**IMPLANTAÇÃO DE PIPELINE DE BIG DATA**

**Curso: Sistemas de Informação**

**1.** Objetivo

O objetivo deste projeto é criar um pipeline de Big Data baseado em recursos da AWS (Amazon Web Services) ou *on premise*, para realizar análises estatísticas descritivas em dados armazenados em um *datalake* ou um *data warehouse*. Além disso, busca-se a criação de um infográfico com base nos resultados do estudo estatístico, de forma genérica para a melhor entrega de valor ao cliente.

**2.** Definição

O pipeline de Big Data poderá ser executado em 2 cenários distintos:

1 - Infraestrutura da AWS e seus serviços na nuvem para realizar as seguintes etapas:

* **Ingestão de dados**: Os dados serão armazenados em um *datalake* utilizando o serviço AWS S3. Será implementada a ingestão de dados, podendo ser em batch ou em streaming, dependendo dos requisitos do projeto. Os dados fornecidos devem ter uma volumetria que após sua ingestão deve ser maior de 1 gigabyte, para que possamos utilizar as ferramentas de big data.
* **Preparação dos dados:** Utilizando o serviço AWS Glue ou uma ETL feita na AWS Lambda com linguagem Python, será realizada a extração, transformação e carga (ETL) dos dados para prepará-los para análise estatística.
* **Análise estatística:** O processamento e a análise dos dados serão realizados utilizando o serviço AWS EMR (Elastic MapReduce), com o suporte de ferramentas como Apache Spark e Apache Hadoop. Serão aplicados cálculos estatísticos descritivos para obter insights relevantes.
* **Criação do infográfico:** Utilizando o serviço AWS QuickSight ou uma ferramenta open-source em uma máquina virtual (em container), será desenvolvido um infográfico visualmente atrativo com base nos resultados estatísticos, permitindo a apresentação clara e concisa das informações.

2 - Infraestrutura *on premise* para realizar as seguintes etapas:

* **Ingestão de dados**: Os dados serão armazenados em um *datalake* local utilizando uma solução de armazenamento *on premise*. Será implementada a ingestão de dados, podendo ser em batch ou em streaming, dependendo dos requisitos do projeto. Os dados fornecidos devem ter uma volumetria que após sua ingestão deve ser maior de 1 gigabyte, para que possamos utilizar as ferramentas de big data disponíveis.
* **Preparação dos dados:** Utilizando uma ferramenta de ETL (Extração, Transformação e Carga) desenvolvida na linguagem Python, será realizada a extração, transformação e carga (ETL) dos dados para prepará-los para análise estatística.
* **Análise estatística:** O processamento e a análise dos dados serão realizados utilizando uma solução *on premise*, como o EMR (Elastic MapReduce) com suporte às ferramentas Apache Spark e Apache Hadoop. Serão aplicados cálculos estatísticos descritivos para obter insights relevantes.
* **Criação do infográfico:** Utilizando uma solução *on premise*, como uma ferramenta de visualização de dados em uma máquina virtual (em container), será desenvolvido um infográfico visualmente atrativo com base nos resultados estatísticos, permitindo a apresentação clara e concisa das informações.

**3.** Possíveis contribuições

Alguns exemplos de possíveis contribuições deste projeto incluem:

* Identificação de padrões e tendências nos dados.
* Descoberta de insights relevantes para tomada de decisões.
* Análise comparativa entre diferentes conjuntos de dados.
* Avaliação do desempenho e monitoramento em tempo real de indicadores chave.
* Identificação de correlações e relações entre variáveis.

**4.** Protótipo

O protótipo será desenvolvido de forma iterativa, seguindo as etapas do pipeline de Big Data descritas, e os resultados de cada etapa serão avaliados e refinados ao longo do processo. Será necessário garantir a integração e o correto funcionamento de cada componente do pipeline, além de assegurar a escalabilidade e o desempenho adequado para lidar com grandes volumes de dados.

O protótipo poderá ser desenvolvido em 2 cenários distintos:

1 - Infraestrutura da AWS:

O protótipo consistirá na implementação de um pipeline de Big Data utilizando recursos da AWS, conforme descrito no objetivo do projeto. Serão utilizados os seguintes serviços da AWS: S3 para armazenamento dos dados no *datalake*, AWS Glue ou AWS Lambda para a preparação dos dados, EMR com Apache Spark e Apache Hadoop para a análise estatística dos dados, e AWS QuickSight ou uma ferramenta open-source em container para a criação do infográfico.

2 – On premise:

O protótipo consistirá na implementação de um pipeline de Big Data utilizando recursos *on premise*, conforme descrito no objetivo do projeto. Serão utilizadas as seguintes tecnologias: um servidor de armazenamento local para o *datalake*, uma ferramenta de ETL (Extração, Transformação e Carga) criada em Python com framework em flask para a preparação dos dados, um cluster com Apache Spark, Apache Hadoop e Apache Airflow em containers para a análise estatística dos dados, e uma ferramenta open-source em container para a criação do infográfico.

**5.** Premissas

Para o desenvolvimento do projeto, considera-se as seguintes premissas:

* O prazo para a conclusão da implementação do pipeline de Big Data é de 70 dias corridos (10 semanas), a partir de 10/16/2023.
* Será fornecido um conjunto de dados com uma volumetria mínima de 1 gigabyte no total para ser utilizado no pipeline de Big Data, podendo ser logs via streaming ou batches via diferentes protocolos a serem analisados.
* O conjunto de dados devem preferencialmente ser de origem não-estruturada.
* Os resultados obtidos a partir da análise estatística dos dados serão validados com o cliente antes da criação do infográfico.

**6.** Restrições

* Não serão desenvolvidas integrações com outros sistemas ou serviços além dos mencionados no objetivo do projeto.
* Não será escopo do projeto o desenvolvimento de funcionalidades avançadas de processamento ou análise de dados além das análises estatísticas descritivas, como redes neurais e machine learning.
* O acesso à infraestrutura da AWS ou on-premisse será limitado aos membros da equipe do projeto e aos docentes e coordenadores responsáveis pelo acompanhamento do projeto.
* E caso de solução *on-premise*, o pipeline de dados será restrito à capacidade de processamento do laptop de cada estudante.
* Não será escopo do projeto o desenvolvimento de ambiente front-end para apoio ao processamento do pipeline de dados.
* Não será escopo do projeto o desenvolvimento de relatórios complementares de análise de dados, restrito apenas ao Infográfico.
* Não será escopo do projeto disponibilizar os dados para acesso externo.

**7.** Entregáveis (Artefatos)

* Sprint 1 – Entendimento do usuário e entendimento do negócio
* Sprint 2 – Sistema de ingestão de dados
* Sprint 3 – *datalake* ou *data warehouse*
* Sprint 4 – Análise estatística sem refinamento
* Sprint 5 – infográfico a partir da análise estatística