Introdução a engenharia de dados na Azure

Decola Tech

Digital Innovation One

Introdução a engenharia de dados na Azure

**Engenharia de dados Azure**

**Objetivo do curso**

Conhecer e executar tarefas comuns à engenharia de dados utilizando os recursos fornecidos pela nuvem da Azure

**Percurso**

* Visão Geral Sobre a Engenharia de Dados
* Conhecendo a Plataforma Azure
* Ingestão de Dados

**Requisitos**

* Fundamentos da Arquitetura de Sistemas
* Curso SQL Server – Criando suas primeiras consultas
* Curso Aprenda o que são Estrutura de Dados e Algoritmos.
* Cadastro no Portal da Azure

**Visão Geral Sobre Engenharia de Dados**

Objetivos

* Conhecer o que é ciência de dados
* Identificar a necessidade de soluções de dados
* Descrever as áreas de conhecimentos relacionadas
* Familiarizar-se com os provedores em nuvem
* Conhecer as carreiras e certificações na área de dados

**Ciência de Dados**

A ciência de dados já existe a 30 anos e combina técnicas da ciência da computação com ferramentas estatísticas para analisar dados.

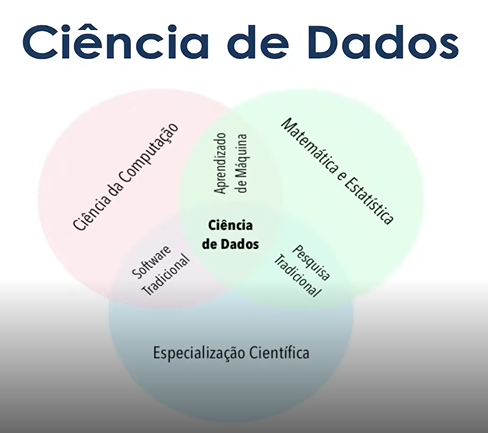
Tudo que realizamos é uma ação, e isso pode ser transformado em dados.

Utiliza a análise preditiva, que tenta descobrir o que vai e o que pode acontecer

**O que é dado?**

Dado é tudo aquilo que a gente pode realizar através de uma ação e isso pode ser transformado em um dado.

**Cientista de Dados**

O cientista de dados é o profissional especialista em analisar processos e volume de dados usando técnicas computação, matemática e estatística para gerar valor ao negócio das empresas e das organizações. 

**O que faz um cientista de dados?**

Responsável por extrair grandes volumes de dados de múltiplas fontes internas e externas, empregar tecnologias de análise, aprendizado de máquina e métodos estatísticos para preparar os dados para uso de modelagem preditiva e prescritiva, além de explorar fraquezas escondidas, tendências e oportunidades de um negócio.

**Modelagem Preditiva**

Utiliza modelos estatísticos para apontar como serão os dados no futuro, ou como se comportam em condições diversas.

Exemplo: um painel de **check list** preditivo aonde estamos querendo ver a probabilidade de um cliente com um perfil escolhido efetuar o pagamento. Vai se levar em consideração o número de parcelas pagas, o valor da compra, a idade do cliente, dias em atraso, vai levar no exemplo 4 variáveis ou informações para fazer o cálculo da probabilidade do pagamento desse cliente.

**Modelagem Prescritiva**

Enquanto a análise preditiva se limita a dizer como provavelmente sera o futuro, a análise prescritiva fornece subsídios para tomar decisões que irão alterar o futuro.

Exemplo: gera recomendações de compras, trocas e outras ações que produzem valor ao negócio.

O cientista de dados ele divide hoje parte de suas atividades com engenheiro de dados. Com o passar do tempo essa área foi se especializando. Antes tarefas que eram feitas por cientistas de dados, hoje é feita por pessoas especializadas como o engenheiro de dados.

**O que faz um engenheiro de dados?**

Cuida de todo o fluxo de dados desde a extração, passando pelo armazenamento, até servir aos dados em qualquer formato: estruturado, não-estruturado ou em tempo real.

Obs: esses dados podem ser estruturados, não estruturados, semi-estruturados ou em tempo real.

Obs 2: hoje a parte de extração e preparação, hoje não precisa ser mais executada por um cientista de dados, mas sim por um engenheiro de dados. E o cientista de dados vai se concentrar na modelagem prescritiva e preditiva já descritos.

**O que faz um cientista de dados?**

Responsável por fazer modelos de modelagem preditiva e prescritiva, identificando fraquezas escondidas, tendências e oportunidade de negócios.

**Para saber mais**

[www.aquare.la/o-que-e-ciencia-de-dados-data-science-para-negocios](http://www.aquare.la/o-que-e-ciencia-de-dados-data-science-para-negocios)

**Soluções de dados**

**O que são** **dados**?

Dados são coleção de fatos, como números, descrições e observações usadas na tomada de decisão. Podemos classificar os dados como **estruturados, semiestruturados ou não estruturados.**



Figura 1 Exemplo de Dados Estruturados

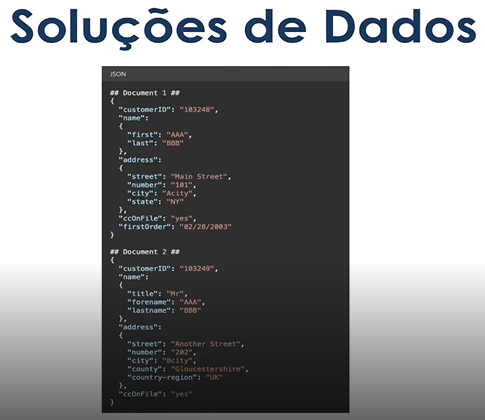


Figura 2 Exemplo de Dados Semiestruturados



Figura 3 Exemplo de Dados Não Estruturados

Exemplo de dados não estruturados: e-mail, mp3, twite, pesquisa, wifi, chat, etc...

**Como os dados são definidos, armazenados e acessados na computação em nuvem?**

* **Dados Estruturados** são armazenas dos em Banco de Dados Relacional. Exemplo: SQL Server, SQL Azure (nuvem)
* **Provisionamento** é o ato de configurar um banco de dados.
* **Dados semiestruturados** são armazenados em serviços como Azure Cosmos DB
* Depois que o serviço é provisionado, é preciso configurar esse serviço para que os usuários possam receber acesso a esses dados. Geralmente vc pode definir vários níveis de acesso como só leitura, leitura e gravação, privilégio de proprietário, definir quais usuários terão permissão de acesso aos dados em primeiro lugar.
* Os usuários individuais não precisaram acessar o sistema diretamente mas farão edições por meio de um aplicativo. Já os analistas e os gerenciadores de dados terão privilégios de proprietários porque precisam gerenciar o acesso de outros usuários e administrar o sistema.

