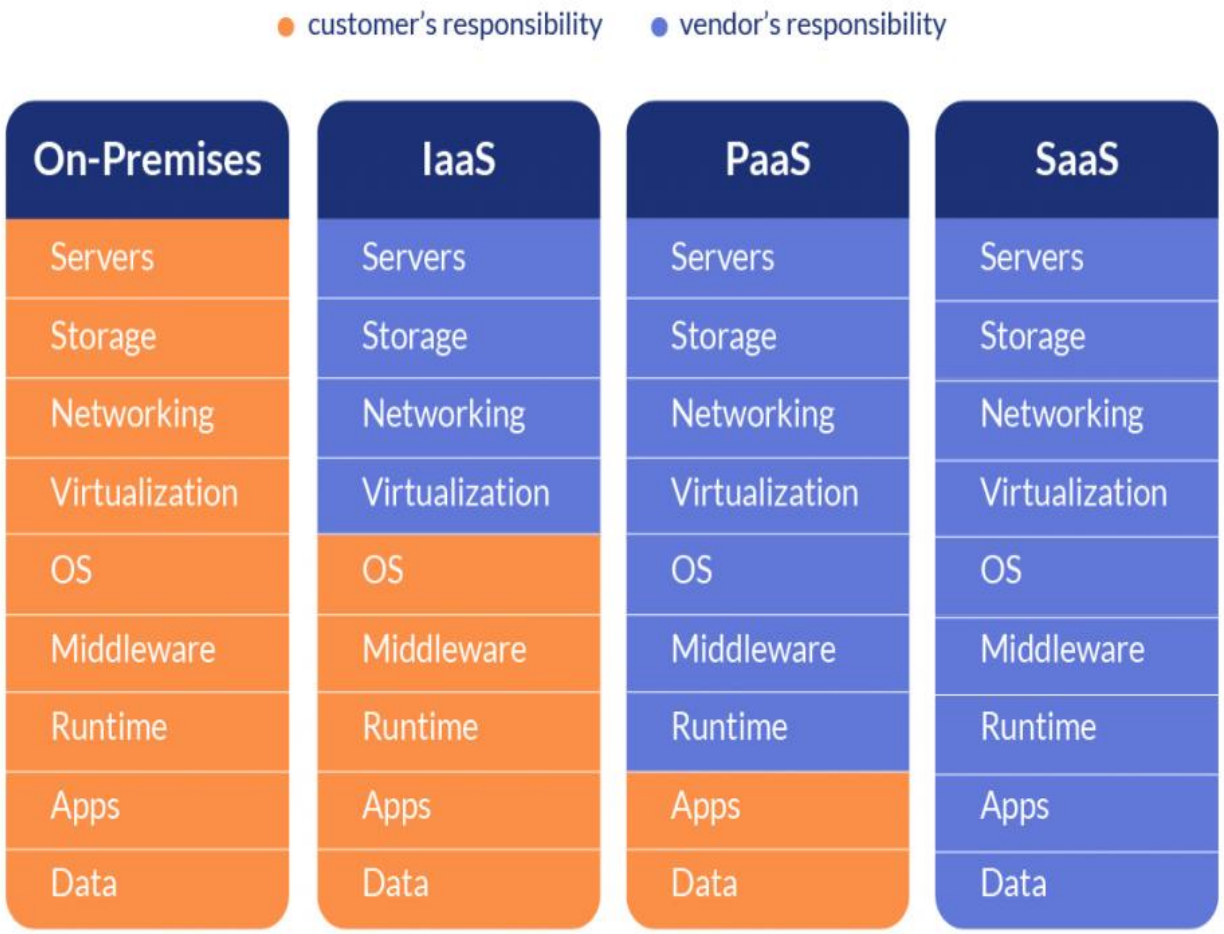
SaaS

Software como Serviço

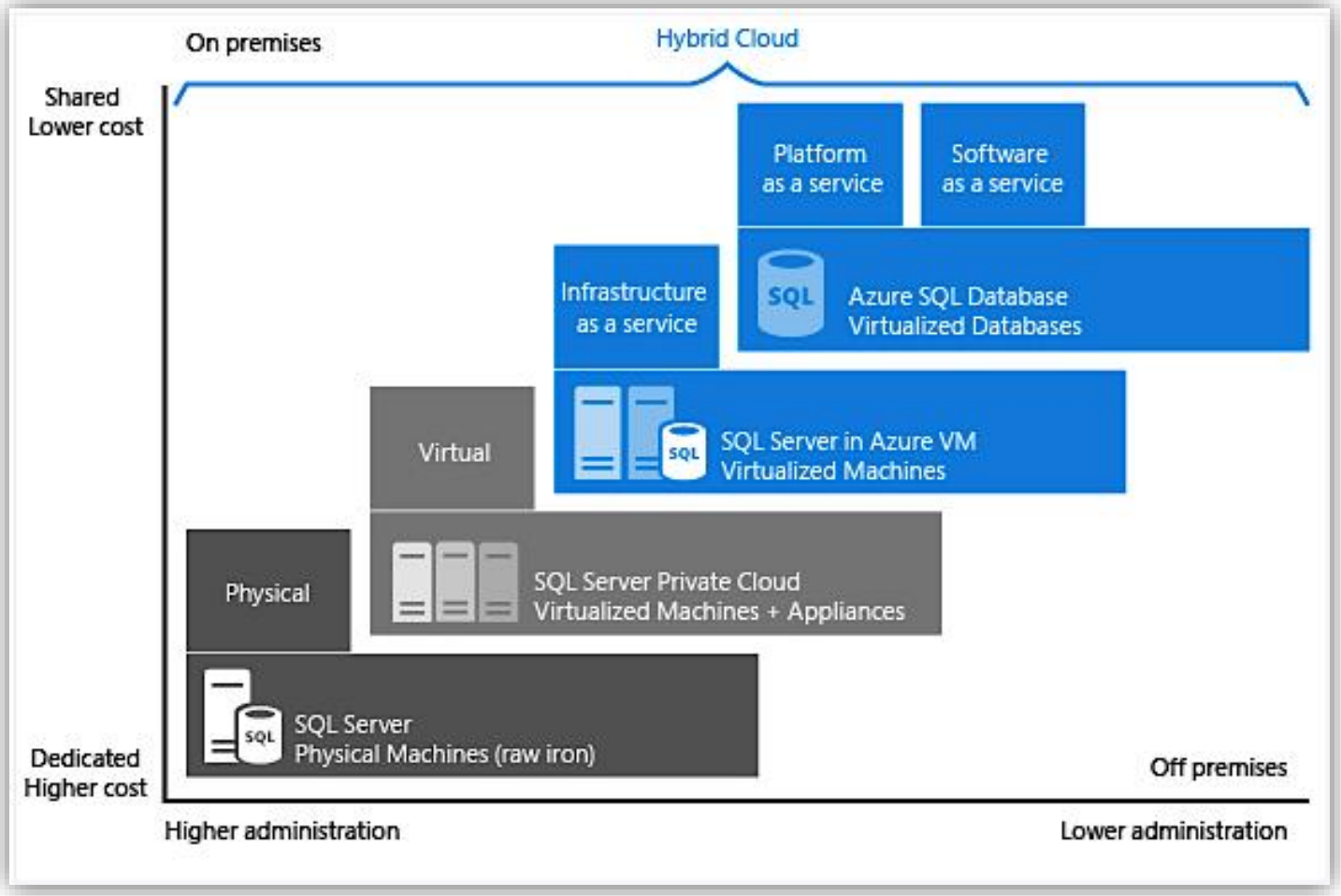
Apps / aplicativos hospedados

Office 365, Webmails, Redes Sociais, etc.

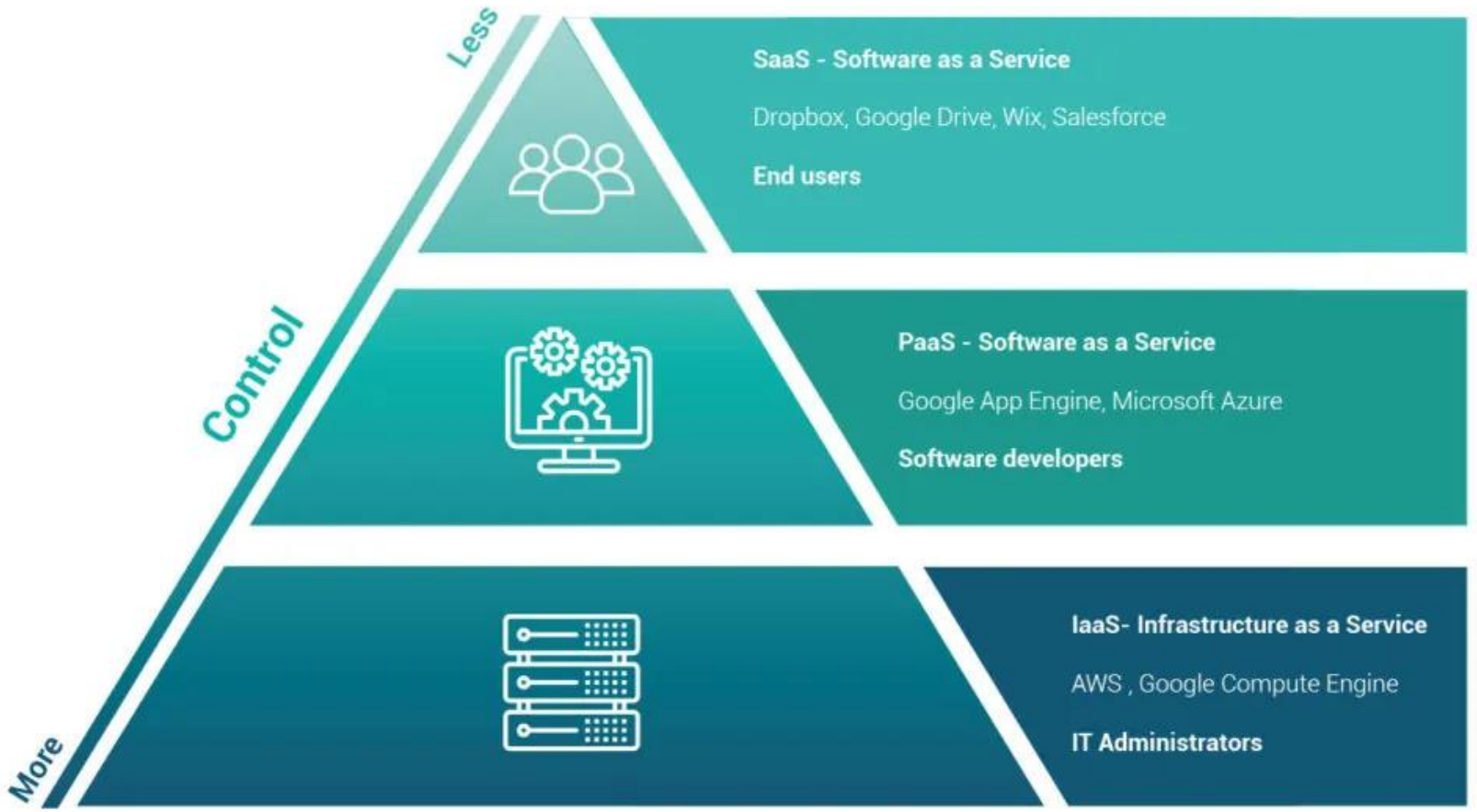
Responsabilidades



Modalidades de Serviços



Modalidades de Serviços



Quando Usar IaaS

- Versões mais antigas do SQL Server;
- Uso de outros serviços do SQL Server;
- Necessidades dos aplicativos x Recursos de PaaS;
- Não desejar atualizações automáticas;
- Facilidade de migração: VM on premise ➔ VM na nuvem.

Próxima Aula

- Tipos de Dados

Soluções de Dados, Big Data e Machine Learning

AULA 1.2. TIPOS DE DADOS

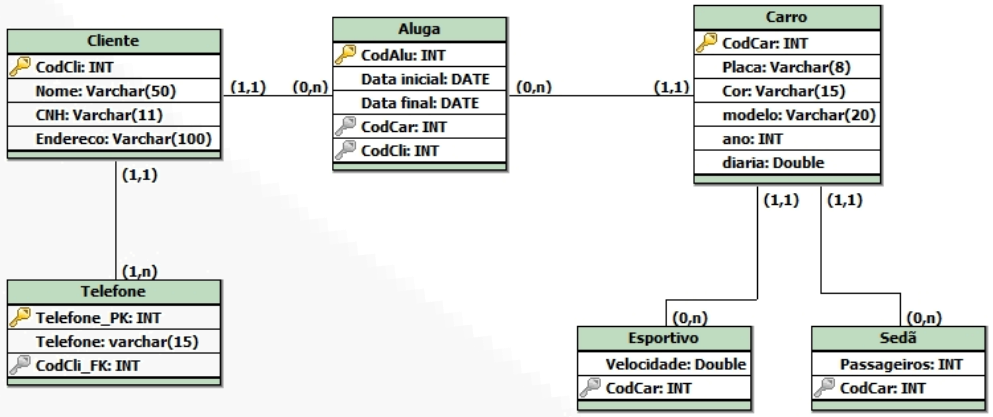
PROF. GUSTAVO AGUILAR

Nesta Aula

- Dados Estruturados
- Dados Semiestruturados
- Dados Não Estruturados
- Dados na Era da Informação

Dados Estruturados

- Organizados e representados com uma estrutura rígida
  - Aderem a um esquema (schema físico dos dados).
- Definição prévia de um modelo de dados
  - Todos os dados têm os mesmos campos ou propriedades.
- Em geral, armazenados no formato tabular (em tabelas)
  - Linha + Coluna(s) ➔ tupla ➔ dados relacionais.



Dados Semiestruturados

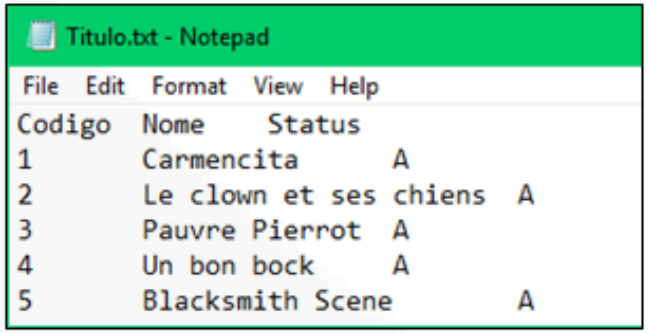
- Não contém toda a rigidez requerida na definição dos tipos de dados estruturados;
- Procuram manter certa uniformidade no armazenamento das informações
  - Mantém tags e marcações internas que identificam elementos de dados separados.
- Comumente chamados de dados não relacionais ou NOSQL.

Dados Não Estruturados

- Não são estruturados por meio de modelos ou esquemas de dados predefinidos;
  - Significa que não há restrições quanto aos tipos de dados que podem conter.
  - **Campo Blob** pode conter**:**
    - Documento PDF
    - Imagem JPEG
    - Áudio MP3
    - Vídeo MPEG
    - Etc.

Dados na Era da Informação

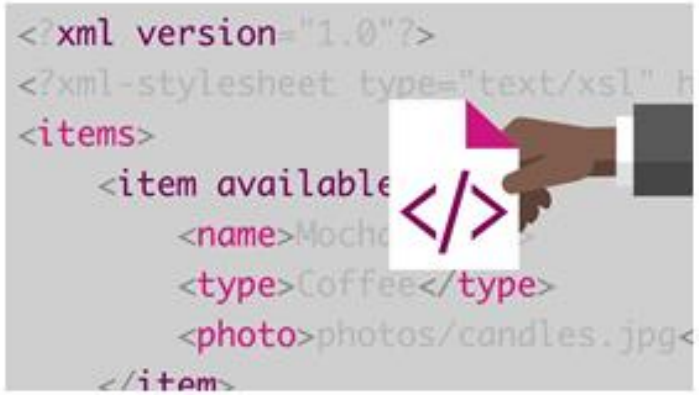
- “Início da Era da Internet”:
  - Volume de dados não muito significativo;
  - Grande parte dos dados de tipo textual;
  - Quase a totalidade dos sistemas trabalhando com **dados estruturados**.



Codigo	Nome	Status
1	Carmencita	A
2	Le clown et ses chiens	A
3	Pauvre Pierrot	A
4	Un bon bock	A
5	Blacksmith Scene	A

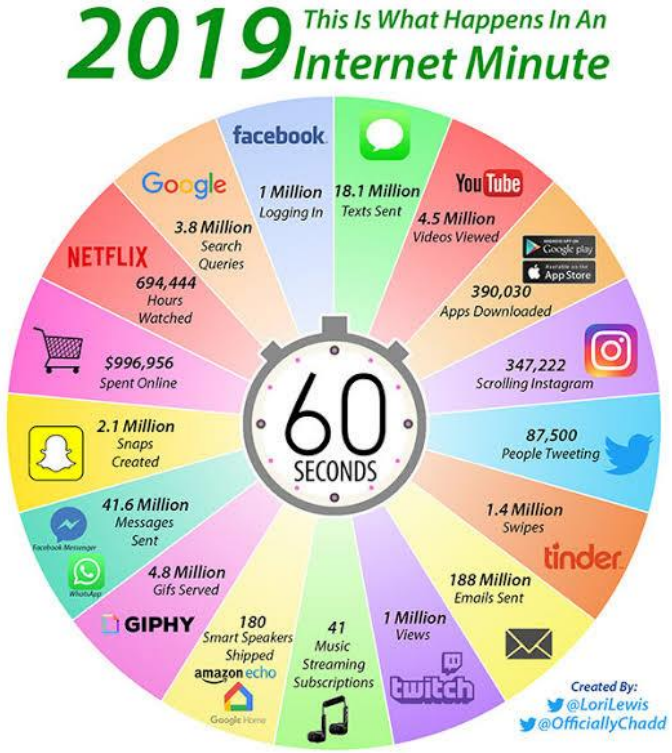
Dados na Era da Informação

- “Pré-era da Informação”:
  - Expansão da Internet;
  - Volume de dados iniciando um crescimento exponencial;
  - Sistemas trabalhando também com dados semiestruturados.



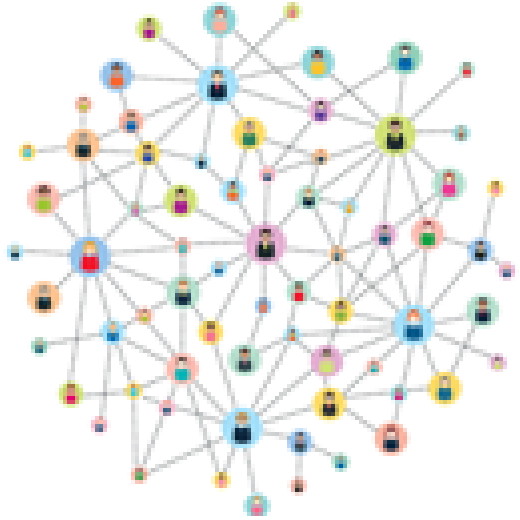
Dados na Era da Informação

- Era da Informação:
  - Uso massivo de Internet banda larga;
  - Computação Obíqua;
  - Internet das Coisas (IOT);
  - Uso intensivo de redes sociais para divulgação de informações e conteúdo;



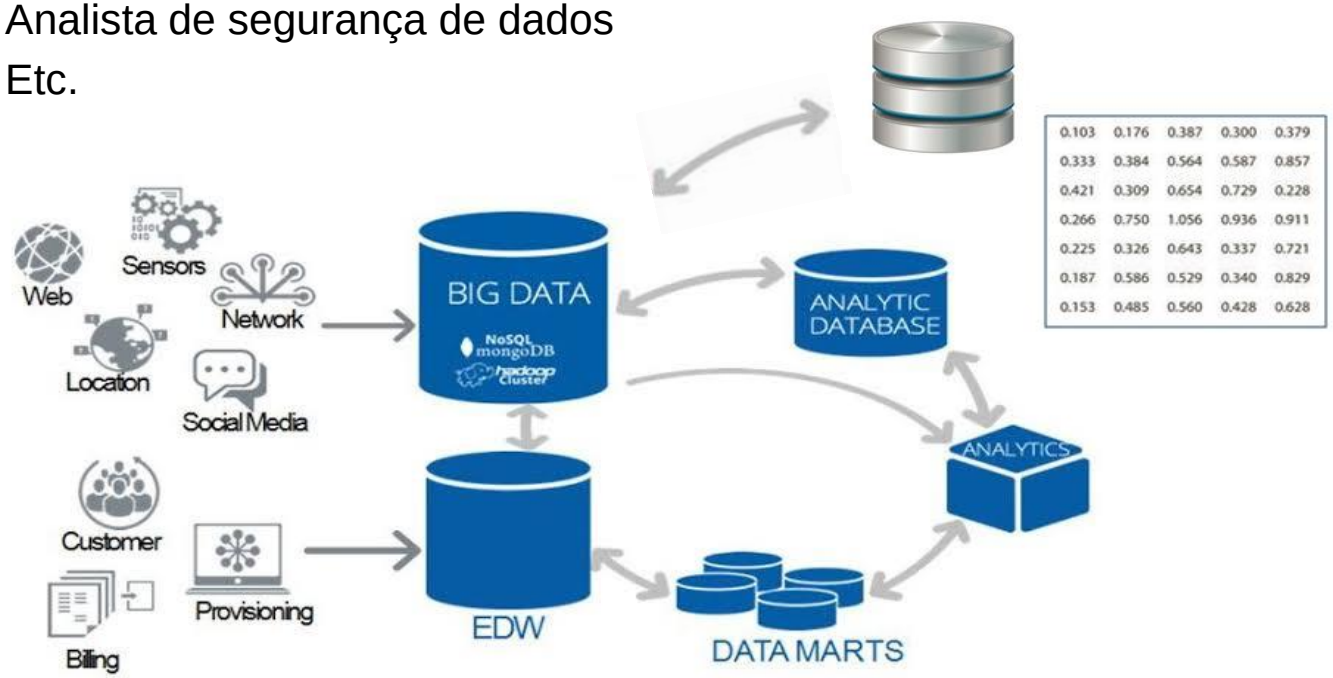
Dados na Era da Informação

- Era da Informação:
  - Crescimento exponencial do volume de dados, principalmente não estruturados
  - Fotos, vídeos, posts, likes, snaps, etc.;
  - Proliferação de soluções de Big Data para lidar com esse alto volume de dados não estruturados;
  - Coexistência com soluções legadas e/ou baseadas em dados estruturados / semiestruturados.



Dados na Era da Informação

- Papel fundamental do **Profissional de Dados**.
  - Arquiteto de dados
  - Engenheiro de dados
  - Analista de dados
  - Analista de segurança de dados
  - Etc.



Próxima Aula

- Perfis de Profissionais de Dados

Soluções de Dados, Big Data e Machine Learning

AULA 1.3. PERFIS DE PROFISSIONAIS DE DADOS

PROF. GUSTAVO AGUILAR

Nesta Aula

- ❑ Overview do Processo de Soluções de Dados
- ❑ Perfis de Profissionais de Dados

Overview do Processo de Soluções de Dados

Projetar a Solução

Implantar a Infraestrutura para a Solução

Implantar o Pipeline de Dados

Análise de Dados
Ciência de Dados
Aprendizado de Máquina
Etc.

Perfis de Profissionais de Dados

- Arquiteto de Soluções
- Arquiteto de Dados
- Administrador de Banco de Dados
- Engenheiro de Dados
- Analista de Dados
- Cientista de Dados
- Engenheiro de Inteligência Artificial

Perfis de Profissionais de Dados

- **Arquiteto de Soluções (Solutions Architect)**
  - Visão holística de toda a solução;
  - Integração das soluções;
  - Recursos e capacidade para a solução;
  - Estratégias de backup, monitoramento;
  - Disponibilidade e escalabilidade da solução;
  - Plano de Continuidade de Negócio (PCN);
  - Custo da solução;
  - Etc.

Perfis de Profissionais de Dados

- **Arquiteto de Dados (Data Architect)**
  - Estruturas de armazenamento dos dados;
  - Modelos de dados;
  - Administração de dados corporativos;
  - Backup de dados de negócio.

Perfis de Profissionais de Dados

- **Administrador de Banco de Dados (DBA)**
  - Instalação / aprovisionamento
    - SGBDs
    - Plataformas de armazenamento de dados
    - Bancos de dados
    - Repositórios / quotas
  - Aspectos operacionais;
  - Tunning e troubleshoting;
  - Disponibilidade dos SGBDs / plataformas;
  - Segurança de acesso.

Perfis de Profissionais de Dados

- **Engenheiro de Dados (Data Engineer)**
  - Projeto do Pipeline (Fluxo) de Dados:
    - Extração de dados
    - Transformação de dados
    - Ingestão (carga) de dados
  - Implementação e gerenciamento do fluxo de dados estruturados e não estruturados de diversas origens (fontes de dados).

Perfis de Profissionais de Dados

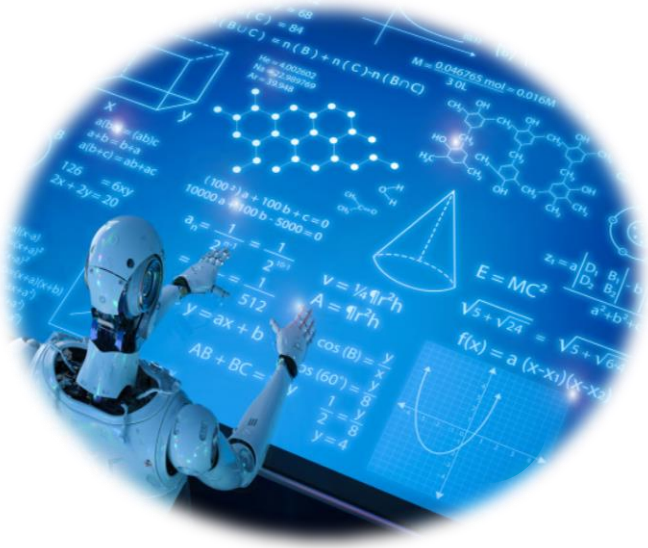
- **Analista de Dados (Data Analyst)**
  - Projetar e construir modelos de dados analíticos;
  - Transformar dados em informações analíticas com valor comercial;
  - Planejamento e gerenciamento de dashboards.

Perfis de Profissionais de Dados

- **Cientista de Dados (Data Scientist)**
  - Análise avançada para ajudar a gerar valor a partir dos dados;
  - Análise Descritiva: Análise Exploratória de Dados (EDA);
  - Análises Preditivas: com Machine Learning;
  - Suporte a decisões orientadas a dados (Data-Driven Decision).

Perfis de Profissionais de Dados

- **Engenheiro de Inteligência Artificial**
  - Arquitetar e implementar soluções de IA
    - Serviços Cognitivos
    - Machine Learning
    - Mineração de conhecimento
  - Suporte a soluções de processamento de linguagem natural, reconhecimento de voz e imagem;
  - Bots e agentes virtuais.



Overview do Processo de Soluções de Dados

Arquiteto de Soluções
Arquiteto de Dados

Adm. Banco de Dados

Engenheiro de Dados

Analista de Dados
Cientista de Dados
Engenheiro de IA

Próxima Aula

- Plataforma de Dados do Azure

Soluções de Dados, Big Data e Machine Learning

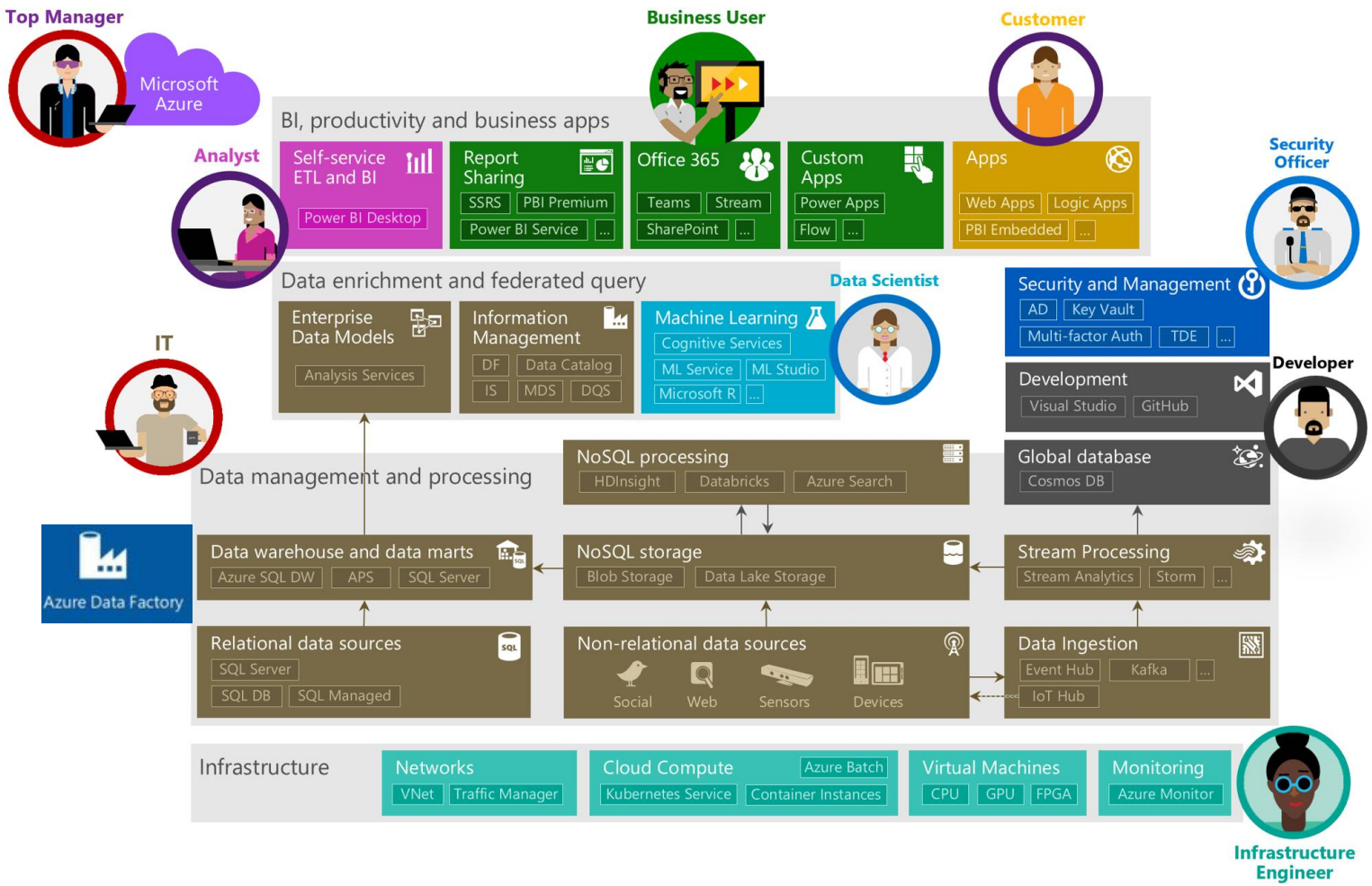
AULA 1.4. PLATAFORMA DE DADOS DO AZURE

PROF. GUSTAVO AGUILAR

Nesta Aula

- Overview da Plataforma de Dados do Azure

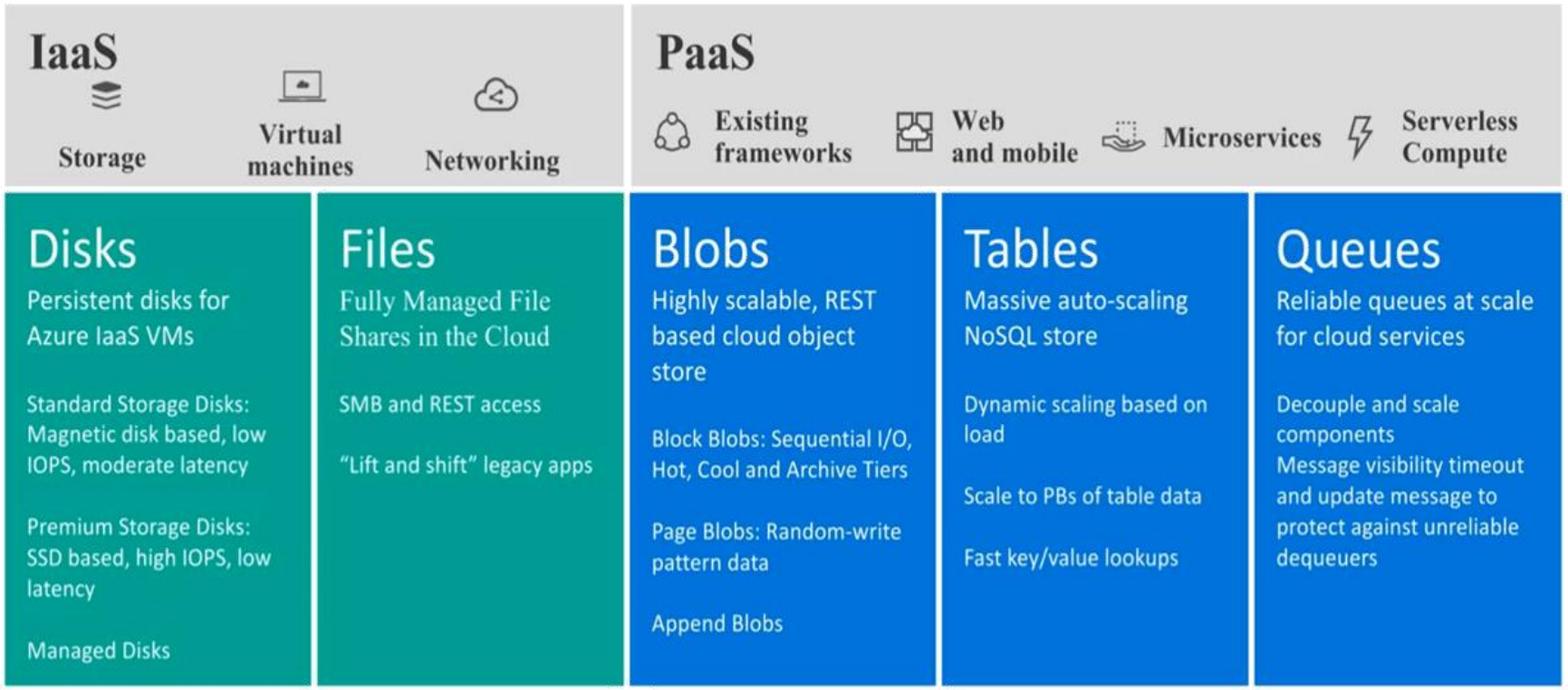
Plataforma de Dados do Azure



Azure Storage

- Tipo de armazenamento básico usado no Microsoft Azure;
- Armazenamento de arquivos, dados textuais / binários / não estruturados e de fila de mensagens;
- Pode ser usado com máquinas virtuais, via SMB ou API;
- Cada serviço é acessado através de uma conta de armazenamento (Storage Account);
- Possuem cinco opções:
  - Azure Blobs;
  - Azure Files;
  - Azure Queues;
  - Azure Tables;
  - Azure Disks.

