# Módulo 1 – Azure Para Engenheiros de Dados

Diferença entre ETL e ELT

ETL => Já transforma direto.

ELT => A Carga, no Cosmos DB ou no Azure Data Lake Storange, vem primeiro.

Vantagens de usar o ELT => Diminuir a sobrecarga do sistema de origem. Transformação leva muito tempo.

Traz flexibilidade na arquitetura quando há mais de uma organização que utilizará os dados. Por exemplo, o pessoa de Marketing usa os dados de um jeito e o pessoal do financeiro de outro.

Armazenamento de Dados no Azure **Data Lake Storange**

Compatível com o Hadoop.

Armazenar grandes quantidades de dados para análise de Big Data.

Para ingerir dados em seu sistema, use o Azure Data Factory, o Apache Sqoop, o Gerenciador de Armazenamento do Azure ou a Ferramenta AzCopy, o PowerShell ou o Visual Studio. Para usar o recurso de Upload de Arquivo para importar tamanhos de arquivos acima de 2 GB, use o PowerShell ou o Visual Studio. O AzCopy dá suporte a um tamanho máximo do arquivo de 1 TB e divide automaticamente os arquivos de dados que excedem 200 GB.

**Azure Synapse Analytics** => plataforma de dados baseada em nuvem, que combina um armazenamento de dados corporativo e uma análise de Big Data.

Os pools de SQL usam MPP (Processamento Paralelo em Massa) para executar consultas em petabytes de dados rapidamente

. O Azure Synapse Analytics dá suporte a três tipos de tabelas distribuídas: hash, round robin e replicado. Use essas tabelas para ajustar o **desempenho**.

Azure Stream Analytics => Processar dados de streaming e responder às anomalias de dados em tempo real

Em tempo real, os dados são ingeridos de aplicativos ou gateways e dispositivos de IoT em um hub de eventos ou hub IoT. Em seguida, o hub de eventos ou hub IoT transmite os dados para o Stream Analytics para análise em tempo real.

Sistemas de lote processam grupos de dados que são armazenados em um Armazenamento de Blobs do Azure.

Para processar dados de streaming, configure trabalhos do Stream Analytics com pipelines de entrada e de saída. Entradas são fornecidas por Hubs de Eventos, Hubs IoT ou o Armazenamento do Azure.

**Azure HDInsight** =>  Ingerir, processar e analisar Big Data. Ele dá suporte a processamento em lotes, data warehousing, IoT e ciência de dados. Ele inclui Apache Hadoop, Spark, Kafka, HBase, Storm e Consulta Interativa.

* O Hadoop inclui Apache Hive, HBase, Spark e Kafka. O Hadoop armazena dados em um sistema de arquivos (HDFS). O Spark armazena dados na memória. Essa diferença no armazenamento torna o Spark aproximadamente 100 vezes mais rápido.
* O HBase é um banco de dados NoSQL criado no Hadoop. Normalmente, ele é usado para mecanismos de pesquisa. O HBase oferece failover automático.
* O Storm é uma solução de análise de streaming em tempo real distribuída.
* O Kafka é uma plataforma de software livre usada para compor pipelines de dados. Ele oferece a funcionalidade de fila de mensagens, que permite aos usuários publicar ou assinar fluxos de dados em tempo real.

Como engenheiro de dados, use o Hive para executar operações de ETL com os dados que você está ingerindo. Ou então, organize consultas de Hive no Azure Data Factory

No Hadoop, use Java e Python para processar Big Data.

O Hadoop dá suporte às linguagens Pig e HiveQL. No Spark, os engenheiros de dados usam o Spark SQL.

## Databricks

plataforma sem servidor otimizada para o Azure.  workspace interativo para aplicativos baseados no Spark.

## Data Factory

serviço de integração de nuvem. Ele orquestra a movimentação de dados entre vários armazenamentos de dados.

 criar e agendar fluxos de trabalhos orientados a dados (chamados pipelines) que podem ingerir dados provenientes de armazenamentos de dados.

Catálogo de Dados do Azure

O Catálogo de Dados do Azure é um local central em que os usuários de uma organização podem contribuir com seus conhecimentos. Juntos, eles criam uma comunidade de fontes de dados de propriedade da organização.

Tarefas de um engenheiro de dados do Azure:

* Projetar e desenvolver soluções de armazenamento e processamento de dados para a empresa.
* Configurar e implantar serviços de dados baseados em nuvem, como serviços de blob, bancos de dados e análise.
* Proteger a plataforma e os dados armazenados. Garantir que somente os usuários necessários possam acessar os dados.
* Garantir a continuidade dos negócios em condições incomuns usando técnicas para alta disponibilidade e recuperação de desastre.
* Monitorar os sistemas para garantir que sejam econômicos e executados corretamente.

<https://docs.microsoft.com/pt-br/learn/modules/data-engineering-processes/4-architecturing-project> => Importante.

* Provisionar o Azure Data Lake Storage Gen2.
* Provisionar o Azure Synapse Analytics.
* Provisionar o Hub IoT do Azure.
* Provisionar o Azure Stream Analytics.
* Provisionar o Azure Machine Learning.
* Provisionar o Azure Data Factory.
* Provisionar o Power BI.

# Módulo 2 – Armazenamento de Dados no Azure

Uma empresa de varejo online tem diferentes tipos de dados. Cada tipo de dados pode se beneficiar de uma solução de armazenamento diferente.

Dados Estruturados => Dados Relacionais.

Dados Semi-Estruturados => NoSql. Chave/Valor. Ex: XML, Json.

Dados Não Estruturados => Fotos, Vídeos. Vem com metadatos Semi-Estruturados.

**Azure Cosmos DB** => Suporte a dados SemiEstruturados ou dados NoSql.

Suporte a SQL para consulta.

Conformidade com o ACID.

Permite replicar os dados para outro Data Center, para que, a consulta de alguém seja mais rápida.

**Azure de Armazenamento de Blobs** => Para vídeos e imagens.

Permite o CDN (Rede de Distribuição de Conteúdo).

**Banco de Dados SQL do Azure** => Para dados Estruturados.