O que é um sistema transacional

Um sistema transacional é geralmente o que a maioria das pessoas considera a principal função da computação empresarial. O Sistema transacional registra transações. Uma transação pode ser financeira, de varejo como acompanhar pagamentos de bens e serviços prestados. Pense na transação como uma unidade de trabalho pequena e discreta.

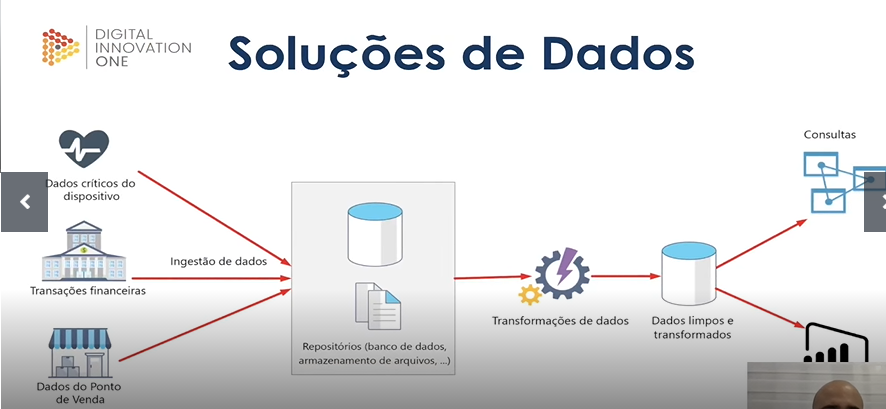
Sistemas transacionais são geralmente de alto volume, as vezes manipulando muitos milhões de transações em um dia. Os dados que estão sendo processados têm de estar acessíveis com rapidez. O trabalho executado por sistemas transacionais é geralmente conhecido por OLTP, ou seja, processamento de transações online.

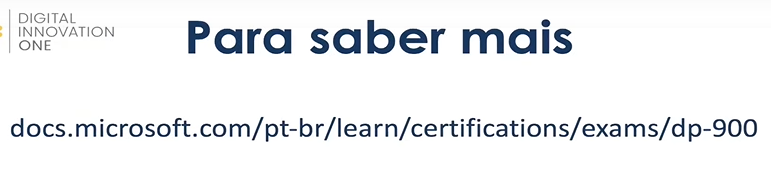
Para dar suporte ao processamento rápido, os dados de um sistema transacional geralmente são divididos em pequenas partes. Por exemplo, se você estiver usando um sistema relacional, cada tabela envolvida em uma transação, contem apenas as linhas necessárias para realizar a tarefa transacional. No exemplo de transferência bancária, uma tabela que contém informações sobre os fundos existentes na conta pode conter apenas o numero da conta e o saldo atual. Outras tabelas não envolvidas na operação de transferência armazenaram informações como nome endereço do cliente e histórico da conta. Dividir a tabela em grupos separados de colunas dessa maneira é um pratica conhecida como normalização. A normalização pode permitir que um sistema transacional armazene em cash grande parte das informações necessárias para executar transações na memoria acelerando a taxa de transferência. Embora normalização permita uma taxa de transferência rápida para as transações ela pode deixar as consultas mais complexas. As consultas que envolvem tabelas normalizadas geralmente precisam reunir de volta dados armazenado em várias tabelas distintas. Isso pode dificultar a análise de usuários empresariais que precisam examinar os dados.

**O que é um sistema analítico**

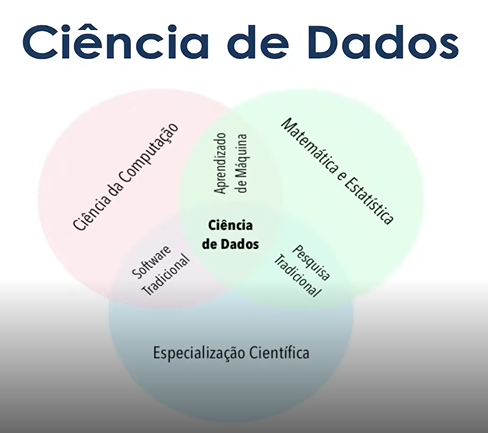
Sistema analítico em contraste com sistemas projetados para dar suporte ao OLTP, um sistema analítico foi projetado para dar suporte a usuários empresariais que precisam consultar dados e obter uma visão do panorama geral das informações mantidas em um banco de dados.

Os sistemas analíticos se preocupam em captura de dados brutos e o seu uso para gerar insights, idéias. Uma organização pode usar essas ideias para tomar decisões empresariais. Por exemplo, os insights detalhados de uma empresa de manufatura podem indicar tendências permitindo que elas determinem em quais linhas de produtos se concentrar para aumentar a rentabilidade. A maioria dos sistemas de dados analíticos precisam executar tarefas semelhantes que seriam **ingestão de dados, transformação de dados, consulta de dados e visualização de dados**.





**Áreas de conhecimento**



**Matemática e estatística:** a armadilha aqui é que muitas coisas na estatística e na matemática são contra intuitivas, e sem uma formação específica, podemos cometer erros grandes. Como cientistas de dados buscam por correspondências e associações, é mito importante calcular corretamente essas probabilidades. Por isso a formação em matemática é importante na ciência de dados e também na engenharia de dados.

**O segundo elemento** é o conhecimento do campo. O pesquisador deve conhecer a área que está trabalhando. Aprender **Radoupi**(não sei como se escreve isso), que é um software comum para lidar com big data, é mais fácil que aprender sobre os negócios da empresa. Isso ressalta a necessidade do conhecimento de campo na ciência de dados.

O terceiro elemento do diagrama é a programação de computadores. Mas não precisa fazer doutorado em ciência da computação, não precisa complicar.

Pesquisa tradicional: quando um pesquisador trabalha na sua área de especialização, usando ferramentas comuns para trabalhar com formatos de dados conhecidos. É extremamente produtivo e quase todas as pesquisas existentes foram feitas assim.

