

Apresentação



I MVP


MINIMUM VIABLE PRODUCT



INOVAÇÃO

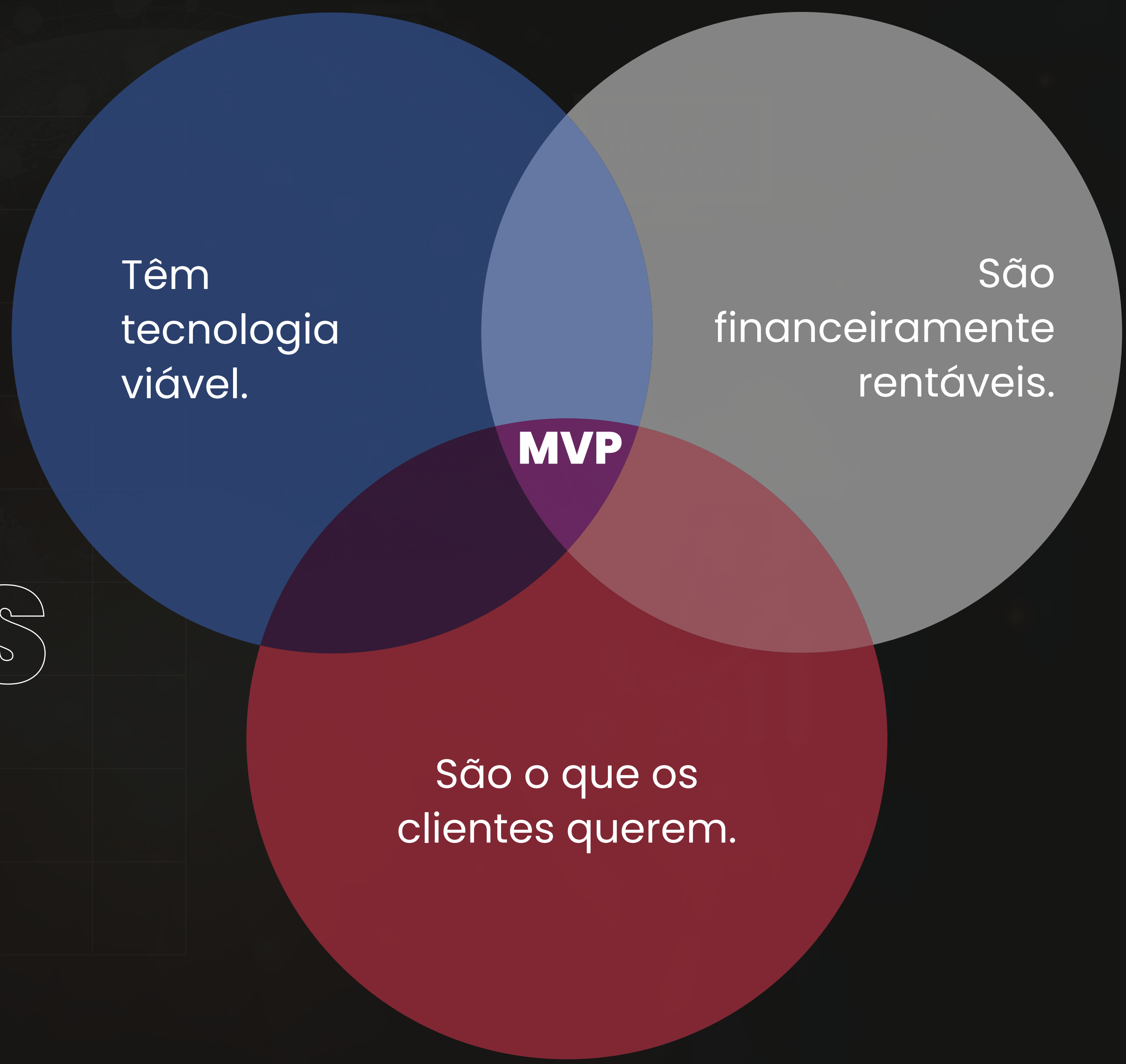
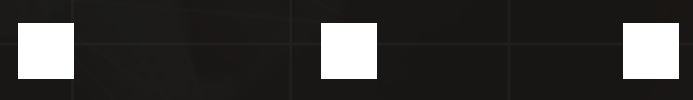
+ Com o avanço tecnológico, a inovação passou a fazer parte do nosso dia a dia. Pensando nisso, a PUC-Rio trouxe para o aprendizado a experimentação por meio do desenvolvimento do *minimum viable product* (MVP).

