1. VISÃO GERAL DO PROJETO

1.1 Contextualização

A PoD Bank acabou de criar o departamento de Business Intelligence e o gestor responsável pela área acabou de chegar na empresa.

A PoD Bank concede crédito para clientes com pouco histórico de crédito, portanto consome informações de bureaus e dados alternativos. Uma vez que a PoD Bank ainda não tem uma área de modelagem de crédito definida, a área de BI será responsável por atender o negócio e auxiliar na geração de insights e cenários que trarão respostas às estratégias de crescimento da empresa.

1.2 Objetivo

1. Mapeamento e catálogo dos dados que a empresa tem e se há necessidade de comprar informações externas.
2. Disponibilizar um catálogo de indicadores de negócio. Esses indicadores devem estar disponíveis em um “Data Lake” prontos para consumo. Os indicadores devem conter:
    - Taxa de Inadimplência: Calcula a porcentagem de clientes que estão inadimplentes em relação ao total de clientes.
    - Valor Médio de Empréstimo: Calcula o valor médio dos empréstimos concedidos aos clientes.
    - Taxa de Aprovação de Crédito: Mede a porcentagem de solicitações de crédito aprovadas em relação ao total de solicitações recebidas.
    - Distribuição de Idade dos Clientes: Mostra a distribuição da idade dos clientes para identificar se há alguma faixa etária com maior risco de inadimplência.
    - Taxa de Refinanciamento: Mede a porcentagem de clientes que refinanciam seus empréstimos em relação ao total de empréstimos concedidos.
    - Tempo Médio de Atraso no Pagamento: Calcula a média de dias de atraso no pagamento dos clientes inadimplentes.
    - Taxa de Recuperação: Avalia a porcentagem de empréstimos inadimplentes que foram recuperados através de ações de cobrança.

- Variação da Taxa de Inadimplência ao Longo do Tempo: Monitora as flutuações da taxa de inadimplência ao longo do tempo para identificar tendências e padrões.
- Performance de Segmentos de Clientes: Analisa a taxa de inadimplência e o valor médio de empréstimo por segmento de clientes (por exemplo, faixa de renda, localização geográfica, profissão/ocupação, escolaridade) para identificar grupos de maior risco e tomar ações direcionadas.
- Número de Créditos Ativos: Contabilize o número de créditos ativos de cada cliente, fornecendo insights sobre a carga de dívida e a capacidade de pagamento.
- Tipos de crédito que os clientes tomam no mercado, em volume e percentual

3. Construção de um ou mais dashboards que contemple esses indicadores de negócio.
4. Estudo dos custos que a solução vai trazer de acordo com a escolha das ferramentas que serão utilizadas.

2. FONTES DE DADOS

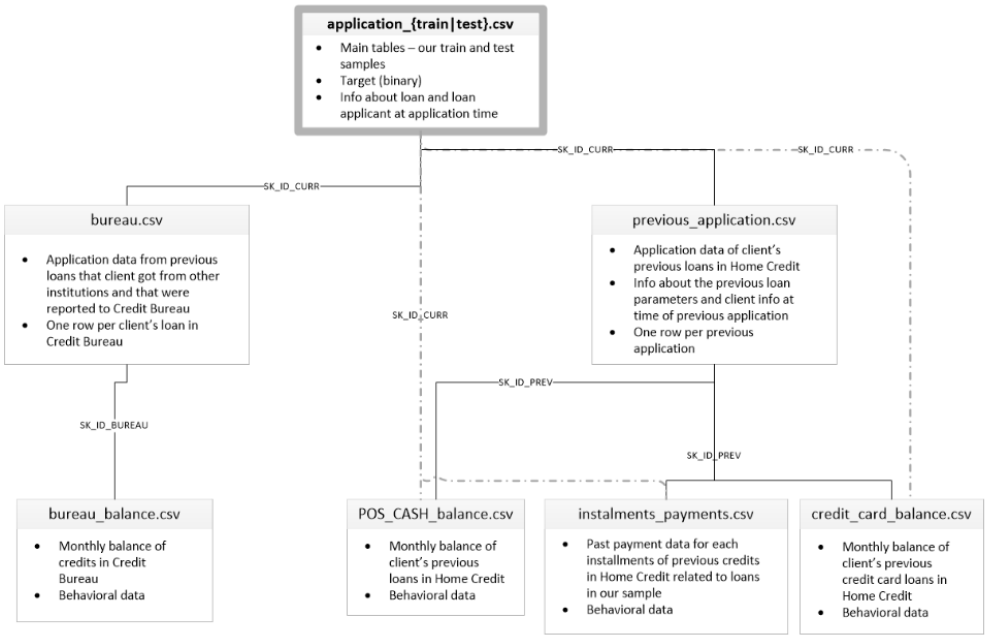
Dentro da PoDBank, existem diretórios. Dentro dos diretórios estão os dados e documentos da empresa. Estes dados são disponibilizados em formato .csv pelo time de negócio.