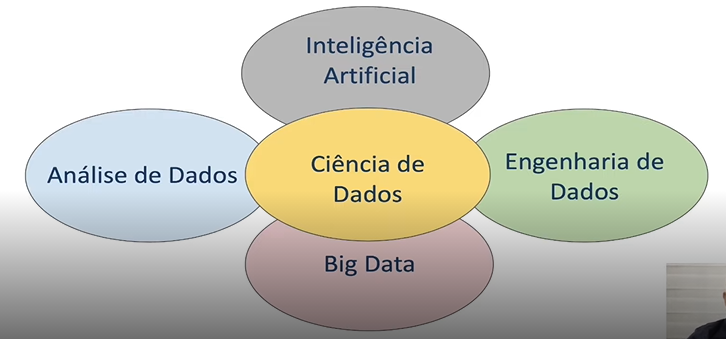
Métodos tradicionais são importantes, mas não são suficientes para trabalhar com Big Data.

A segunda combinação é a estatística com a programação sem a especialização científica. É o aprendizado de máquina, a inteligência artificial.

**Aprendizado de máquina** é quando o algoritmo ou um programa se atualiza e evolui para realizar uma tarefa analítica específica. Exemplo: filtro de spam.

A terceira combinação é o conhecimento do campo e programação sem estatística. Esse caso raramente ocorre por outro lado traz contribuições importantes para a ciência de dados como a contagem de palavras, é simples, são procedimentos que não exigem estatísticas sofisticadas. Apenas contabilizamos as ocorrências de items e obtendo percepções importantes.

E por último tem a combinação de todas, que forma a ciência de dados.



Consumidores geram e usam dados e os usam para tomar decisões.

**Big Data**

Big Data é um conceito que descreve o grande volume de dados estruturados e não estruturados que são gerados a cada segundo. Para muitos, o conceito de Big Data é algo novo mas mesmo antes de existir qualquer meio digital ou tecnologias computacionais, os dados já eram gerados. A diferença é que nos dias de hoje geramos muito mais dados do que no passado. Hoje já é realidade dispositivo distintos(carro, geladeira, relógios) interconectados entre si, demandando ainda mais dados a serem processados e transformados em informações úteis. O diferencial do Big Data esta atrelado a possibilidade de cruzar esses dados por meio de diversas fontes para obter insights, idéias rápidas e precisas. A exigência dos consumidores e o aumento da competitividade em todos os mercados nos força a inovar e ter esse tipo de caminho como premissa básica dos negócios.

Por isso que o Big Data é tão importante nos dias de hoje. Através dele conseguimos saber informações de mercado por meio de nossos consumidores extraindo o que eles estão dizendo, sobretudo o que você faz, insatisfações, satisfações, desejos, necessidades, são possíveis de captar em mídias sociais e cruzar dados internos de sua empresa e assim cria propagando, modelo de negócios, serviços. A essência do conceito está em gerar valor para os negócios.

**Análise de dados**

O que é análise de dados e por quê ela é importante?

Análise de dados é o ato de transformar dados em informação. O grande acúmulo de dados sem o tratamento adequado não garante competitividade, assim, a relevância da análise de dados se encontra justamente em interpretar as informações e conseguir elaborar estratégias a partir delas. Entender e saber estruturar uma análise de dados garante a empresa uma boa tomada de decisão. Dessa forma, a capacidade de prever, mesmo que parcialmente o comportamento do consumidor, é muito apreciada. Da mesma forma há como vislumbrar o comportamento do mercado, usando como base padrões do passado. Além disso é possível incrementar sistemas computacionais para analisar dados em tempo real.

Para executar a análise de dados corretamente, precisamos seguir alguns passos. Esses passos são:

* **Coletar os dados** – A coleta pode acontecer de diversas formas, como questionário e sistema de estoque da empresa.
* **Agrupar os dados por tópicos** – É importante para facilitar a correlação dos dados.
* **Utilizar ferramentas que buscam correlação entre esses dados**.
* **Montar gráficos com informações encontradas pela correlação** - Os gráficos devem ser montados tanto dos dados sem correlação quanto dos relacionados. Dessa forma fica mais fácil de entender o comportamento do sistema.
* **Interpretar os gráficos de acordo com o sistema analisado** – Com os dados devidamente separados é possível fazer uma análise mais precisa.
* **Montar um relatório de análise** – A criação e interpretação do relatório são fundamentais para os gestores da empresa. Tudo deve ficar bem explicado, indicando resultados importantes e possíveis usos estratégicos.

**Inteligência Artificial**

Também conhecido como **IA**, refere-se a sistemas ou maquinas, que imitam a inteligência humana para executar tarefas que podem se aprimorar continuamente com base nas informações que coletam.

A **IA** se manifesta de várias formas, exemplo: os **chatbots** usam inteligência artificial para entender os problemas dos clientes mais rapidamente e fornecer respostas mais eficientes. Os assistentes inteligentes usam IA para analisar informações críticas de grandes conjuntos de dados de texto livre para melhorar a programação. Os mecanismos de recomendação podem fornecer recomendações automatizadas para programas de tv com base nos hábitos de visualização dos usuários. Ela está rapidamente se tornando o pilar da inovação, se alimentando de diversas formas de Machine Learning e reconhece padrões e dados para fazer previsões. A IA pode agregar valor ao negócio fornecendo uma compreensão mais abrangente sobre a abundância de dados disponíveis e contar com previsões para automatizar tarefas excessivamente complexas.

Para saber mais **docs.microsoft.com/pt-br/learn/certifications/exams/dp-203**

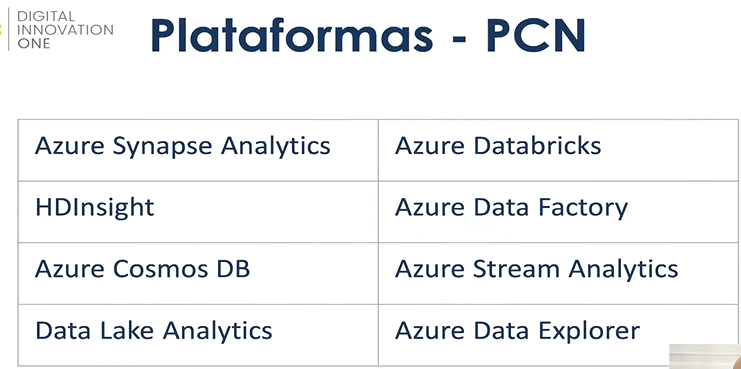
**Provedores CSP**

CSP (Cloud Service Providers)

**Plataformas – PCN**

Plataformas de Computação em Nuvem são destinadas à execução de aplicativos e serviços baseados nos conceitos de computação remota.

* Amazon Web Services – AWS
* Microsoft Azure
* Google Cloud Platform –CGP
* Alibaba Group
* Oracle Cloud
* IBM Cloud



**O que é computação em nuvem?**

Computação em nuvem é o fornecimento de serviços de computação, incluindo servidores, armazenamento, banco de dados, rede, software, análise e inteligência, pela internet para oferecer recursos flexíveis e economias em escala.