Nosso propósito é preparar você para desenvolver uma solução e testá-la para atender as necessidades dos clientes, representando uma versão que é estritamente para cumprir sua função.



```
var val = randomInt(1, 7);
if (val === 1) {
  println("Signs point to yes.");
} else if (val === 2) {
  println("Outlook good.");
} else if (val === 3) {
  println("It is decidedly so.");
} else if (val === 4) {
  println("Reply hazy, ask again.");
} else if (val === 5) {
  println("Concentrate and ask again.");
} else if (val === 6) {
  println("My sources say no.");
} else {
  println("Don't count on it.");
}
```

PROJETOS INOVADORES

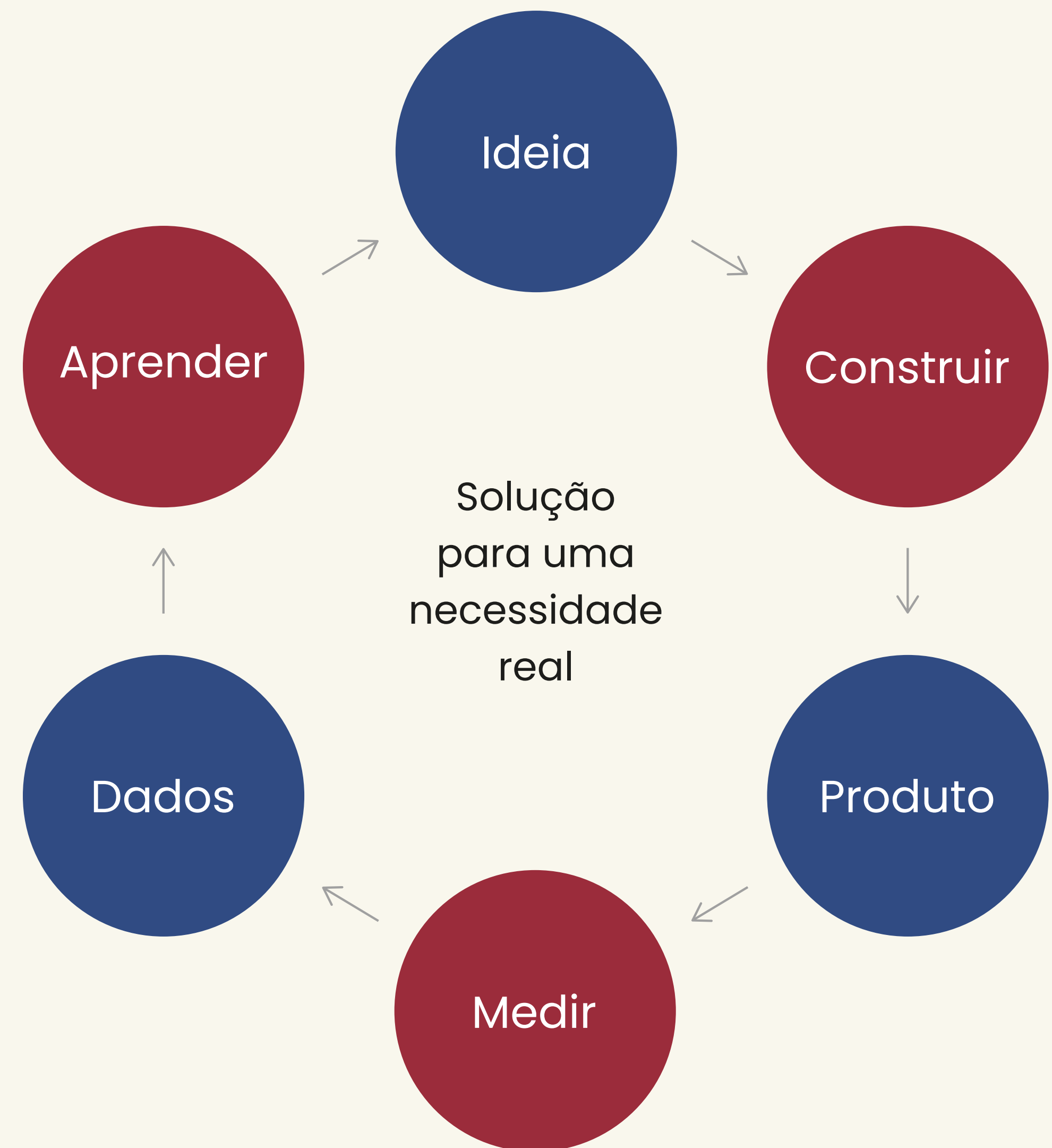


BENEFÍCIOS DE DESENVOLVER UM MVP

- Testar uma ideia com poucos recursos.
- Verificar tendências de mercado.
- Lançar produtos mais rapidamente.
- Atrair investidores.



O MVP é a peça central de uma estratégia de experimentação. Para isso, você deve desenvolver uma solução real com base nos conhecimentos adquiridos em sua *sprint*.



