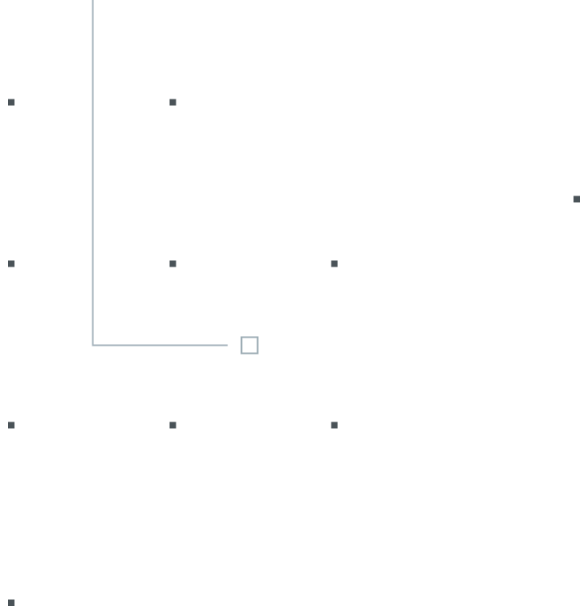


FIAP

NBA



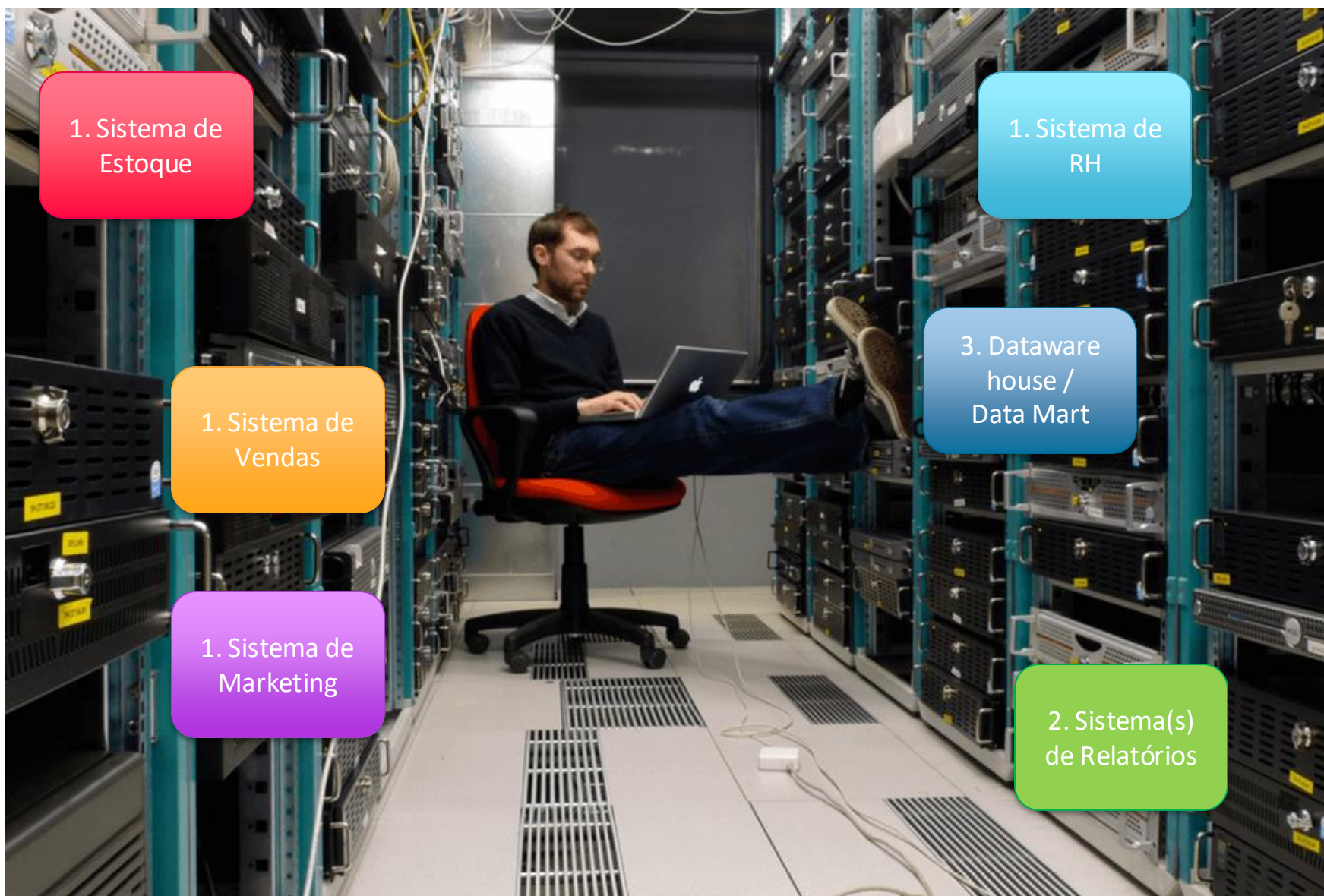
# CLOUD & COGNITIVE ENVIRONMENTS

MBA EM DATA SCIENCE & IA

# Big Data Analysis on Cloud

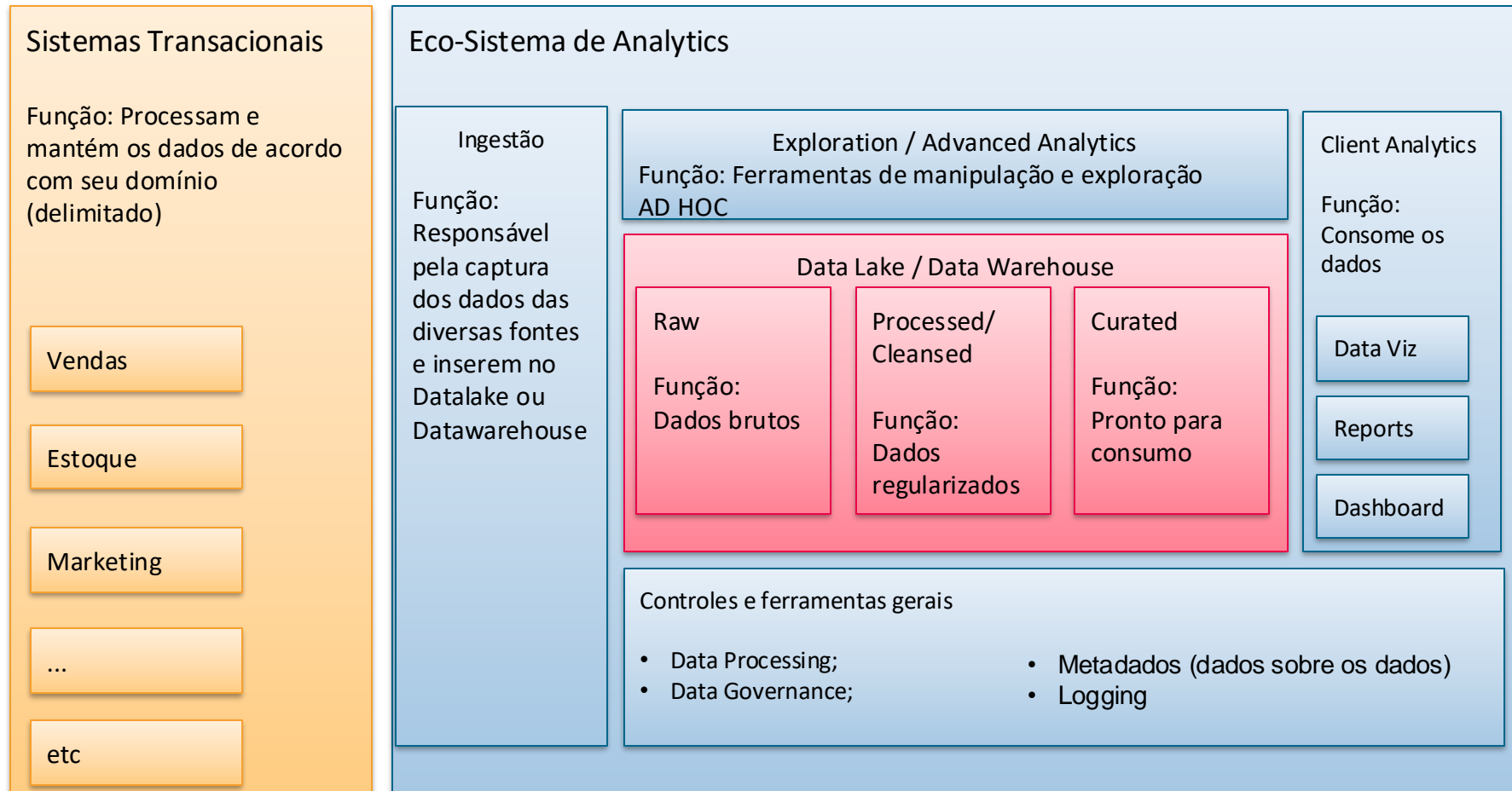
Utilização de nuvem para análise de dados gigantes  
Big Data Data Science on Cloud