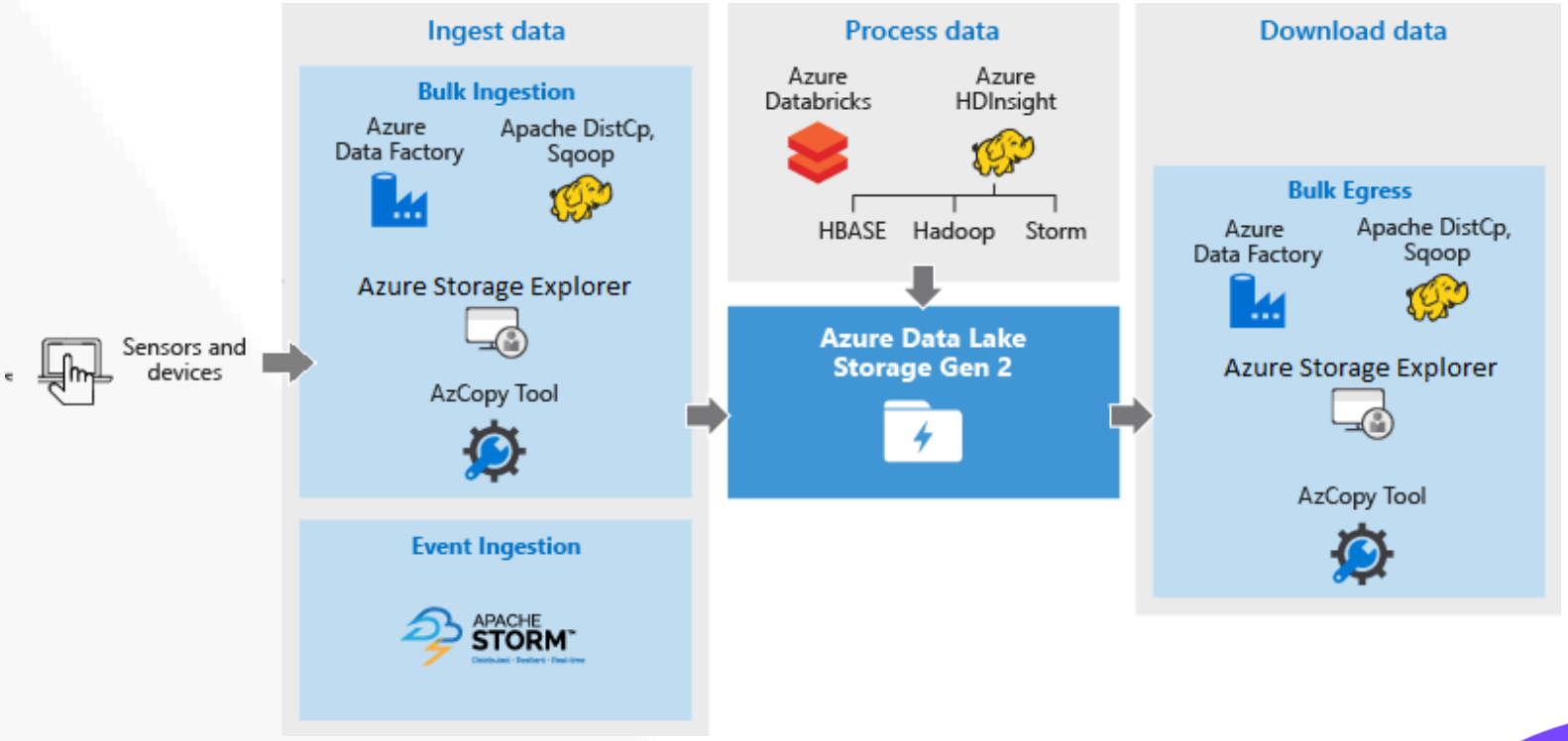
Azure Storage



Azure Data Lake Storage

- Repositório de dados compatível com HDFS
  - HDFS (Hadoop Distributed File Sysyem) ➔ Sistema de Arquivos Distribuído;
  - Pode armazenar qualquer tamanho ou tipo de dados.
- Geração 1 (Gen1) e Geração 2 (Gen2)
  - Gen2 combina os serviços de armazenamento do Gen1 com os benefícios do Azure Blob Storage;
  - Desempenho ajustado para o processamento de soluções de análise de big data.

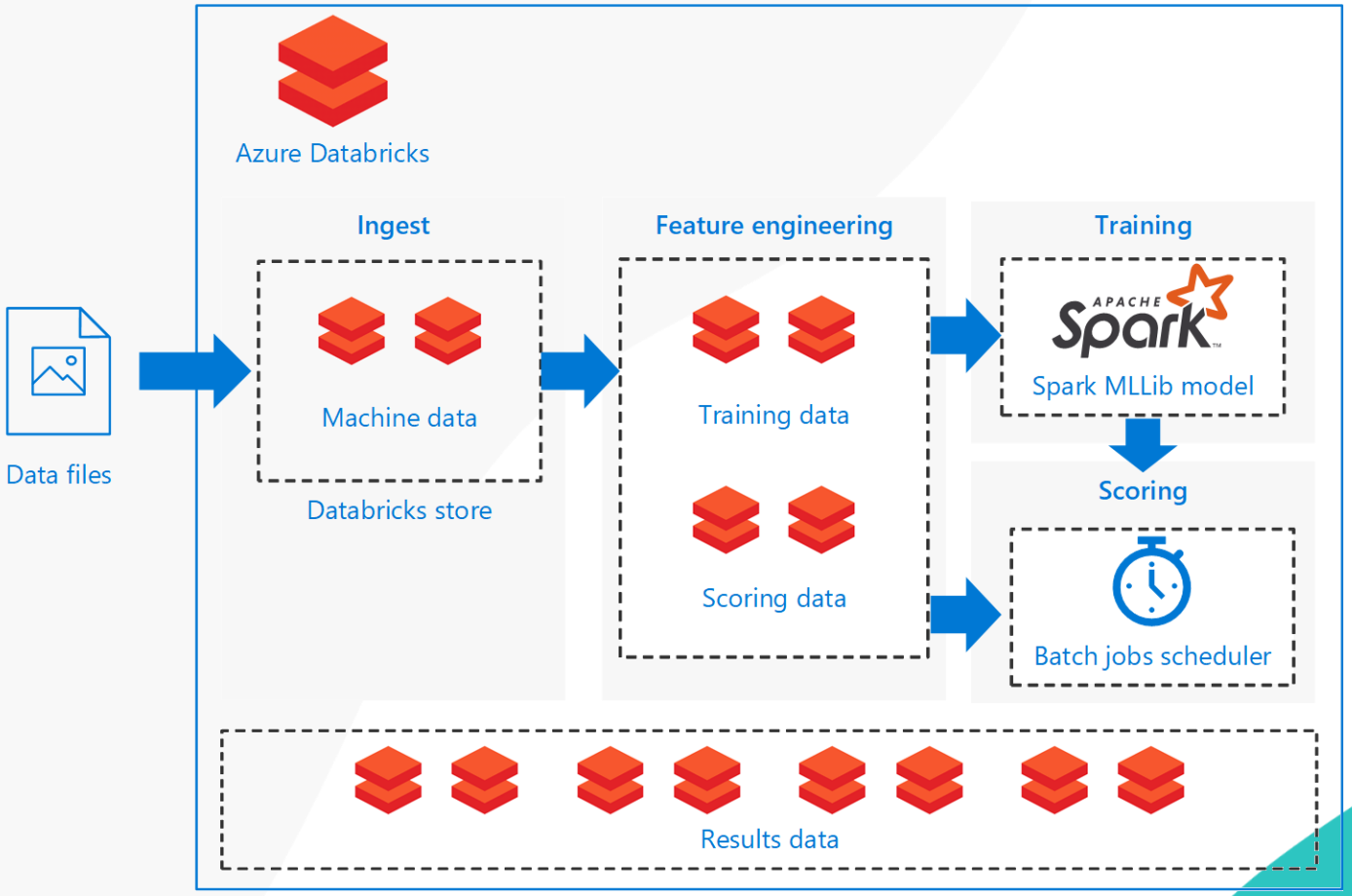
Azure Data Lake Storage



Azure Databricks

- O Databricks é uma versão do popular mecanismo de análise e processamento de dados Apache Spark;
- O Azure Databricks é a versão totalmente gerenciada do Databricks;
- Oferece uma plataforma de Big Data e Aprendizado de Máquina;
- Suporte a Python, Scala, R, Java e SQL, além de estruturas e bibliotecas de ciência de dados, incluindo TensorFlow, PyTorch e Scikit-learn.

Azure Databricks



Azure Cosmos DB

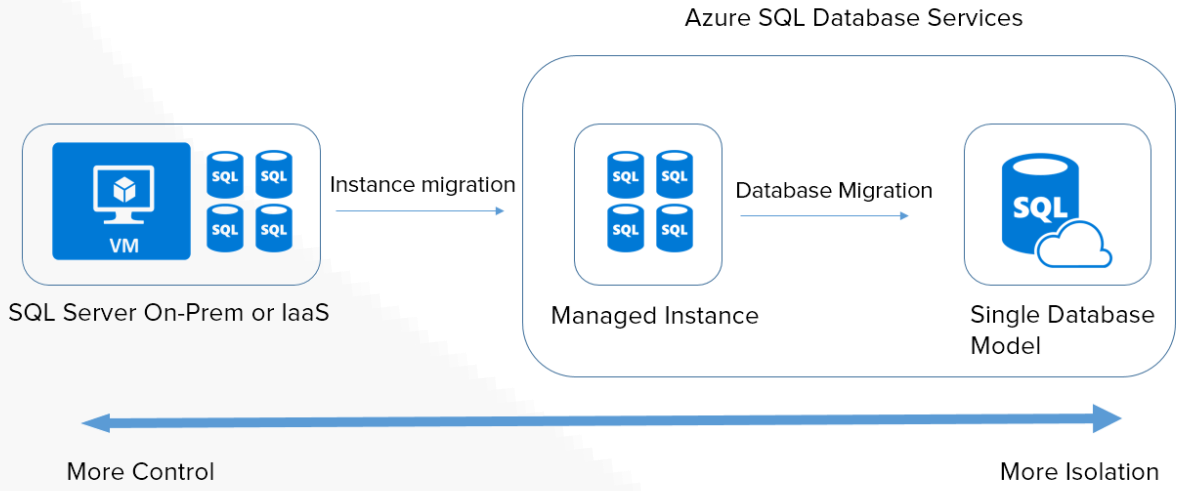
- Serviço de banco de dados multimodelo e distribuídos globalmente;
- Database as a Service ➔ PaaS;
- Suporte a 5 APIs (modelos de DB Engine)
  - SQL (relacional);
  - Mongo DB (documento)
  - Cassandra (colunar)
  - Gremlin (grafo)
  - Table

Azure SQL Database

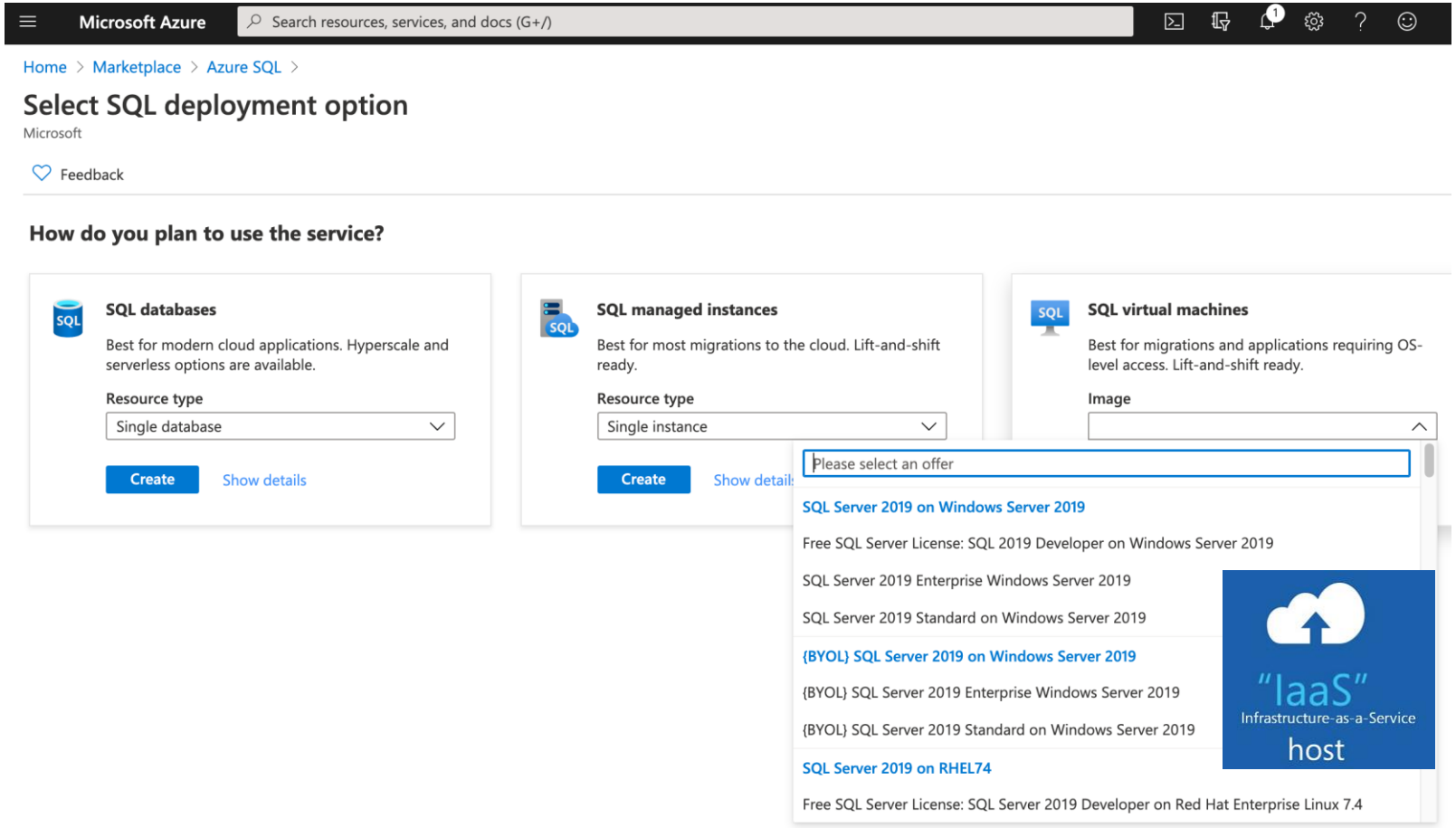
- Serviço de banco de dados relacional gerenciado no Azure;
- Database as a Service ➔ PaaS;
- Suporta estruturas como dados relacionais e formatos não estruturados, como dados espaciais e XML.

Azure SQL Managed Instance

- Fornece uma instância inteira do SQL Server ➔ PaaS
  - Até 100 bancos de dados;
- Fornece outros recursos que não estão disponíveis no Azure SQL Database:
  - Consultas entre bancos de dados;
  - CLR (Common Language Runtime);
  - SQL Agent (jobs e scheduler).

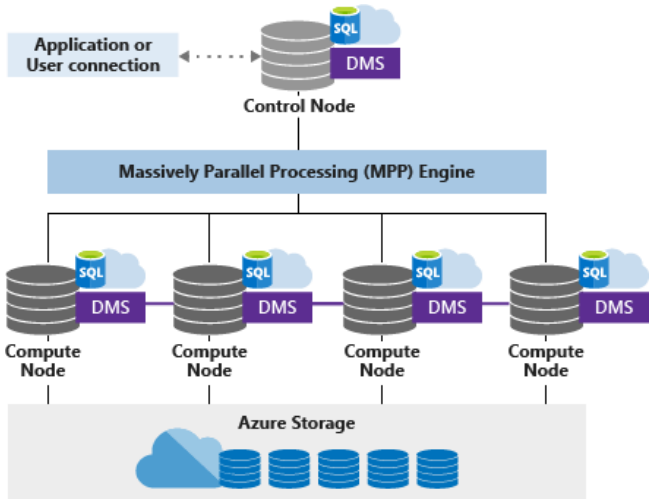


SQL Virtual Machines



Azure Synapse Analytics

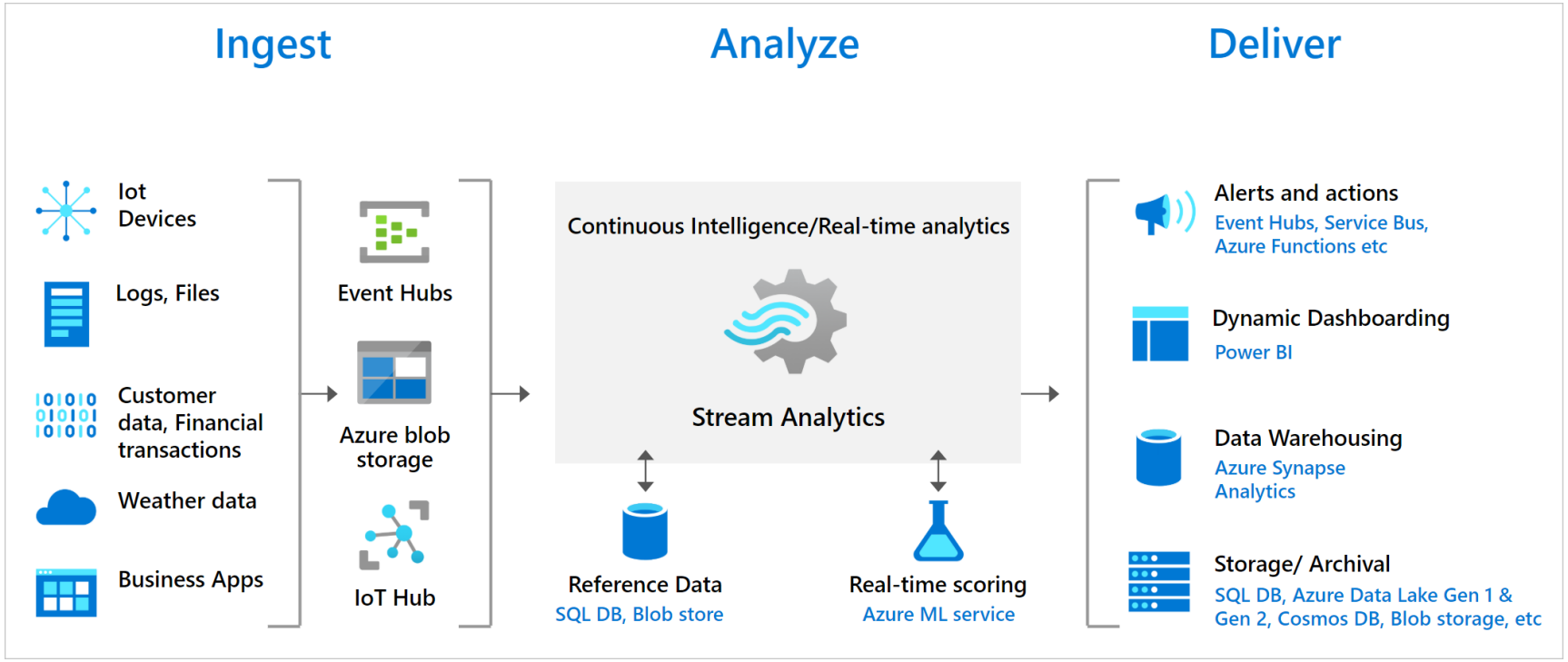
- Anteriormente conhecido como SQL DW;
- Fornece um ambiente unificado combinando data warehouse e os recursos de análise de Big Data do Spark;
- Azure Synapse tem quatro componentes:
  - Synapse SQL: análise completa baseada em T-SQL;
    - Pool de SQL: pagamento por DWU provisionado;
    - SQL sob demanda: pagamento por TB processado.
  - Apache Spark;
  - Data Integration: Integração de dados híbridos;
  - Studio: experiência de usuário unificada.
- MPP ➔ processamento massivo paralelo.



Stream Analytics

- Mecanismo de análise em tempo real e de processamento de eventos;
- Analisar e processar grandes volumes de streaming de dados de várias fontes simultaneamente;
- Padrões e relacionamentos podem ser identificados;
- Respostas a anomalias de dados em tempo real;
  - Monitoramento de Internet das Coisas (IoT) / Gadgets;
  - Logs da WEB;
  - Monitoramento remoto de pacientes;
  - Acompanhamento de sistemas de ponto de venda (PDV).

Stream Analytics



Próxima Aula

- Outros Serviços da Plataforma de Dados do Azure

Soluções de Dados, Big Data e Machine Learning

AULA 1.5. OUTROS SERVIÇOS DA PLATAFORMA DE DADOS DO AZURE

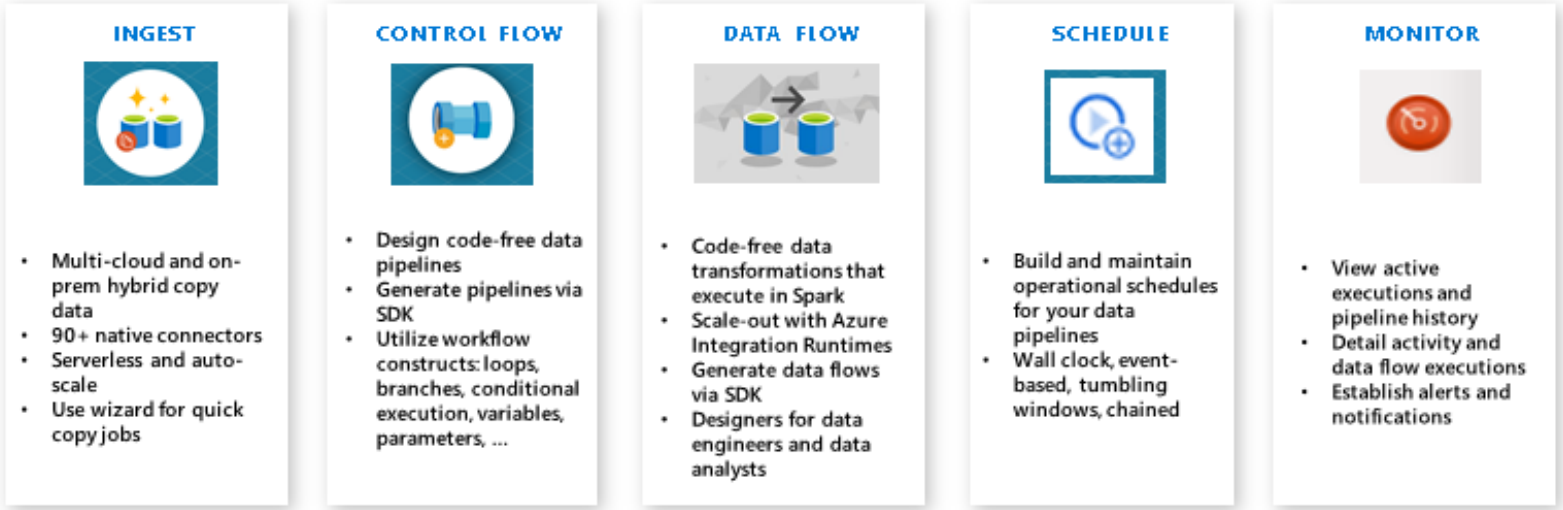
PROF. GUSTAVO AGUILAR

Nesta Aula

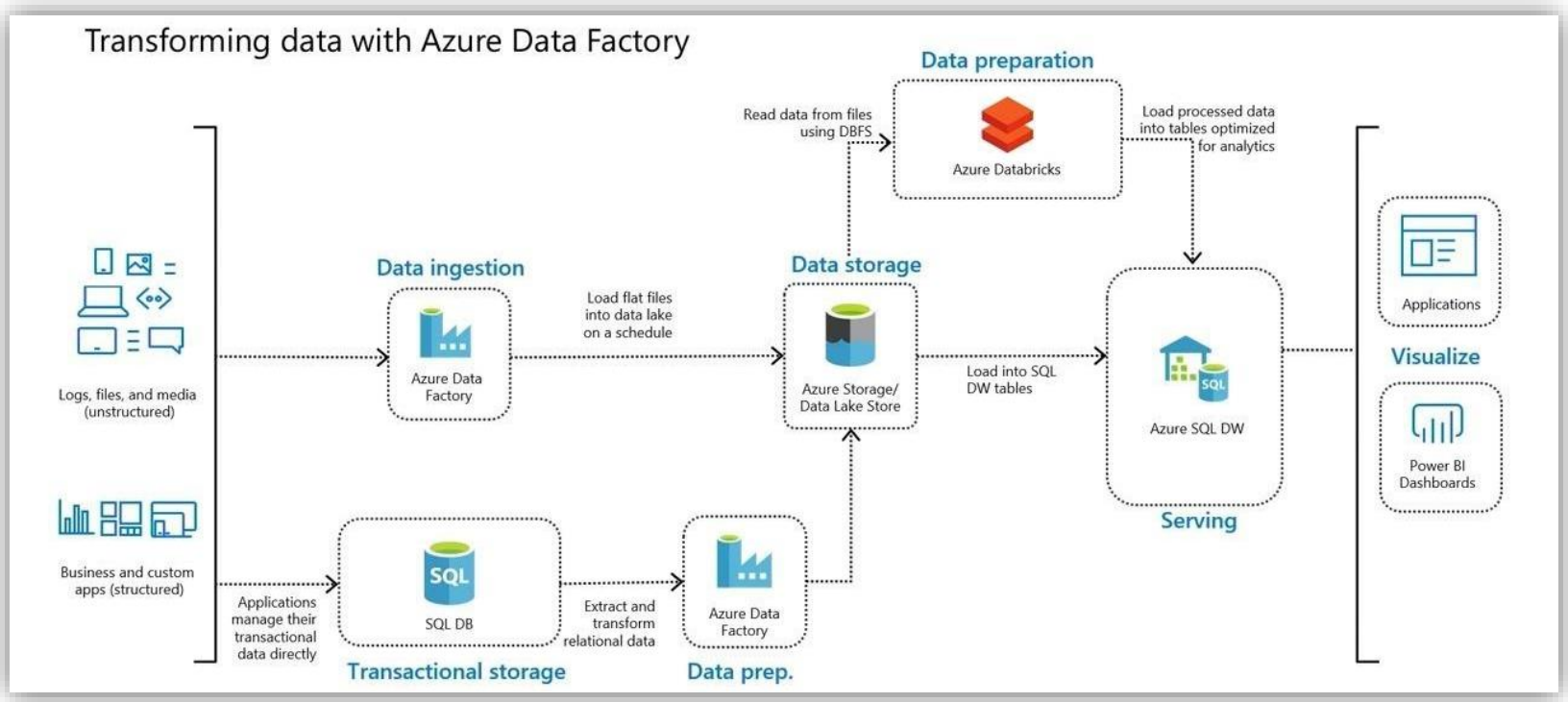
- Azure Data Factory
- Azure HDInsight
- Azure Data Catalog

Azure Data Factory

- Serviço de integração de dados que orquestra a movimentação de dados entre várias fontes de dados;
- ETL baseado em nuvem
  - Similar ao que o Integration Services (SSIS) faz no on premises.



Azure Data Factory



Azure HDInsight

- Processamento e análise de big data para dar suporte ao processamento em lote, de data warehousing, IoT e Data Science;
- Solução em nuvem de baixo custo que contém várias tecnologias:

Apache Hadoop

Apache Spark

Apache Kafka

Apache HBase

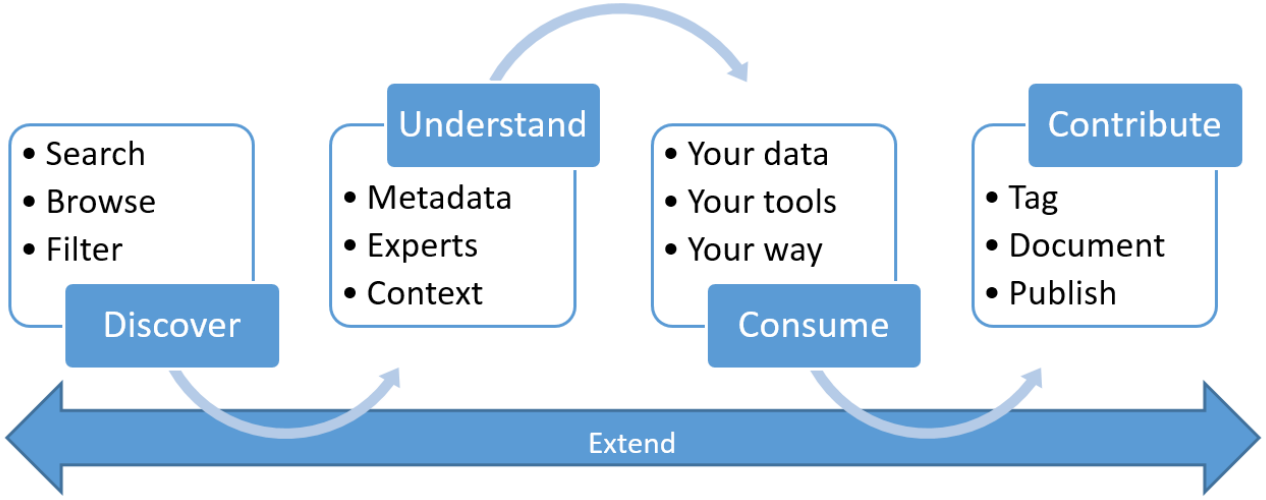
Apache Hive LLAP

Apache Storm

Machine Learning

Azure Data Catalog

- Serviço para catalogar dados e fontes de dados;
- Inclui um modelo de alimentação de metadados;
- Os usuários podem descobrir as fontes de dados de que precisam e entender as fontes de dados que encontram, além de usar o Catálogo de Dados para documentar as informações sobre suas fontes de dados.



Próxima Aula

- Capítulo 2. Armazenamento de Dados

Soluções de Dados, Big Data e Machine Learning

Capítulo 2. Armazenamento de Dados

PROF. GUSTAVO AGUILAR

Soluções de Dados, Big Data e Machine Learning

AULA 2.1. STORAGE ACCOUNT

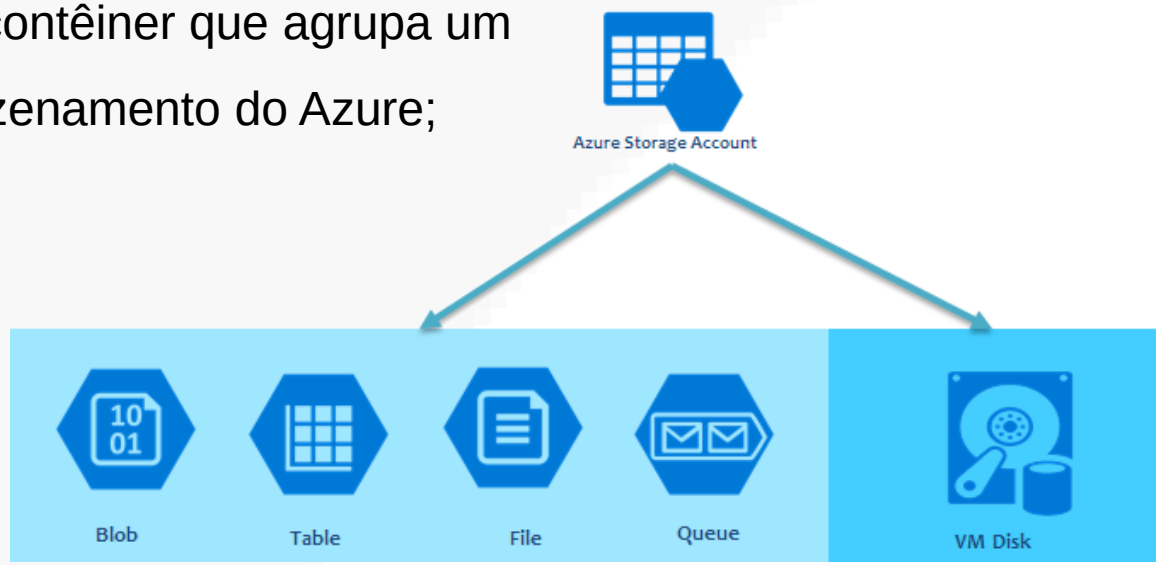
PROF. GUSTAVO AGUILAR

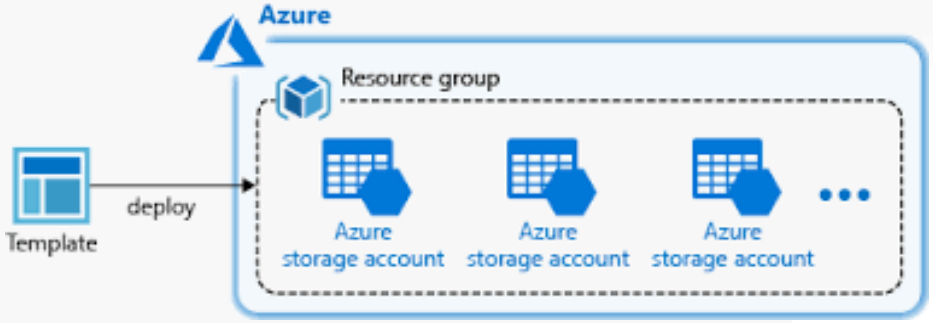
Nesta Aula

- ❑ Storage Account
- ❑ Configurações
- ❑ Ferramentas para Criação

Storage Account

- **Conta de Armazenamento:** contêiner que agrupa um conjunto de serviços de armazenamento do Azure;
- Recurso do Azure, de uma assinatura (subscription), que está incluída em um grupo de recursos (Resource Manager).