**E a contratação?**

Normalmente apenas se paga pelos serviços em nuvem em uso, ajudando a reduzir os custos operacionais, conforme as necessidades da empresa.

**IaaS (infraestrutura como serviço)**

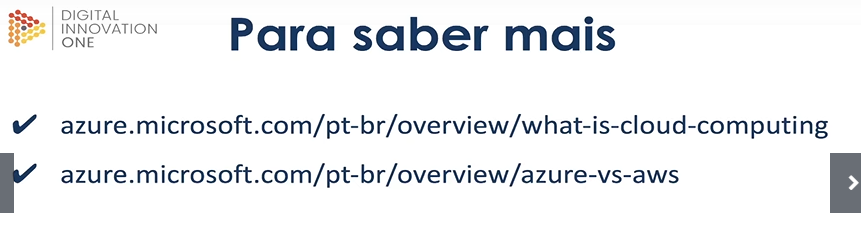
A categoria mais básica de serviços de computação em nuvem. Com a IaaS, você aluga a infraestrutura de TI, (que inclui servidores e máquinas virtuais, armazenamento (VM’s), redes e sistemas operacionais), de um provedor em nuvem, com pagamento conforme o uso.

**PaaS (Plataforma como serviço)**

Oferece serviços de computação em nuvem sobre demanda para desenvolvimento, teste e gerenciamento de aplicativos de software. A PaaS cuida de toda a configuração e gerencia da infraestrutura dos servidores, armazenamento, rede e banco de dados necessários para desenvolvimento.

**SaaS (Software como Serviço)**

É um método para a distribuição de aplicativos de software pela internet sob demanda normalmente baseado em assinaturas. O SaaS permite que os usuários conectem o aplicativo pela internet, normalmente com um navegador da web em seu telefone, tablete ou pc.



Carreira e Certificações

Há três funções principais nas quais os engenheiros de dados podem se encaixar:

* Generalista
* Centralizado em pipeline
* Centralizado no Banco de Dados

Algumas das tarefas e responsabilidades do Engenheiro de Dados

* Desenvolvimento, criação, teste e manutenção de banco de dados e estruturas de dados.
* Alinhamento da arquitetura de dados com requisitos empresariais.
* Aquisição de dados
* Desenvolver processos para criar e recuperar informações de conjunto de dados
* Usar linguagem e ferramentas de programação para examinar os dados.
* Examinar maneiras de aprimorar a confiabilidade, eficiência e qualidade dos dados.
* Conduzir pesquisas para questões relacionadas aos negócios ao setor
* Implantar programas de análise sofisticados, machine learning e métodos estatísticos.
* Preparar os dados para uma modelagem prescritiva e preditiva

Engenheiro de dados tem formação em Ciência da Computação, Engenharia, Matemática Aplicada ou diplomas relacionados a outras áreas de TI.

Salário de Engenheiro de dados (Ano base de 2021)

* Junior - R$ 5 000,00
* Pleno – R$ 7 000,00
* Senior – R$ R$ 11 000,00

**Certificações Azure**

DP -900: Azure Data Fundamentals

DP -900: Azure Data Engineer Associate

**Por que usar a Microsoft Azure?**

Os engenheiros de dados precisam se envolver em uma integração eficaz de aplicativos para atingir seus objetivos de negócios. Plataformas em nuvems como a da Microsoft são construídas para se envolverem especificamente em dados e análises. A Microsoft Azure pode ser descrita como uma solução de serviços em nuvem em contínua expansão, evoluindo esses serviços ajudando empresas a enfrentar seus desafios de negócios com eficácia.

Você pode construir qualquer tipo de aplicativo e implantá-lo utilizando os conjuntos de habilidades e ferramentas existentes de sua equipe e implantá-lo em qualquer lugar.

**Serviços não relacionais do Azure**: o armazenamento de blobs é certamente a oferta mais madura: altamente disponível, durável, estrutura plana (sem arvore ou estrutura hierárquica).

**Armazenamento Azure Data Lake**: para quem procura uma estrutura hierárquica, ele usa o armazenamento de blob como sua base. Essa solução é extremamente útil ao lidar com sistemas de processamento de big data

**Azure Cosmos DB**: Outra opção de banco de dados não relacional disponível no Azure, é o cosmos db. Essa solução é impressionante pois pode dimensionada tanto quanto forem necessário sem nunca comprometer a flexibilidade, ou o desempenho da solução. Ele oferece suporte a diferentes tipos de modelos de dados. Desde gráficos a colunas largas. Valores chave e modelos de documentos. Além disso ele oferece suporte a 5 níveis de consistência: de forte a eventual.

**Azure Data Factory**: facilita a cópia de dado sem armazenamentos de dados, como de blob, para bancos de dados SQL. Ele também permite que você transforme dados com ajuda de outras ferramentas

Azure Data Briks: Com um serviço de análise gerenciado, o Azure Data Bricks é criado no Apache Spark, ferramenta conhecida na área de ciência de dados.

Plataforma de dados e ferramentas de análise

**HD Insights** – é utilizado para gerenciar clusters de nuvens usando o radoupi, o Hbase

**Machine Learning Studio** – é pra construir, implantar e gerenciar perfeitamente soluções de análise preditiva.

**Azure Stream Analytics** – é pra processamento de fluxo de dados em tempo real de milhões de **dispositivos.**

**Azure Data Lake Analytics** – é um serviço de análise distribuída que facilita o big data.

Algumas ferramentas relacionadas

* **Serviços cognitivos**, para permitir a interação contextual e adicionando recursos de API
* **Serviços de bot do Azure**
* **Serviço de bot** **sem servidor** que é inteligente e pode ser ampliado ou diminuído sob demanda
* Ferramentas do desenvolvedor do Azure – criar, implantar, diagnosticar e gerenciar aplicativos e serviços escalonáveis em várias plataformas



**Conhecendo a Plataforma Azure**

Objetivos

* Conhecer o Portal Azure
* Cadastra-se no Portal
* Explorar os recursos disponíveis
* Localizando os cursos

Recursos disponíveis

* Data Factories
* Reservas
* Data Catalog
* Data Lake Analytics
* Data Lake Storage Gen1
* Data Box
* On-Premises Data Gateways
* Datalog
* Azure Databricks
* LiveDataMigrators
* Azure Synapse Analytics

Cursos **Azure**

Site: <https://docs.microsoft.com/pt-br/>

**Ingestão de dados**

Objetivos

* Conhecer o que é a ingestão de dados
* Entender o funcionamento da ingestão de dados
* Explorar a ingestão de dados na Azure

**Ingestão de Dados**

É o processo usado para carregar ou importar dados de uma ou mais fontes para um serviço de armazenamento de dados.