# Como são (eram) os servidores das empresas



# Arquitetura tradicional de analytics

Em uma empresa data-driven, uma arquitetura de analytics geralmente é representada desta forma:

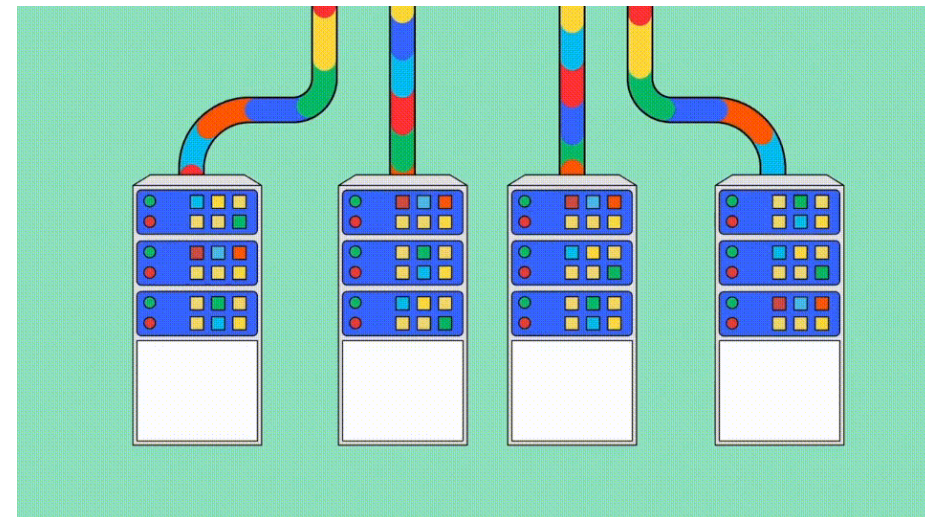


# O que é Cloud Based Analytics

É a capacidade de conseguir fazer **análises de dados** através de softwares e procedimentos baseados em nuvem\*.



Considerando o ecossistema, essa aderência à nuvem por ser total ou parcial.



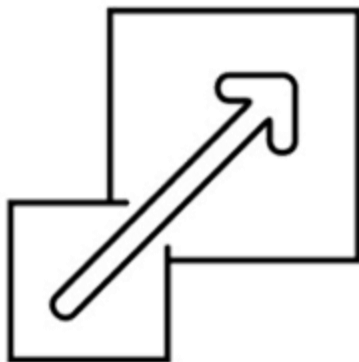
\* Geralmente por assinatura



# Principais motivadores de Cloud Based Analytics



- ☐ Empresa diminui o investimento em infra-estrutura de servidores
- ☐ Menor gasto de energia
- ☐ Sustentabilidade, data centers na nuvem precisam de selo de alta-eficiência energética
- ☐ Menor risco de perda de conteúdo por falhas físicas  
Nota: derrubar café no servidor
- ☐ Discos de Backup” não circulam por aí, com todas as informações da empresa



- ☐ Escalabilidade, se a empresa aumentar de tamanho, basta aumentar o consumo da nuvem
- ☐ Flexibilização do modelo de traba
- ☐ Acesso remoto



# Atividade Prática - Preparação

## Vamos começar analisando dados!!!

Usando a tecnologia que você preferir, vamos analisar os dados para responder algumas perguntas objetivas. Vamos considerar os dados do ENEM dos anos 2018, 2019 e 2020 para nossa análise...

Os dados do enem estão disponíveis no site do INEP em:

<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>

E como em qualquer análise, vamos começar obtendo conhecimento sobre nossos dados



Amostras aleatórias de 10 mil alunos:

Formato Excel:

[https://storage.googleapis.com/cloud-based-analytics\\_public/BI-2024/enem\\_sample10k.xlsx](https://storage.googleapis.com/cloud-based-analytics_public/BI-2024/enem_sample10k.xlsx)

Formato CSV:

[https://storage.googleapis.com/cloud-based-analytics\\_public/BI-2024/enem\\_sample10k.csv](https://storage.googleapis.com/cloud-based-analytics_public/BI-2024/enem_sample10k.csv)



# Atividade Prática: ENEM 2018-2019-2020

Conjunto de dados (dicionário)



Veja os conjuntos  
de dados com atenção

- **DADOS DO PARTICIPANTE**
  - Idade, sexo, estado civil, cor/raça, residência, etc.
- **DADOS DA ESCOLA**
  - Localização, Pública/Privada, etc.
- DADOS DOS PEDIDOS DE ATENDIMENTO ESPECIALIZADO
  - Surdez, cegueira, dislexia, autismo, etc.
- DADOS DOS PEDIDOS DE ATENDIMENTO ESPECÍFICO
  - Gestante, lactante, idoso, hospitalar
- PEDIDOS DE RECURSOS ESPECIALIZADOS E ESPECÍFICOS PARA REALIZAÇÃO DAS PROVAS
  - Braile, libras, fonte ampliada, cadeira especial, etc.
- LOCAL DE APLICAÇÃO DA PROVA
  - Município, UF
- **DADOS DA PROVA OBJETIVA**
  - Presença, tipo de prova, nota, respostas, gabarito
- **DADOS DA REDAÇÃO**
  - Situação, notas de competências, nota final
- **DADOS DO QUESTIONÁRIO SOCIOECONÔMICO**
  - Instrução dos pais, profissão dos pais, renda, pessoas na residência, casa própria, carros, etc.