Configurações

- Uma storage account define uma política que se aplica a todos os serviços de armazenamento na conta;
- Exemplo:
  - Datacenter Brazil South;
  - Acessíveis apenas por https;
  - Cobrados na assinatura do departamento de TI.
- **Tipo de conta:** conjunto de políticas que determinam quais serviços de dados você pode incluir na conta e os preços desses serviços.
  - **StorageV2 (general purpose v2):** oferta atual que suporta todos os tipos de armazenamento e todos os recursos mais recentes;
  - **Storage (general purpose v1)**: legado; suporta todos os tipos de armazenamento, mas pode não ter todos os recursos do V2;
  - **Blob Storage**: legado; permite apenas block blobs e append blobs, não suportando file share, tables e queues.

Configurações

- **Assinatura** e **Localização;**

- **Desempenho**: determina os serviços de dados e o tipo de disco;
  - **Standard**: qualquer serviço de dados (Blob, Arquivo, Fila, Tabela) e usa unidades de disco magnético;
  - **Premium:** tipo específico de blob chamado Page Blob e usa disco SSD.

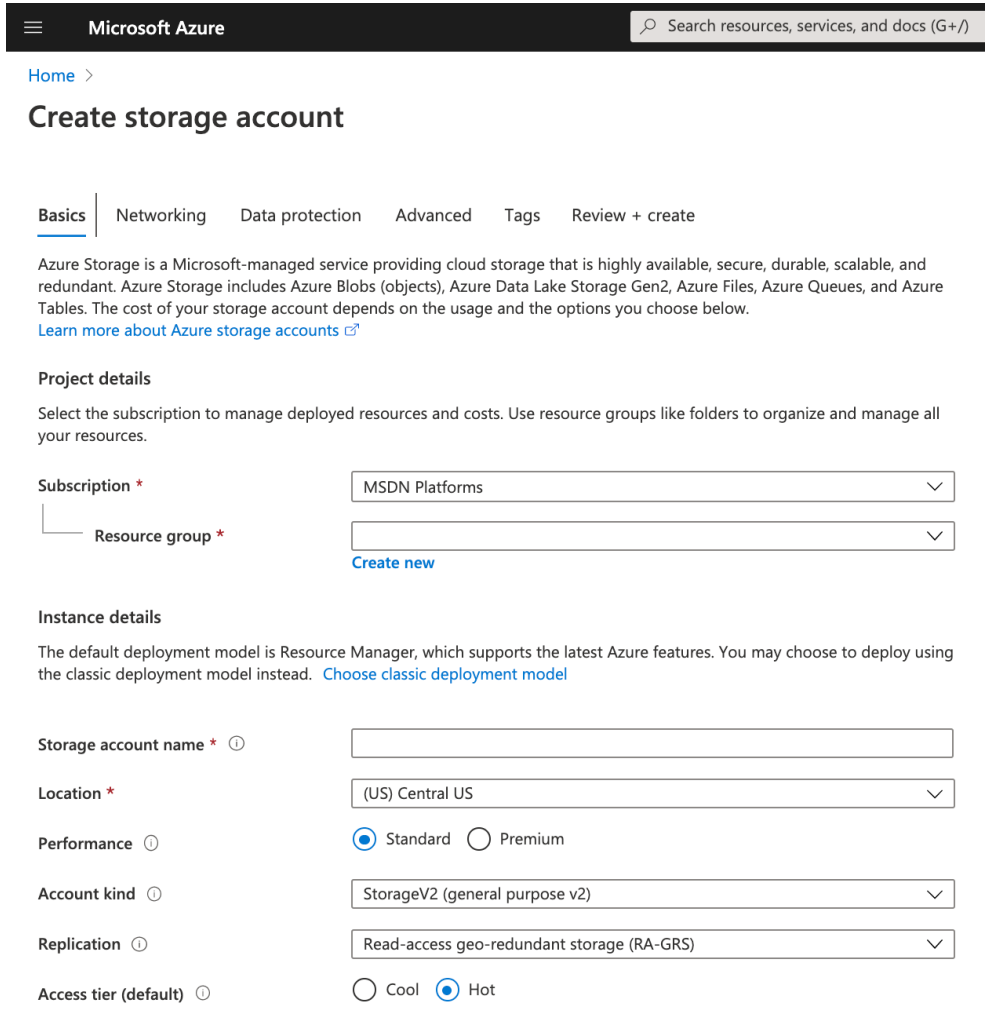
- **Replicação**: cópia dos dados para proteger contra falhas de hardware ou desastres
  - No mínimo: **armazenamento localmente redundante (LRS)**
    - Protege contra falhas de hardware;
    - Não protege de um evento que indisponibilizada todo o datacenter.
  - **Armazenamento georredundante (GRS):** replicação em diferentes datacenters em todo o mundo.

Configurações

- **Camada de acesso**: velocidade que você poderá acessar os blobs
    - **Hot** fornece acesso mais rápido que **Cool**, mas a um custo maior.
- **Redes virtuais**: método de conectividade (público / privado).
- **Proteção dos Dados**
    - Modelo de “lixeira” com expurgo automático ➔ blob e fileshare;
    - Opção de versionamento.
- **Transferência segura necessária (TLS)**: ativado requer HTTPs, enquanto desativado permite HTTP.

Ferramentas para Criação

- Portal do Azure;
- Azure CLI
  - Command-line interface.
- Azure PowerShell;
- Management client libraries
  - Incorporar a criação dentro de um app.



Próxima Aula

- Criando uma Storage Account

Soluções de Dados, Big Data e Machine Learning

AULA 2.2. CRIANDO UMA STORAGE ACCOUNT

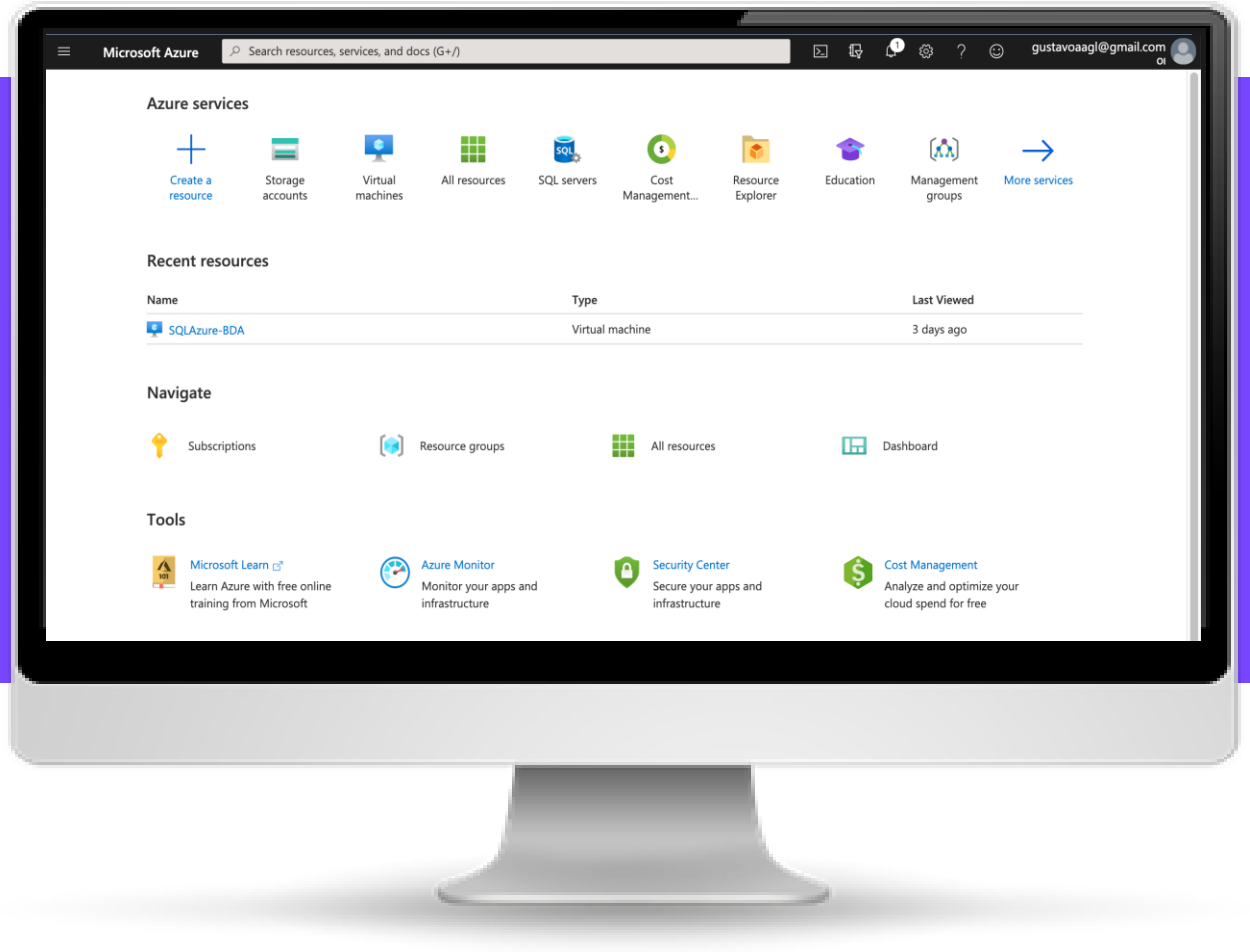
PROF. GUSTAVO AGUILAR

Nesta Aula

- Criando uma Storage Account

Criando uma Storage Account

Demo



Próxima Aula

- Ingestão de Dados

Soluções de Dados, Big Data e Machine Learning

AULA 2.3. INGESTÃO DE DADOS

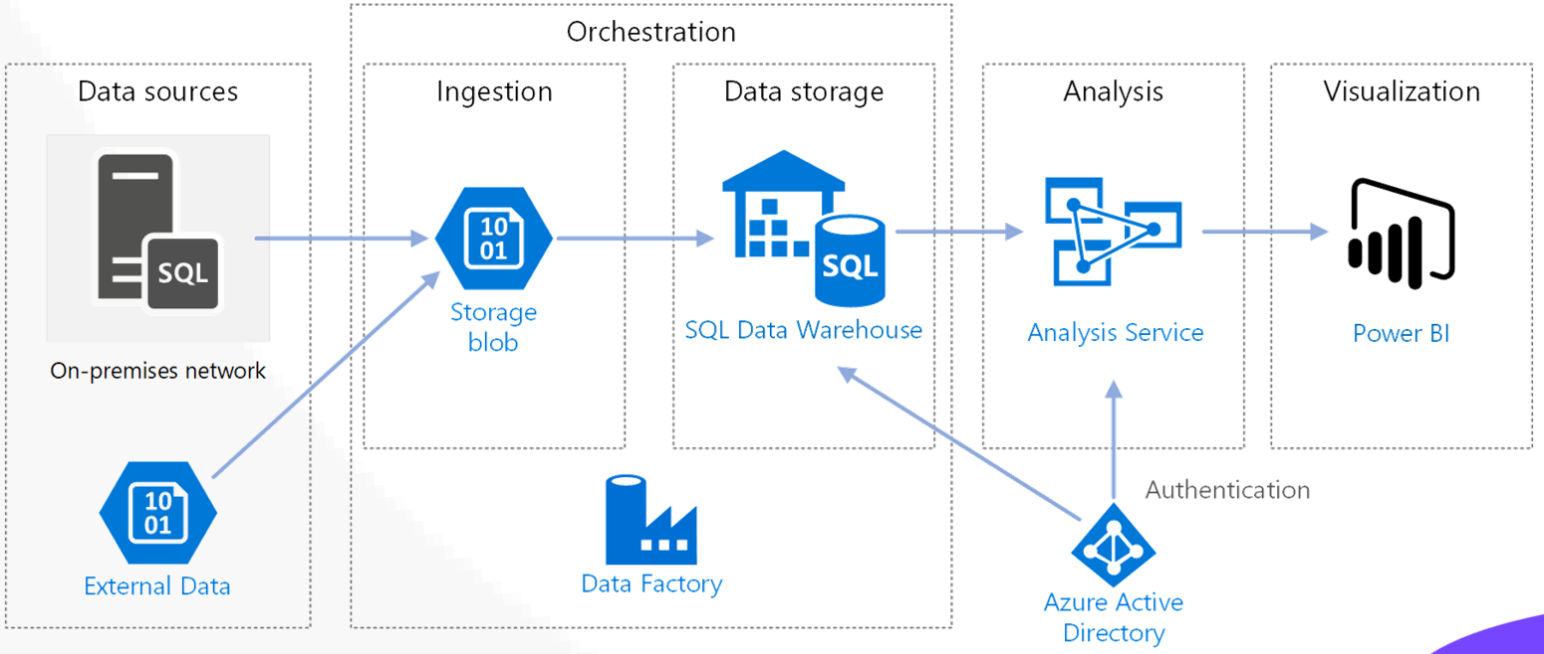
PROF. GUSTAVO AGUILAR

Nesta Aula

- Ingestão de Dados
- Azure Storage Explorer

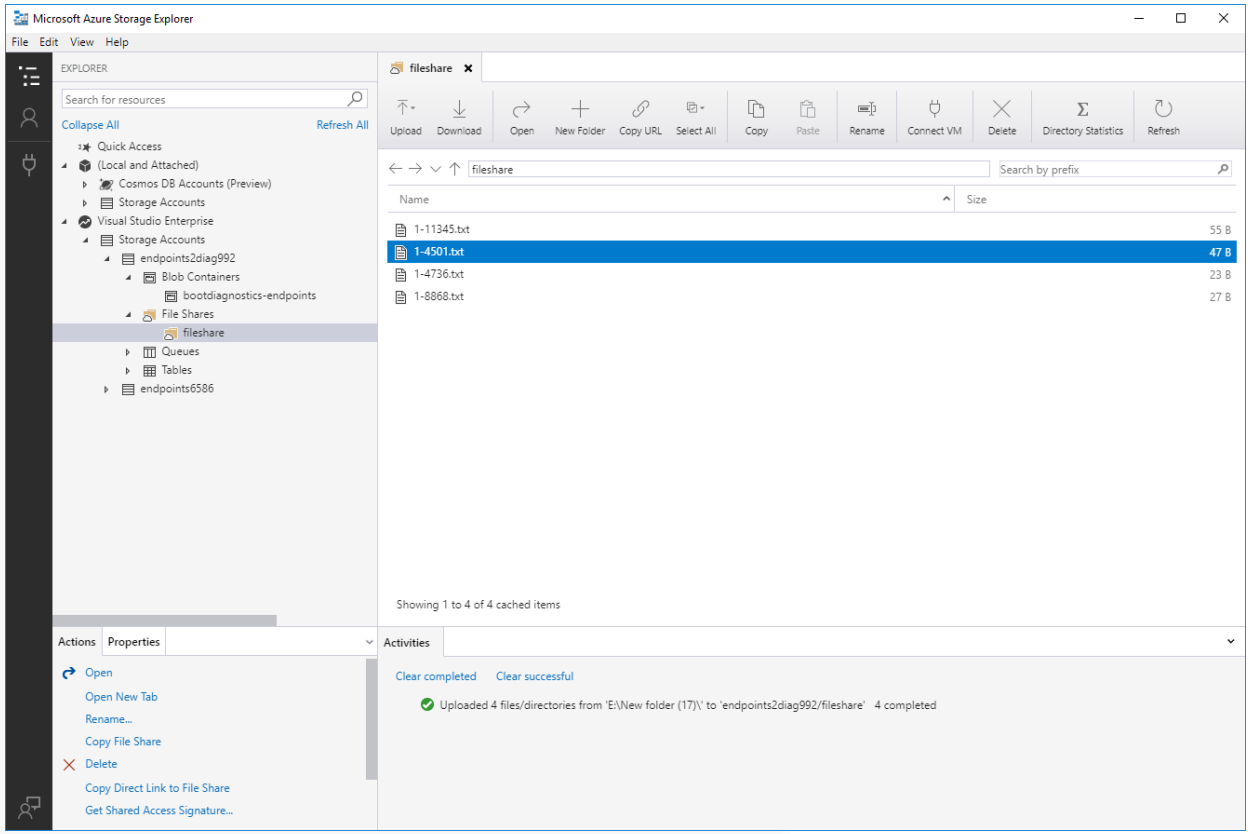
Ingestão de Dados

- **Data Factory**



Ingestão de Dados

- **Azure Storage Explorer:** www.storageexplorer.com

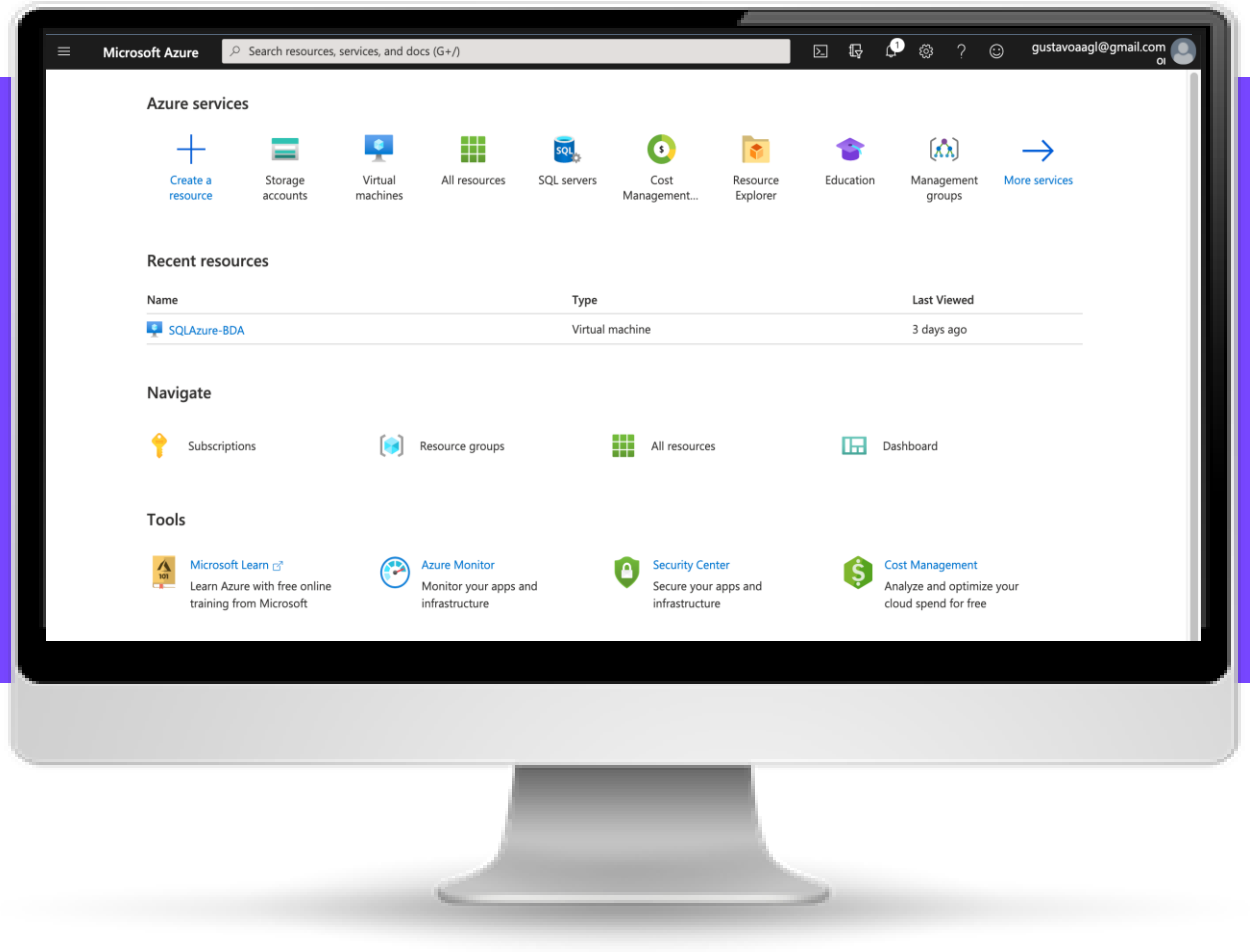


Ingestão de Dados

- **AzCopy;**
- **PowerShell;**
- **Visual Studio;**
- Arquivos acima de 2 GB ➔ PowerShell ou Visual Studio;
- AzCopy ➔ tamanho máximo de arquivo de 1 TB
    - Automaticamente dividido em vários arquivos se o arquivo exceder 200 GB.

Storage Explorer

Demo



Próxima Aula

- Criando um Data Lake Storage Gen2

Soluções de Dados, Big Data e Machine Learning

AULA 2.4. CRIANDO UM DATA LAKE STORAGE GEN2

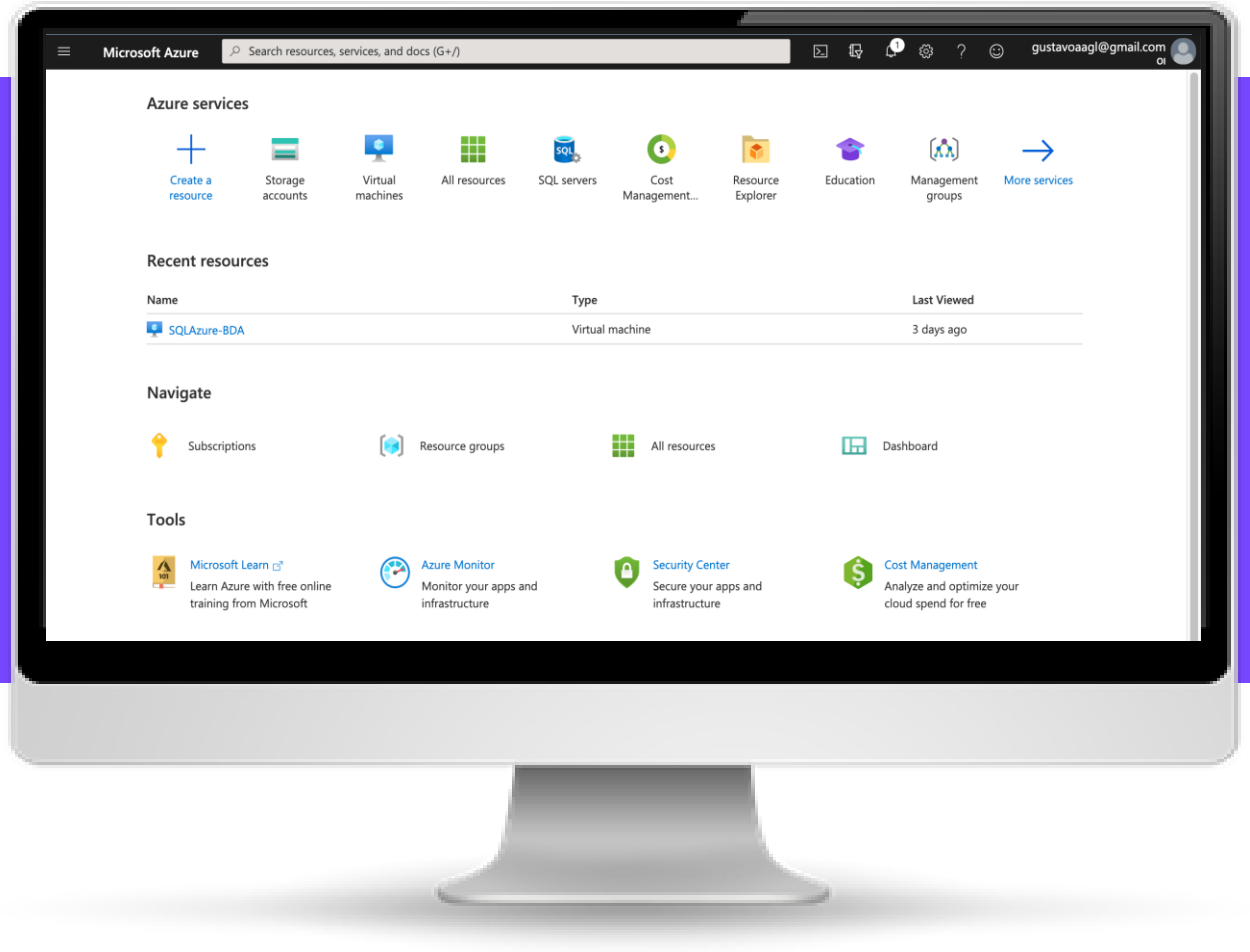
PROF. GUSTAVO AGUILAR

Nesta Aula

- Criando um Data Lake Storage Gen2
- Usando um Data Lake Storage Gen2

Criando e Usando um Data Lake Storage Gen2

Demo



Próxima Aula

- Capítulo 3. Armazenamento de Dados Relacionais no Azure

Soluções de Dados, Big Data e Machine Learning

Capítulo 3. Armazenamento de Dados Relacionais no Azure

PROF. GUSTAVO AGUILAR

Soluções de Dados, Big Data e Machine Learning

AULA 3.1. BANCOS DE DADOS RELACIONAIS EM IAAS

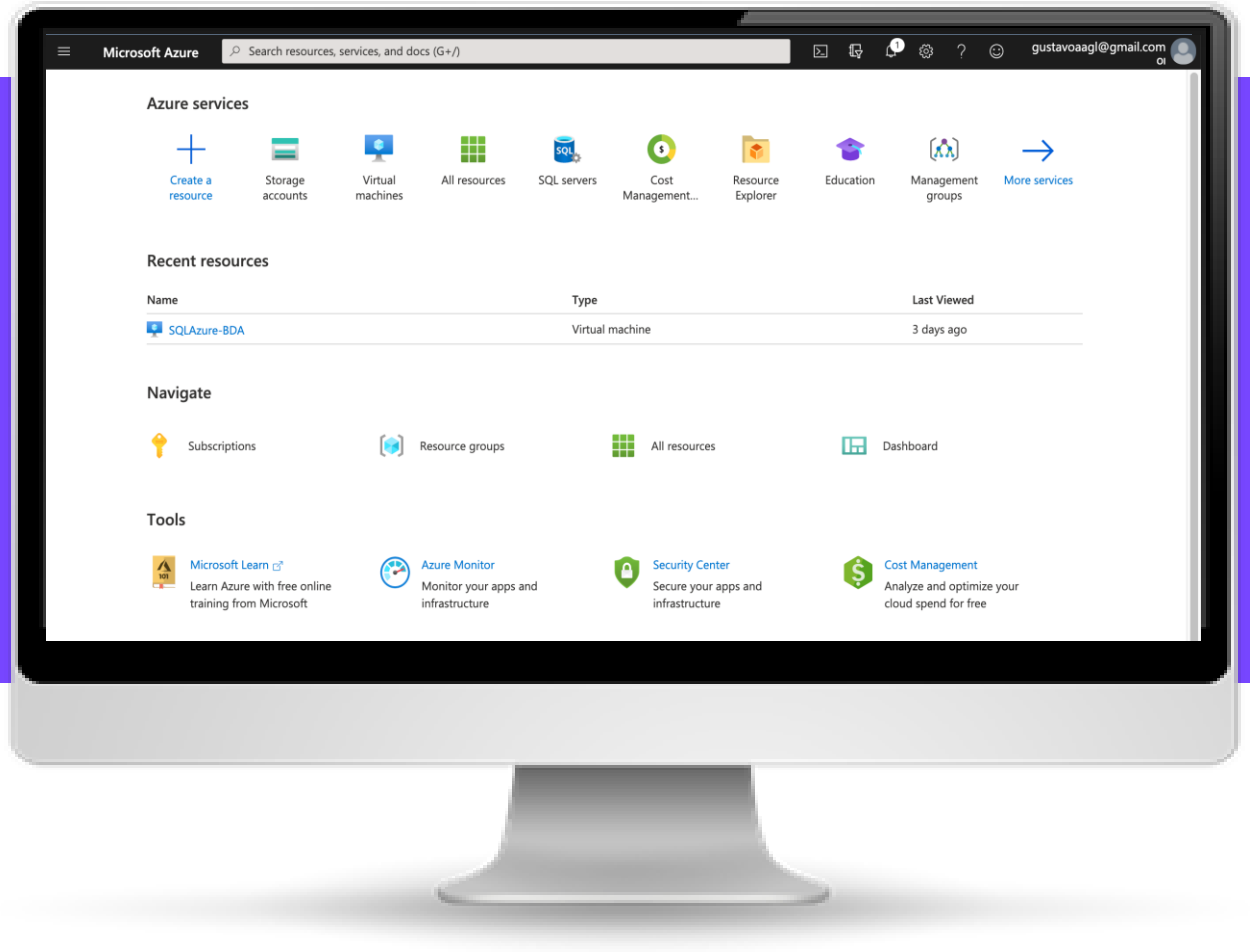
PROF. GUSTAVO AGUILAR

Nesta Aula

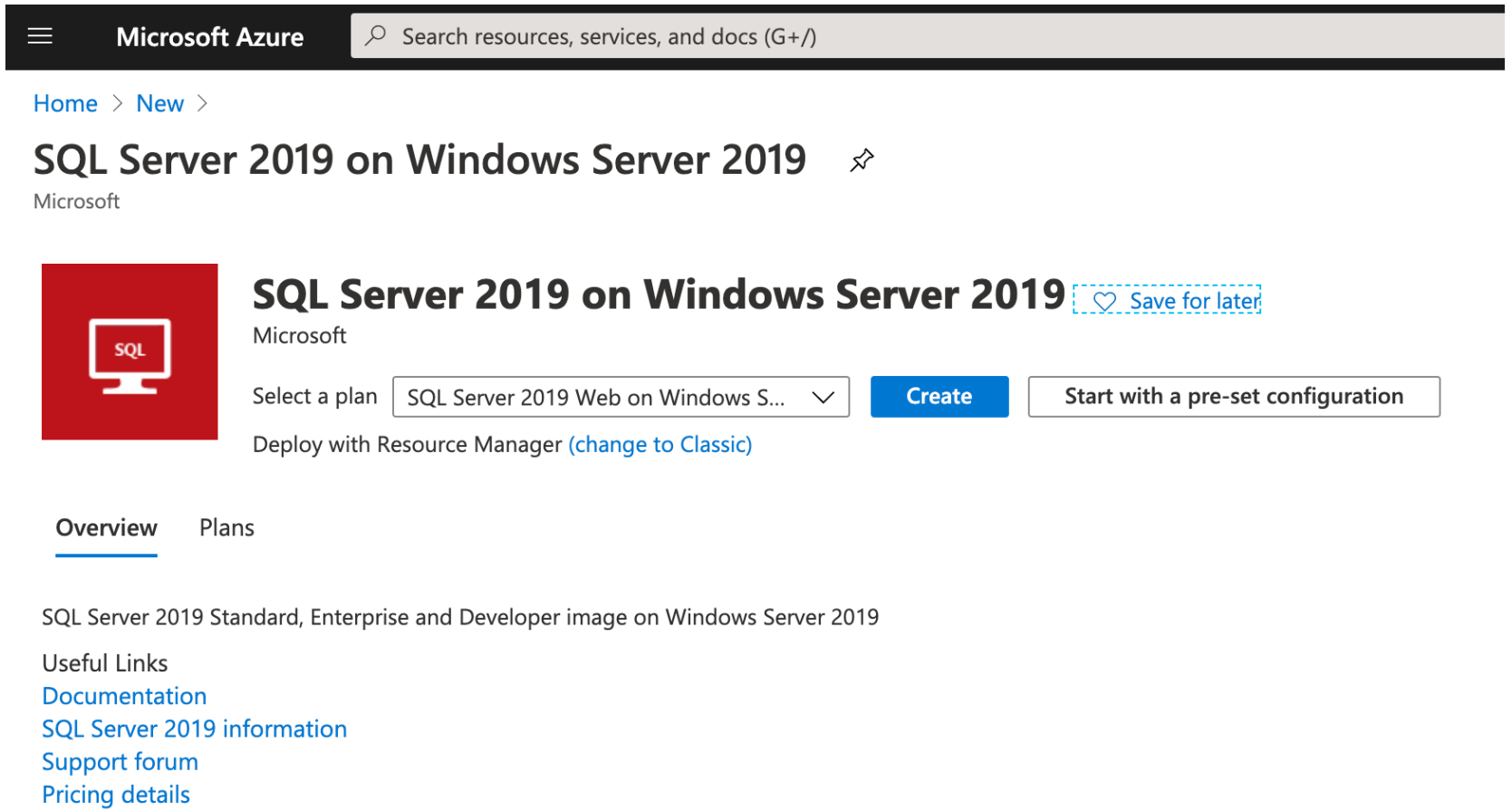
- Bancos de Dados Relacionais em IaaS
- SQL Server do Azure Marketplace

Bancos de Dados Relacionais em IaaS

Demo



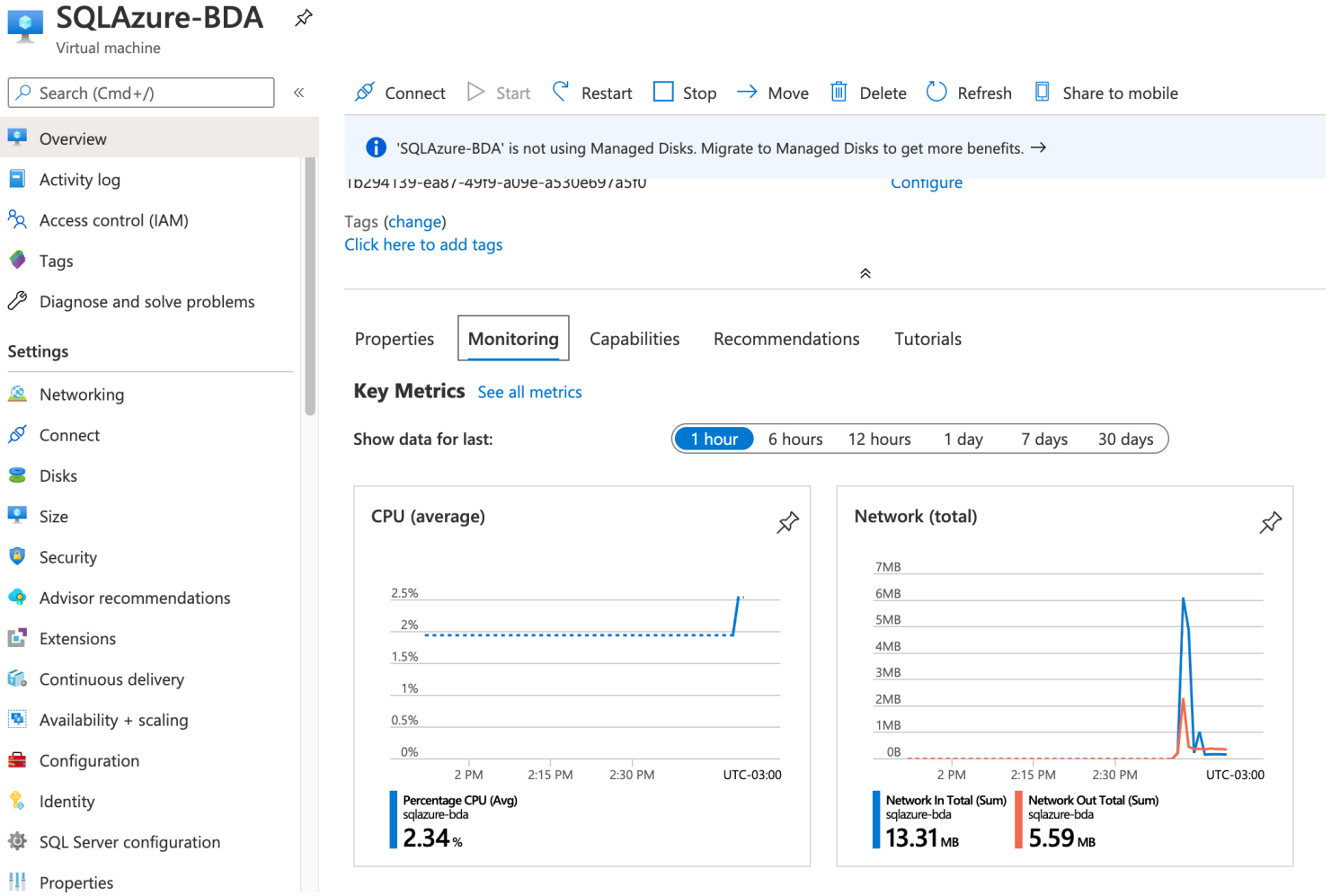
SQL Server do Azure Marketplace



SQL Server do Azure Marketplace

- Quando aprovisiona uma VM do Marketplace com SQL Server: parte do processo instala o **SQL Server IaaS Agent Extension:**
  - Extensões são códigos executados na pós-implantação da VM;
  - Instalação de antivírus ou instalação de um recurso do Windows;
  - Três recursos principais que podem reduzir a sobrecarga administrative:
    - Backup automatizado do SQL Server;
    - Patches automatizados do SQL Server;
    - Integração com Azure Key Vault.
  - Informações sobre a configuração e utilização do SQL Server.