Em um sistema de Big Data, a ingestão de dados precisa ser rápida o suficiente para capturar as grandes quantidades de dados que chegam e para ter capacidade de computação suficiente para processar esses dados em tempo hábil.

Quais são os serviços de Ingestão de Dados que a Azure oferece?

* **Azure Data Factory** – é um serviço de ingestão e de transformação de dados que permite carregar dados brutos das mais diversas fontes, tanto locais como na nuvem.
* PolyBase
* AQL Server Integration Services
* Azure Databricks

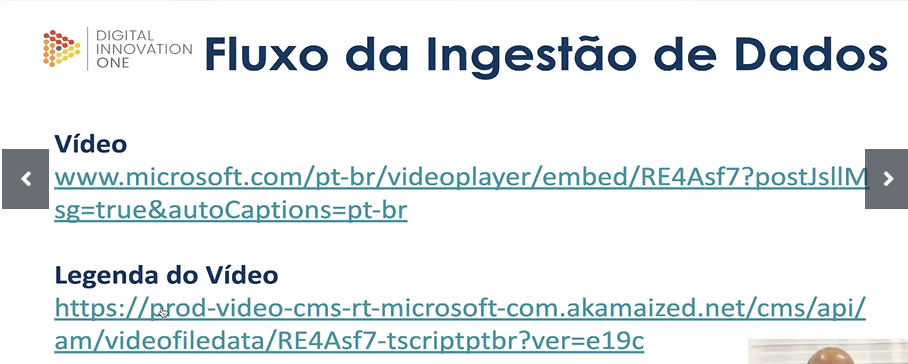
O que é pipeline?

Um pipeline é um agrupamento lógico de atividades que, juntas executam uma tarefa. As atividades de um pipeline definem as ações a serem executadas nos seus dados.

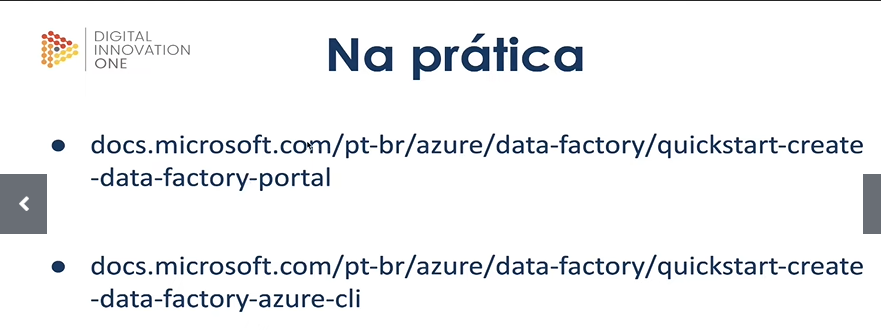
Fluxo da Ingestão de Dados

Os fluxos de dados são transformações visualmente projetadas no Visual Factory. Eles permitem que os engenheiros de dados desenvolvam a lógica de transformação de dados sem escrever código.

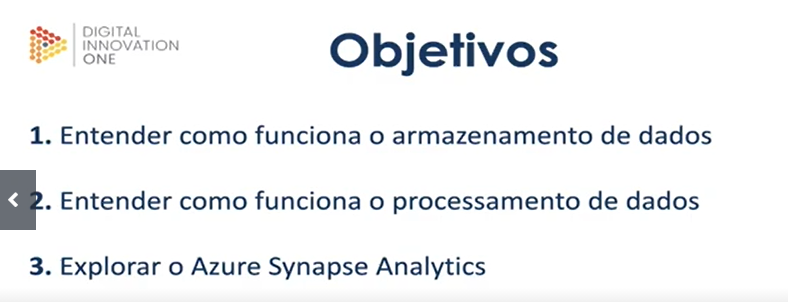
As atividades de fluxo de dados incluem agendamento, controle, fluxo e monitoramento do Azure Data Factory.