CN – Ciências da Natureza  
CH – Ciências Humanas  
LC – Linguagens e Códigos  
MT – Matemática  
REDAÇÃO

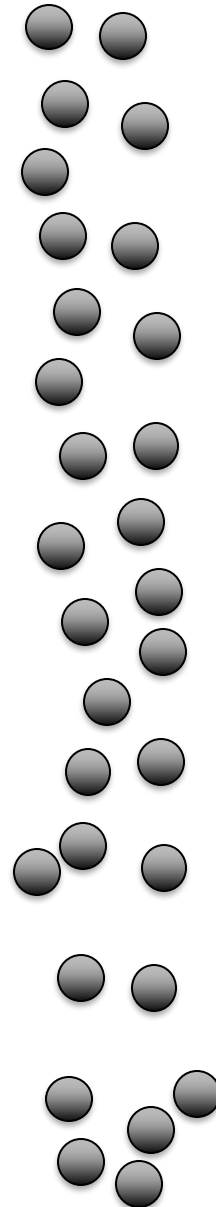
Dicionário detalhado em:

[https://storage.googleapis.com/cloud-based-analytics\\_public/BI-2024/DicionarioEnem.xlsx](https://storage.googleapis.com/cloud-based-analytics_public/BI-2024/DicionarioEnem.xlsx)

# Nossa análise

Imagine que cada bolinha dessas  
é um participante do ENEM

Cada uma dessas pessoas  
possui um “Background”,  
uma história!



Após fazerem a prova, cada  
participante obtém uma  
nota.

Se ordenarmos as notas,  
da menor nota à maior  
nota teremos :

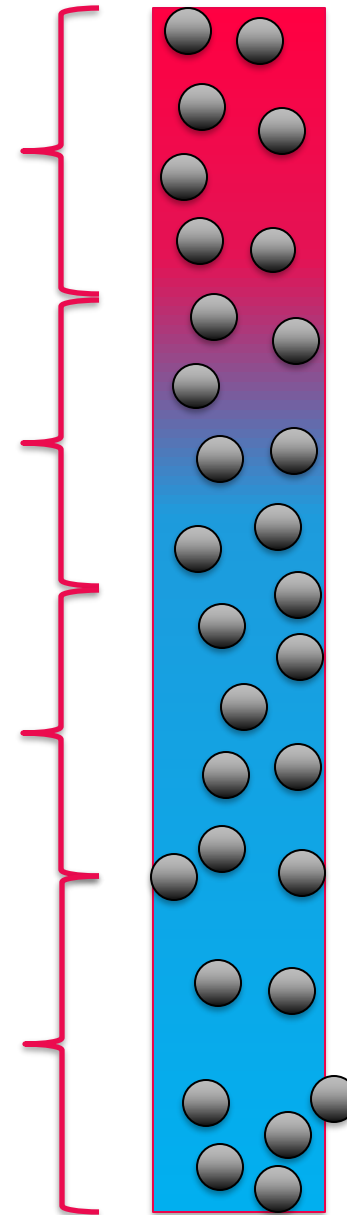
Melhores Resultados  
em uma área ou no geral

Resultados médios  
superiores

Resultados médios  
inferiores

Piores resultados

## Nossa análise



Hipótese: Será que o  
histórico influencia na nota?

Vamos responder essa  
hipótese fazendo análises  
nos dados

# Atividade Prática: ENEM 2018-2019-2020

## Tarefa

Crie um notebook simples (pode usar Colab) em Python para responder às perguntas (15 minutos):

1. Qual a distribuição dos participantes por estado onde a prova foi realizada em cada ano? E no total do Brasil?  
**Questão teste:** Qual ano e estado que tiveram mais provas aplicadas?
2. Considerando apenas 2018, quais as métricas globais de média, mediana, primeiro quartil (25%), terceiro quartil (75%) dos participantes em matemática (NU\_NOTA\_MT)?  
Obs.: Considerar apenas quem participou da prova, TP\_PRESENCA\_MT = 1  
**Questão teste:** Qual o terceiro quartil Nacional em matemática?
3. Faça a mesma análise de 2018, porém considerando quebras por estado (local da prova). Indique quais estados tem métricas superiores e inferiores do que as métricas nacionais.  
**Questão teste:** Qual o terceiro quartil de Roraima? Como está RR comparado com BR?
4. Faça a mesma análise de 2018, porém considerando escolaridade do pai (Q001). Indique quais casos possuem métricas superiores e inferiores do que as métricas nacionais.  
**Questão teste:** Qual a mediana dos participantes cujos pais completaram o ensino médio e qual a mediada quando os pais concluíram a faculdade?

Nota: Submeter o notebook gerado pelo portal

# Nossa arquitetura

## O que acabamos de fazer?



Fonte dos dados  
(Site/URL)  
5,4MB  
9913 registros (linhas)

Internet



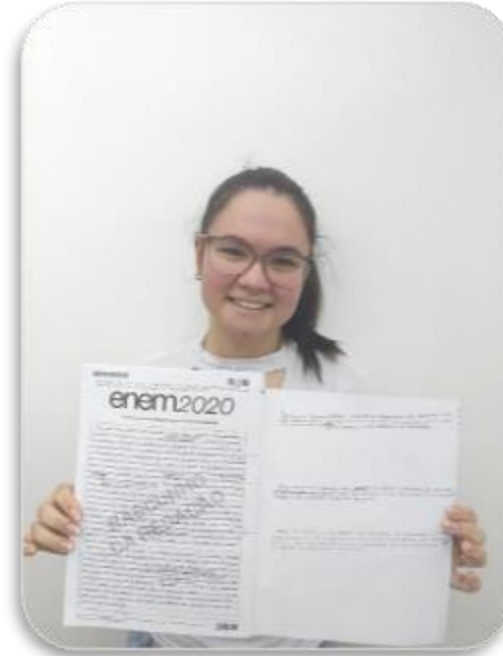
Seu computador físico

# Chega de brincadeira, agora vamos trabalhar sério!

ENEM 2018  
5.513.747 participantes  
5,64 GB

ENEM 2019  
5.095.270 participantes  
5,25 GB

ENEM 2020  
5.783.109 participantes  
3,58 GB



Boa sorte pra  
vocês!

Marianna Caldas, 980 na  
redação do ENEM 2020

[https://storage.googleapis.com/cloud-based-analytics\\_public/BI-2024/MICRODADOS\\_ENEM\\_2018.csv](https://storage.googleapis.com/cloud-based-analytics_public/BI-2024/MICRODADOS_ENEM_2018.csv)  
[https://storage.googleapis.com/cloud-based-analytics\\_public/BI-2024/MICRODADOS\\_ENEM\\_2019.csv](https://storage.googleapis.com/cloud-based-analytics_public/BI-2024/MICRODADOS_ENEM_2019.csv)  
[https://storage.googleapis.com/cloud-based-analytics\\_public/BI-2024/MICRODADOS\\_ENEM\\_2020.csv](https://storage.googleapis.com/cloud-based-analytics_public/BI-2024/MICRODADOS_ENEM_2020.csv)

Download  
Opcional



# Chega de brincadeira, agora vamos trabalhar sério!



Power BI:  
Problemas de performance a partir de 20000 registros

Limites:

- 1GB para conta free
- 10GB para contas Pró



SQL / Oracle

Necessita de servidores maiores  
para grandes análises



Excel:

Limite de 1.048.576 linhas



Python e R:

Memória do computador para os  
dados + overhead dos objetos

# Chega de brincadeira, agora vamos trabalhar sério!

ENEM 2018  
5.513.747 participantes  
5,64 GB

ENEM 2019  
5.095.270 participantes  
5,25 GB

ENEM 2020  
5.783.109 participantes  
3,58 GB

## Mas como?



# Análise de dados em universo de dados gigantes (BigData)?

Para muitos problemas das empresas, o volume dados impede que a arquitetura usada até o momento funcione.