SQL Server do Azure Marketplace



Próxima Aula

- Azure SQL Database Managed Instance

Soluções de Dados, Big Data e Machine Learning

AULA 3.2. AZURE SQL DATABASE MANAGED INSTANCE

PROF. GUSTAVO AGUILAR

Nesta Aula

- ❑ Ofertas de Instâncias Gerenciada
- ❑ Recursos da Instância Gerenciada
- ❑ Criando uma SQL Managed Instance
- ❑ Usando uma SQL Managed Instance

Ofertas de Instâncias Gerenciada

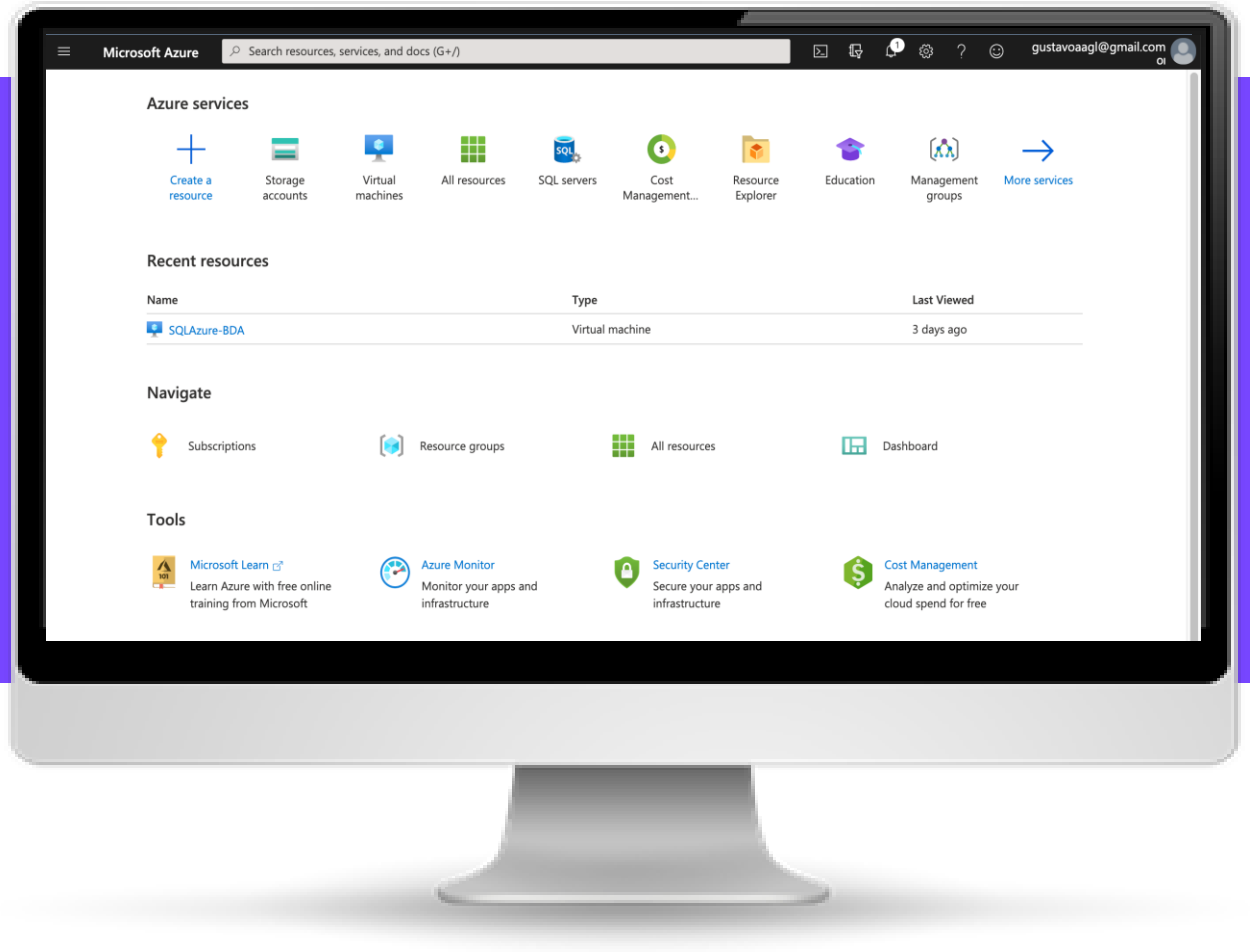
- **General Purpose**
- **Business Critical**
  - Maior desempenho e disponibilidade;
  - Suporta recursos In-Memory OLTP do SQL;
  - Leitura em réplicas secundárias;
  - Mais memória RAM por vCore;
  - Usa storage atachado diretamente (sem NAS)
    - Menor latência nas operações de I/O.
- Uso das licenças de SQL já compradas para reduzir o custo.

Recursos da Instância Gerenciada

- 99.99% de disponibilidade;
- Atualizações automáticas de patches Windows / SQL;
- Backup gerenciado pelo Azure
  - Possível gerar um backup manual “copy only” para o Azure Storage.
- Automatic tuning:
  - Identificar queries onerosas;
  - Forçar o último plano execução bom;
  - Adicionar índices;
  - Remover índices.

Criando e Usando uma SQL Managed Instance

Demo



Próxima Aula

- Azure SQL Database e Cosmos DB

Soluções de Dados, Big Data e Machine Learning

AULA 3.3. AZURE SQL DATABASE

PROF. GUSTAVO AGUILAR

Nesta Aula

- Azure SQL Database
- Criando um Azure SQL Database
- Usando um Azure SQL Database

Azure SQL Database

- Possui 2 modelos de implantação;
- **Single Database**
  - Banco de dados isolado;
  - Totalmente gerenciado pelo Azure.
- **Elastic Pool**
  - Coleção de single databases;
  - Com um conjunto compartilhado de recursos (CPU / RAM).

Azure SQL Database

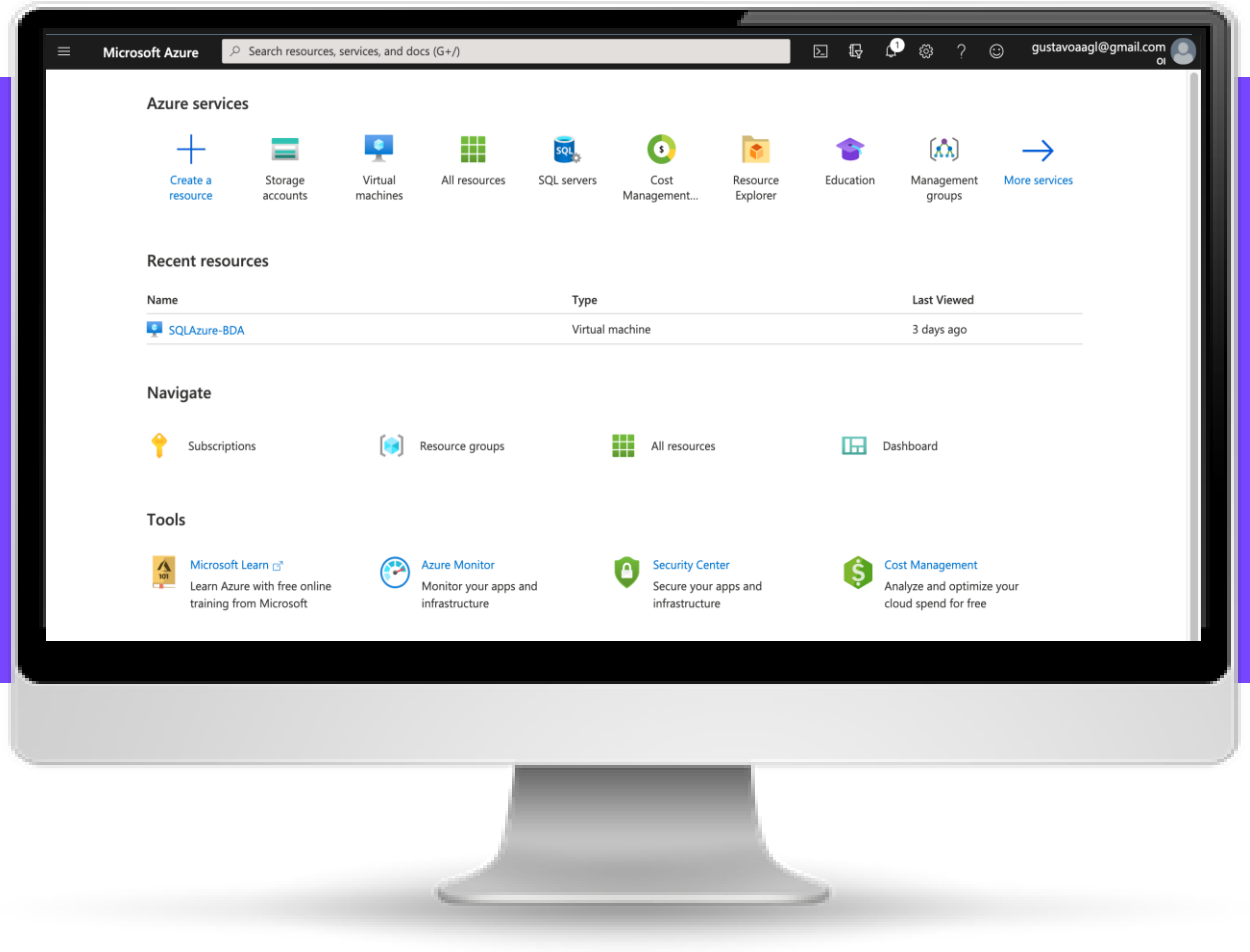
- Modelos de contratação
  - **Baseado em vCore**: permite selecionar a quantidade de CPU, memória RAM e velocidade do storage.
  - **Baseado em DTU (Database Transaction Units):** combinação de recursos de computação, memória e I/O em três camadas de serviço, para suportar cargas de trabalho de banco de dados leves a pesadas.
  - **Serverless:**
    - Dimensiona automaticamente os recursos com base na demanda da carga de trabalho e cobra pela quantidade de recursos usados por segundo.
    - Pausa automaticamente os bancos de dados durante os períodos inativos ➔ cobra o storage.

Azure SQL Database

- Disponível em 3 camadas de serviço (service tiers);
- **General Purpose / Standard**: para cargas de trabalho comuns.
- **Business Critical / Premium:**
  - Para aplicações OLTP com alta taxa de transações;
  - Cenários que necessitam de baixa latência de I/O;
  - Oferece a maior resiliência a falhas usando várias réplicas isoladas.
- **Hyperscale:**
  - Para bancos de dados OLTP muito grandes;
  - Cenários que necessitem de dimensionamento automático de armazenamento.

Criando e Usando um Azure SQL Database

Demo



Próxima Aula

- Azure Cosmos DB

Soluções de Dados, Big Data e Machine Learning

AULA 3.4. **AZURE COSMOS DB**

PROF. GUSTAVO AGUILAR

Nesta Aula

- ❑ Pré-requisitos para Criação
- ❑ Criando um Azure Cosmos DB (SQL)
- ❑ Usando um Azure Cosmos DB (SQL)

Pré-requisitos para Criação

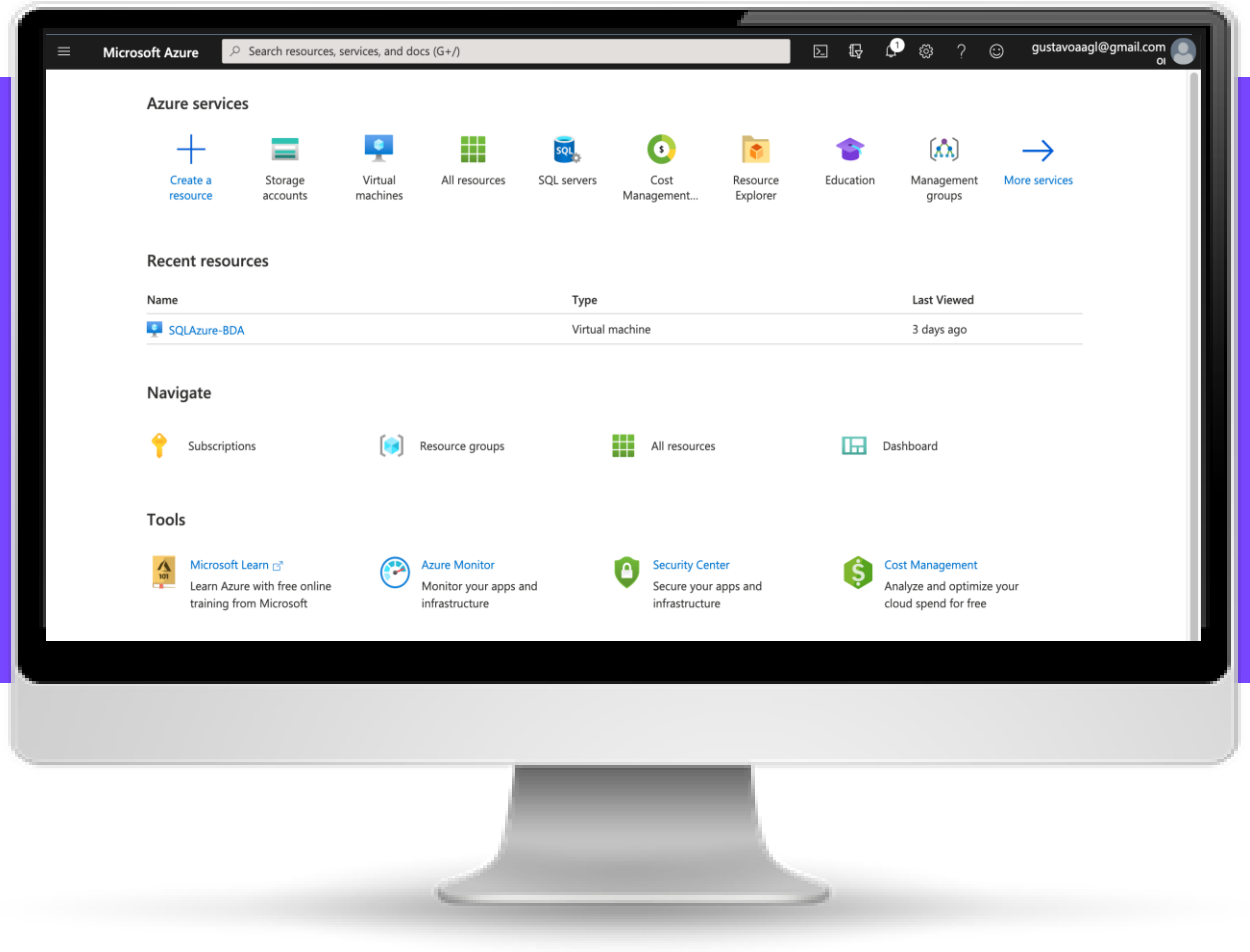
- **Azure Cosmos DB Account**
  - Recurso que atua como uma entidade organizacional;
  - Cada conta está associada a um dos vários modelos de dados aos quais o Azure Cosmos DB oferece suporte;
  - Quantas contas precisar.

Pré-requisitos para Criação

- **Request Unit (RU)**
  - Medida da taxa de transferência (throughput) por segundo;
  - Reservar o número de RU/s que deseja que o Azure Cosmos DB provisione com antecedência, para que ele possa lidar com a carga estimada e você possa aumentar ou diminuir sua RU/s a qualquer .

Item size	Reads/second	Writes/second	Request units
1 KB	500	100	(500 * 1) + (100 * 5) = 1,000 RU/s
1 KB	500	500	(500 * 1) + (500 * 5) = 3,000 RU/s
4 KB	500	100	(500 * 1.3) + (100 * 7) = 1,350 RU/s
4 KB	500	500	(500 * 1.3) + (500 * 7) = 4,150 RU/s
64 KB	500	100	(500 * 10) + (100 * 48) = 9,800 RU/s
64 KB	500	500	(500 * 10) + (500 * 48) = 29,000 RU/s

Criando e Usando um Azure Cosmos DB (SQL)



Demo

Próxima Aula

- Bancos de Dados Open Source no Azure

Soluções de Dados, Big Data e Machine Learning

AULA 3.5. BANCOS DE DADOS OPEN SOURCE NO AZURE

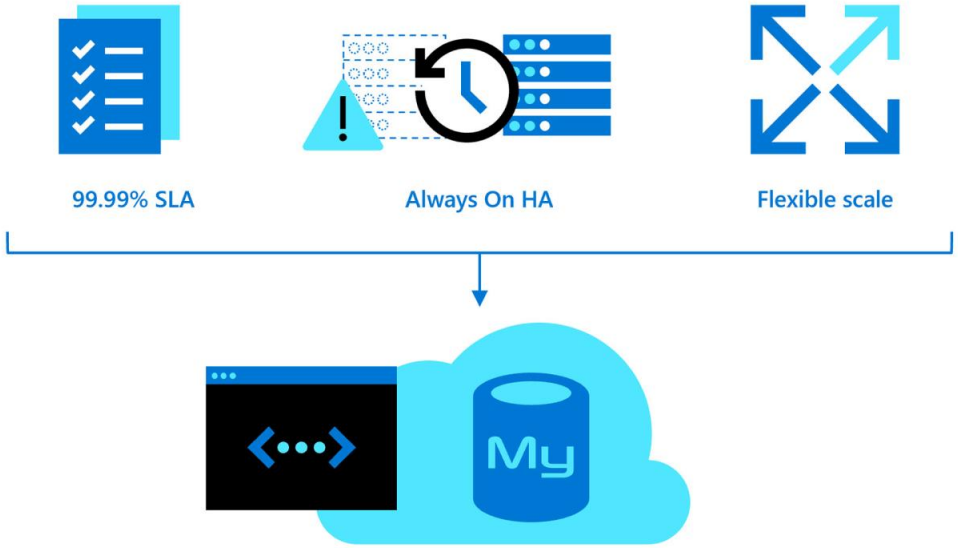
PROF. GUSTAVO AGUILAR

Nesta Aula

- Azure Database for MySQL
- Azure Database for MariaDB
- Azure Database for PostgreSQL

Bancos de Dados Open Source no Azure

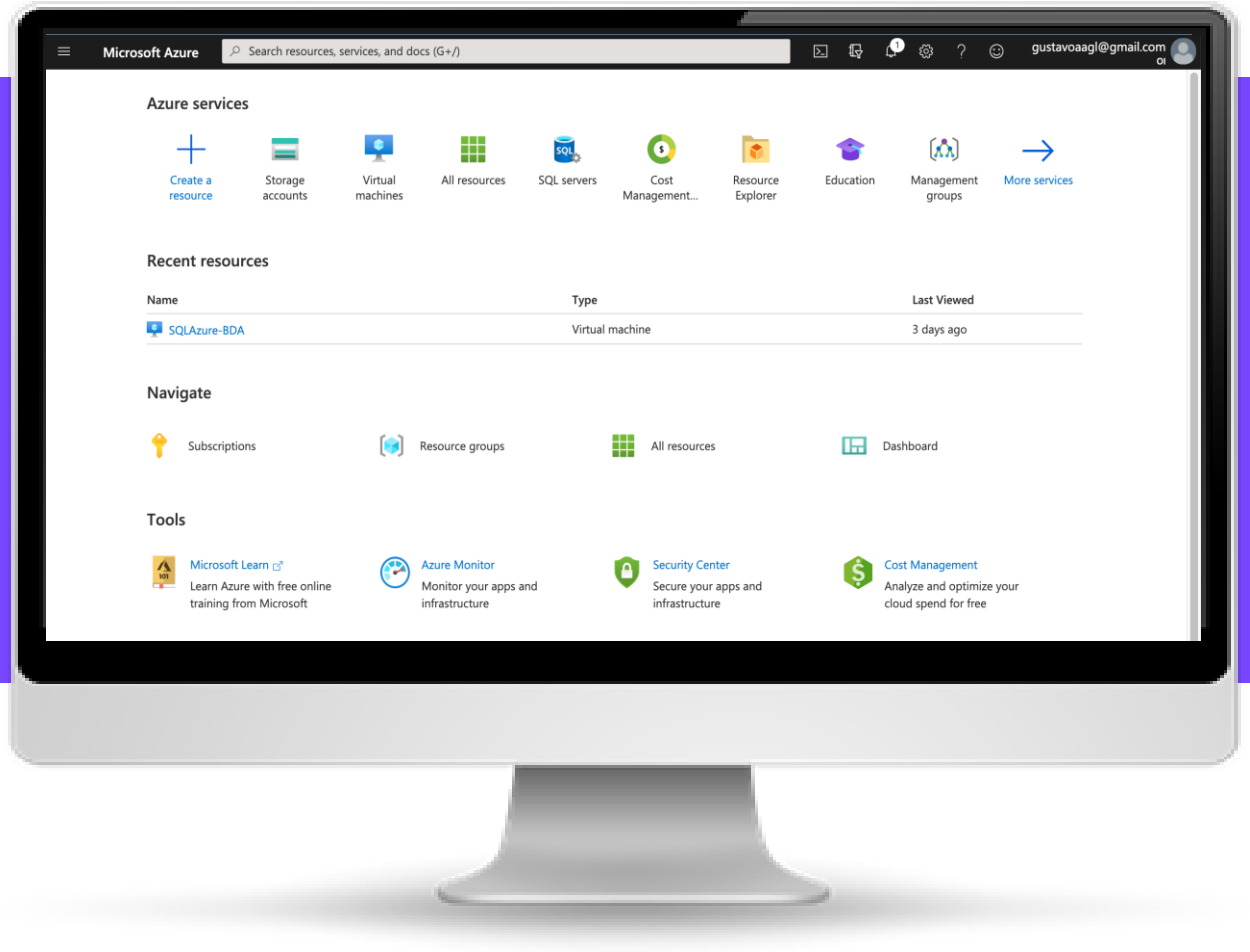
- Mecanismos para deploy:
  - Azure Portal;
  - PowerShell;
  - ARM templates;
  - Azure CLI.
- Azure Database for MySQL / MariaDB
  - Single Database / Replicação.
- Azure Database for PostgreSQL
  - Single Database / Hyperscale.



Criando Azure Database for MySQL

e for PostgreSQL

Demo



Próxima Aula

- Azure Synapse Analytics

Soluções de Dados, Big Data e Machine Learning

AULA 3.6. AZURE SYNAPSE ANALYTICS

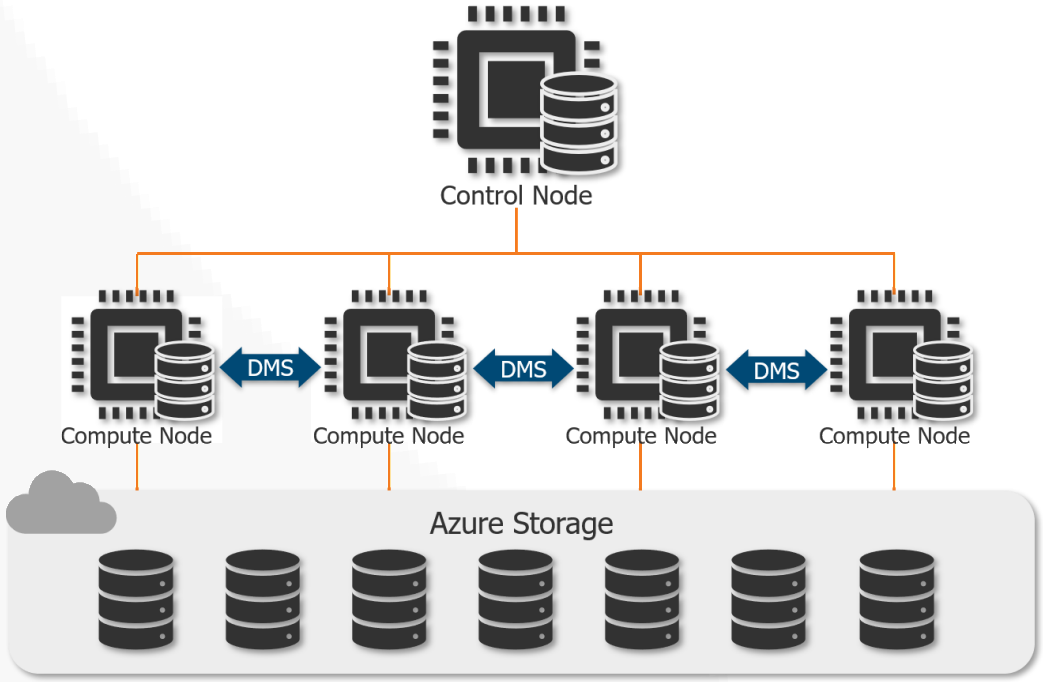
PROF. GUSTAVO AGUILAR

Nesta Aula

- SQL Pools
- Workload Groups
- Demo Azure Synapse Analytics

SQL Pools

- Agrupamento de recursos (CPU, RAM e I/O);
- Tamanho é determinado por DWU (Data Warehousing Units).



