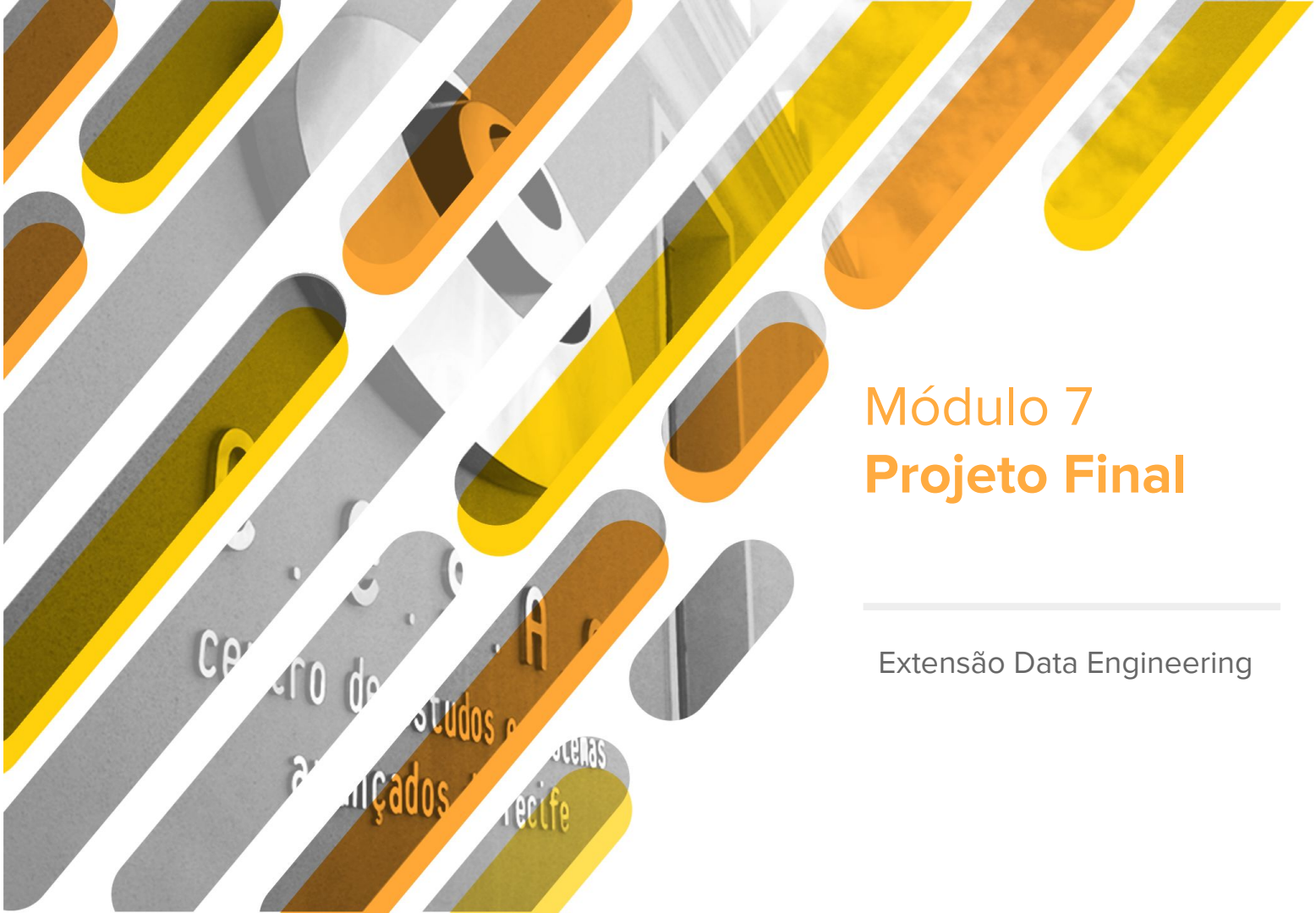

Módulo 7
Projeto Final

Extensão Data Engineering

Contexto

- Módulo final da extensão Data Science
- Turma dividida em grupos de no máximo 5 pessoas
- Cada grupo resolverá um problema no contexto de aplicação da In Loco Media
    - A solução do problema deve ter o formato de um Jupyter Notebook na linguagem Python, contendo o código e os elementos gráficos correspondentes
- Serão avaliados itens técnicos, de comunicação e organizacionais