É necessário ter:

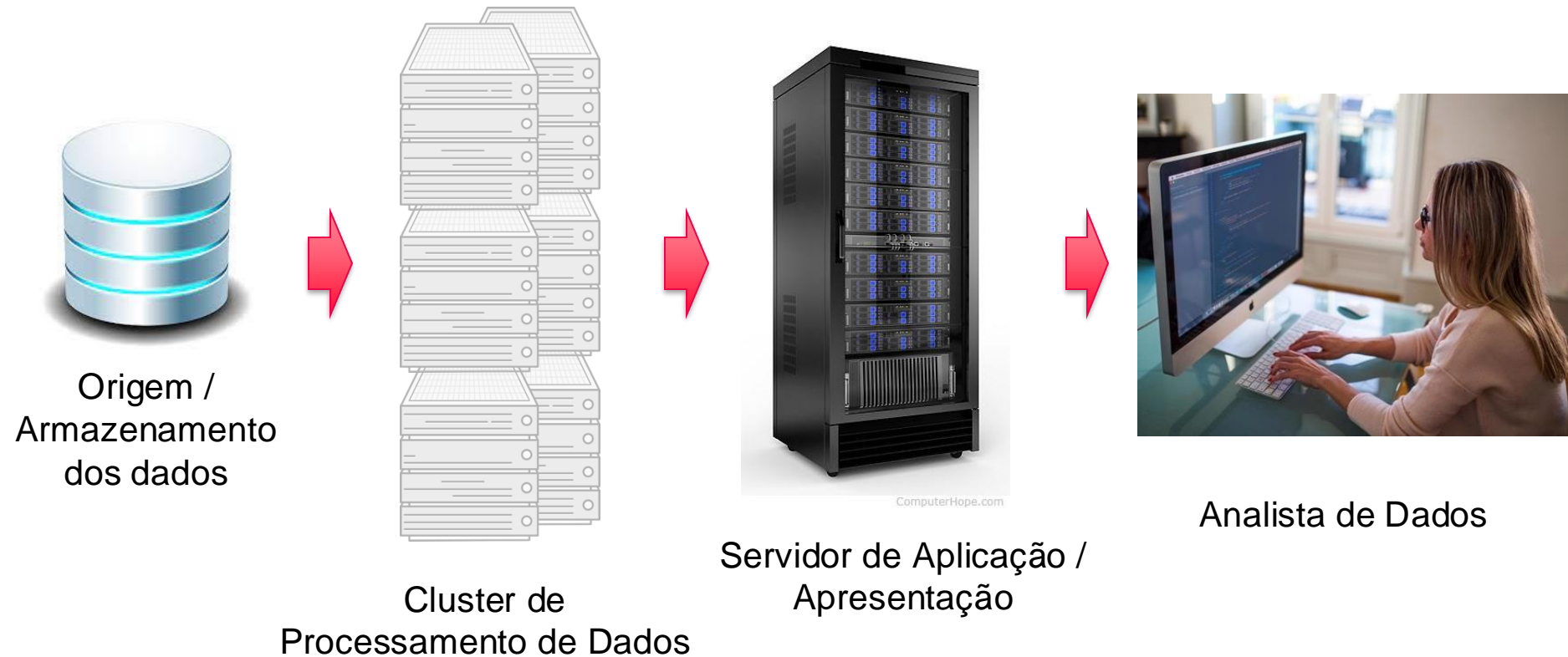
1. Muitos computadores disponíveis para o armazenamento;
2. Muitos computadores disponíveis para fazer o processamento;
3. O processamento das análises “**mais perto**” dos dados.



## Data Analytics

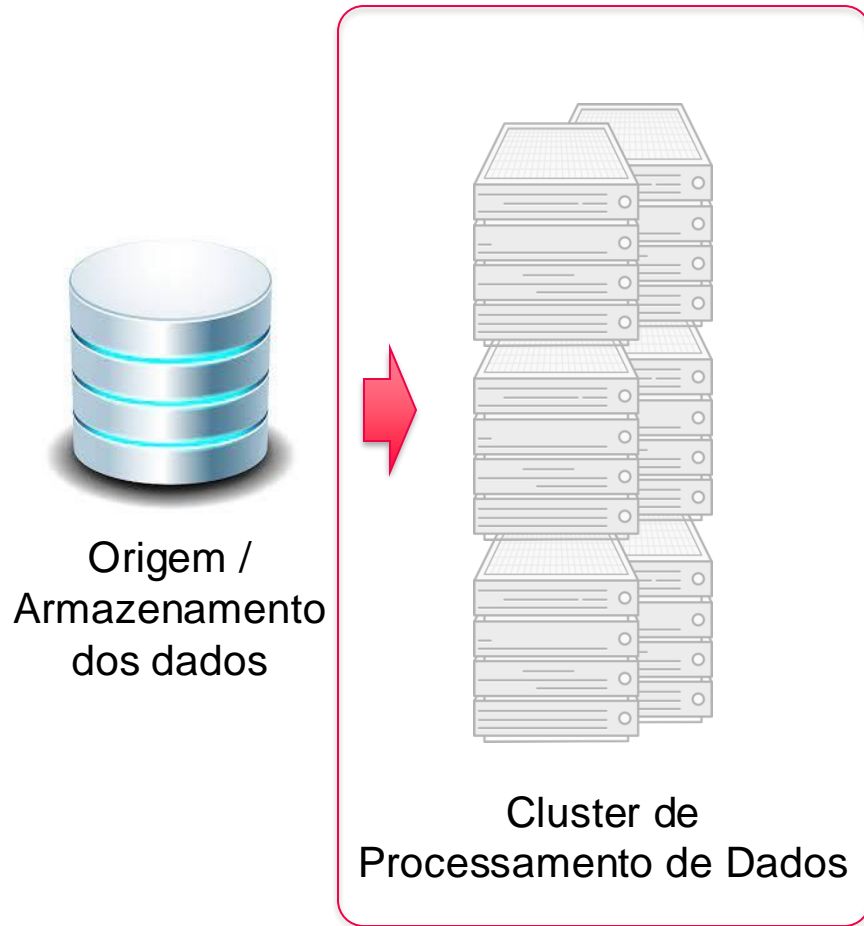
Nota: Apesar deste exercício termos alguns GB para processar, a arquitetura é a mesma para TB e PB de dados

# Análise de dados em universo de dados gigantes (BigData)?



# Análise de dados em universo de dados gigantes (BigData)?

Vamos buscar alternativas



# Análise de dados em universo de dados gigantes (BigData)?

Quais as principais alternativas utilizadas pelas empresas atualmente?