Workload Groups

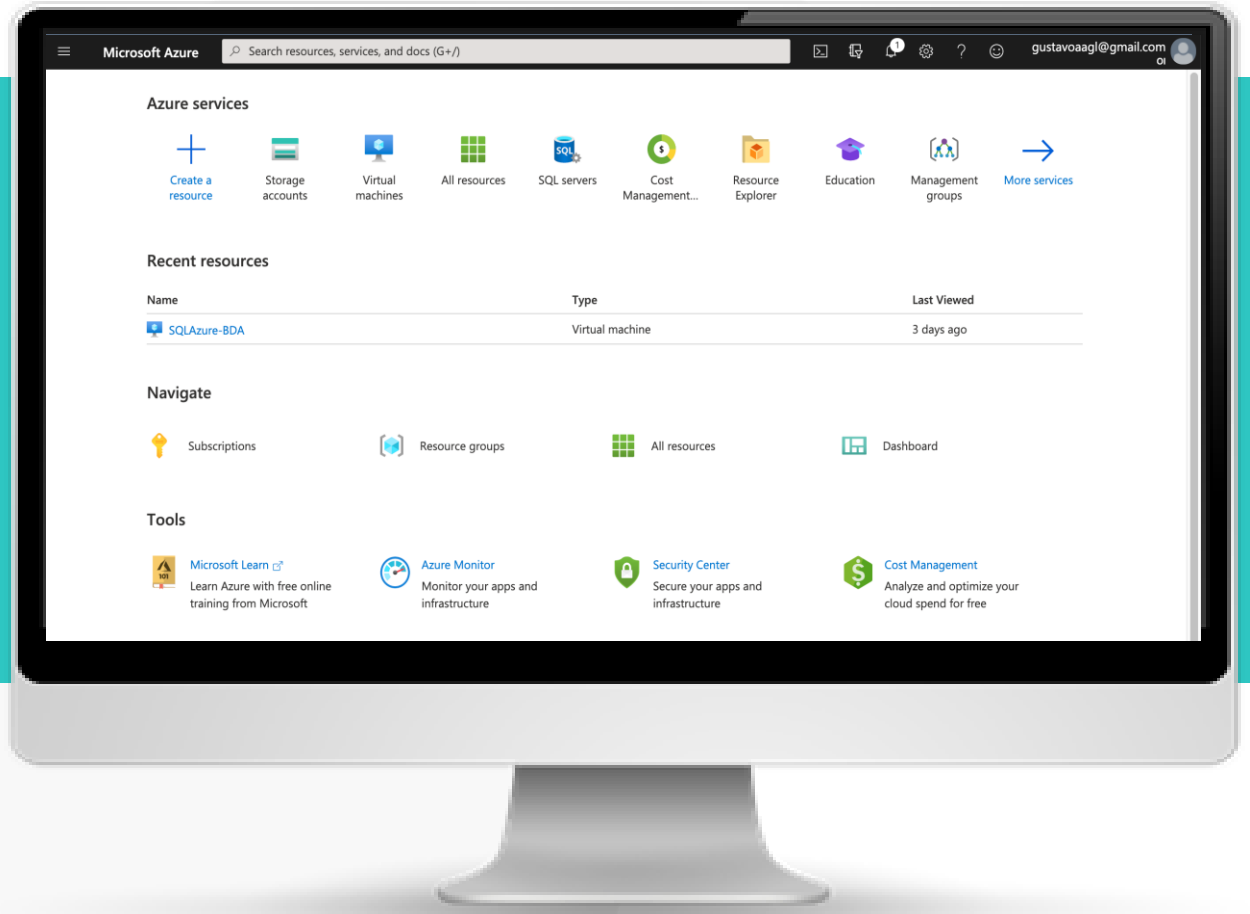
- Definir os recursos para isolar e reservar recursos para uso:
  - Reserva de recursos para um grupo de solicitações;
  - Limite da quantidade de recursos que um grupo de solicitações pode consumer;
  - Recursos compartilhados acessados com base no nível de importância do workload;
  - Definição do valor do tempo limite da consulta.

```
CREATE WORKLOAD GROUP group_name
WITH…..
(
MIN_PERCENTAGE_RESOURCE = value
, CAP_PERCENTAGE_RESOURCE = value
, REQUEST_MIN_RESOURCE_GRANT_PERCENT = value
[ [ , ] REQUEST_MAX_RESOURCE_GRANT_PERCENT = value ]
[ [ , ] IMPORTANCE = {LOW | BELOW_NORMAL | NORMAL | ABOVE_NORMAL | HIGH} ] [ [ ,
] QUERY_EXECUTION_TIMEOUT_SEC = value ]
)[ ; ]
```

```
CREATE WORKLOAD CLASSIFIER classifier_name
WITH
(
WORKLOAD_GROUP = 'name'
, MEMBERNAME = 'security_account'
……..
```

Demo Azure Synapse Analytics

Demo



Próxima Aula

- Capítulo 4 – Armazenamento de Dados Não Relacionais no Azure

Soluções de Dados, Big Data e Machine Learning

Capítulo 4. Armazenamento de Dados Não Relacionais no Azure

PROF. GUSTAVO AGUILAR

Soluções de Dados, Big Data e Machine Learning

AULA 4.1. BANCOS DE DADOS NÃO RELACIONAIS NO AZURE

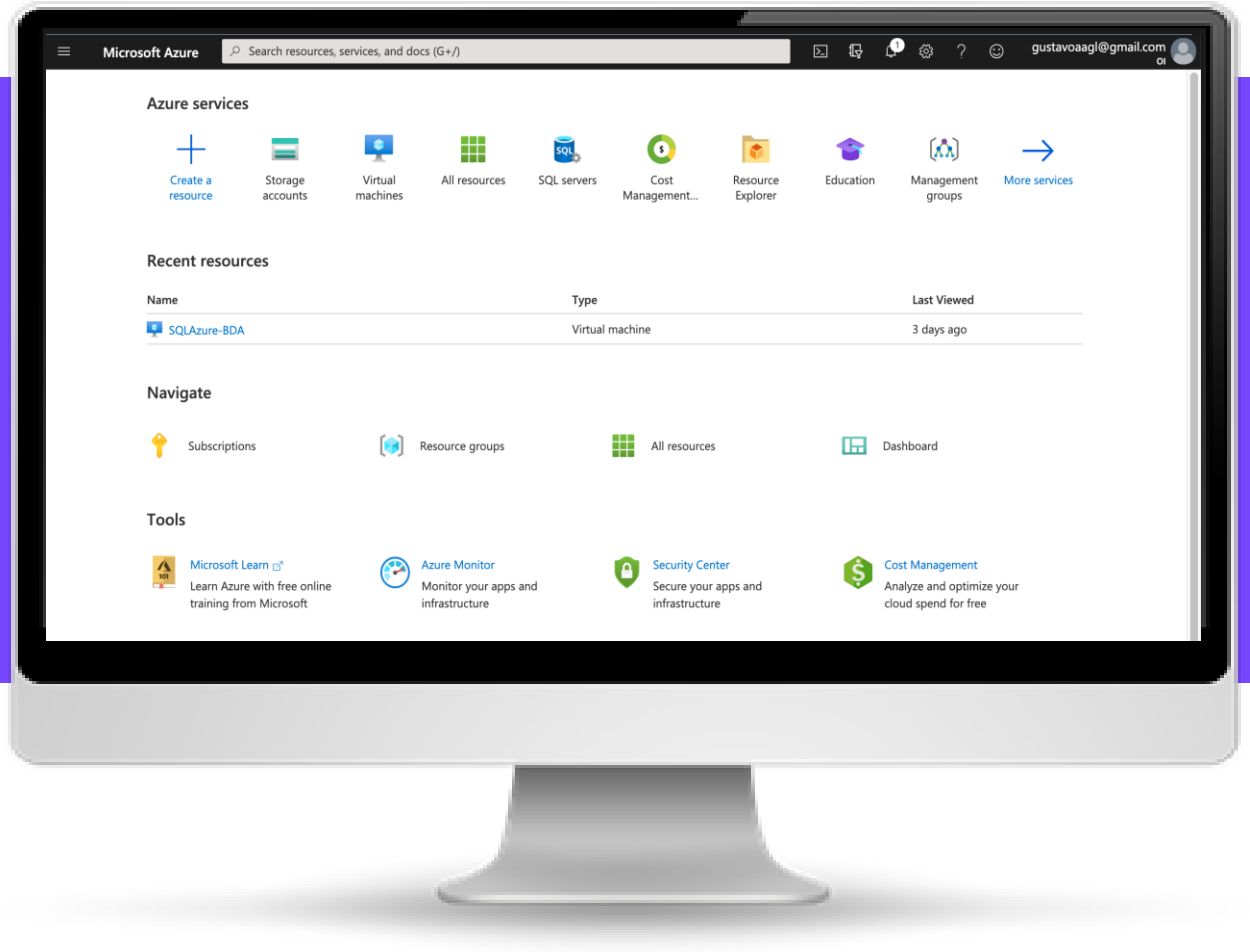
PROF. GUSTAVO AGUILAR

Nesta Aula

- Bancos de Dados Não Relacionais com IaaS
- Bancos de Dados Não Relacionais com PaaS

Banco de Dados Não Relacional com IaaS e PaaS

Demo



Próxima Aula

- Capítulo 5 - Soluções de Big Data

Soluções de Dados, Big Data e Machine Learning

Capítulo 5. Soluções de Big Data

PROF. GUSTAVO AGUILAR

Material das videoaulas do Módulo 3 – Parte 2

Soluções de Dados, Big Data e Machine Learning

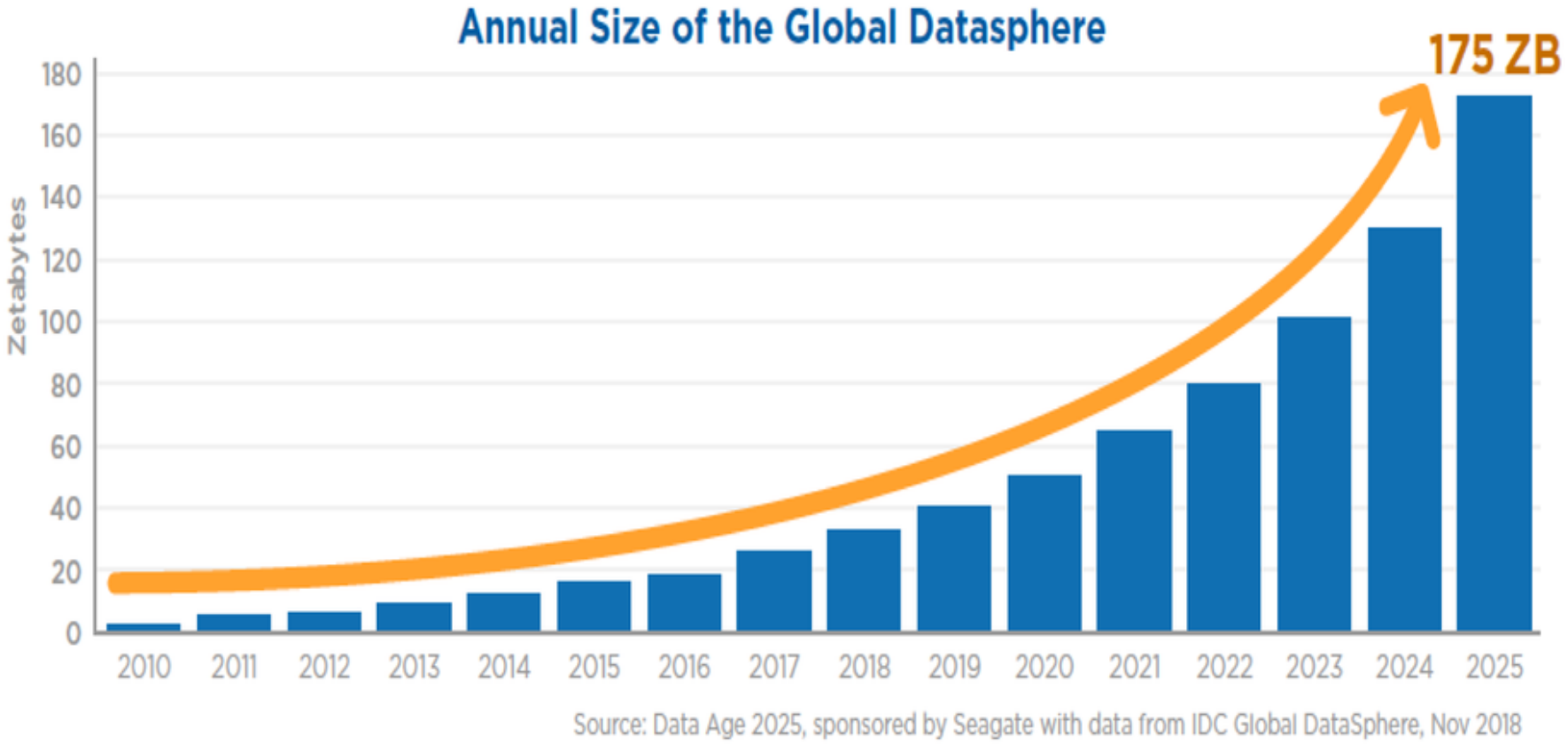
AULA 5.1. INTRODUÇÃO AO BIG DATA

PROF. GUSTAVO AGUILAR

Nesta Aula

- O Que É Big Data?
- Fundamentos de Big Data.

O Que É Big Data



O Que É Big Data

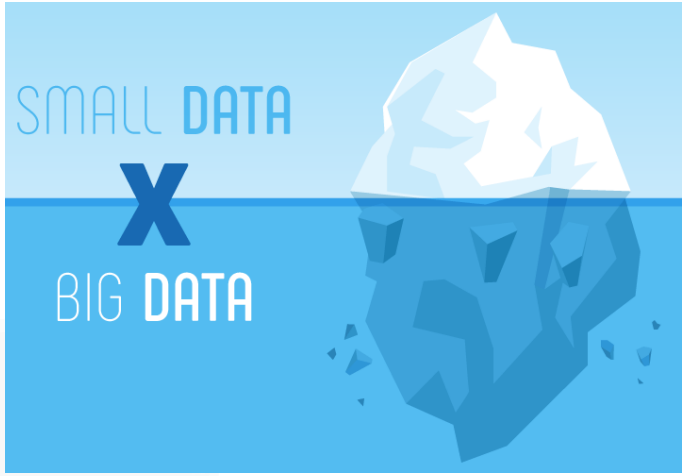
- Termo adotado pelo mercado para descrever problemas no gerenciamento e processamento de informações extremas, as quais excedem a capacidade das tecnologias de informações tradicionais ao longo de uma ou várias dimensões.
- Big Data está focado principalmente em questões de volume de conjunto de dados extremamente grandes gerados a partir de práticas tecnológicas, tais como mídia social, tecnologias operacionais, acessos à Internet e fontes de informações distribuídas. Big Data é essencialmente uma prática que apresenta novas oportunidades de negócios. (Gartner Group)

O Que É Big Data

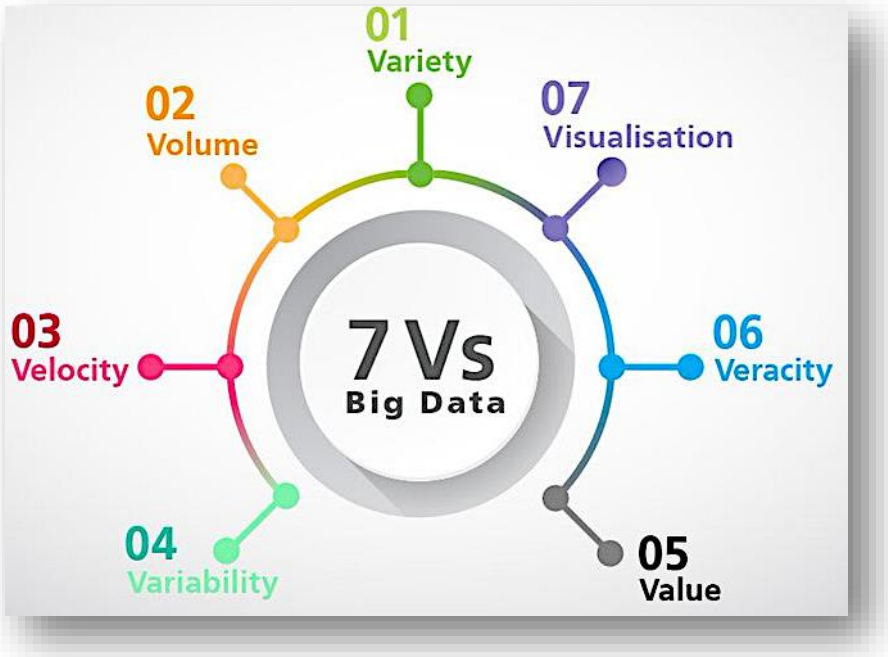
- A intensa utilização de redes sociais on-line, de dispositivos móveis para conexão à Internet, transações e conteúdos digitais, e também o crescente uso de computação em nuvem tem gerado quantidades incalculáveis de dados. O termo Big Data refere-se à este conjunto de dados cujo crescimento é exponencial e cuja dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados. (McKinsey Global Institute)

O Que É Big Data

- Campo que trata de maneiras de analisar, extrair sistematicamente informações ou, de outra forma, lidar com conjuntos de dados que são muito grandes ou complexos para serem tratados por softwares de aplicativos de processamento de dados tradicionais.



Fundamentos de Big Data



Fundamentos de Big Data

- **VOLUME:** a quantidade de dados gerados, que costumava ser medida em Gigabytes agora é medida em Zettabytes (ZB), caminhando para Yottabytes (YB).
- **VELOCIDADE:** velocidade em que os dados são gerados e se tornam acessíveis.
- **VARIEDADE:** de formatos, de tipos (estruturado, não estruturado e semiestruturado), e da natureza (numérica, data, caractere, etc.).
- **VARIABILIDADE:** mesma combinação de dados cujo significado muda constantemente.

Fundamentos de Big Data

- **VERACIDADE:** garantir que os dados sejam precisos, verdadeiros, fidedignos.
- **VISUALIZAÇÃO:** formas mais aderentes para visualizar grandes quantidades de dados complexos.
- **VALOR:** transformar dados em valor, vantagem competitiva ou otimização operacional.

Conclusão

- Big Data é mais que um produto de software ou hardware;
- É um conjunto de tecnologias, processos e práticas que permitem às empresas analisarem dados que antes não tinham acesso e tomar decisões, ou mesmo gerenciar atividades de forma muito mais eficiente.

Próxima Aula

- Introdução ao HDInsight.

Soluções de Dados, Big Data e Machine Learning

AULA 5.2. INTRODUÇÃO AO HDINSIGHT

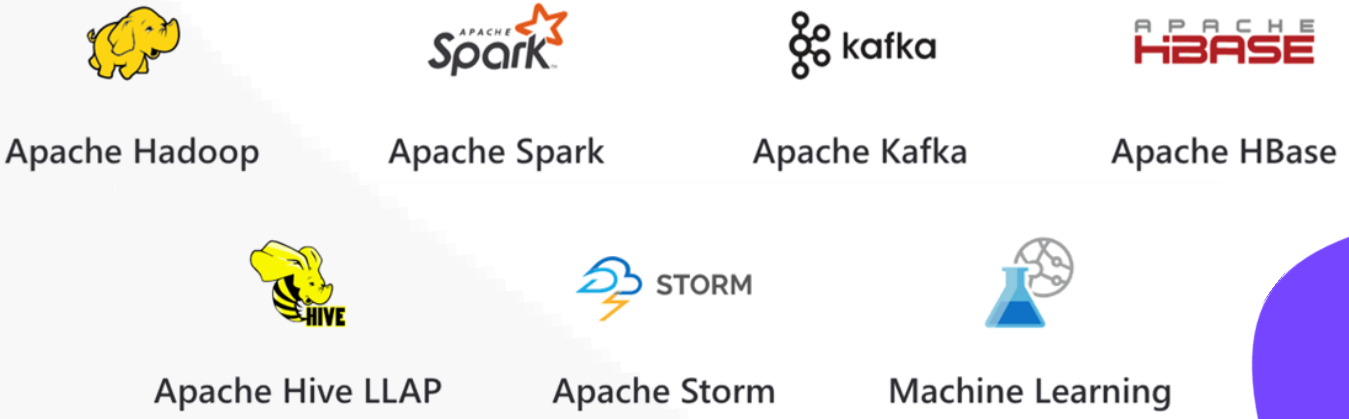
PROF. GUSTAVO AGUILAR

Nesta Aula

- Overview do HDInsight.
- Principais Componentes.
- Ecossistema do Azure HDInsight.

Overview do HDInsight

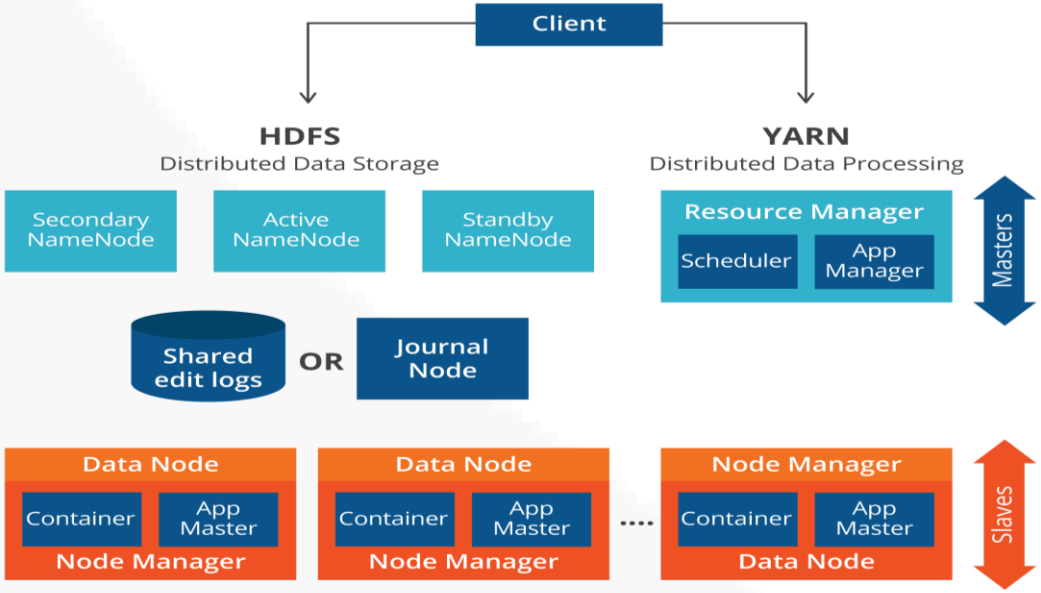
- Recurso para processamento e análise de Big Data;
- Criação rápida de clusters de Big Data sob demanda, com escalabilidade horizontal e/ou vertical;
- Solução em nuvem de baixo custo para projetos de Big Data;
- Possui os principais softwares livres da Apache para projetos de Big Data:



Principais Componentes

APACHE HADOOP

- Estrutura que usa HDFS, gerenciamento de recursos YARN e um modelo de programação MapReduce simples para processar e analisar, paralelamente, dados em lote.

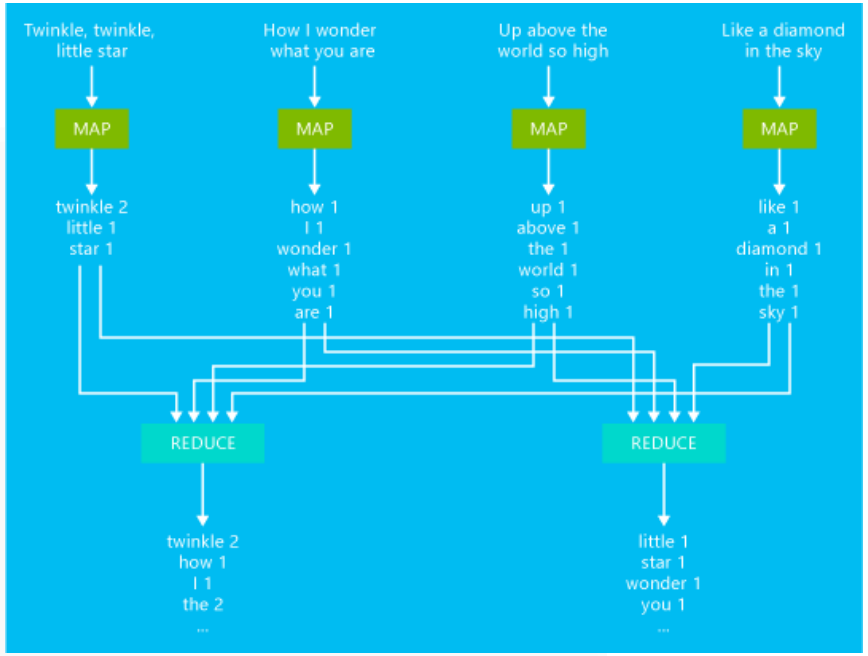


Apache Hadoop

Principais Componentes

APACHE HADOOP

- Utiliza um modelo de programação MapReduce simples para processar e analisar, paralelamente, dados em lote.



Apache Hadoop

Principais Componentes

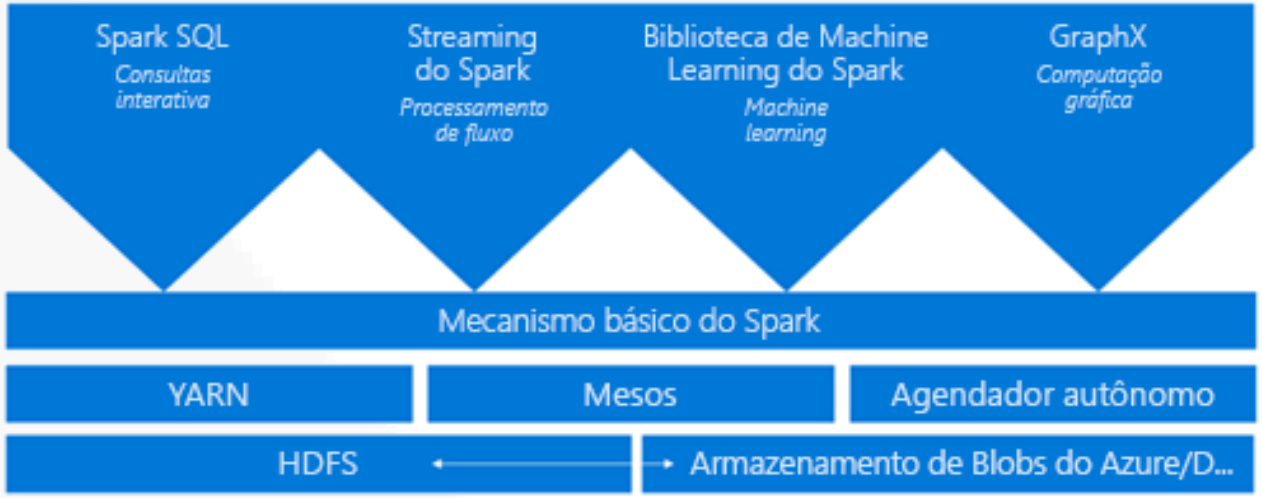
APACHE SPARK

- Estrutura de processamento paralelo, de software livre, que dá suporte ao processamento em memória para melhorar o desempenho dos aplicativos de análise de Big Data.

Apache Spark

Principais Componentes

APACHE SPARK



Principais Componentes

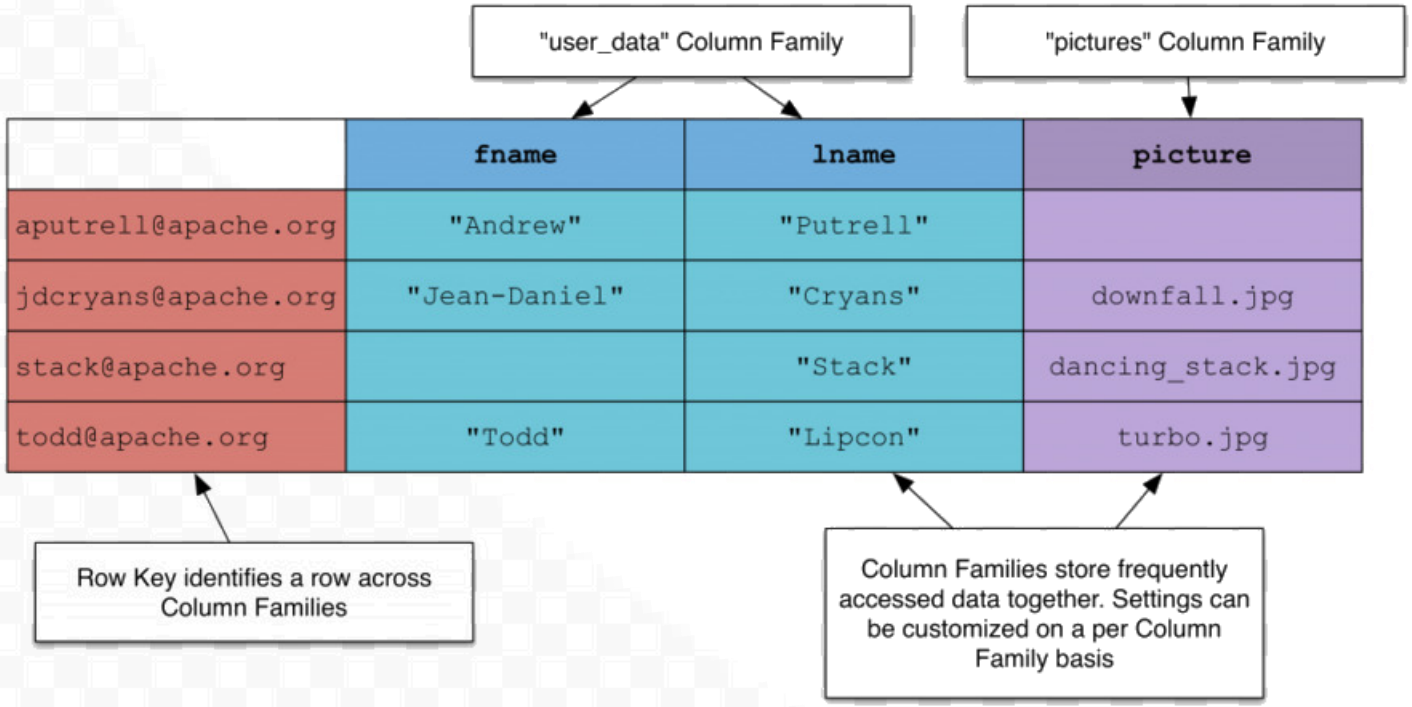
APACHE HBASE

- Banco de dados NOSQL baseado em Hadoop que fornece acesso aleatório e forte coerência para grandes quantidades de dados sem esquema;
- Construído com base no Google BigTable;
- Os dados são armazenados nas linhas e colunas de uma tabela, e os dados em uma linha são agrupados por família de colunas.

Apache HBase

Principais Componentes

APACHE HBASE



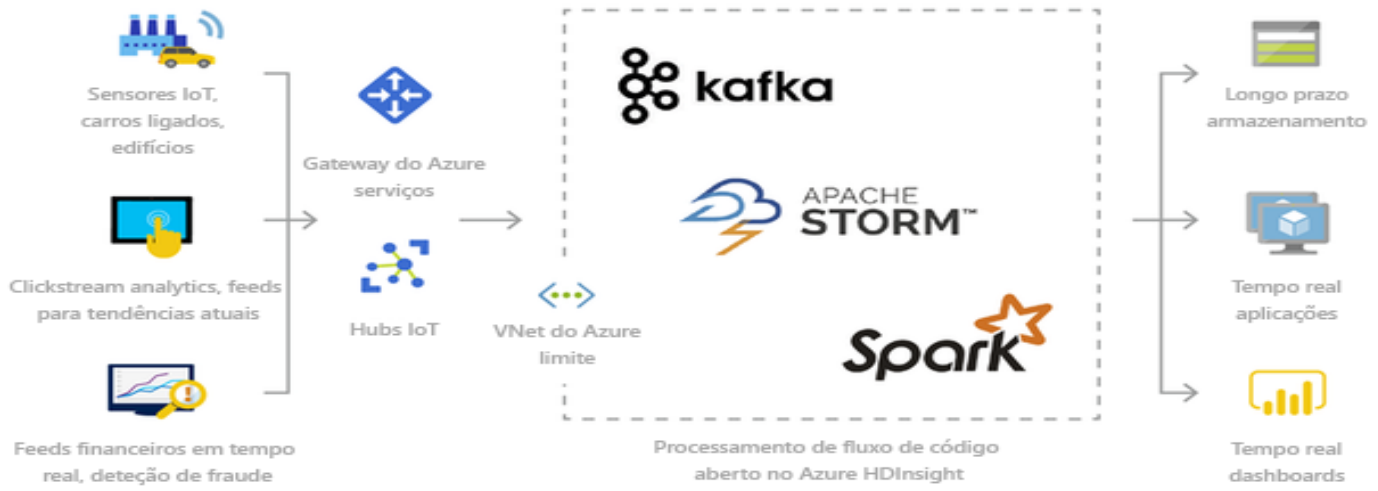
Principais Componentes

APACHE KAFKA

- Plataforma de código fonte aberto usada para criar aplicativos e pipelines de streaming de dados;
- Também fornece funcionalidade de fila de mensagens, o que permite publicar e consumir pipelines de dados.

kafka

Apache Kafka



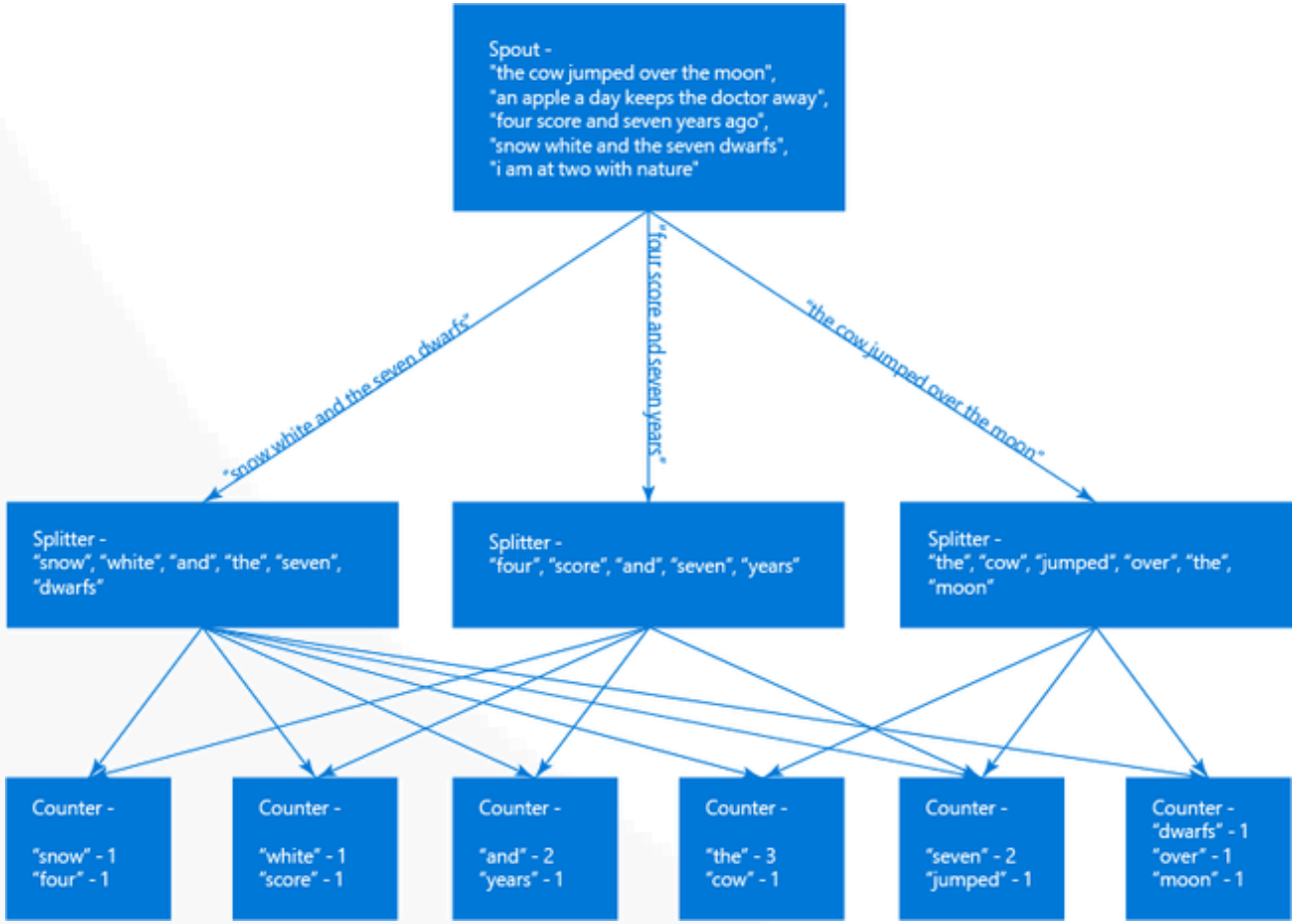
Principais Componentes

APACHE STORM

- Sistema de computação distribuído e em tempo real para processar rapidamente grandes fluxos de dados;
- Processa topologias ao invés de trabalhos MapReduce.

Apache Storm

Principais Componentes



STORM

Apache Storm

Principais Componentes

APACHE HIVE LLAP

- O Hive oferece uma interface semelhante à SQL para consulta de dados em diferentes bancos de dados e sistemas de arquivos integrados ao Hadoop;
- Comandos tradicionais de SQL são implementados na API Java (HiveQL) para serem executados em dados distribuídos.
- O Hive LLAP é um recurso para cache de dados em memória para consultas do Hive.