O recebimento dos dados é feito a cada doze hora, em dias úteis, quando os bancos de dados fazem uma varredura no sistema e os dados transacionados naquele dia são compilados e adicionados em um bucket no S3, da AWS.

Atualmente, devido a imaturidade na área de dados, estes dados são recebidos apenas pelo time de BI da Pod Bank.

3. MODELAGEM DOS DADOS

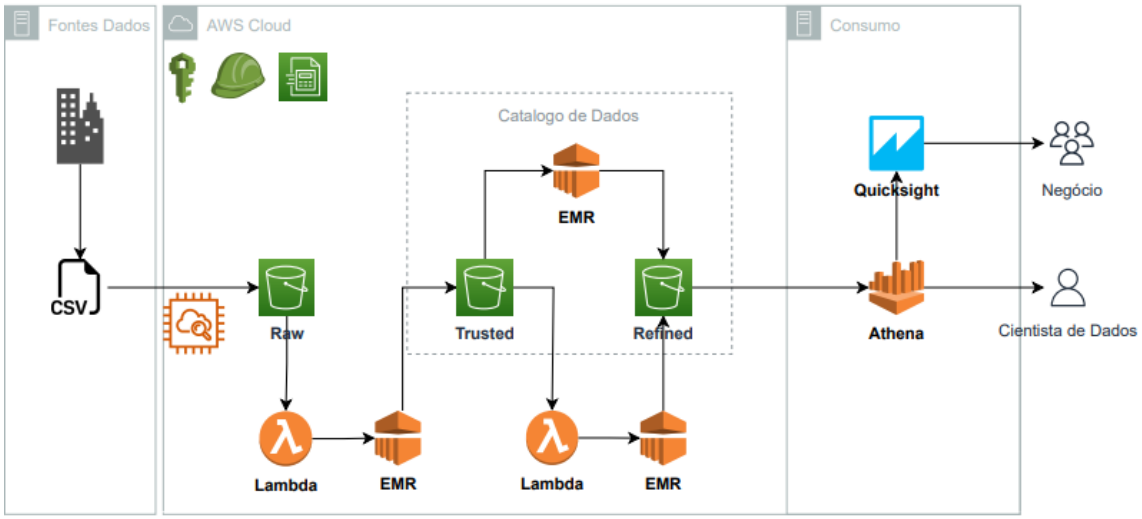
Os dados da empresa são armazenados no S3 da AWS, conta com o modelo entidade relacionamento dos dados, conforme o esquema abaixo.



4. ARQUITETURA DOS DADOS

4.1 Visão geral dos componentes.

Para a resolução do problema, foi construído uma estrutura de Data Lake, conforme arquitetura abaixo:



Para entender o funcionamento da arquitetura é preciso entender os pontos a seguir.

4.1.1 Origem dos dados

Os dados são adquiridos pelos times de negócio da Pod Bank, “on-premise”. Após um período de novas informações coletadas, é feito o upload desses dados é realizado para um bucket S3, na AWS.