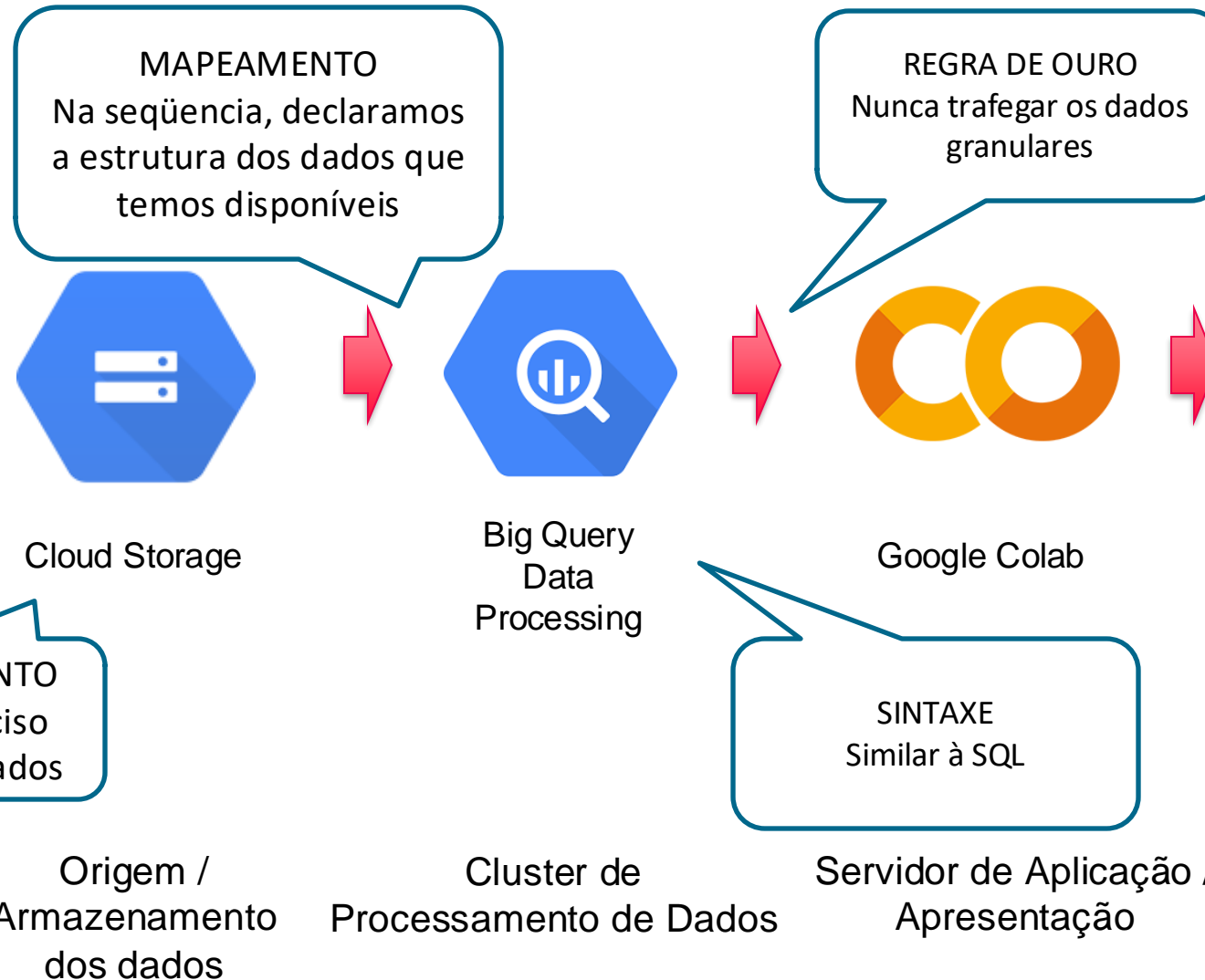
Quais dessas soluções de processamento são

On Premises e On Cloud?



# Proposta de Solução

Então: Vamos fazer **análises de dados** através de softwares e procedimentos baseados em nuvem.



Nós aqui na FIAP

Analista de Dados

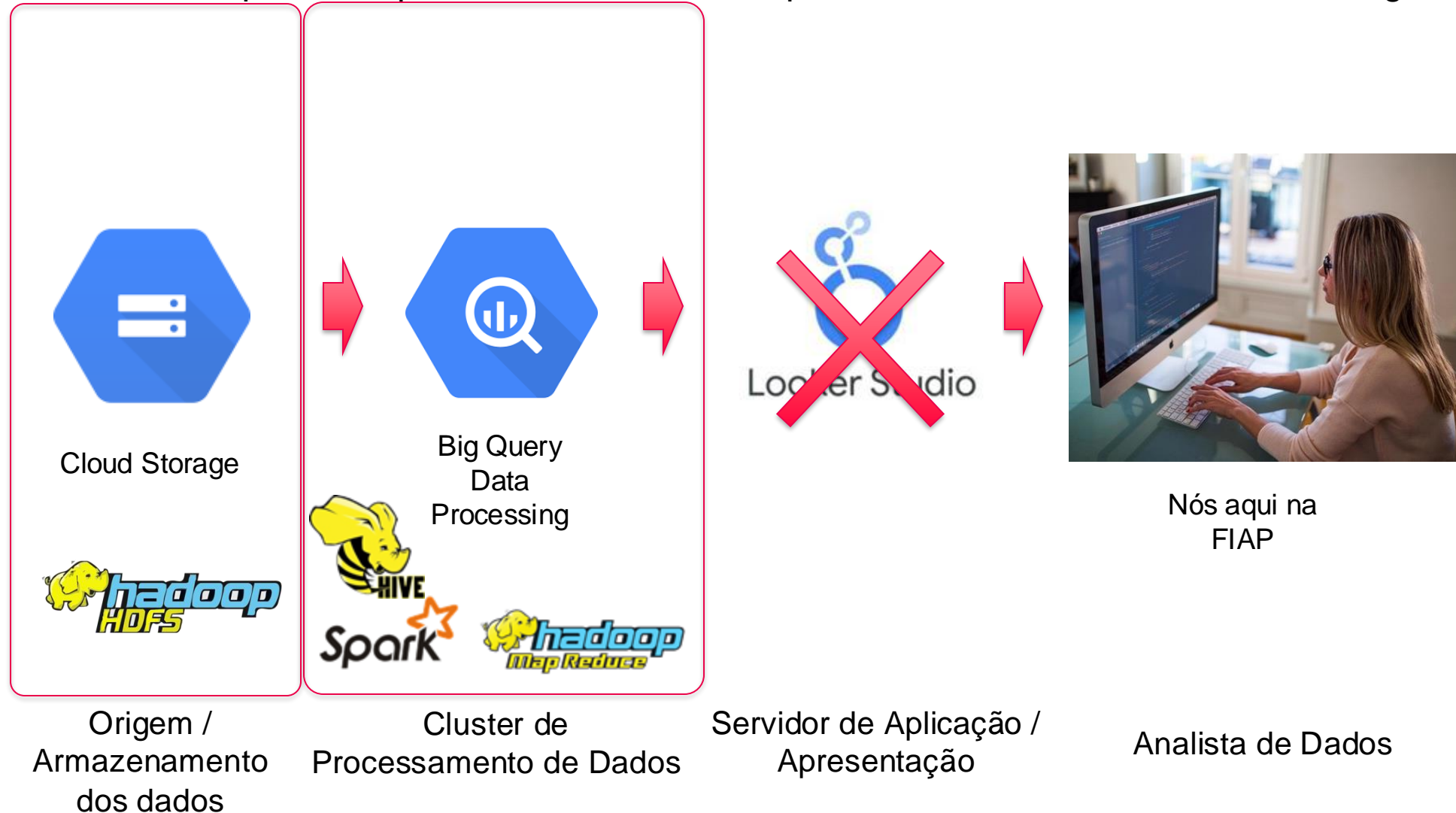
# Proposta de Solução

Então: Vamos fazer **análises de dados** através de softwares e procedimentos baseados em nuvem.