Apache Hive LLAP

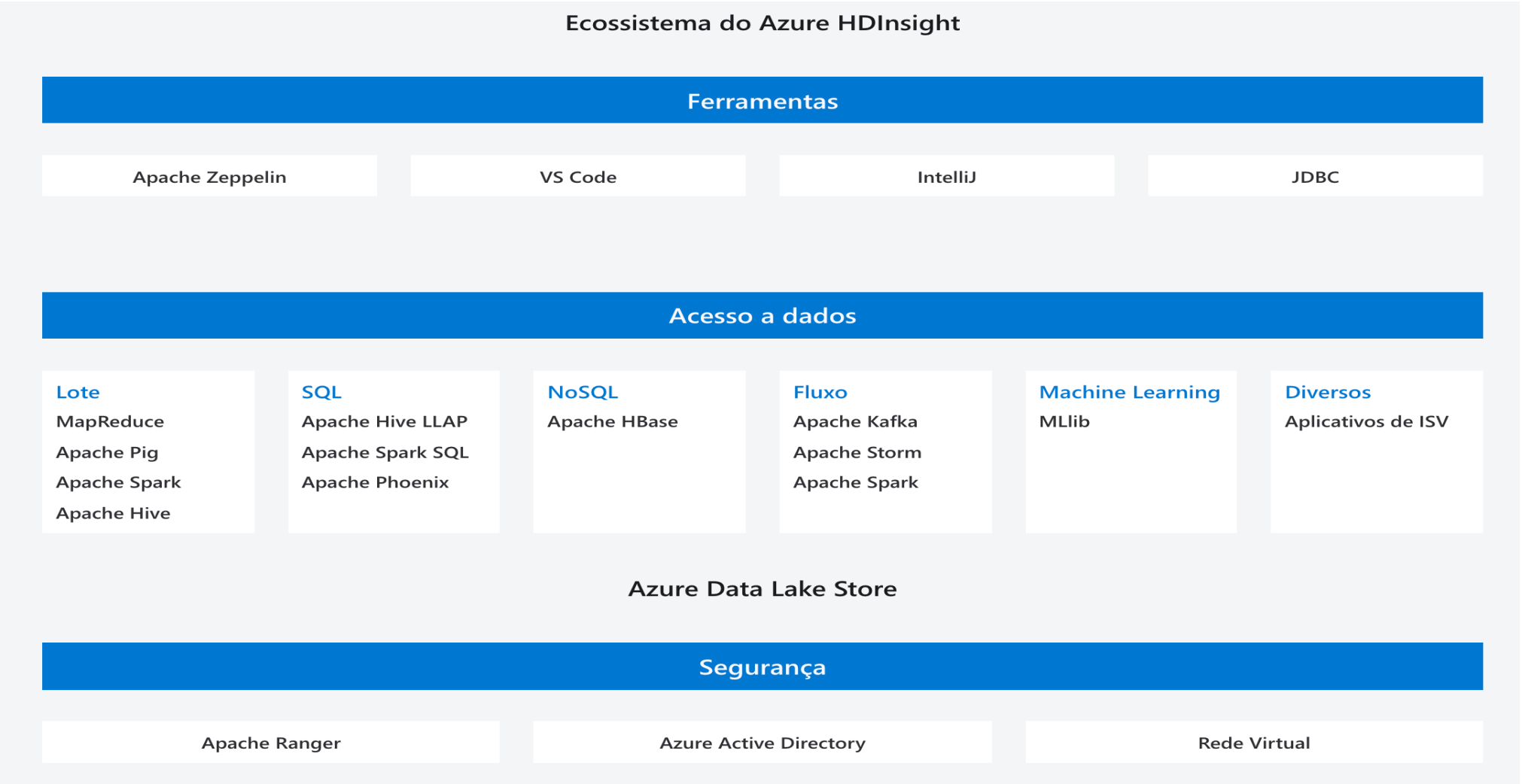
Principais Componentes

APACHE ML SERVICES

- Cluster para soluções de aprendizagem de máquina;
- Fornece aos cientistas de dados, estatísticos e programadores de R/Python o acesso sob demanda a métodos escalonáveis e distribuídos de análise no HDInsigh;
- Possui um conjunto de modelos e algoritmos de machine learning que podem ser adaptados.

Machine Learning

Ecossistema do Azure HDInsight



Próxima Aula

- Aprovisionando um Ambiente do HDInsight.

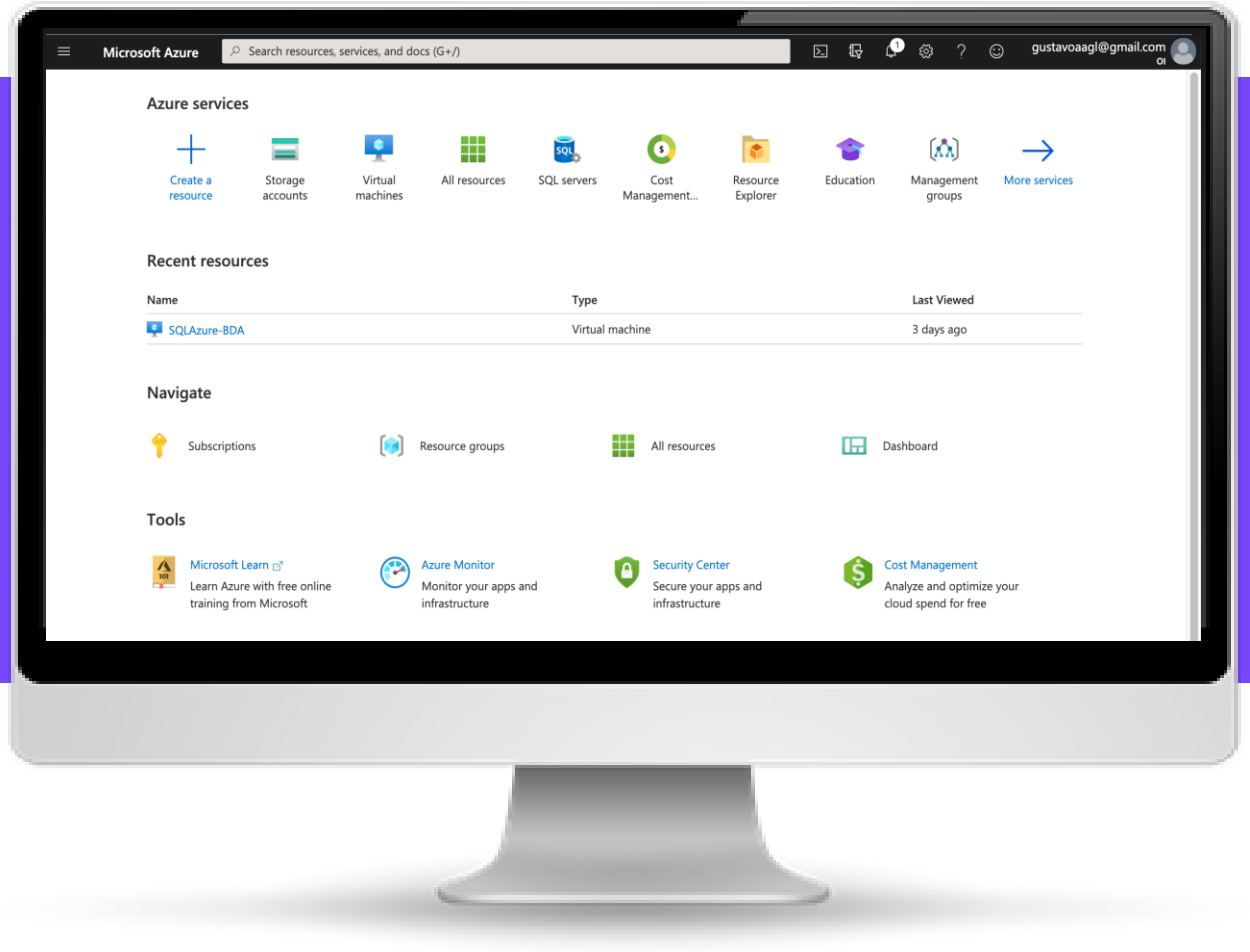
Soluções de Dados, Big Data e Machine Learning

AULA 5.3. APROVISIONANDO UM AMBIENTE DO HDINSIGHT

PROF. GUSTAVO AGUILAR

Aprovisionando um Ambiente de HDInsight

Demo



Próxima Aula

- Introdução ao Azure DataBricks.

Soluções de Dados, Big Data e Machine Learning

AULA 5.4. INTRODUÇÃO AO AZURE DATABRICKS

PROF. GUSTAVO AGUILAR

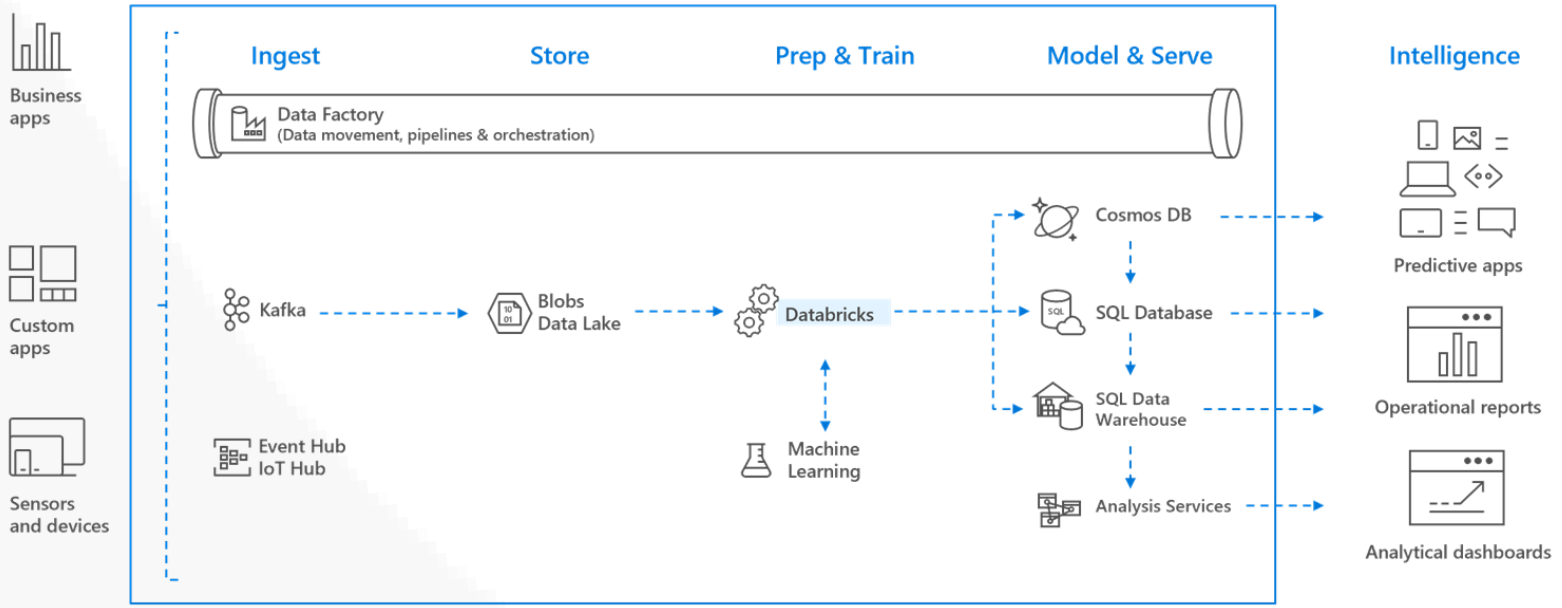
Nesta Aula

- Overview do Azure Databricks.
- Componentes do Azure Databricks.

Overview do Azure Databricks

- Plataforma de análise baseada no Apache Spark e otimizada para a plataforma de serviços de nuvem do Microsoft Azure;
- Configuração rápida e automatizada do cluster Spark;
- Fluxos de trabalho simplificados;
- Workspace interativo que permite a colaboração entre os cientistas de dados, os engenheiros de dados e os analistas de dados/negócios.

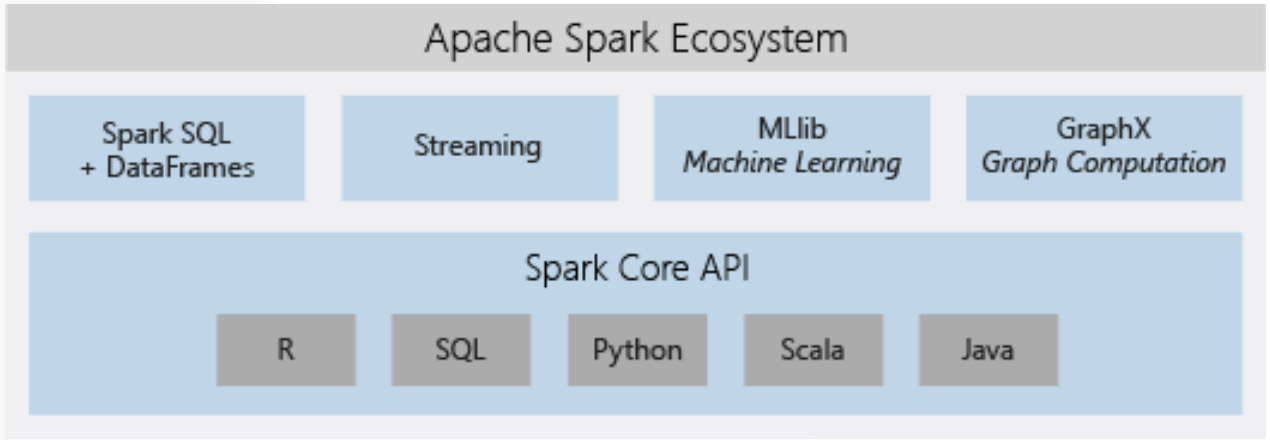
Overview do Azure Databricks



Componentes do Azure Databricks

SPARK SQL E DATAFRAMES

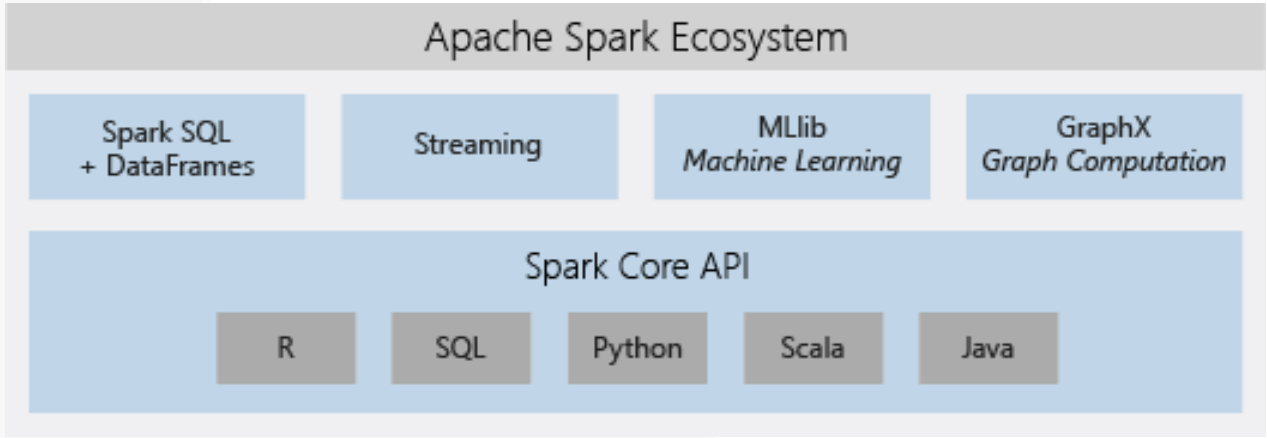
- Módulo Spark para trabalhar usando dados estruturados;
- DataFrame é uma coleção distribuída de dados organizados em colunas nomeadas (conceitualmente equivalente a uma tabela em um banco de dados relacional ou uma estrutura de dados em R/Python).



Componentes do Azure Databricks

STREAMING

- Módulo para processamento de dados em tempo real e análise para aplicativos analíticos e interativos;
- Integra-se com HDFS, Flume e Kafka.

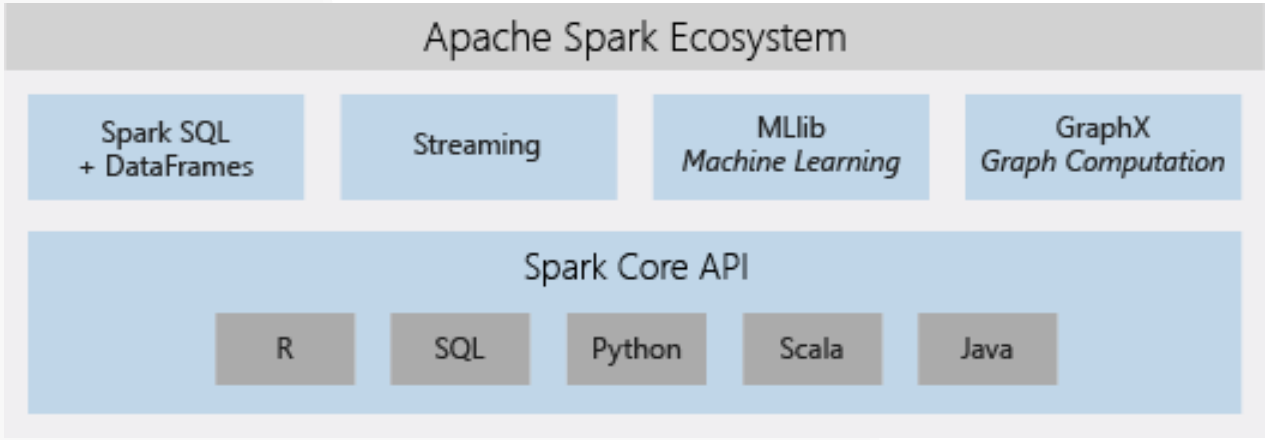


Componentes do Azure Databricks

MLIB: biblioteca Machine Learning que consiste em algoritmos e utilitários de aprendizado comuns, incluindo classificação, regressão, clustering, etc.

GRAPHX: módulo para tarefas de análise de gráficos e operações em grafos.

SPARK CORE API: suporte para R, SQL, Python, Scala e Java.



Próxima Aula

- Demonstração do Azure DataBricks.

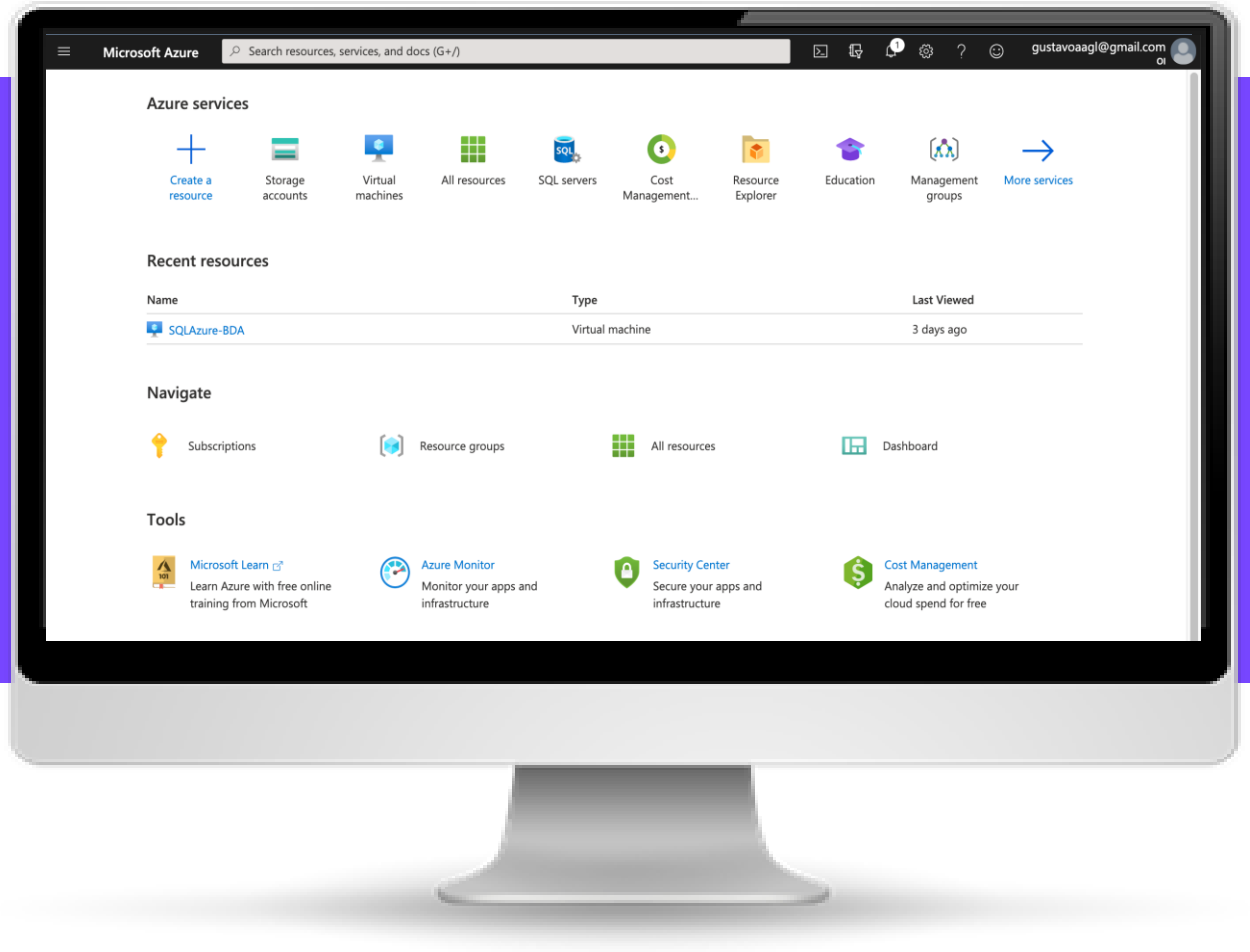
Soluções de Dados, Big Data e Machine Learning

AULA 5.5. DEMONSTRAÇÃO DO AZURE DATABRICKS

PROF. GUSTAVO AGUILAR

Azure Databricks

Demo



Próxima Aula

- Capítulo 6 – Soluções para Pipeline de Dados.

Soluções de Dados, Big Data e Machine Learning

Capítulo 6. Soluções para Pipeline de Dados

PROF. GUSTAVO AGUILAR

Soluções de Dados, Big Data e Machine Learning

AULA 6.1. INTRODUÇÃO AO AZURE DATA FACTORY

PROF. GUSTAVO AGUILAR

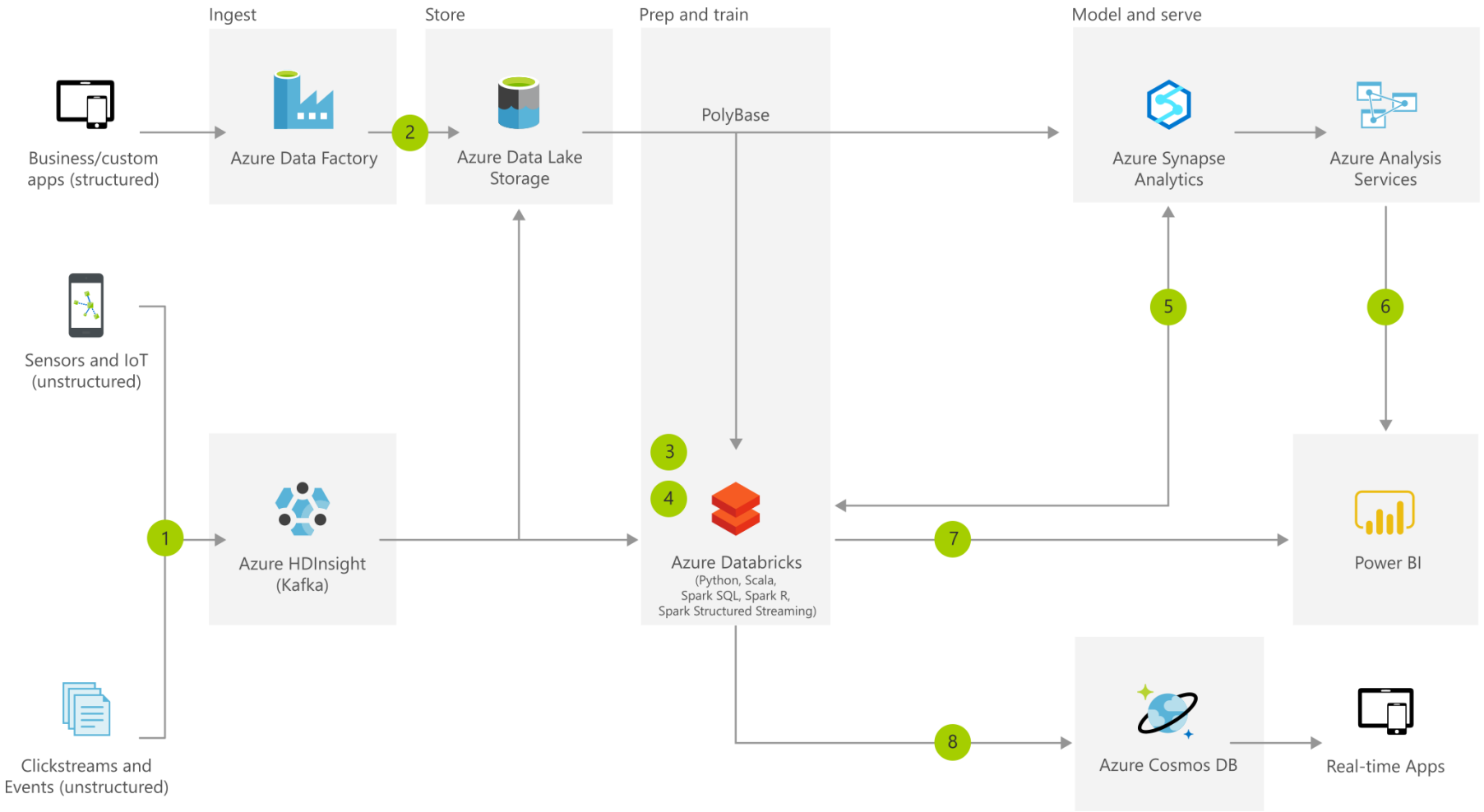
Nesta Aula

- Overview do Azure Data Factory.
- Componentes do Azure Data Factory.

Overview do Azure Data Factory

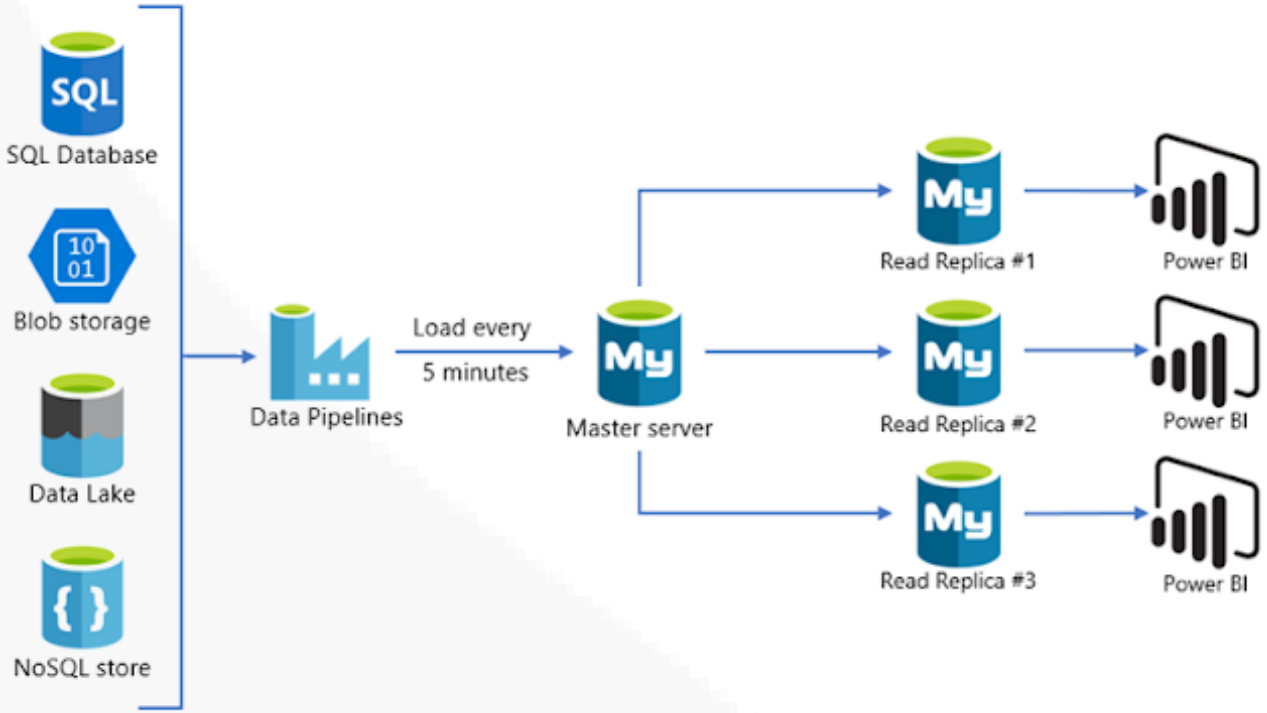
- Serviço de integração de dados e ETL baseado em nuvem;
- Conectores para mais de 90 tipos de fontes de dados diferentes;
- ETLs simples à complexos, com integração com Azure HDInsight, Azure Databricks ou Azure Synapse Analytics.

Overview do Azure Data Factory

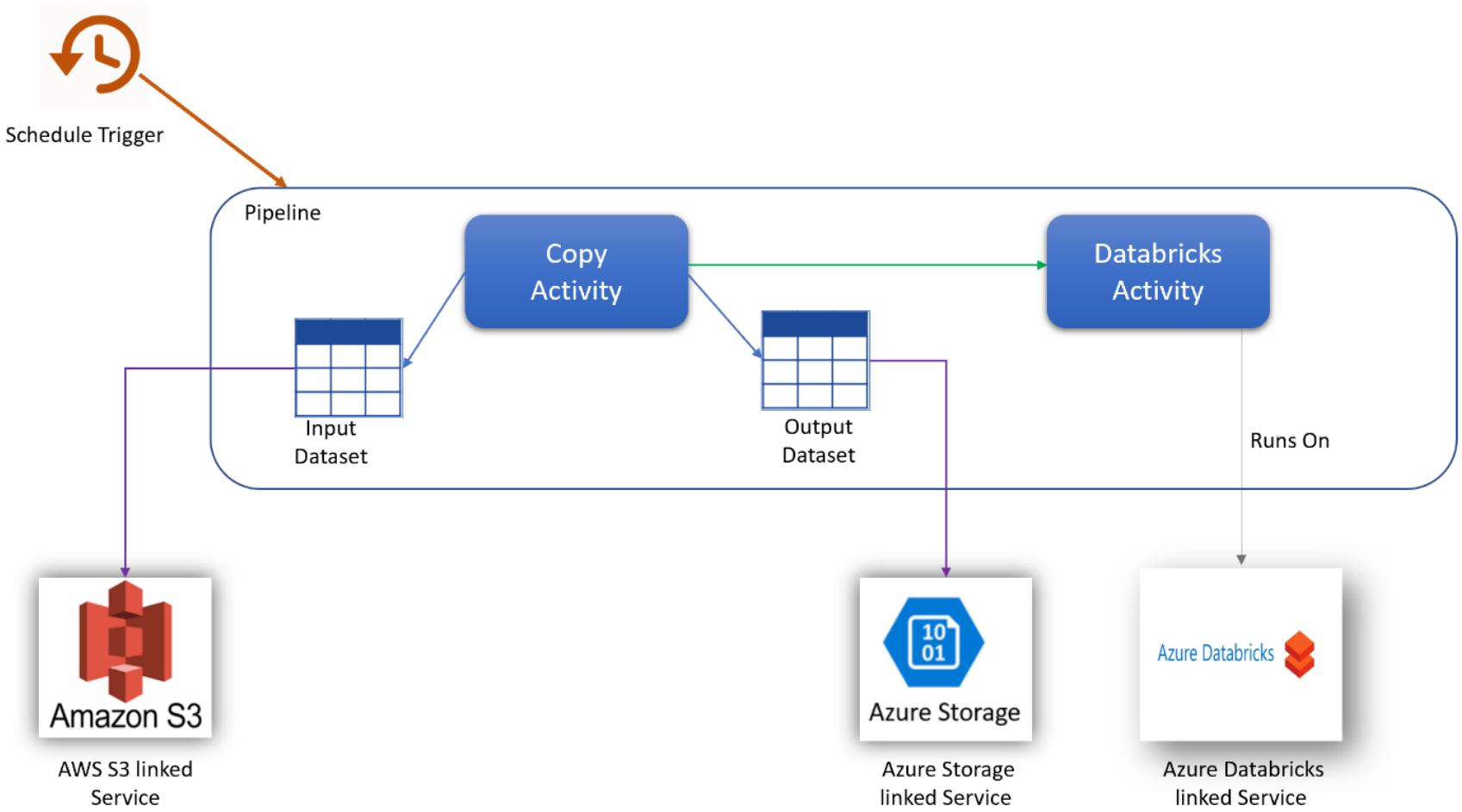


Overview do Azure Data Factory

- Permite criar e agendar fluxos de trabalho orientados a dados para orquestrar a movimentação de dados e transformá-los;

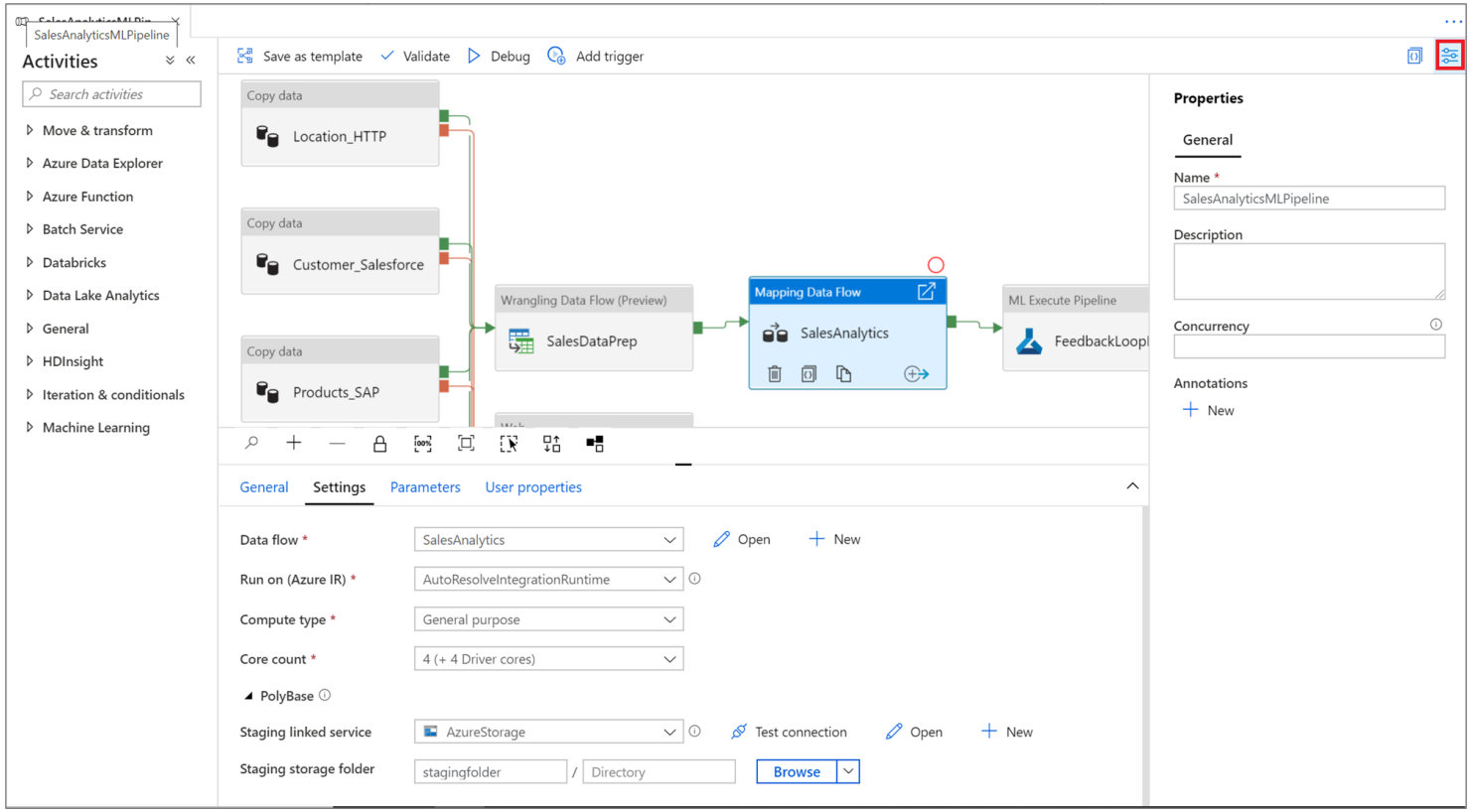


Overview do Azure Data Factory



Overview do Azure Data Factory

- Possibilita a criação de pipelines de forma gráfica ou via código.