Uma vez que esses dados chegam no bucket, a estrutura do Data Lake irá identificar esses dados periodicamente, através de rotinas implementadas pelo CloudWatch. Essas rotinas identificam os novos dados e direcionam esses arquivos .csv para a camada “transient” do Data Lake.

4.1.2 Tratamentos e Manipulações.

No Data Lake, os dados passam por um processo de tratamento e manipulação ao longo de várias camadas, para garantir que estejam prontos para uso pelo time de negócios. Neste tópico, iremos abordar quais transformações são executadas nos dados em cada etapa do processo.

Na camada "transient", é aonde os dados chegam após serem disponibilizados. Uma vez que os dados se encontram nesta camada e a periodicidade do CloudWatch é executada, é iniciado a primeira transformação dos dados. Neste momento, o Amazon EMR é utilizado para transformar os dados do formato CSV para o formato Parquet, uma otimização que reduz o espaço de armazenamento necessário, tornando-os mais eficientes e econômicos. Após esta transformação, os dados são carregados na camada “raw”.

Na camada "raw", os dados permanecem em seu estado bruto e ainda não foram interpretados. Aqui, os dados ainda não são conhecidos – apenas se sabe a origem do dado – e ainda não foram catalogados. Com a chegada dos arquivos nesta camada, uma AWS Lambda é sensibilizada e um novo processamento no Amazon EMR é iniciado e os dados são direcionados para a camada “trusted”.

À medida que os dados progridem para a camada "trusted", ocorrem transformações mais significativas. Nesta camada, os tratamentos feitos incluem a tipificação e a interpretação das colunas das tabelas. Este é o ponto em que os dados começam a ser compreendidos e categorizados. Outra AWS Lambda é sensibilizada e executa outra transformação no Amazon EMR e os dados são descarregados na camada “refined”.

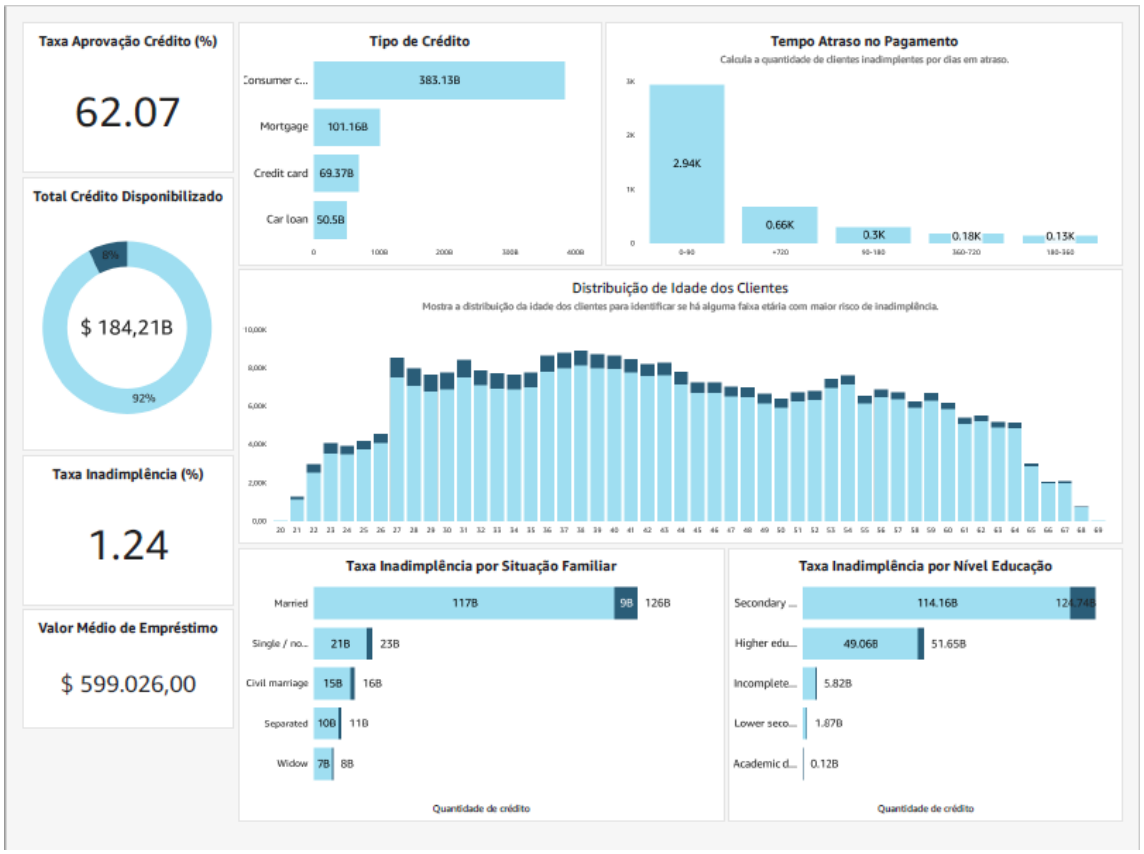
A camada "refined", é onde os dados passam por tratamentos mais refinados, focados na preparação dos indicadores de negócios. Nesta fase, os dados estão quase prontos para