Contexto

- **Técnicos:**
  - Uso de conceitos e ferramentas apresentados durante os módulos anteriores para configuração do ambiente cloud e desenvolvimento de soluções funcionais

- **Organizacionais:**
  - Participação nos encontros, trabalho em equipe, conformidade aos requisitos do problema

Contexto

- **Comunicação:**
  - Apresentação dos resultados desenvolvidos pelo grupo durante o projeto como uma apresentação de no máximo 30 min, descrevendo a solução desenvolvida por meio de um conjunto de slides e demonstrando em tempo real sua funcionalidade.
  - 14 e 15 de setembro
  - Ambiente colaborativo, Slack: https://bit.ly/2N9KnDv

Orientações

1. Use as credenciais no ambiente AWS para criar a infraestrutura de cluster necessária para o projeto
    a. Um único cluster pode ser instalado com Dask e Spark
2. Use Jupyterhub e Jupyter notebooks para o desenvolvimento
3. Instale de antemão todas as dependências (e.g. bibliotecas de geolocalização)
    a. Dica: no caso do EMR, uma vez que todas as suas dependências estiverem instaladas, criem um arquivo bash para executar a instalação dessas dependências.
        i. Dúvida: https://ryanstutorials.net/bash-scripting-tutorial/bash-script.php
4. Comece o desenvolvimento usando pequenas amostras dos dados, sem preocupação inicial com volumes

Orientações

1. Cada cluster terá **uma quantidade de horas disponível** para cada usuário (a.k.a cada grupo).
2. Lembre-se de utilizar o **S3** para armazenar dados e resultados de processamentos
    a. Lembre-se-2, as instâncias são efêmeras (no caso do EMR/Spark), ou seja, resultados podem ser perdidos.
    b. Sejam organizados, se por algum motivo vocês subscreverem (acidentalmente) algum arquivo, não teremos como recuperar em tempo hábil.
        i. **Trabalhem dentro dos seus buckets!!**
3. Quando criarem uma instância (caso do Dask) lembrem de:
    a. **Colocar um nome na instância identificando o nome do seu grupo. Instâncias sem nome serão apagadas!**
    b. Sempre que não estiverem usando as instâncias, lembrem de desligar.
4. Criem clusters com no máximo 4 instâncias (1 master 3 slaves)
    a. m4.large
    b. Lembre-se, faça pequenos processamentos antes de processar grandes bases.

Problemas

1. Desempenho das campanhas publicitárias
2. Desempenho dos estabelecimentos
3. Correlação entre interações e visitas
4. Atratividade dos anúncios
5. Perfil de interação com de anúncios
6. Perfil de visita aos estabelecimentos
7. Recomendação em tempo real

Grupos

1. Desempenho das campanhas publicitárias
    a. Grupo:
2. Desempenho dos estabelecimentos
    a. Grupo:
3. Correlação entre interações e visitas
    a. Grupo:
4. Atratividade dos anúncios
    a. Grupo:

Grupos

5. Perfil de interação com de anúncios
    a. Grupo:
6. Perfil de visita aos estabelecimentos
    a. Grupo:
7. Recomendação em tempo real
    a. Grupo:

Problema 1

Desempenho das campanhas

- O *click-through rate* (CTR) de uma campanha é a razão entre o número de anúncios que foram clicados e o número de anúncios que foram visualizados
  - Exemplo: campanha A tem um único anúncio, visualizado 10.000 vezes e clicado 185 vezes. CTR = 185/10.000 ou 1,85%
- Os dados disponibilizados indicam ações (cliques ou visualizações) feitas para cada anúncio; uma campanha pode ter mais de um anúncio
- O CTR é uma métrica importante para avaliar a atratividade de uma campanha para determinado público alvo

Problema 1

Desempenho das campanhas

- Para os dados estáticos disponibilizados:
  - Com a plataforma Dask, crie um cluster com um master e dois workers, persista os dados em memória e crie rotinas que calculam o **CTR para cada campanha**, agregando por mês e pelo período completo;
  - Crie gráficos para indicar as **campanhas com melhor CTR** em cada mês e para o período completo;

Problema 2

Desempenho de estabelecimentos

- Os anúncios de uma campanha são relativos a estabelecimentos, listados nos dados disponibilizados contendo sua localização geográfica
- Uma visita de um cliente a um estabelecimento é definida como um registro de localização daquele cliente a menos de 50 metros da localização do estabelecimento
- Saber quais estabelecimentos recebem mais visitas é relevante na avaliação de retorno de campanhas

Problema 2

Desempenho de estabelecimentos

- Para os dados estáticos disponibilizados:
  - Com a plataforma Spark, crie um cluster com um master e dois workers, persista os dados em memória e crie rotinas que calculam para cada registro de localização do um cliente **se houve visita** a um estabelecimento. Agregue os número de visitas a cada estabelecimento **por mês e pelo período completo**;
  - Crie gráficos para indicar os estabelecimentos com mais visitas em cada mês e para o período completo;

Problema 3

Correlação entre interações e visitas

- A geração de uma visita a um estabelecimento a partir de uma interação com um anúncio é uma métrica de importância para a avaliação de sua eficácia
- A correlação forte entre interações e visitas sugere que o anúncio tem poder de convencimento significativo
- A correlação é considerada quando um registro de localização do usuário é feito a menos de 50m do estabelecimento relacionado ao anúncio, em até 6 horas depois da interação ter acontecido

Problema 3

Correlação entre interações e visitas

- Para os dados estáticos disponibilizados:
  - Com a plataforma Dask, crie um cluster com um master e dois workers, persista os dados em memória e crie rotinas que calculam para cada estabelecimento a razão entre o número de visitas relacionadas a interações e o número total de visitas, agregando por mês e pelo período completo;
  - Crie gráficos para indicar por mês os estabelecimentos com maior correlação entre interações e visitas;

Problema 4

Atratividade de anúncios

- Um potencial cliente pode interagir com os anúncios de uma campanha por meio de visualização e cliques
- Cada anúncio pertence a uma categoria, que agrupa áreas de atuação similares, tais como games, restaurantes e hotéis
- O desempenho das categorias de anúncios ao longo do dia pode ser medido pela quantidade de interações registradas

Problema 4

Atratividade de anúncios

- Para os dados estáticos disponibilizados:
  - Com a plataforma Spark, crie um cluster com um master e dois workers, persista os dados em memória e crie rotinas que calculam para cada hora do dia a **quantidade de interações** registradas para cada categoria de anúncios;
  - Crie gráficos para indicar a **distribuição de interações** durante as horas do dia nas categorias de anúncios, agregando por mês e pelo período completo;

Problema 5

Perfil de interação com anúncios

- A interação de potenciais clientes com anúncios de campanhas publicitárias cria um histórico que pode indicar os interesses pessoais
- Uma predominância de interação com determinadas categorias de anúncios, em determinadas horas do dia, podem indicar preferências e orientar futuras ofertas de produtos e serviços
- A identificação deste perfil, relacionado ao período do dia, é valioso para aumentar a eficiência das campanhas