Componentes do Azure Data Factory

ATIVIDADE

- Representa uma etapa de processamento em um pipeline.
  - Atividade para copiar dados de um repositório de dados para outro;
  - Atividade que executa uma consulta de Hive em um cluster do Azure HDInsight para transformar ou analisar dados;
  - Etc.
- O Data Factory dá suporte a três tipos de atividades:
  - Atividades de movimentação de dados;
  - Atividades de transformação de dados;
  - Atividades de controle.

Componentes do Azure Data Factory

PIPELINE

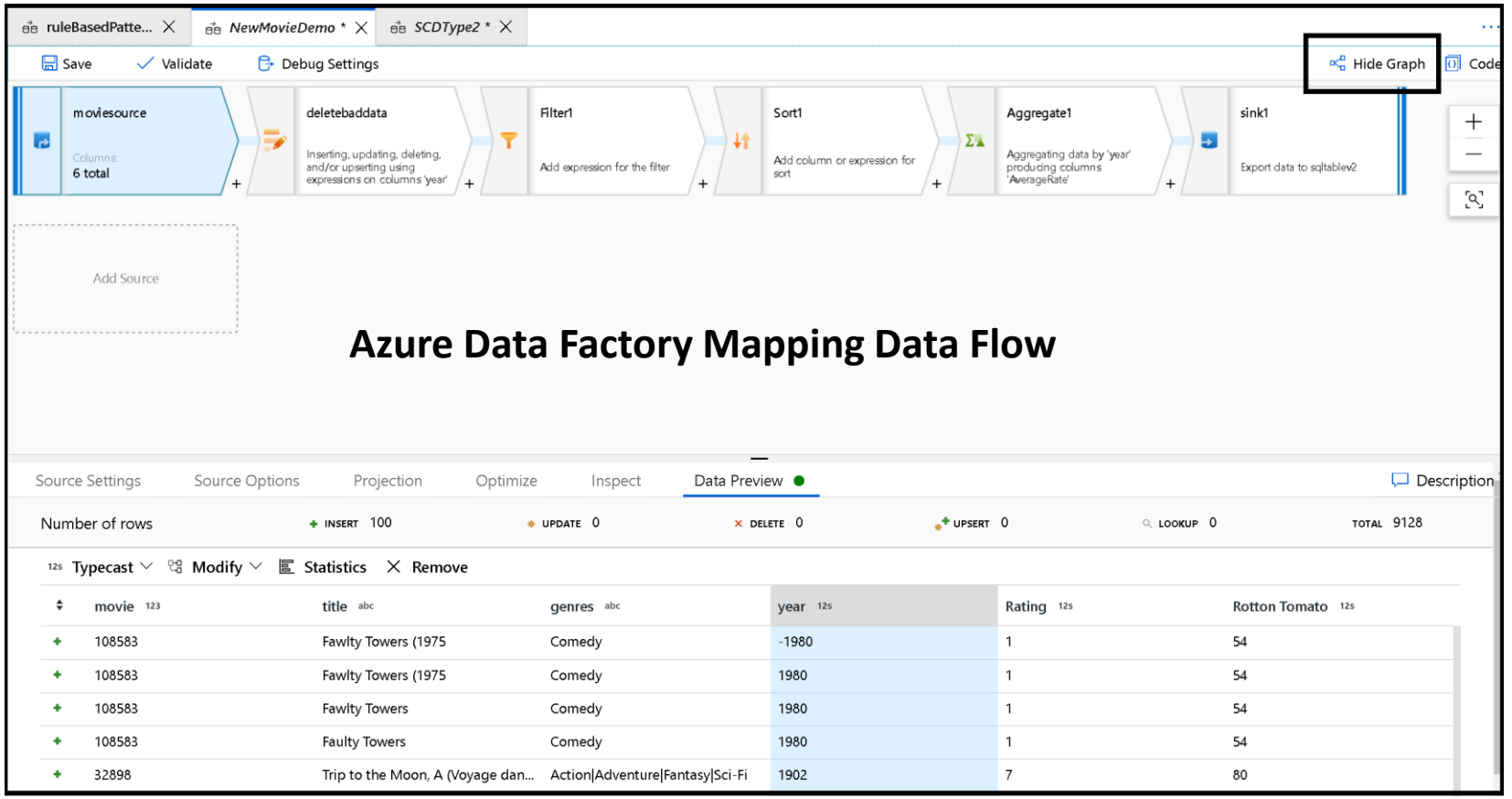
- Agrupamento lógico de atividades que realiza uma unidade de trabalho. Juntas, as atividades em um pipeline executam uma tarefa.
- Exemplo: pipeline contém um grupo de atividades que ingere dados provenientes de um blob do Azure e, em seguida, executa uma consulta Hive em um cluster HDInsight para particionar os dados.
- Pipeline permite gerenciar atividades como um conjunto, em vez de gerenciar cada uma individualmente.
- Atividades podem operar de modo sequencial ou de forma independente, em paralelo.

Componentes do Azure Data Factory

MAPEAMENTO DE FLUXO DE DADOS (MAPPING DATA FLOW)

- São transformações de dados visualmente projetadas no Azure Data Factory;
- Permitem que os engenheiros de dados desenvolvam lógicas de transformação de dados sem escrever código;
- O Data Factory executará a lógica em um cluster Spark, autogerenciado pelo Azure, que será ativado e desativado quando necessário.

Componentes do Azure Data Factory



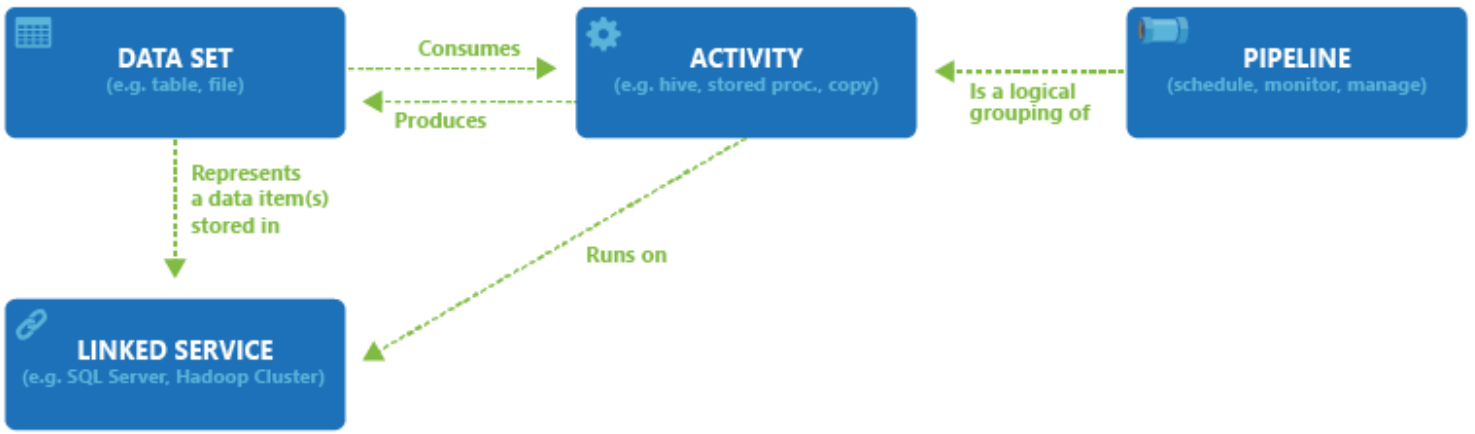
Componentes do Azure Data Factory

- **CONJUNTO DE DADOS (DATASET):** representam as estruturas de dados nos repositórios de dados, que simplesmente apontam para ou fazem referência aos dados que deseja-se usar em atividades, seja como entrada ou saída.
- **SERVIÇO VINCULADO (LINKED SERVICE):** define as informações de conexão necessárias para que o Data Factory se conecte aos recursos externos. Duas finalidades:
  - Para representar um **armazenamento de dados:** ex. banco SQL / Oracle;
  - Para representar um **recurso de computação** que pode hospedar a execução de uma atividade: ex. um cluster Hadoop do HDInsight, onde a atividade HDInsightHive é executada.

Componentes do Azure

Data Factory

- Um serviço vinculado define a conexão à fonte de dados e um conjunto de dados representa a estrutura dos dados.
  - Por exemplo, um serviço vinculado de armazenamento do azure especifica a string de conexão para conectar-se à conta de Armazenamento do Azure (Storage Account), e um conjunto de dados de blob do Azure especifica o contêiner de blob e a pasta que contém os dados.



Próxima Aula

- Criando um Pipeline de Dados com o Azure Data Factory.

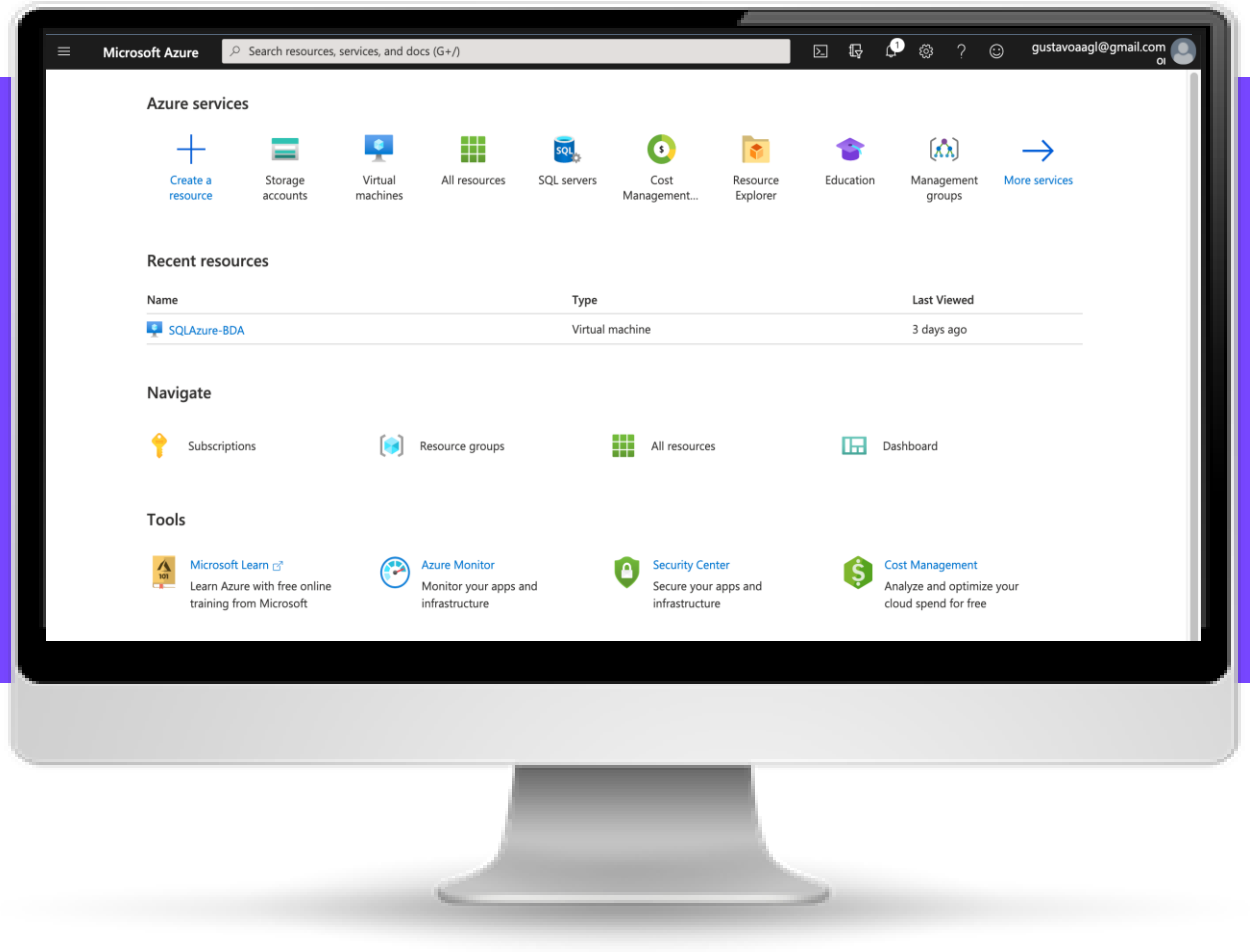
Soluções de Dados, Big Data e Machine Learning

AULA 6.2. CRIANDO UM PIPELINE DE DADOS COM O AZURE DATA FACTORY

PROF. GUSTAVO AGUILAR

Criando um Pipeline de Dados com o Azure Data Factory

Demo



Próxima Aula

- Capítulo 7 – Soluções de Machine Learning.

Soluções de Dados, Big Data e Machine Learning

Capítulo 7. Soluções de Machine Learning

PROF. GUSTAVO AGUILAR

Soluções de Dados, Big Data e Machine Learning

AULA 7.1. OVERVIEW DO AZURE MACHINE LEARNING

PROF. GUSTAVO AGUILAR

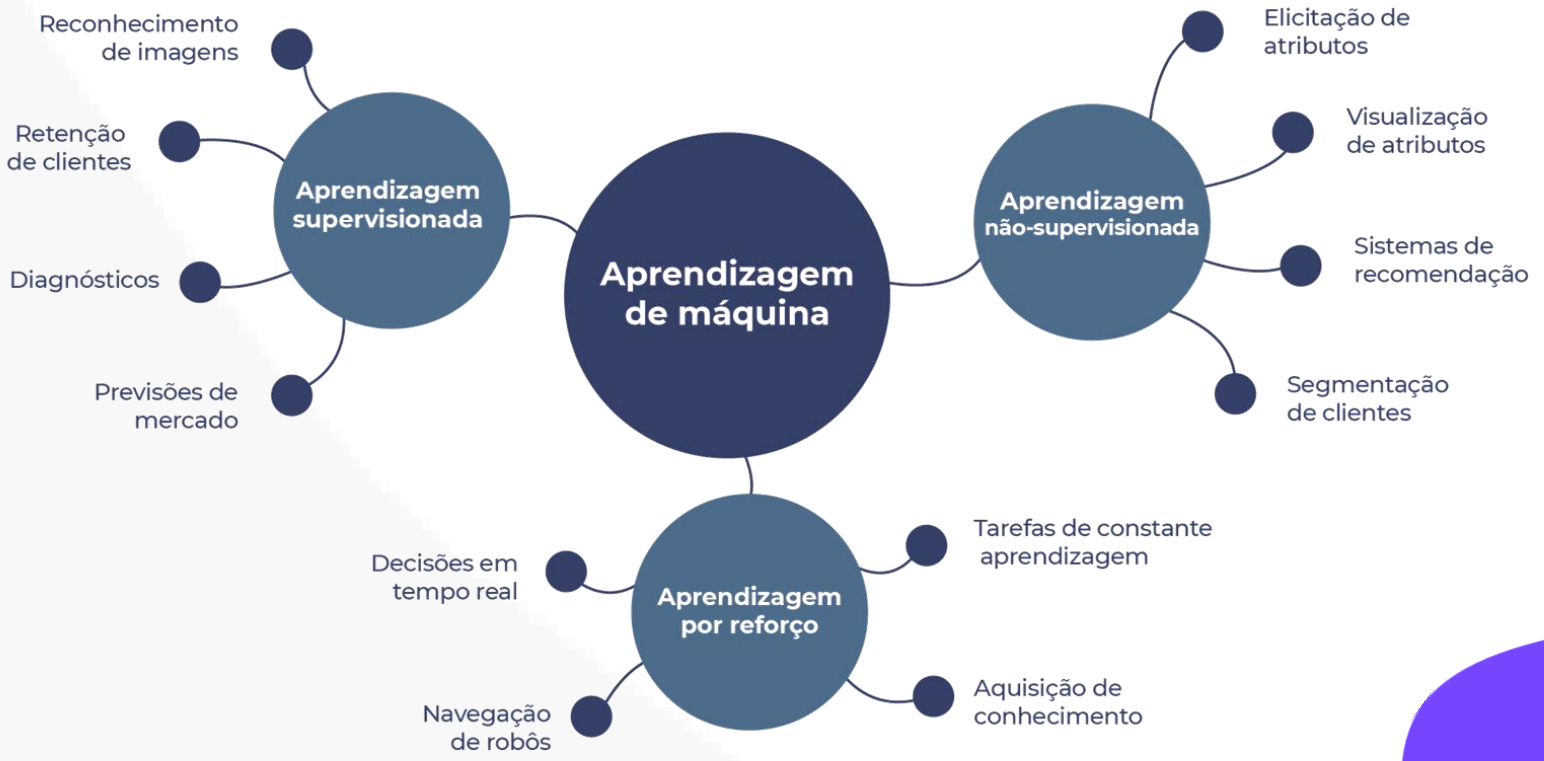
Nesta Aula

- Introdução ao Aprendizado de Máquina.
- Azure Machine Learning.

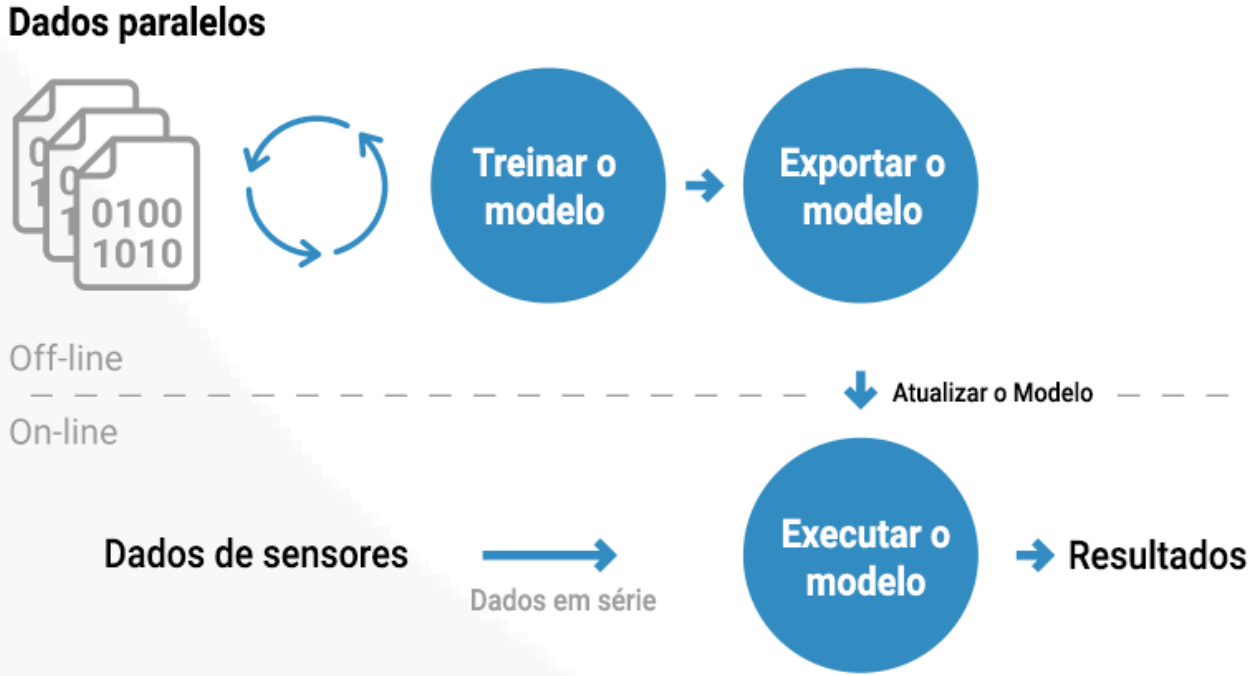
Introdução ao Aprendizado de Máquina

- Aprendizado de máquina (Machine Learning - ML) é uma técnica da ciência de dados que permite que os computadores usem os dados existentes para prever tendências, resultados e comportamentos futuros.
- Usando ML, os computadores têm a capacidade de aprender de acordo com as respostas esperadas por meio das associações de diferentes dados, os quais podem ser imagens, áudio, números, etc.

Introdução ao Aprendizado de Máquina

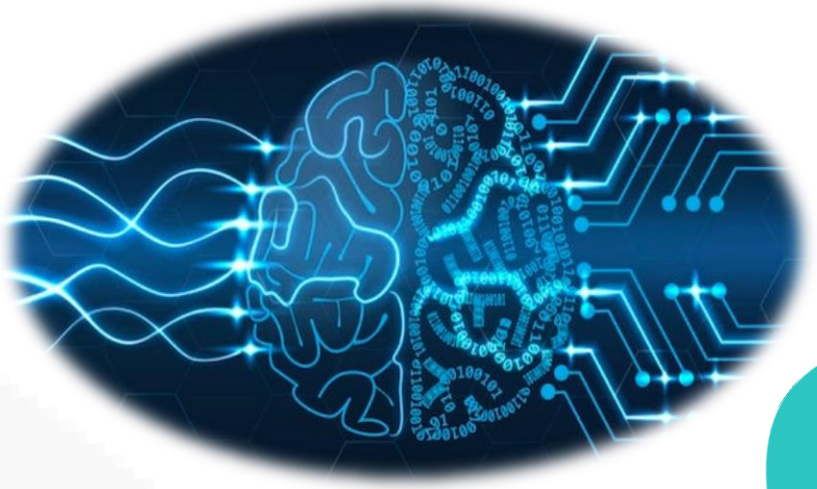


Introdução ao Aprendizado de Máquina

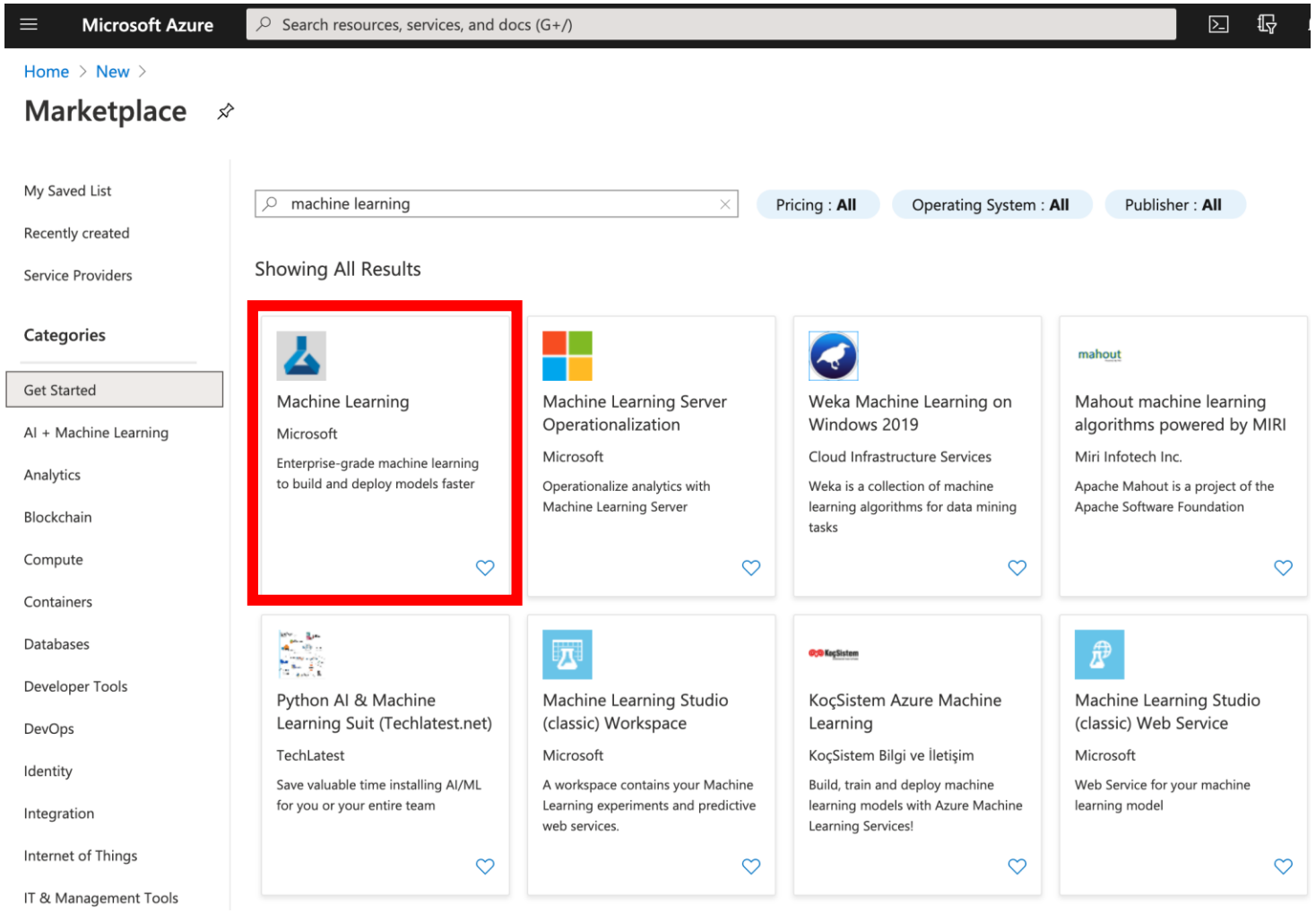


Azure Machine Learning

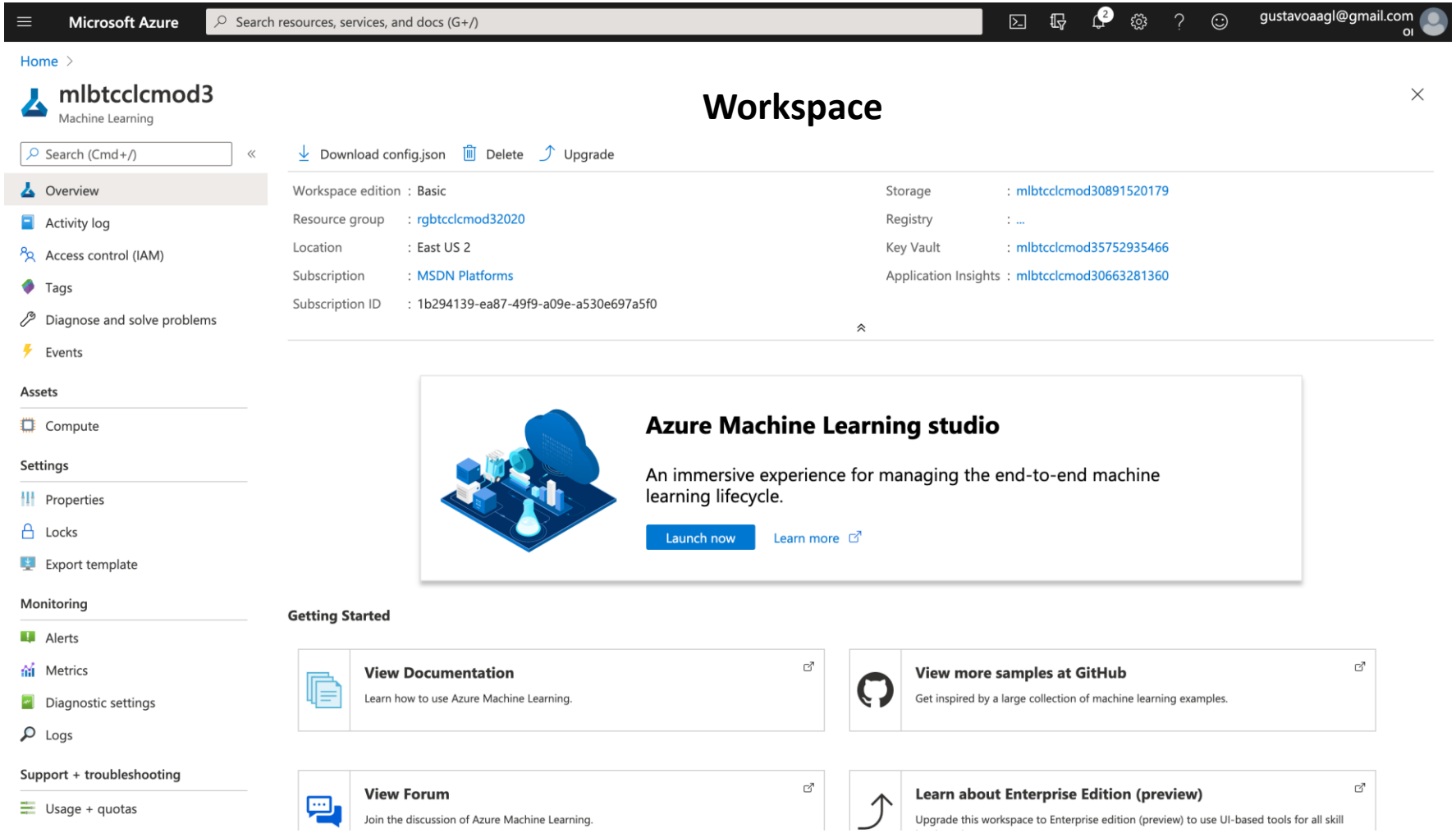
- Ambiente baseado em nuvem que pode ser usado para treinar, implantar, automatizar, gerenciar e rastrear modelos de ML;
- Pode ser usado para qualquer tipo de aprendizado de máquina, desde ML clássico até aprendizado profundo, aprendizado supervisionado e não supervisionado.