serem usados em análises e geração de insights, através do AWS Athena e Amazon Quicksight.

É importante mencionar que as informações na camada "trusted" e "refined" são catalogadas para rastrear a origem e o destino de cada informação, mapeando de forma detalhada as relações entre os dados. Essa catalogação inclui informações como tipos de dados, formatos, descrições de variáveis, além de, explicitar o processo de construção de cada indicador da camada “refined”. A elaboração dos metadados é fundamental para fornecer clareza e transparência em todo o ciclo de vida dos dados no Data Lake.

Outro fator relevante durante as transformações mencionadas é a estruturação de variáveis e tabelas de controle do processo. Em cada etapa do fluxo dos dados, durante a transição entre as camadas do Data Lake, é essencial manter um controle detalhado das operações realizadas. Para isso, é implementada uma estratégia de registro de logs, como os tempos de processamento, a quantidade de dados processados e outras métricas relevantes relacionadas a cada etapa do processo. Esses registros são armazenados de forma segura em um bucket dedicado ao controle e monitoramento.

4.1.3 Consumo dos Dados



A disponibilização dos dados no Data Lake é planejada para atender às necessidades de diferentes tipos de usuários. Os dados provenientes da camada "trusted", juntamente com os indicadores refinados da camada "refined", estão prontos para consulta pelos times de negócio. Esses dados são disponibilizados de maneira acessível por meio de dashboard, gráficos e tabelas no Amazon QuickSight, permitindo que os membros da equipe de negócios, sem conhecimento de programação, realizem consultas e análises de forma eficiente e intuitiva.

Para usuários mais técnicos, os dados processados que chegam à camada "refined" são disponibilizados no Amazon Athena. Isso oferece a flexibilidade de executar consultas SQL diretamente sobre os dados, permitindo análises mais detalhadas e personalizadas.

Além disso, os metadados são apresentados em um dashboard no Amazon QuickSight, proporcionando uma visão interativa das informações.

Essa variedade de opções de acesso aos dados possibilita que os stakeholders tenham as ferramentas necessárias para extrair insights valiosos do Data Lake de acordo com suas necessidades e habilidades técnicas.

4.2 Detalhes sobre configuração, capacidade, escalabilidade e segurança da infraestrutura.

A infraestrutura utilizada no processamento dos dados é através do Amazon EMR, com máquinas “m5xlarge” equipadas com 2 núcleos de processamento. Essas instâncias oferecem um desempenho consistente e confiável, garantindo que as transformações de dados sejam executadas de maneira eficiente, com baixo custo.

A configuração do console é facilmente escalável, permitindo que mais instâncias sejam adicionadas conforme a necessidade, garantindo que o Data Lake possa lidar com volumes crescentes de dados sem comprometer o desempenho.

Em termos de segurança, todas as medidas recomendadas pela AWS são implementadas. Isso inclui o uso de grupos de segurança, políticas de IAM, criptografia de dados em repouso e em trânsito, bem como a segmentação adequada da rede para proteger os recursos.