Problema 5

Perfil de interação com anúncios

- Para os dados estáticos disponibilizados:
  - Com a plataforma Dask, crie um cluster com um master e dois workers, persista os dados em memória e crie rotinas que calculam para cada usuário as distribuições de interações com categorias de anúncio período do dia. Crie também uma tabela com a categoria preferida de cada usuário em cada período;
  - Crie gráficos para ilustrar, para alguns usuários, a distribuição de frequências das categorias ao longo do dia e as categorias preferidas;

Problema 6

Perfil de visita aos estabelecimentos

- O padrão de visitas de um potencial cliente aos estabelecimentos conveniados cria um histórico que pode indicar os interesses pessoais
- Uma predominância de interação com determinadas categorias de estabelecimentos, em determinados dias da semana, podem indicar preferências e orientar futuras ofertas de produtos e serviços
- A identificação deste perfil, relacionado ao período da semana, é valioso para aumentar a eficiência das campanhas

Problema 6

Perfil de visita aos estabelecimentos

- Para os dados estáticos disponibilizados:
  - Com a plataforma Spark, crie um cluster com um master e dois workers, persista os dados em memória e crie rotinas que calculam para cada usuário as distribuições de visitas a categorias de estabelecimentos por dia da semana. Crie também uma tabela com a categoria preferida de cada usuário em cada dia da semana;
  - Crie gráficos para ilustrar, para alguns usuários, a distribuição de frequências das categorias ao longo da semana e as categorias preferidas;

Problema 7

Recomendação em tempo real

- Os registros de posição em tempo real recebidos pela plataforma permitem acompanhar o deslocamento de clientes potenciais dos estabelecimentos conveniados
- Ao calcular em tempo real a distância de cada cliente para os estabelecimentos é possível mostrar de ofertas fisicamente próximas, aumentando a conveniência do cliente
- Ofertas em tempo real baseadas em geolocalização criam novas oportunidades para campanhas publicitárias

Problema 7

Recomendação em tempo real

- Para os dados **dinâmicos** disponibilizados:
  - Com a plataforma Spark, crie um cluster com um master e dois workers, conecte-se ao serviço de mensagens que envia as posições dos clientes em tempo real e crie rotinas que calculam a distância de cada cliente para cada um dos estabelecimentos. Com base nisso, crie uma rotina que sugira em tempo real anúncios de estabelecimentos dos quais o cliente está se aproximando;
  - Crie mapas georeferenciados para ilustrar, para alguns usuários, quando a aproximação de um estabelecimento resultou em uma sugestão de anúncio;

Sobre Spark e Dask

Sobre Spark e Dask

- Spark (EMR)
  - Ver slides do último módulo

- Dask (SCHEDULER)
  - Para instalar:http://dask.pydata.org/en/latest/install.html
  - Instalar o Bokeh : https://bokeh.pydata.org/en/latest/docs/installation.html
  - Gerar uma imagem.
  - Gerar novas instâncias a partir dessa instância.
    - Uma delas será o scheduler e as demais workers

Sobre Spark e Dask

- Dask-scheduler

```
distributed.scheduler - INFO - -----------------------------------------------
distributed.scheduler - INFO -   Scheduler at:    tcp://172.31.34.3:8786
distributed.scheduler - INFO - Local Directory:    /tmp/scheduler-f2wvtl5w
distributed.scheduler - INFO - -----------------------------------------------
distributed.scheduler - INFO - Register tcp://172.31.28.205:41188
distributed.scheduler - INFO - Starting worker compute stream, tcp://172.31.28.205:41188
distributed.scheduler - INFO - Register tcp://172.31.25.47:41100
distributed.scheduler - INFO - Starting worker compute stream, tcp://172.31.25.47:41100
```

- Dask-work
  - dask-work 172.31.34.3:8786 (IP do scheduler)
- Geral:
  - Jupiter hub pode ser instalado da mesma forma que no EMR
  - Pacotes específicos também precisam ser instalados

C.E.S.A.R

inovação
é a gente