Etapas para o desenvolvimento do MVP



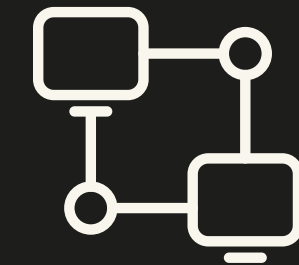
Ideia

Qual será a principal funcionalidade do MVP?



Estrutura

Que ferramentas serão utilizadas para o desenvolvimento?



Protótipo

Entrega do protótipo final





MUITO BEM!

Agora que você conheceu um pouco mais sobre o MVP,
está pronto para iniciar o desenvolvimento do seu.

Fique atento às etapas e aos prazos de entrega.

Requisitos para o Desenvolvimento do MVP

Descrição

Neste trabalho, você deverá ser capaz de construir um pipeline de dados utilizando tecnologias na nuvem. O pipeline irá envolver a busca, coleta, modelagem, carga e análise dos dados.

Objetivo

Comece pelo objetivo do seu trabalho. Antes de iniciar sua busca pelos dados, pense e descreva claramente qual problema deseja resolver com este MVP. Enumere as perguntas que deseja responder.

É de extrema importância que esta etapa seja feita antes de iniciar qualquer outra etapa.

Uma vez traçado o objetivo e conhecendo bem qual problema se deseja resolver, quais perguntas se deseja responder, é hora de iniciar a busca pelos dados.

Não é necessário atingir todos os objetivos desenhados nesta seção. Assim, não remova perguntas as quais não se conseguiu responder. Deixe a documentação do objetivo intacta e faça uma boa discussão do atingimento deste ao final do trabalho (vide Autoavaliação).

Plataforma

Vamos direcionar os esforços de apoio na Plataforma Databricks. A Databricks possui uma versão de uso chamada Databricks Community Edition, que é de uso gratuito, com limitação na qualidade e quantidade de máquinas no cluster.

Não haverá limitação de utilização de outras plataformas de dados e provedores de nuvem. Todos os trabalhos, em qualquer plataforma escolhida pelo aluno, serão devidamente avaliados. Entretanto, conforme dito acima, nossos esforços estarão voltados para a solução e apoio com problemas dos alunos na Plataforma Databricks.

Detalhamento

1. Busca pelos dados

Escolha uma base de dados para utilizar em seu MVP de forma que se possa atingir os objetivos traçados na etapa anterior.

Há inúmeras bases de dados gratuitas disponíveis na web, por exemplo:

- Google Cloud Public Datasets (<https://cloud.google.com/datasets>)
- Kaggle (<https://www.kaggle.com/datasets>)
- Portal da Transparência (<https://portal.datransparencia.gov.br/>)
- IMDB (<https://datasets.imdbws.com/>)
- Tableau (<https://www.tableau.com/learn/articles/free-public-data-sets>)
- Stanford Large Network Dataset Collection (<https://snap.stanford.edu/data/index.html>)
- Yelp Open Dataset (<https://www.yelp.ca/dataset>)

Discutiremos sobre bases de dados abertas disponíveis para o MVP no Discord e iremos montar colaborativamente um repositório de possibilidades.

Caso haja uma licença de uso para o conjunto de dados escolhido, isto deve constar na documentação do MVP.

2. Coleta

Uma vez definido o conjunto de dados, devemos coletar e armazená-los na nuvem.

É possível que, a partir de sua escolha do conjunto de dados, seja necessária uma etapa de construção de robôs de coleta, e.g. via Web Scraping. Neste caso, atente-se para questões éticas sobre se é possível utilizar os robôs de coleta de informação nos sites escolhidos.

Caso tenha optado por utilizar um conjunto de dados real da empresa onde trabalha, tenha bastante cuidado com a confidencialidade destes dados e/ou das análises que serão feitas em sequência.

3. Modelagem

Você deve construir um modelo de dados em Esquema Estrela ou Snowflake, como em um Data Warehouse, ou flat por cada conceito, como em um Data Lake.

Independentemente do modelo, deve ser construído um Catálogo de Dados contendo minimamente uma descrição detalhada dos dados e seus domínios, contendo valores mínimos e

máximos esperados para dados numéricos, e possíveis categorias para dados categóricos.

Este modelo deve também descrever a linhagem dos dados, de onde os mesmos foram baixados e qual técnica foi utilizada para compor o conjunto de dados, caso haja.

4. Carga

Nesta etapa, será feita a carga dos dados para o Data Warehouse/Data Lake. Na avaliação, nesta etapa, será dado valor pela utilização da pipelines de ETL (Extração, Transformação e Carga) na plataforma de dados utilizada. Iremos discutir pipelines de ETL na Plataforma Databricks durante nossos encontros pelo Zoom e no Discord.

Deve-se documentar os processos de transformação e carga, principalmente os de transformação, e.g. a junção e conciliação de dois conjuntos de dados diferentes.