# Alternativa – Looker Studio (atalho, não vale como entrega) FIAP MBA+

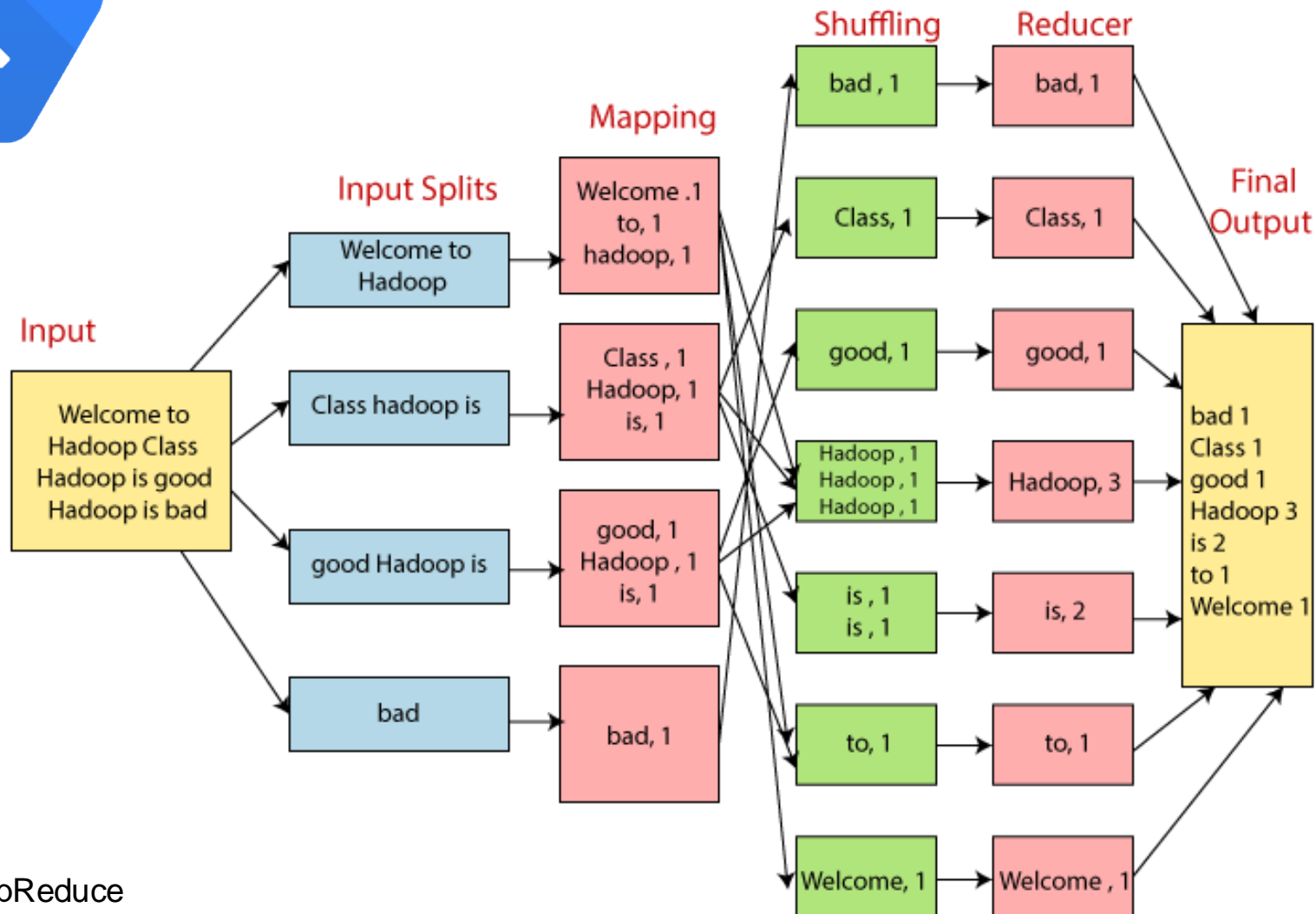
Existem novas alternativas em que a camada de apresentação realiza o processamento com instruções ao Cluster. Esta solução não será utilizada pois **mascara os principais conceitos de tráfego de dados**, competência muito importante que o cientista de dados precisa conhecer em análises com Big Data.



# Sobre BIG QUERY e MapReduce com Word Count



O processamento do BIG QUERY é baseado no MapReduce, formulado em 1995 pelo time do Google.



# Setup on GCP



VAMOS BOTAR A MÃO NA MASSA



# Nuvem – Google Cloud Platform



Nome do teu Projeto ▼

## GCP Projects

- Agrupador lógico de recursos de nuvem



## Cloud Storage

- Armazenamento de objetos



## Big Query

- Mapeamento de dados em tabelas



## Big Query

- Processamento de dados (Serverless/Clusterless)



Google Cloud



Nome do teu Projeto ▼

Conta na GCP



# Proposta de Solução

Então: Vamos fazer **análises de dados** através de softwares e procedimentos baseados em nuvem.



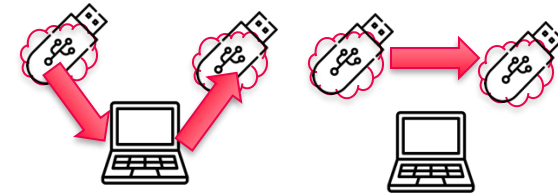
# Big Data Analysis

Setup de um ambiente Serverless/Clusterless utilizando BigQuery

# Atividade – Big Data Analysis – Parte 01 Setup

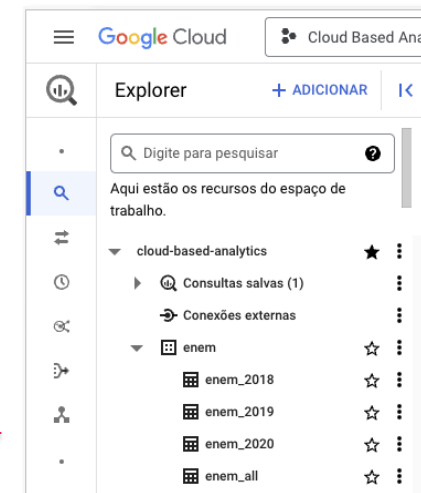
Na primeira parte da atividade vamos configurar o armazenamento e o mapeamento das tabelas

1. Localizar Cloud Storage e criar um bucket de armazenamento (ou bloco)
  - Imagine que você está criando um pendrive global
  - Por isso, o nome deve ser único no mundo
  - Nota: Criar **Multi-region em “US”** para estar na mesma região de processamento
2. Copiar os 3 arquivos informados anteriormente para uma pasta no seu Bucket



```
gsutil cp gs://cloud-based-analytics_public/BI-2024/MICRODADOS_ENEM_*.csv gs://<nome_do_seu_bucket>/fiapbigdata/
```

3. No BigQuery, criar conjunto de dados chamado “**enem**” (Multirregional US)
4. No BigQuery criar 3 tabelas chamadas enem\_2018, enem\_2019 e enem\_2020.
  - Observe que os dados apresentam colunas diferentes, logo, deve-se criar tabelas diferentes também.
  - O slide a seguir mostra uma **colinha** dessa importação
5. Criar uma Tabela Nativa final chamada “enem\_all” com o conteúdo das 3 tabelas
  - Aguardar o professor
6. Submeter no portal o print do Setup, conforme figura exemplo ao lado.



## Criar tabela

## Origem

Criar tabela de  
Google Cloud Storage

Selecione o arquivo do bucket do GCS ou use um padrão de URI [?](#) \*  
☒ nomeunico/fiapbigdata/MICRODADOS\_ENEM\_2018.csv

Formato do arquivo  
CSV

☐ Particionamento de dados de origem

## Destino

Projeto \*  
sacred-truck-402301

Conjunto de dados \*  
enem

Tabela \*  
enem\_2018

Letras Unicode, marcas, números, conectores, traços ou espaços são permitidos.

