

**PROJETO:** (Ex: Criação de uma solução de uma aplicação utilizando linguagem Natural para conscientização da preservação dos meios / recursos naturais)

**EMPRESA:** (Ex: Nome fantasia ou Razão social da empresa)

**BREVE DESCRIÇÃO EMPRESA / MINI BIO:** (Ex: A empresa xxx atua no mercado “Z” desde “aaaa” e é pioneira em...)

**PROFESSOR ORIENTADOR:** RENATO PENHA

**OVERVIEW**

* PRINCIPAL ÁREA DE NEGÓCIO: Serviços
* LÍDER DO PROJETO: (Ex: “João d e Tal” – Diretor comercial)
* PONTO FOCAL BACKUP: Mandatório (sugerimos definir um responsável na ausência do ponto focal)
* LÍDER TÉCNICO: Se existir
* LÍDER DE NEGÓCIO: Se existir
* LÍDER EXECUTIVO [Onboarding Executivo]:

**ESBOÇO DO PROJETO**

● PROBLEMA: (Expor a necessidade de negócio do cliente referente à implantação de Big Data)

● OBJETIVO:

O objetivo deste projeto é criar um pipeline de Big Data baseado em recursos da AWS (Amazon Web Services),, para realizar análises estatísticas 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.

● BENEFÍCIOS ESPERADOS PARA O PARCEIRO:

Desenvolvimento de um método de trabalho para implantação do Pipeline de Big Data, junto aos estudantes do Inteli.

● MATERIAIS DE ESTUDOS ANEXADOS:

**ESCOPO MACRO**

O pipeline de Big Data poderá ser executado no seguinte cenário:

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.

**MVP**

O protótipo será desenvolvido de forma iterativa, seguindo as etapas do pipeline de Big Data descritas no item 2. Os resultados de cada etapa serão avaliados e refinados ao longo do desenvolvimento do projeto. 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 no seguinte cenário:

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.

OBS: Lembre-se a definição de MPV é: MVP é a sigla que representa o Mínimo Produto Viável. De um jeito simples, podemos definir o MVP como uma versão enxuta de uma solução, que contém apenas suas funcionalidades básicas. Pode ser um software, serviço, produto físico ou digital.

**INFRAESTRUTURA NECESSÁRIA**

Os recursos de infraestrutura poderão ser providos pelo Parceiro, conforme alinhamento na produção do TAPI. A listagem acima faz referência aos requisitos necessários que o projeto deverá ter para o desenvolvimento das competências dos alunos. No caso do desenvolvimento contar com recursos do parceiro, o projeto deverá ser desenvolvido totalmente nas dependências do Inteli e será garantido o acesso a toda a equipe (docentes, coordenação e alunos) à infraestrutura necessária.

**RESTRIÇÕES / O PROJETO NÃO CONTEMPLA**

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 será limitado aos membros da equipe do projeto e aos docentes e coordenadores responsáveis pelo acompanhamento do projeto.

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.

**PREMISSAS**

O prazo para a conclusão da implementação do pipeline de Big Data é de 70 dias corridos (10 semanas).

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.

OBS: Nessa seção em específico, haverá durante o refinamento do documento com o corpo docente do INTELI a inclusão de restrições da própria instituição com base no escopo mencionado, alinhando assim o desejo / necessidade do parceiro de projetos com o ensino dos alunos.

**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

**CONTEÚDO RESTRITO**

No caso de materiais compartilhados com a turma, descrever quais deverão ser restritos ao compartilhamento aberto no Github no site do Inteli.

**STAKEHOLDERS**

Definir demais stakeholders envolvidos no projeto, para auxiliar os alunos a compreenderem o domínio. Caso não estejam na relação de participantes.