5. Análise

Vamos dividir a etapa de análise em duas: qualidade de dados e solução do problema.

a. Qualidade de dados

Deve ser feita uma análise da qualidade para cada atributo do conjunto de dados. Existem problemas no conjunto de dados? Caso haja, como esses problemas podem ser resolvidos para que não afetem as respostas das perguntas que quer solucionar?

É possível que não se encontre problemas nos conjuntos de dados, em alguns casos há conjuntos de dados curados e já bem tratados antes de serem disponibilizados. Entretanto, mesmo nestes casos, espera-se que seja feita uma análise de valores por atributo e que se demonstre que não se encontrou problemas.

b. Solução do problema

Chegou o momento de se solucionar o problema em questão, definido preliminarmente nos objetivos. Deve-se buscar respostas para as perguntas elencadas. Para cada resposta obtida tecnicamente através da análise dos dados deve haver uma discussão do seu resultado, conectando os números obtidos às respostas ao problema a ser solucionado.

Ao final, deve haver uma discussão de uma forma geral sobre a solução do problema a partir das discussões de cada resposta.

Aqui, podem ser utilizadas bibliotecas Python vistas na disciplina Análise Exploratória e Pré-Processamento de Dados, ou as ferramentas vistas na disciplina Visualização de Informação. Entretanto, caso não tenha ainda cursado estas disciplinas, é possível responder as perguntas do objetivo somente através da linguagem SQL, objeto da disciplina de Banco de Dados ou através da linguagem de consulta do banco NoSQL escolhido, objeto da disciplina de Data Warehouse.

Entrega

O trabalho é individual.

Deverá ser disponibilizado todo código construído em um repositório público do GitHub. Se tiver dúvidas sobre como criar um repositório público no GitHub, consulte <https://docs.github.com/pt/repositories/creating-and-managing-repositories/creating-a-new-repository>

Algumas tarefas das etapas do trabalho podem ser feitas a partir de componentes visuais da plataforma de nuvem. Desta forma, deve se gerar evidência da execução destes passos através de *screenshots* ou vídeos.

Deve se gerar evidência dos resultados das respostas às perguntas que definem o problema do MVP através de *screenshots* ou vídeos.

Autoavaliação

Ao finalizar o trabalho, é esperado que o aluno faça uma autoavaliação contendo uma discussão sobre se conseguiu atingir os objetivos delineados antes do início das outras etapas, suas dificuldades encontradas na execução do trabalho, bem como trabalhos futuros para enriquecer o problema e sua solução em seu portfólio.

Critérios de avaliação

- **Objetivo (1,0 pt).** O objetivo do trabalho deve ser muito bem detalhado; é um planejamento do trabalho, contendo de forma clara e objetiva o problema a ser resolvido e as perguntas de negócio a serem respondidas. Será avaliada a qualidade desta descrição.
- **Coleta (0,5 pt).** Será avaliada a documentação sobre a coleta dos conjuntos de dados e a persistência dos mesmos na plataforma de nuvem.
- **Modelagem (2,0 pt).** Será avaliada a qualidade da modelagem dos dados (1,0 pt) e documentação do Catálogo de Dados (1,0 pt).

- Carga (**1,0 pt**). Será avaliada a qualidade da documentação da carga dos dados, bem como a corretude e persistência dos dados na plataforma de nuvem após a carga.
- Análise (**3,0 pt**). Serão avaliados a análise de qualidade dos dados (**1,0 pt**) e da solução do problema de forma correta (**0 pt**) e bem analisada pela discussão a partir das respostas obtidas (**1,0 pt**).
- Autoavaliação (**0,5 pt**). Será avaliada a autoavaliação do aluno com as questões pertinentes sobre o atingimento de seus objetivos traçados no início do trabalho.
- Capricho (**2,0 pt**). Aqui serão avaliados o capricho e a qualidade geral do trabalho como um todo de forma subjetiva.

Datas Importantes

12/11: Orientação inicial do MVP

18/11: Encontro sobre Transformação de Mercado

26/11: Sessão de dúvidas do MVP

3/12: Sessão de dúvidas do MVP

10/12: Sessão de dúvidas do MVP

17/12: Sessão de dúvidas do MVP

21/12: Entrega do MVP

12/01: Resultado final do MVP

Poste aqui seu MVP



Atenção!

É neste fórum que você deverá realizar a entrega oficial do seu MVP e para isso é necessário que compartilhe o link do Github. Caso tenha algum material complementar ele poderá ser postado com anexo em sua resposta.

Será permitido o envio de até duas respostas nesse fórum caso a primeira haja algum erro na segunda resposta. Certifique-se de que todas as informações estejam entregues corretamente. Qualquer dúvida o Community Manager poderá lhe auxiliar.