Além disso, a infraestrutura é monitorada de perto por meio de soluções de monitoramento e logging, garantindo a detecção precoce de qualquer atividade suspeita e a conformidade com as melhores práticas de segurança.

Em suma, a configuração da infraestrutura no Amazon EMR oferece um equilíbrio entre desempenho, escalabilidade e segurança, garantindo que o Data Lake seja capaz de processar, armazenar e proteger os dados de forma eficiente e econômica.

5. DESENVOLVIMENTO

5.1 Linguagens e ferramentas utilizadas.

- Linguagens de programação:
    - Python;
    - SQL;
    - PySpark
- Armazenamento e banco de dados:
    - AWS S3;
    - AWS Athena
- Orquestração da ferramenta:
    - Amazon CloudWatch
    - AWS Lambda
- Processamento Spark, mapeamento e catálogo dos dados:
    - Amazon EMR;
    - AWS Glue
- Dashboard:
    - Amazon Quicksight
- Segurança e acessos:
    - IAM
    - Roles

5.2 Como executar tarefas específicas.

Para execução de tarefas específicas é possível acessar aos Jobs, referentes aos tratamentos de cada camada do Data Lake através dos links abaixo:

- Transient

- Raw
- Trusted
- Refined
- Controle

6. AGENDAMENTO E PERIODICIDADE

O Data Lake possui uma estratégia de processamento com uma periodicidade de execução a cada 12 horas, através do AWS CloudWatch, sincronizada com o envio de arquivos pelo time de negócios.

Quando os arquivos são recebidos na camada "transient" do Data Lake, essa rotina é acionada, permitindo que os dados sejam movidos para a camada "raw". Assim que novos arquivos são detectados na camada "raw", uma função Lambda é ativada para processar esses dados, aplicando transformações e preparando-os para a próxima fase.

Esse processo é repetido sequencialmente em cada uma das camadas do Data Lake, garantindo que os dados passem por etapas de tratamento, agregação e transformações necessárias para atualizar os indicadores e negócio apresentados no dashboard a ser visualizado pelo time de negócios.

Dessa forma, a arquitetura assegura que os dados sejam atualizados para fornecer insights à equipe de negócios de forma regular e eficiente.

7. DOCUMENTAR OS METADADOS

Os metadados desempenham um papel fundamental na gestão e na compreensão dos dados nas camadas "trusted" e "refined" do Data Lake.

Para cada tabela e indicador gerado, são mantidos metadados que fornecem informações sobre a estrutura e contexto dos dados. Incluem descrições das tabelas e indicadores, definições de colunas, tipos de dados, formatos, bem como detalhes sobre o processo de construção, transformações aplicadas e fontes de dados.

Essa formalização e documentação possibilita, aos usuários, a capacidade de explorar os dados com maior profundidade e confiança. É possível, facilmente acessar informações sobre a origem dos dados, entender como cada indicador de negócios foi construído e

interpretar as tabelas de forma que a informação se encontra centralizada nestes documentos. Excelente para manter a governança e credibilidade nos dados.

Metadados Trusted Zone - Data Lake POD Bank

Esta tabela representa o conjunto de metadados, da camada trusted, do Data Lake, que representam os dados disponíveis na empresa POD Bank.