Tipo de tabela  
Tabela externa

**i** Os buckets regionais/birregionais do GCS são recomendados para tabelas externas.

☐ Criar uma tabela do BigLake usando uma conexão do Cloud Resource

## Esquema

☒ Detectar automaticamente

**i** O esquema será gerado automaticamente.

Opções avançadas

Discussão:  
Tabela Externa  
vs Tabela Nativa

# Criação de Tabela única

- A instrução à direita cria um JOB de processamento Spark a partir de uma instrução SQL;
- Observe com o professor que trata-se da **união** de 3 buscas de origens distintas, com particularidades para equalização das colunas;
- Observe também que uma nova tabela enem.enem\_all será criada ao final deste processamento, com os resultados da busca.
- Explorar com o processor:
  - Tempo de processamento decorrido e total
  - Detalhes da execução
  - Armazenamento interno (Parquet vs CSV)

A instrução completa pode ser obtida através do link:

[https://storage.googleapis.com/cloud-based-analytics\\_public/BI-2024/CREATE\\_OR\\_REPLACE\\_TABLE\\_enem\\_all.sql.txt](https://storage.googleapis.com/cloud-based-analytics_public/BI-2024/CREATE_OR_REPLACE_TABLE_enem_all.sql.txt)

```
CREATE OR REPLACE TABLE enem.enem_all
AS
SELECT
  farm_fingerprint( GENERATE_UUID()) as HASHID,
  NU_INSCRICAO,
  NU_ANO,
  ...
  Q027,
  CAST( NULL as int) as IN_TEMPO_ADICIONAL,
  CAST(NULL as int) AS TP_FAIXA_ETARIA
FROM enem.enem_2018
```

UNION ALL

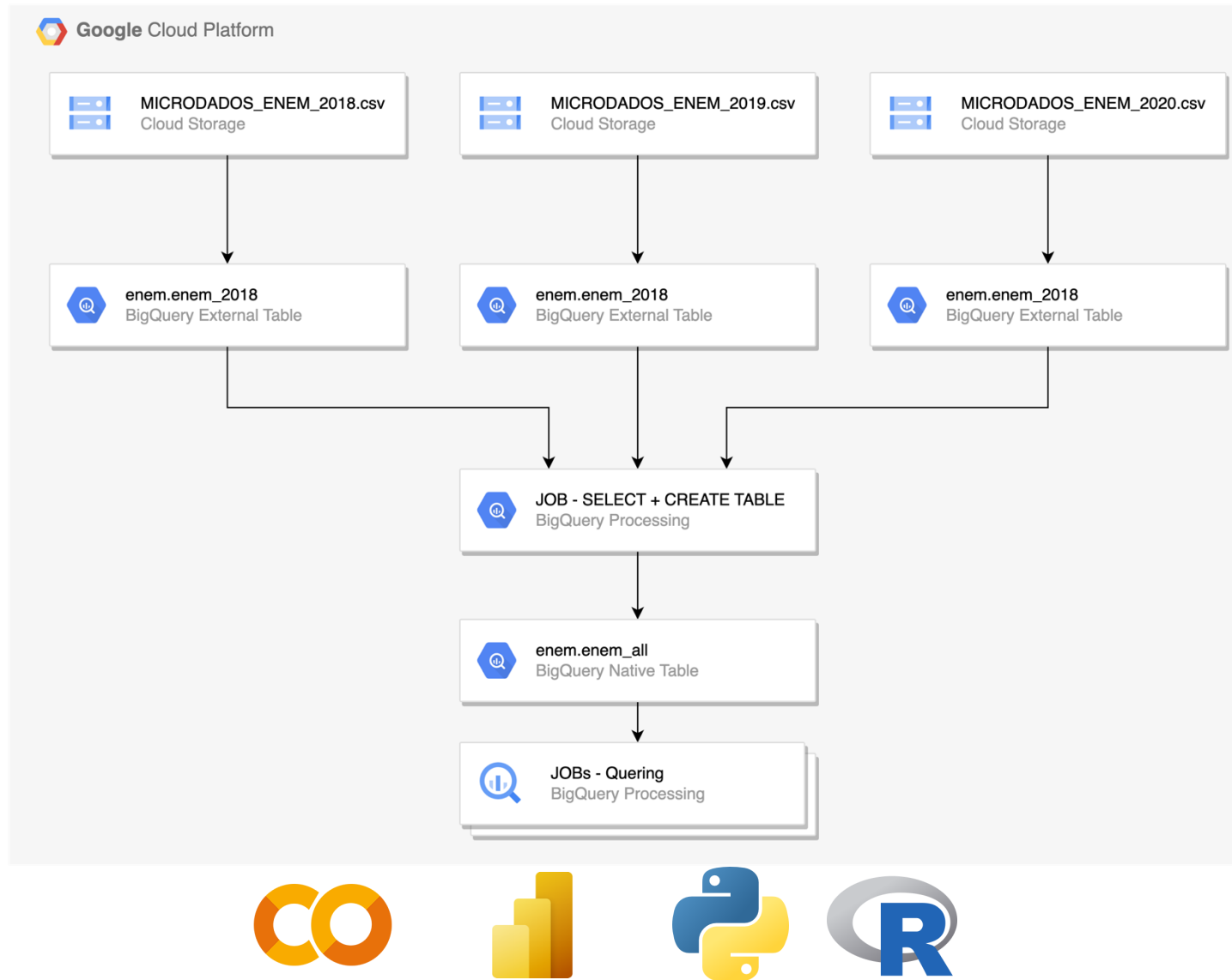
```
SELECT
  farm_fingerprint( GENERATE_UUID()) as HASHID,
  NU_INSCRICAO,
  ...
  CAST (NULL as string) AS Q026,
  CAST (NULL as string) AS Q027,
  IN_TEMPO_ADICIONAL,
  CAST(NULL as int) AS TP_FAIXA_ETARIA
FROM enem.enem_2019
```

UNION ALL

```
SELECT
  farm_fingerprint( GENERATE_UUID()) as HASHID,
  NU_INSCRICAO,
  NU_ANO,
  CAST (NULL as int) AS CO_MUNICIPIO_RESIDENCIA,
  ...
  CAST (NULL as string) AS Q026,
  CAST (NULL as string) AS Q027,
  CAST( NULL as int) as IN_TEMPO_ADICIONAL,
  TP_FAIXA_ETARIA
FROM enem.enem_2020;
```

# Arquitetura

- O diagrama a seguir sumariza a arquitetura da tarefa que acabamos de realizar





# Atividade – Big Data Analysis

Query & Presentation

# Proposta de Solução

Neste material temos uma proposta de solução contemplando:

- Query Analytics
  - Análises através de queries (buscas) nas fontes de dados com cluster de processamento
- Análise de detalhes da execução em múltiplos computadores
  - Demonstração para a turma usando interface gráfica
- Geração de gráficos de análise
  - Com os dados sumarizados no servidor de aplicação, podemos fazer exibições gráficas
- Análise do tempo de execução
- Análises de dados em bancos “tradicionais”
- Aluguel de um cluster de máquinas por poucos segundos

<https://colab.research.google.com/drive/1xuwXTSuRTzs59vRw5VvRkAcEKJVS3K-3>

Conteúdo de referência:

<https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax>

<https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators>



## Exercícios para fazer no material provisionado

- Exercício 1

Faça análise análoga à realizada por estado, porém observe como a escolaridade da mãe influencia nas notas do ENEM (Q002). Comente sua análise!

- Exercício 2

Faça análise análoga à realizada por estado, porém observe como ter computador em casa influencia nas notas do ENEM (Q024). Comente sua análise!