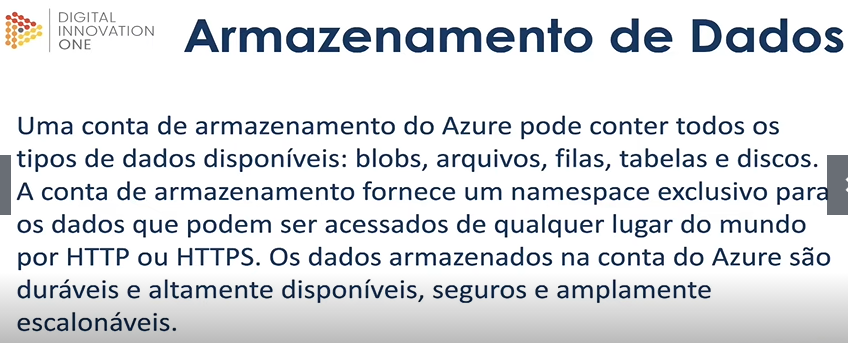
Explorando o funcionamento



**Armazenamento e processamento de Dados**

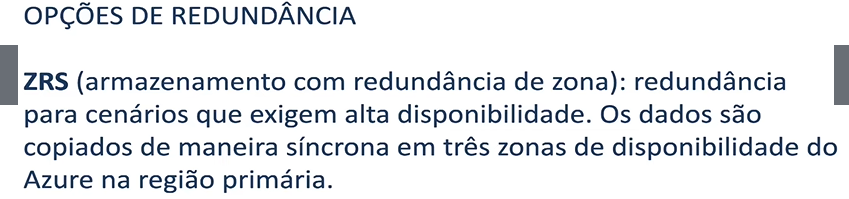


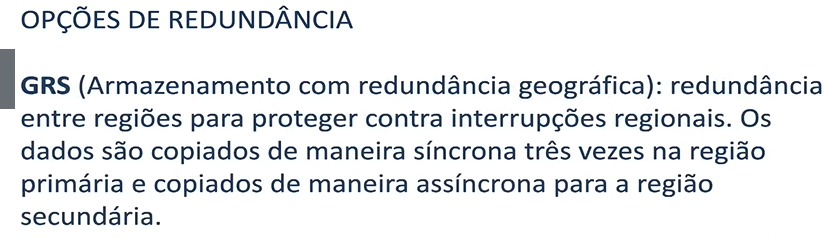
**Armazenamento de Dados**

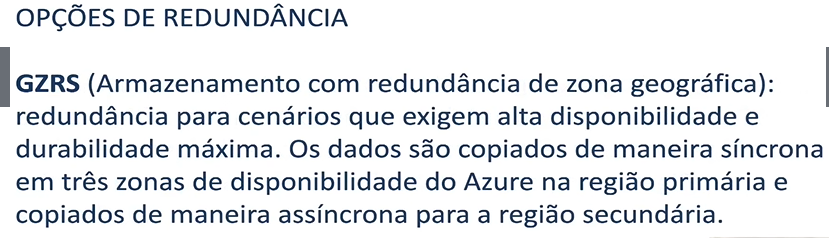








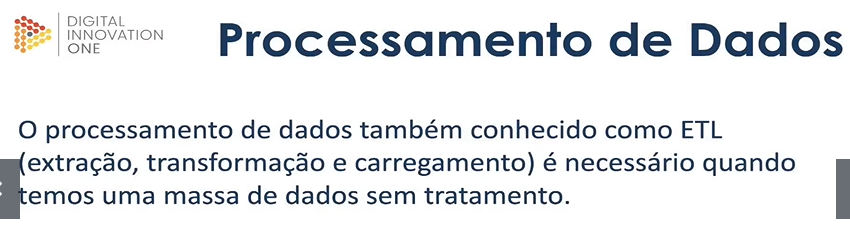


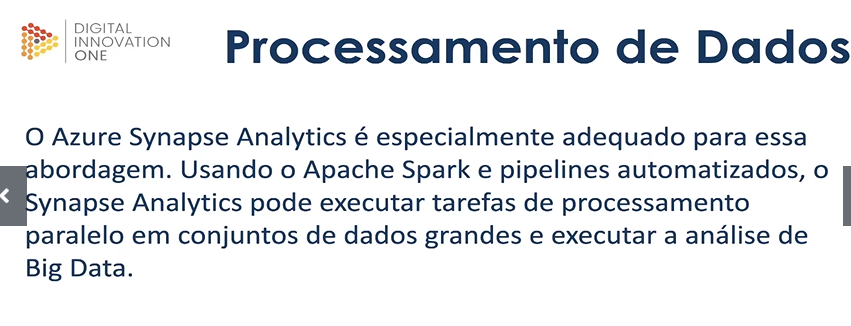


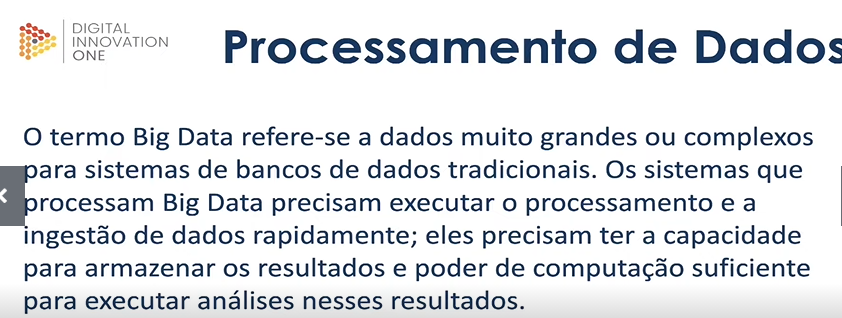
Para saber mais

Acesse: <https://docs.microsoft.com/pt-br/azure/storage/blobs>

**Processamentos de dados**

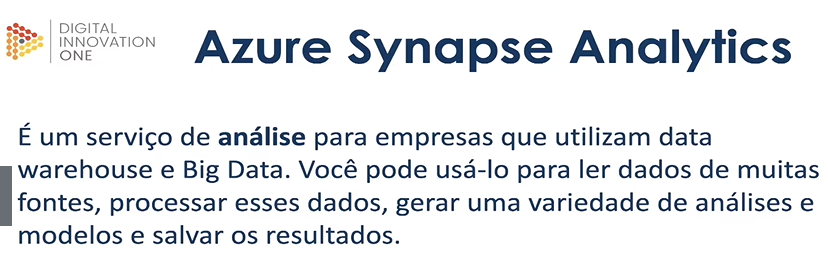