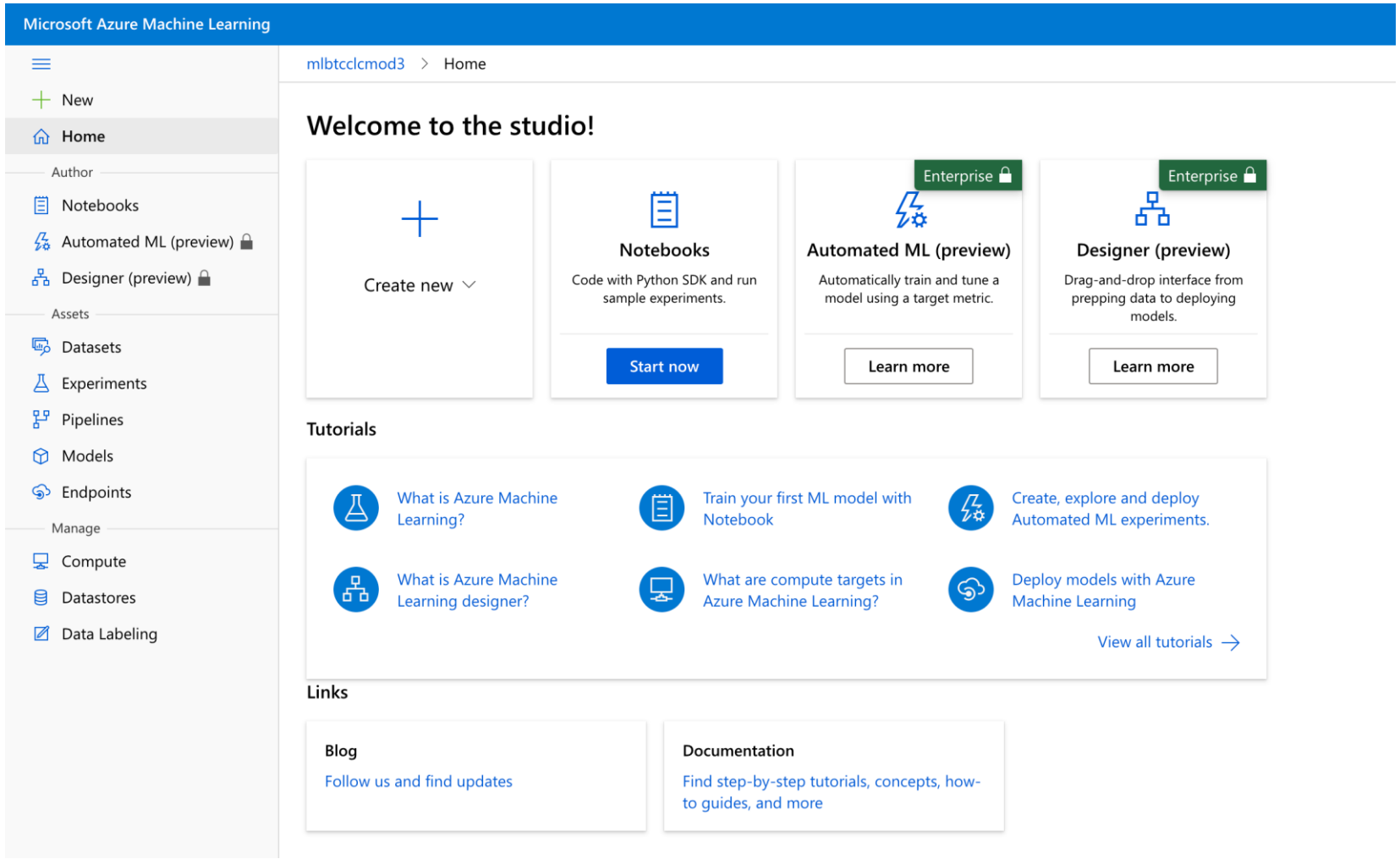
Azure Machine Learning



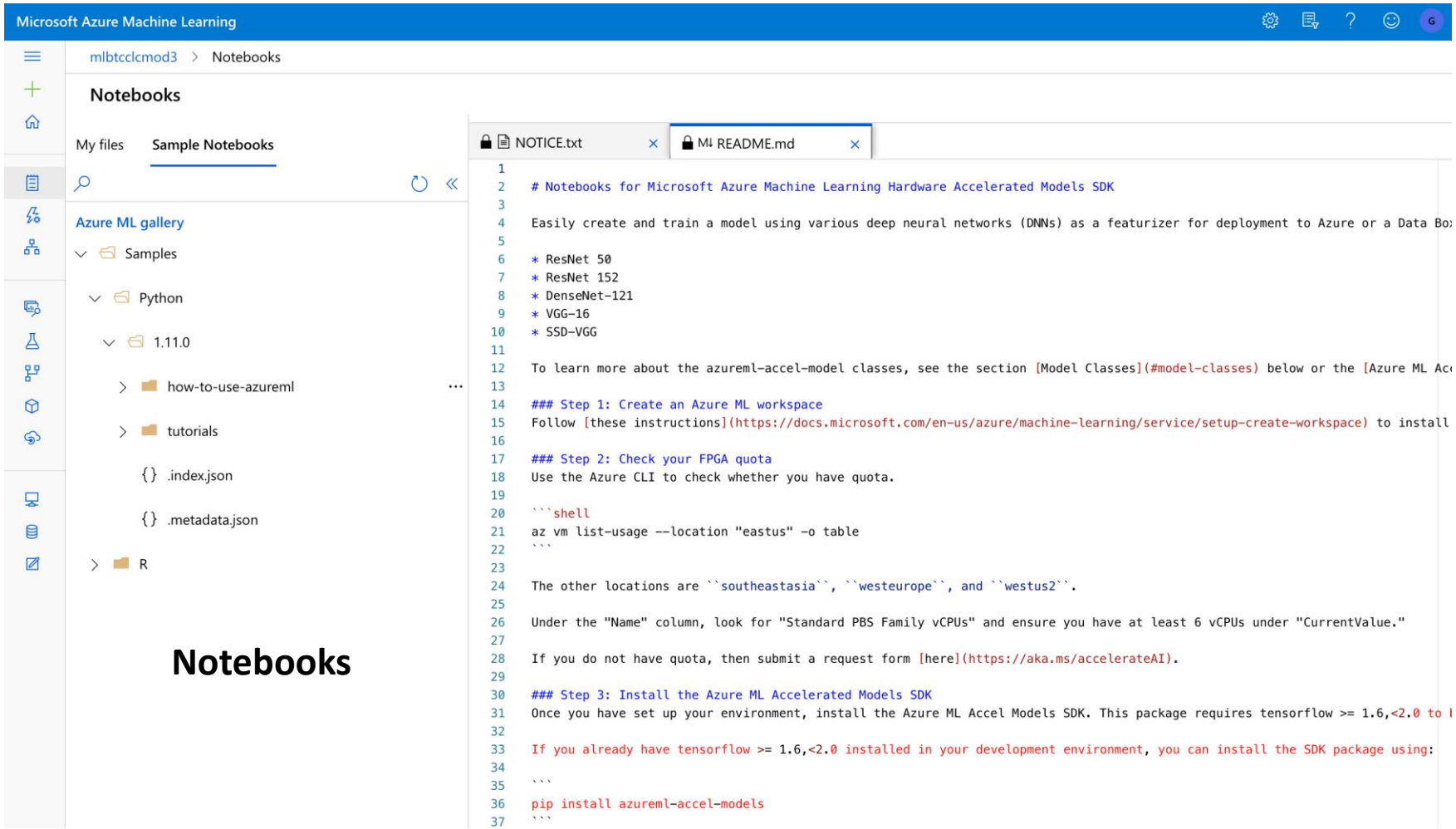
Azure Machine Learning



Azure Machine Learning Studio



Azure Machine Learning Studio

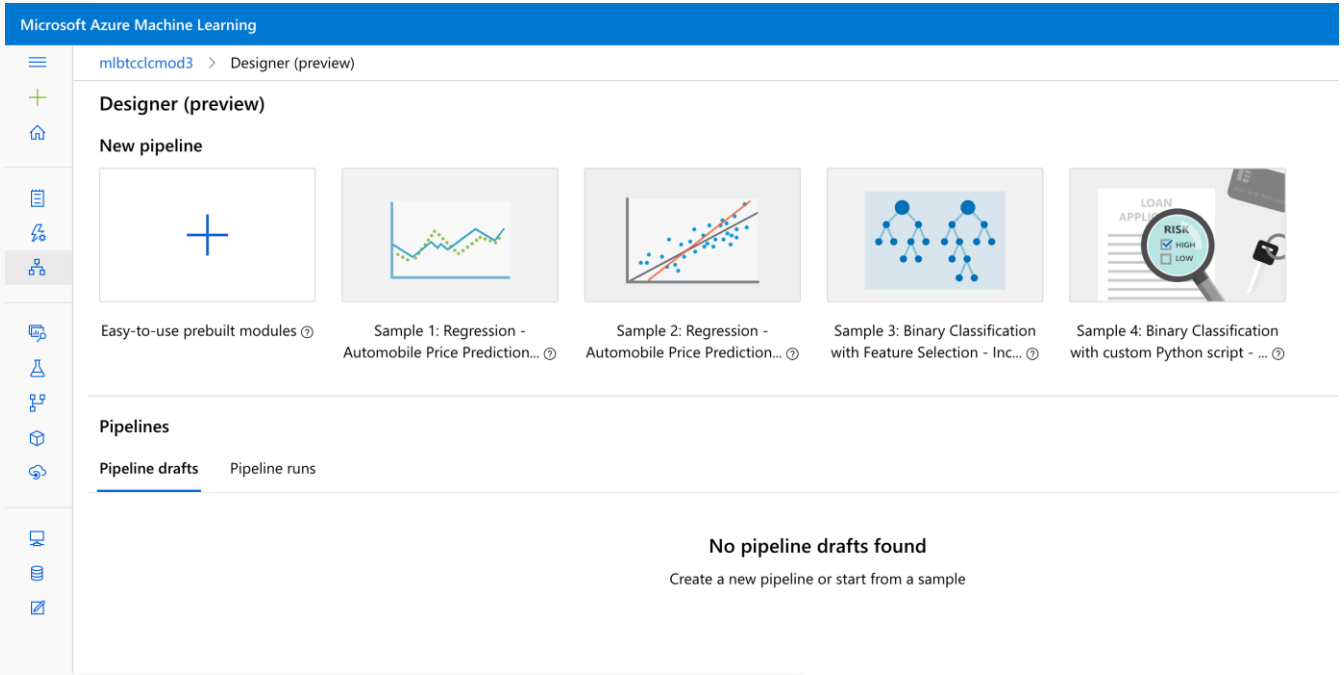


Notebooks

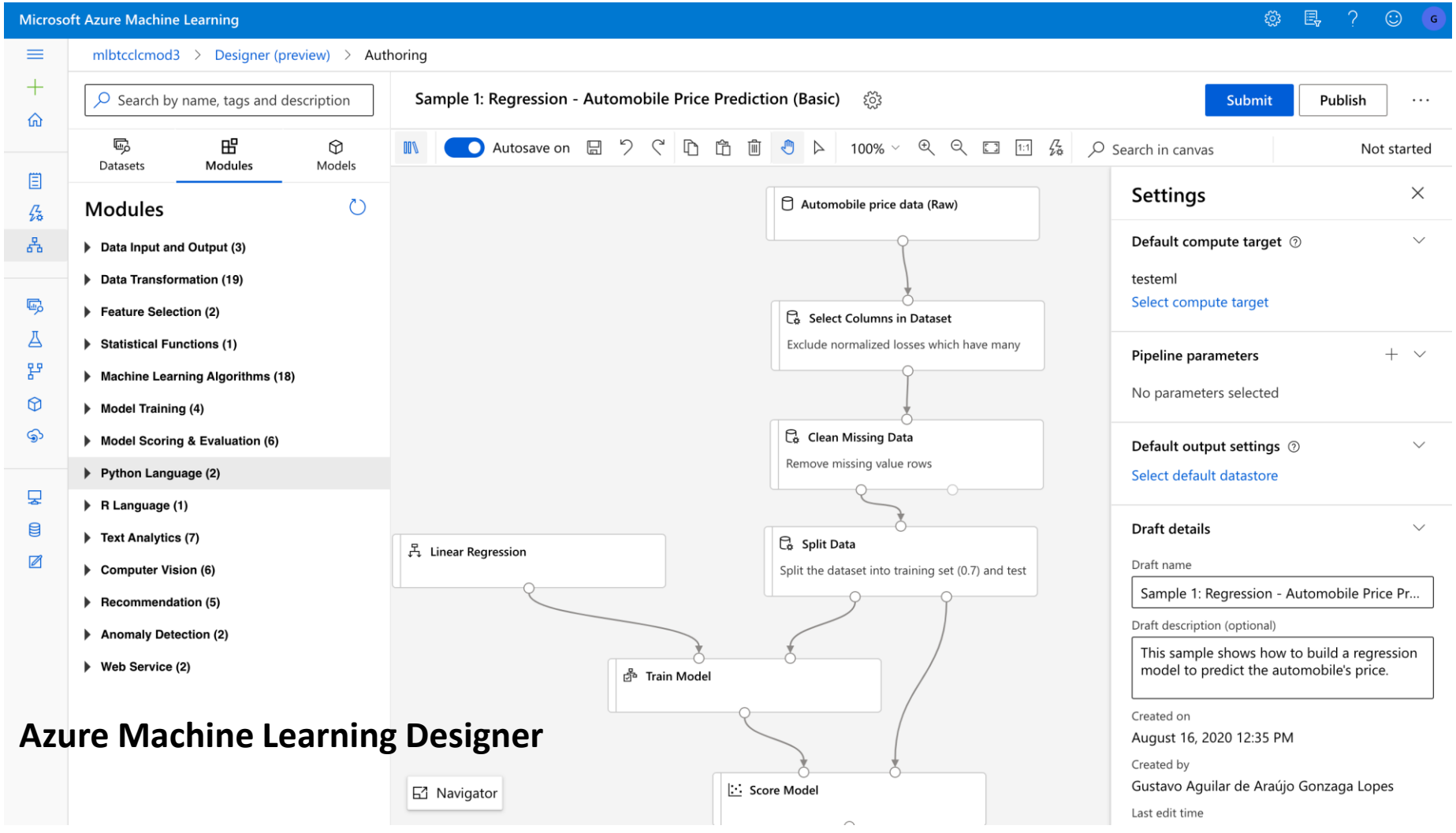
Azure Machine Learning

AZURE MACHINE LEARNING DESIGNER

- Preparar dados, treinar, testar, implantar, gerenciar e rastrear modelos de aprendizado de máquina sem escrever nenhum código.



Azure Machine Learning Studio



Azure Machine Learning Designer

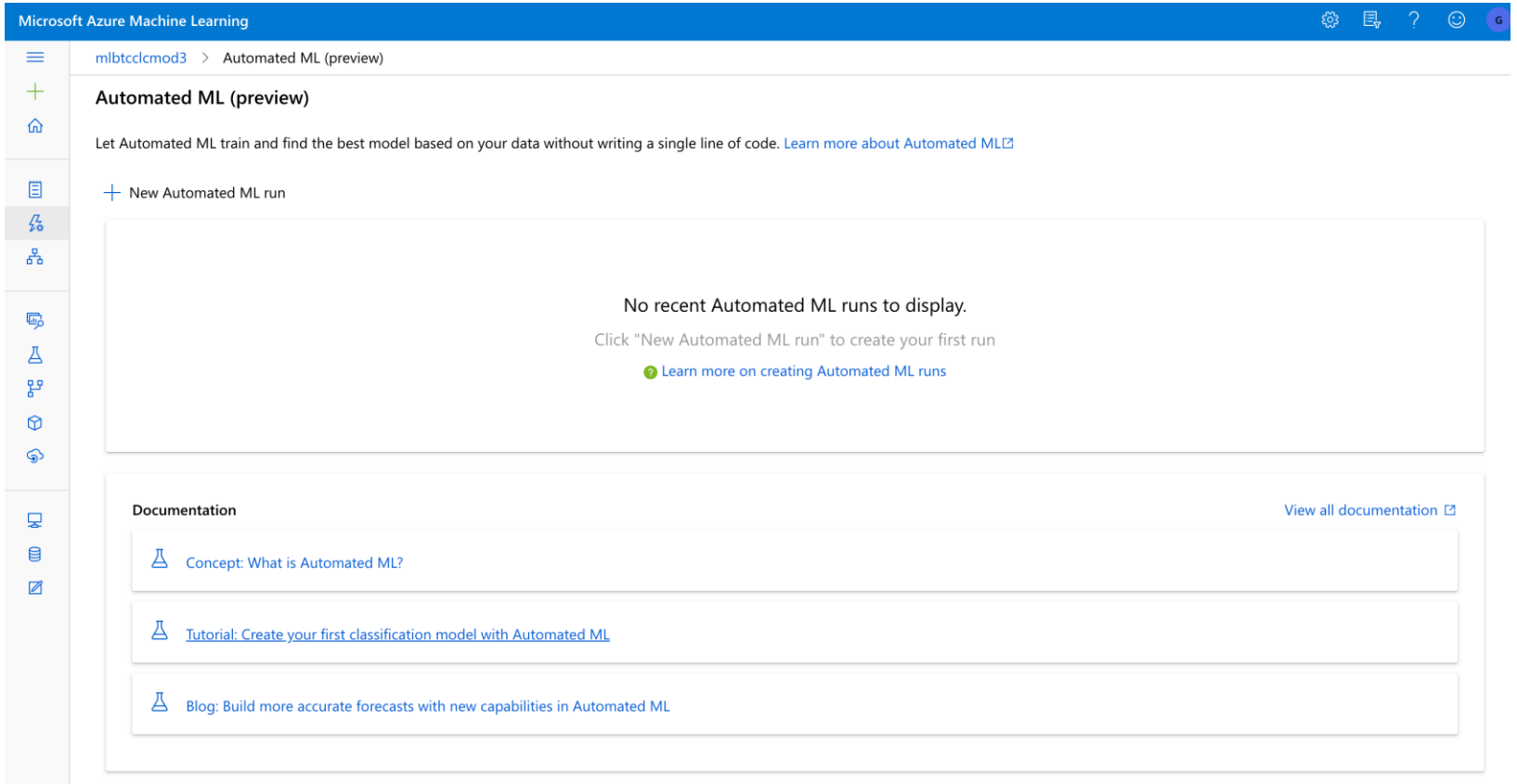
Azure Machine Learning

MACHINE LEARNING AUTOMATIZADO

- Automatizar tarefas intensivas e demoradas;
- Construção drag & drop (interface com componentes prontos);
- Realiza a iteração, de forma rápida, entre várias combinações de algoritmos e parâmetros, para ajudar a encontrar o melhor modelo com base em uma métrica selecionada;
- Somente na assinatura **Enterprise,** assim como o Azure Machine Learning Designer.

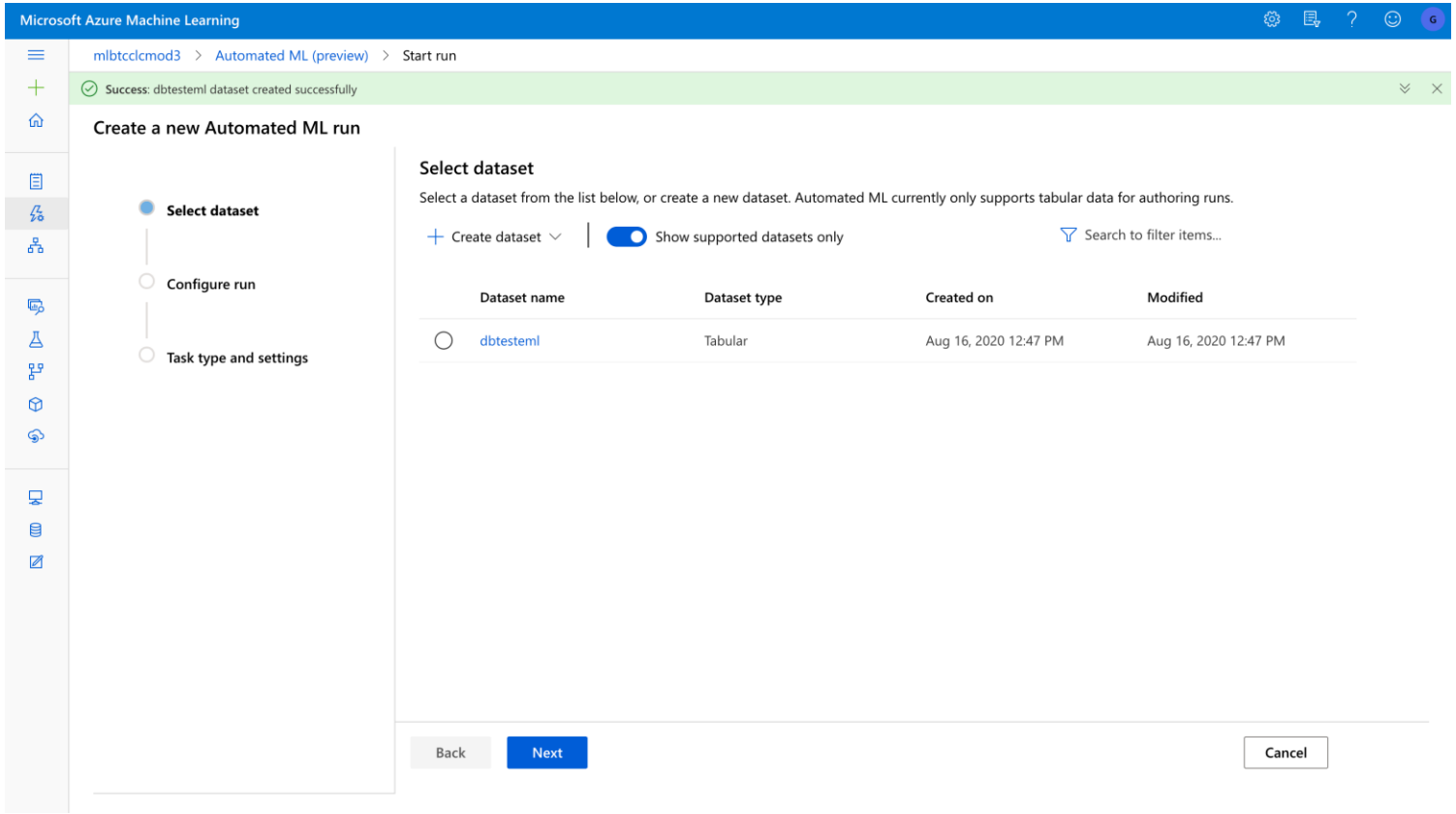
Azure Machine Learning

MACHINE LEARNING AUTOMATIZADO



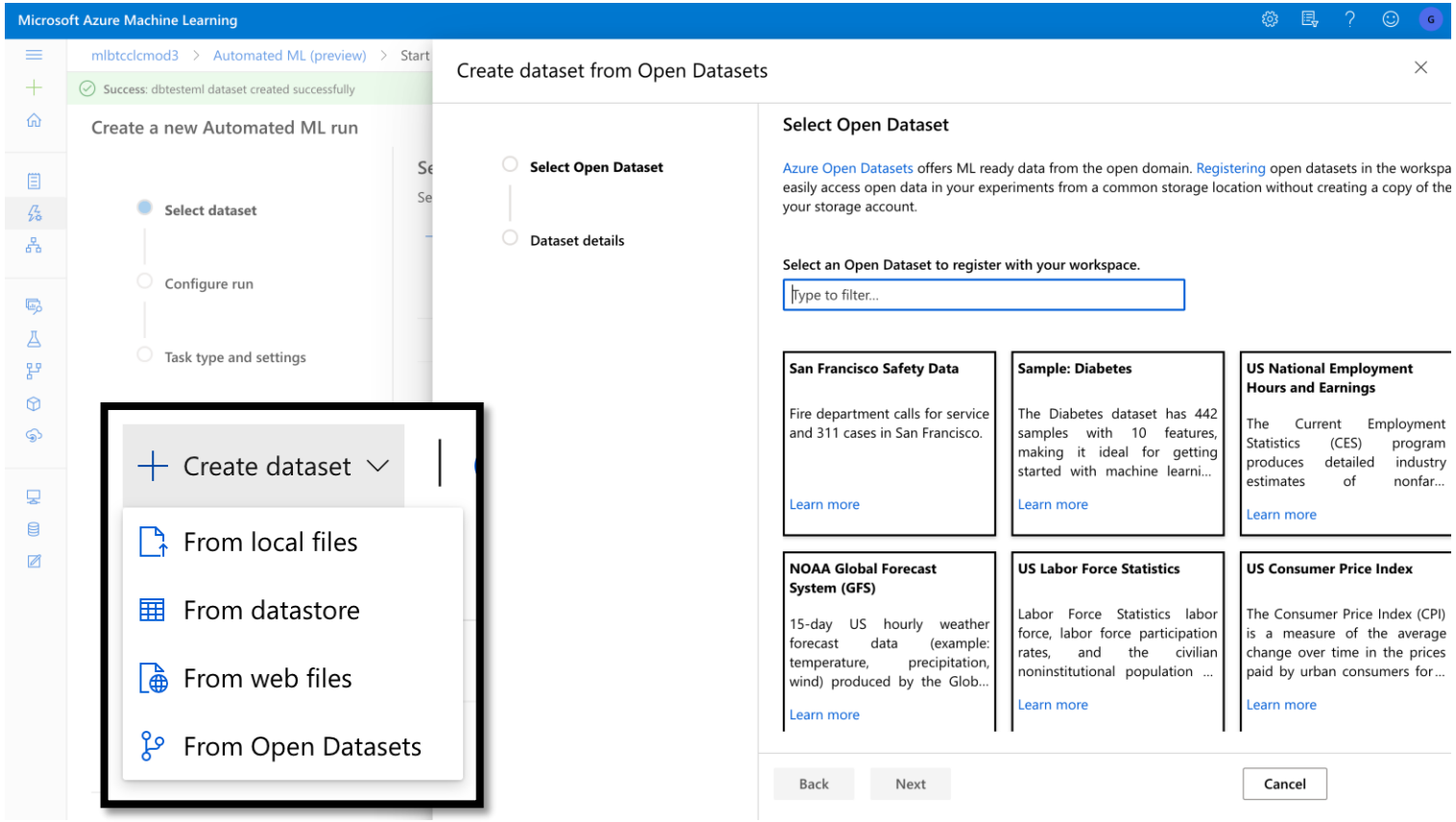
Azure Machine Learning

MACHINE LEARNING AUTOMATIZADO



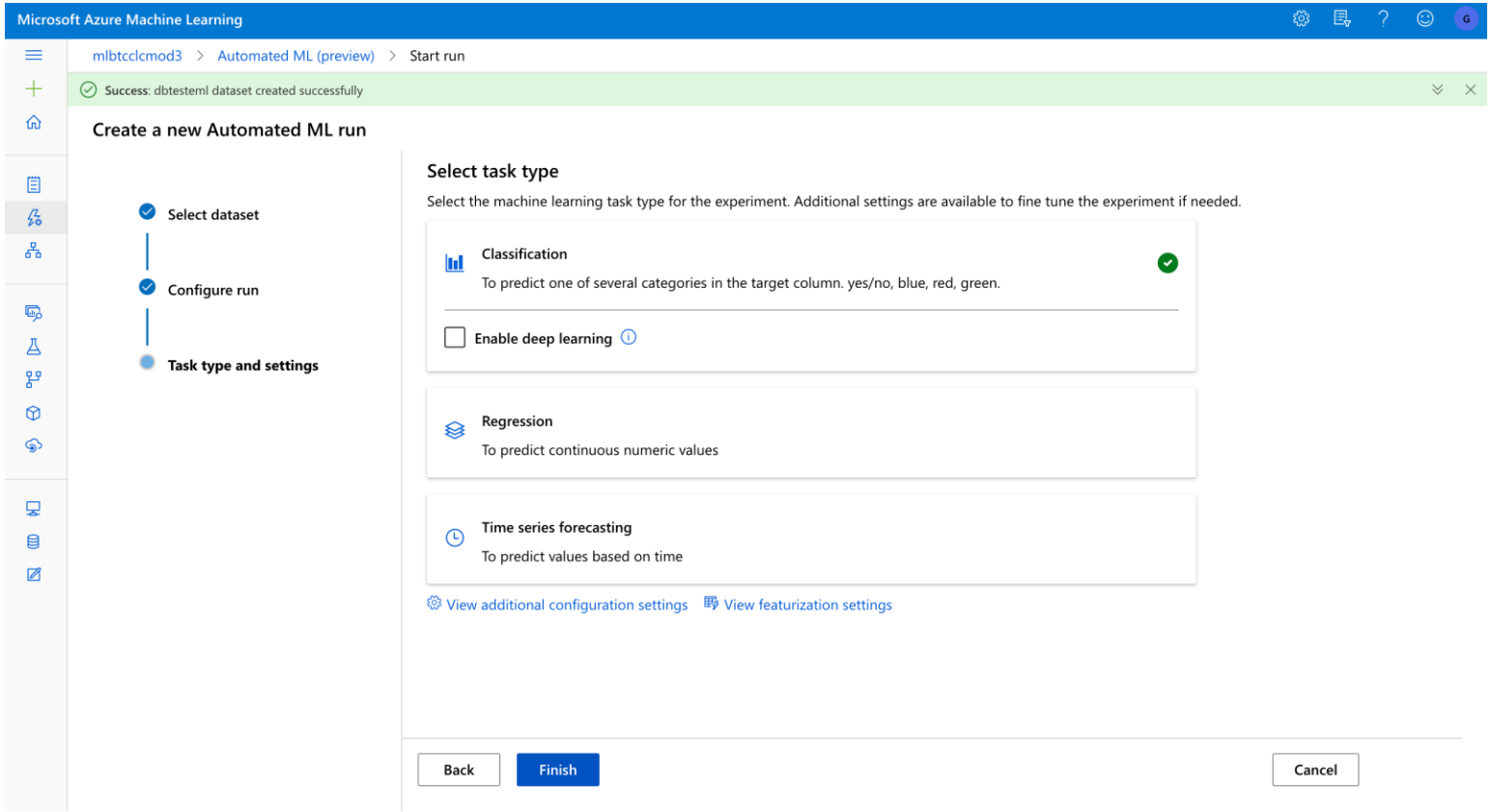
Azure Machine Learning

MACHINE LEARNING AUTOMATIZADO



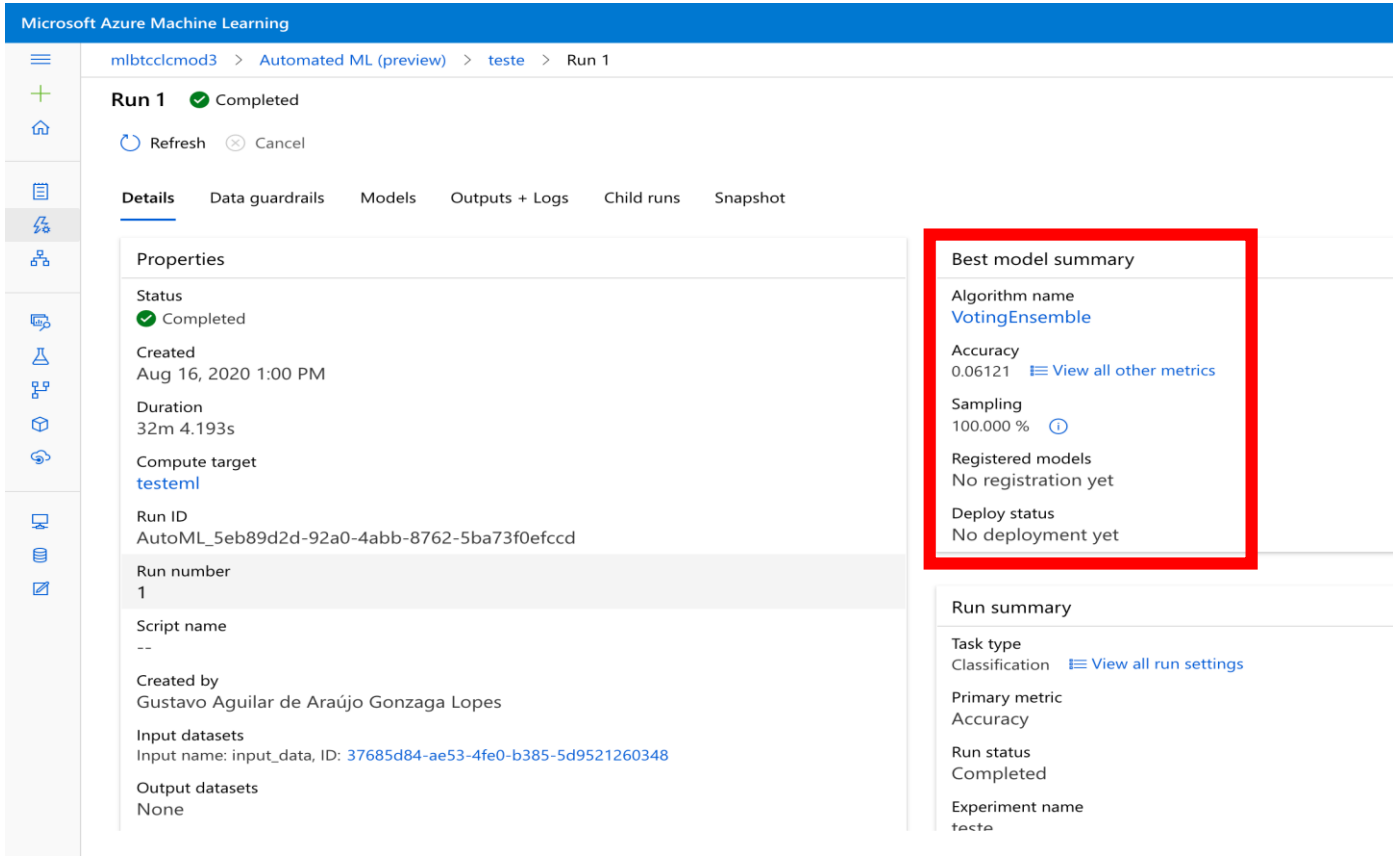
Azure Machine Learning

MACHINE LEARNING AUTOMATIZADO



Azure Machine Learning

MACHINE LEARNING AUTOMATIZADO



Soluções de Dados, Big Data e Machine Learning

FIM

PROF. GUSTAVO AGUILAR