zone	table	column	type	description
trusted	application_train	amt_annuity	double	Anuidade do empréstimo
trusted	application_train	amt_credit	double	Valor do crédito do empréstimo
trusted	application_train	amt_goods_price	double	Para empréstimos ao consumidor, é o preço dos bens para os quais o empréstimo é concedido
trusted	application_train	amt_income_total	double	Renda do cliente
trusted	application_train	amt_req_credit_bureau_day	double	Número de consultas ao Credit Bureau sobre o cliente um dia antes da aplicação (excluindo uma hora antes da aplicação)
trusted	application_train	amt_req_credit_bureau_hour	double	Número de consultas ao Credit Bureau sobre o cliente uma hora antes da aplicação
trusted	application_train	amt_req_credit_bureau_mon	double	Número de consultas ao Credit Bureau sobre o cliente um mês antes da inscrição (excluindo uma semana antes da inscrição)
trusted	application_train	amt_req_credit_bureau_qrt	double	Número de consultas ao Credit Bureau sobre o cliente 3 meses antes da inscrição (excluindo um mês antes da inscrição)
trusted	application_train	amt_req_credit_bureau_week	double	Número de consultas ao Credit Bureau sobre o cliente uma semana antes da inscrição (excluindo um dia antes da inscrição)
trusted	application_train	amt_req_credit_bureau_year	double	Número de consultas ao Credit Bureau sobre o cliente um dia ano (excluindo os últimos 3 meses antes da aplicação)
trusted	application_train	apartments_avg	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	apartments_medi	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	apartments_mode	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	basementarea_avg	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	basementarea_medi	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	basementarea_mode	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	cnt_children	int	Número de filhos que o cliente tem
trusted	application_train	cnt_fam_members	double	Quantos membros da família o cliente tem
trusted	application_train	code_gender	string	Sexo do cliente
trusted	application_train	commonarea_avg	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	commonarea_medi	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	commonarea_mode	double	Informações normalizadas sobre o prédio onde o cliente mora, Qual é a média (sufixo _AVG), modus (sufixo _MODE), tamanho médio do apartamento (sufixo _MEDI), área comum, ár
trusted	application_train	days_birth	int	Idade do cliente em dias no momento da aplicação
trusted	application_train	days_employed	int	Quantos dias antes da aplicação a pessoa começou o emprego atual

8. MANUTENÇÃO E MONITORAMENTO

[Em desenvolvimento...] O Data Lake contará com tabelas de controle das informações e dos dados. É entendido a importância do monitoramento da quantidade e qualidade das informações tratadas na arquitetura.

9. DETALHES DE SEGURANÇA E POLÍTICAS DE ACESSOS

9.1 Quem acessa e por quê? Dados com criptografia?

O acesso às diferentes camadas do Data Lake é controlado para garantir a segurança e a privacidade dos dados armazenados. As políticas de acesso são baseadas no princípio do menor privilégio, onde apenas as equipes e os indivíduos necessários têm permissão para acessar cada camada do Data Lake.

Os dados armazenados no Amazon S3 são criptografados em repouso usando o serviço de criptografia gerenciada pela AWS, garantindo a confidencialidade dos dados.

9.2 Quais são os controles de acesso?

Para garantir a segurança e o controle de acesso, são implementadas políticas de IAM, na AWS, atribuindo funções e permissões específicas aos usuários e aos componentes do sistema.

Isso significa que apenas as equipes autorizadas têm acesso às instâncias do EMR e às funções do Lambda para realizar o tratamento dos dados. Além disso, o acesso ao Athena, onde os dados estão disponíveis para consulta SQL; assim como o dashboard de indicadores, é controlado por meio de credenciais de autenticação seguras e gerenciadas pela AWS.

Essas medidas de segurança e controle de acesso são essenciais para proteger os dados sensíveis e garantir que apenas os usuários autorizados possam interagir com as diversas camadas do data lake.

10. REQUISITOS

10.1 O que é preciso para usar a ferramenta?

É necessário estar conectado a um VPN da Pod Bank e ter os acessos necessário, de acordo com o nível de permissão do usuário.

10.2 Quais as bibliotecas necessárias para rodar os códigos?

As bibliotecas necessárias para rodar os códigos da arquitetura estão disponíveis no arquivo “requirements.txt”.

10.3 Quais linguagens de programação disponível para execução da estrutura?

A linguagem de programação disponível para execução da estrutura é o SQL.

10.4 Como conecto no dashboard de visualização?

O Dashboard pode ser acessado através do link de acesso para os usuários permitidos.

10.5 Como me conecto no AWS Athena para busca SQL?

A base de dados no AWS Athena pode ser acessada através do link de acesso para os usuários permitidos.