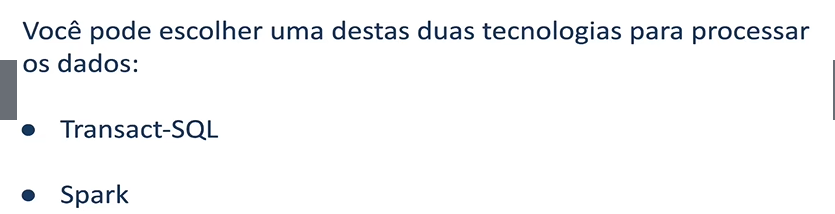


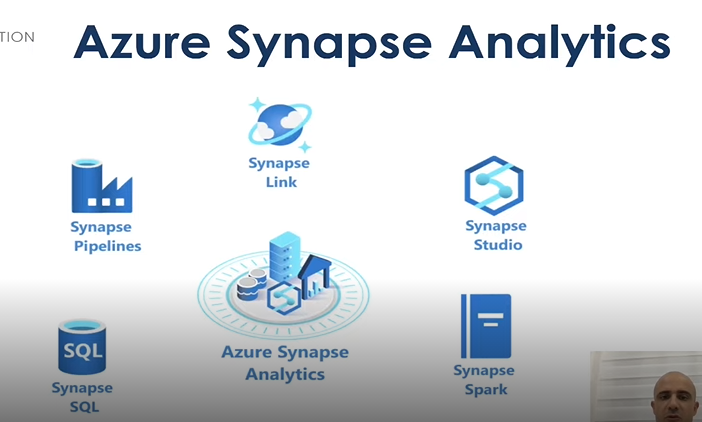


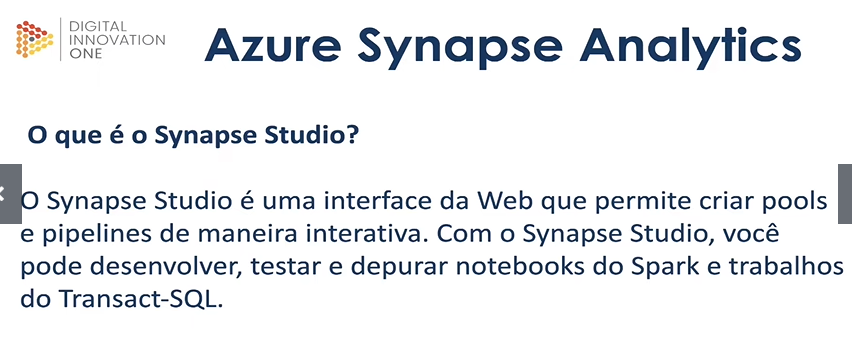


Azure SynapseAnalytics











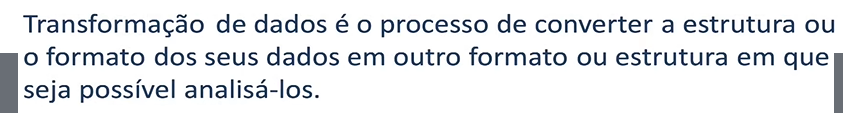
**Azure HDInsight**

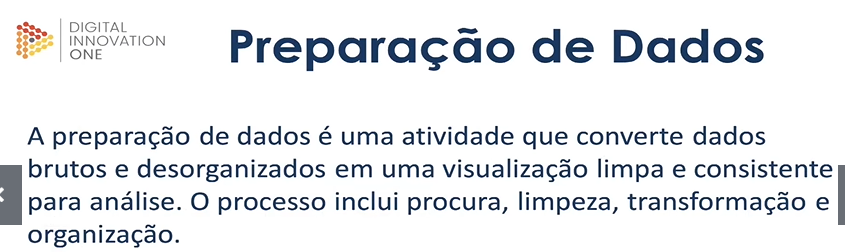
Transformação e preparação de dados

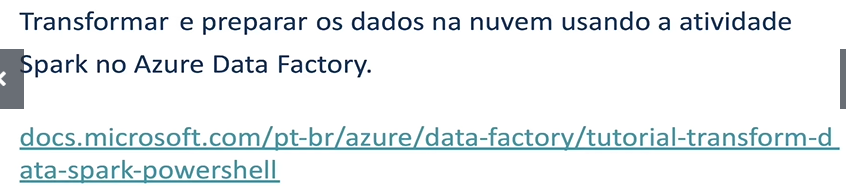
Objetivos

* Revisar a transformação e preparação de dados
* Entender os estágios de processamento de Big Data
* Explorar o Azure HDInsight

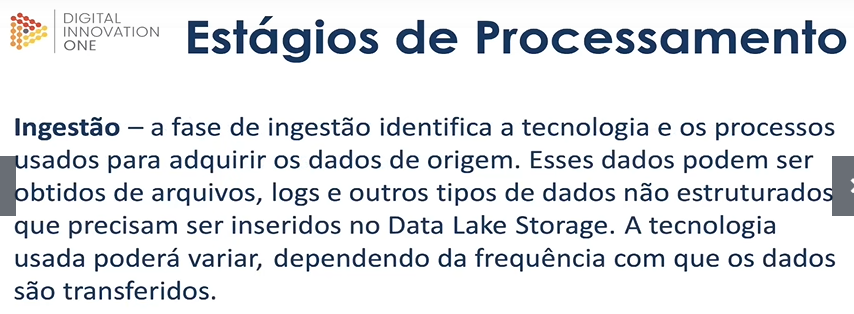
**Transformação de Dados**

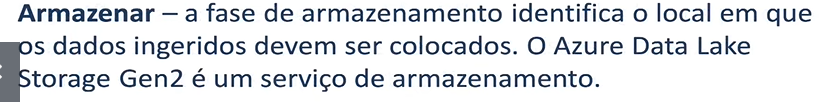


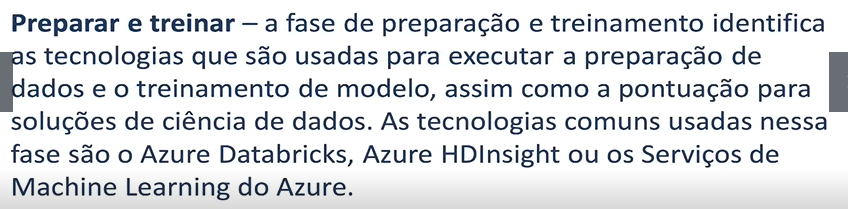


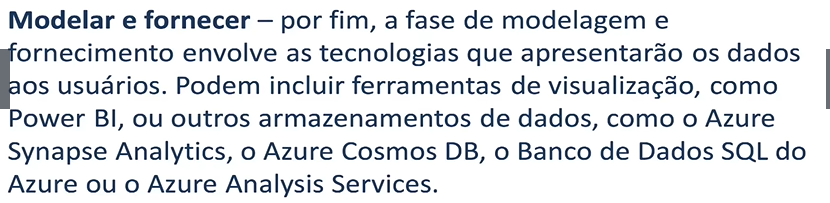


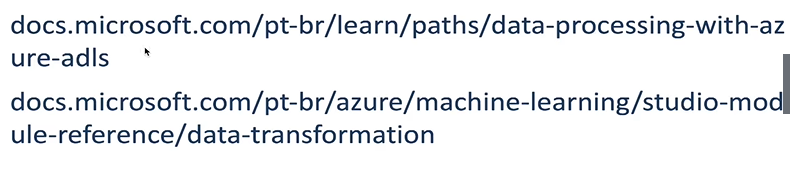
Estágios de processamento



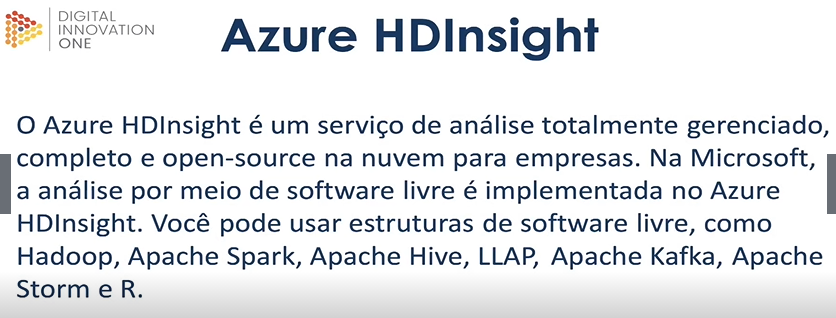


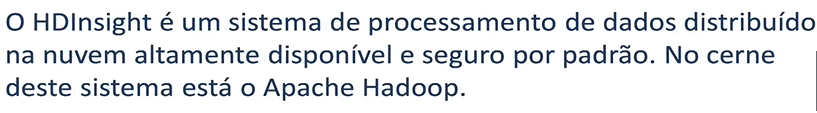


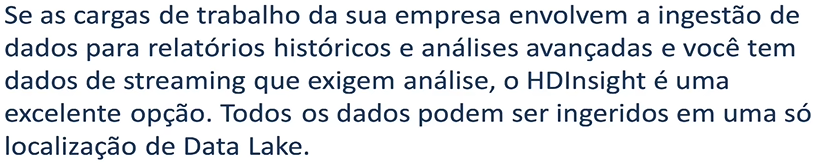


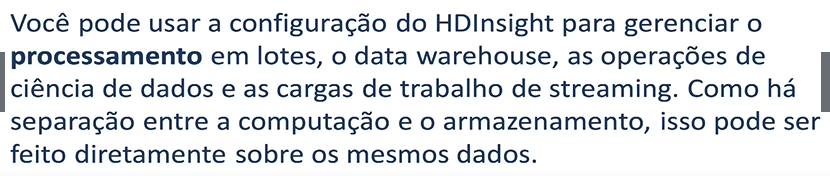


**Azure HDInsight**





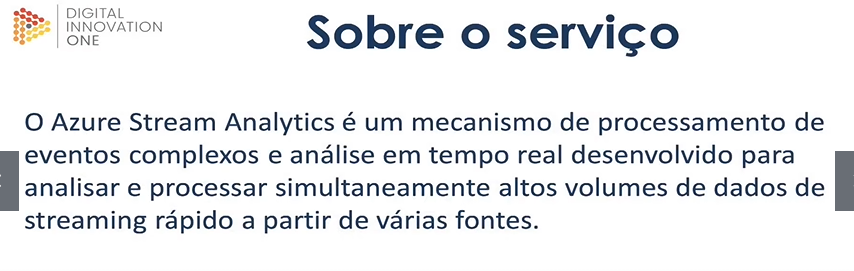


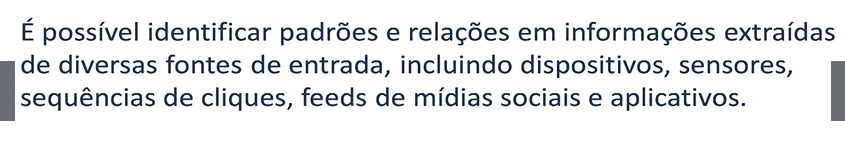


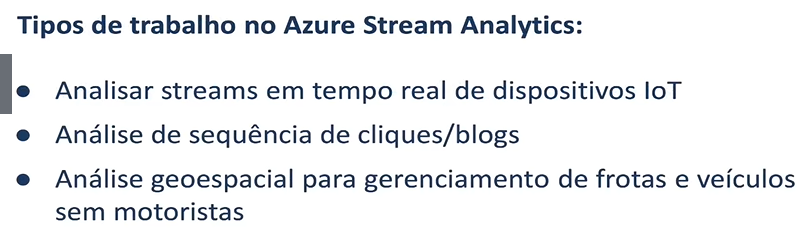
**Azure Stream Analytics**

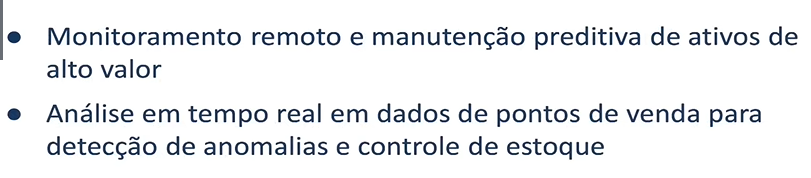
Objetivos

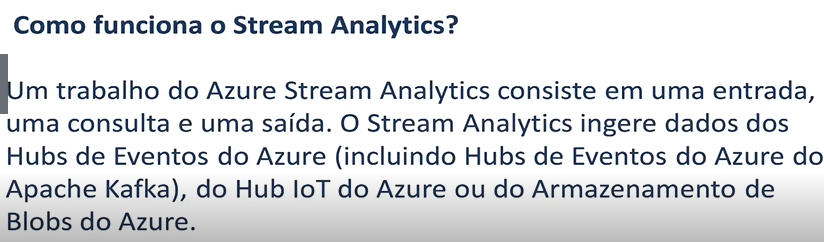
* Conhecer o Azure Stream Analytics
* Identificar a aplicações possíveis
* Explorar o Azure Stream Analytics

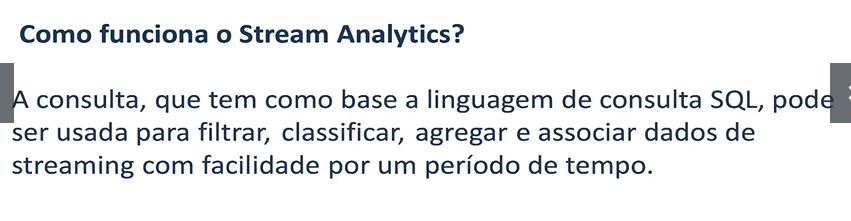


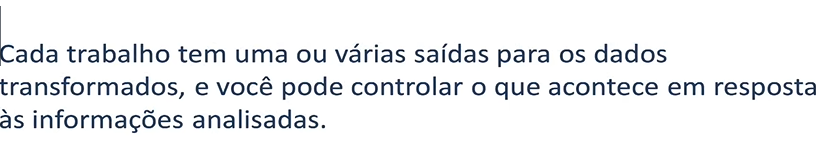


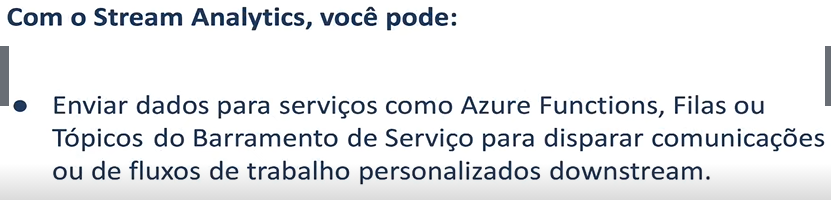


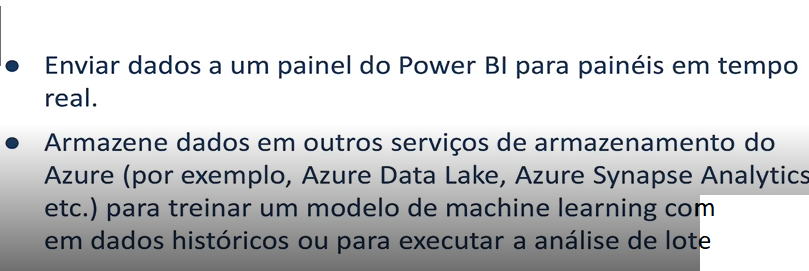












Explorando o Azure Stream Analytics