- Exercício 3

Faça análise análoga à realizada por estado, porém observe como a renda mensal da família influencia nas notas do ENEM (Q006). Comente sua análise!

Submeter pelo portal o Notebook com os exercícios realizados

<https://colab.research.google.com/drive/1xuwXTSuRTzs59vRw5VvRkAcEKJVS3K-3>

Conteúdo de referência:

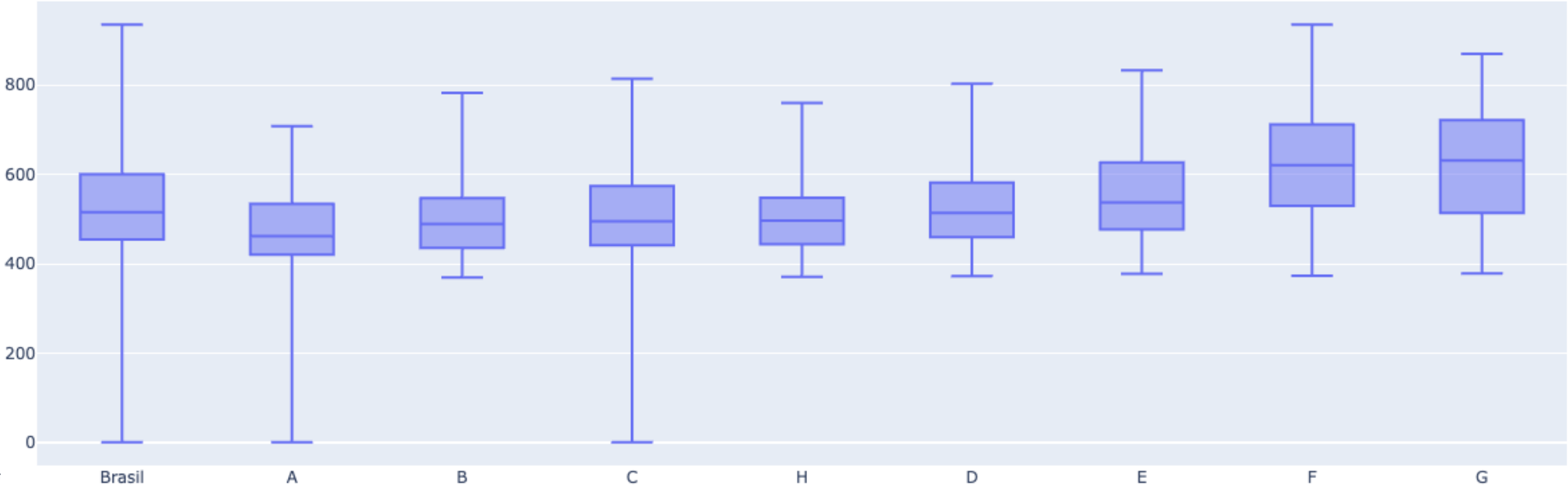
<https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax>

<https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators>



# Conclusões sobre as análises realizadas

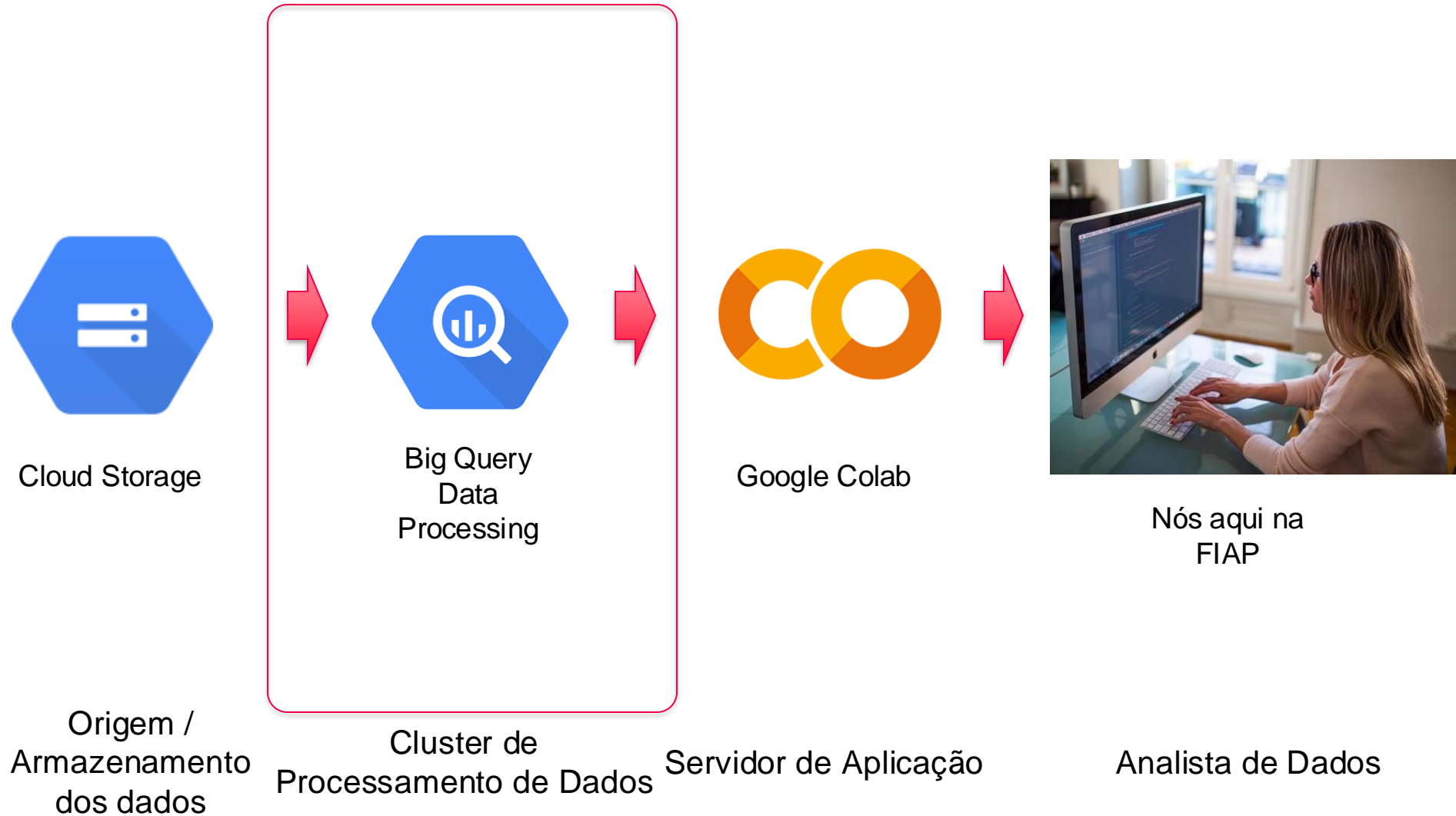
A	Nunca estudou.
B	Não completou a 4ª série/5º ano do Ensino Fundamental.
C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
E	Completou o Ensino Médio, mas não completou a Faculdade.
F	Completou a Faculdade, mas não completou a Pós-graduação.
G	Completou a Pós-graduação.
H	Não sei.



# Discussão sobre a arquitetura



# Discussão sobre a arquitetura



# ATIVIDADE - DESAFIO

Considere que a nota final do ENEM é composta pela nota das quatro áreas mais a nota de redação, responda:

Como a escolaridade de ambos os pais SIMULTANEAMENTE influencia na nota final do ENEM?

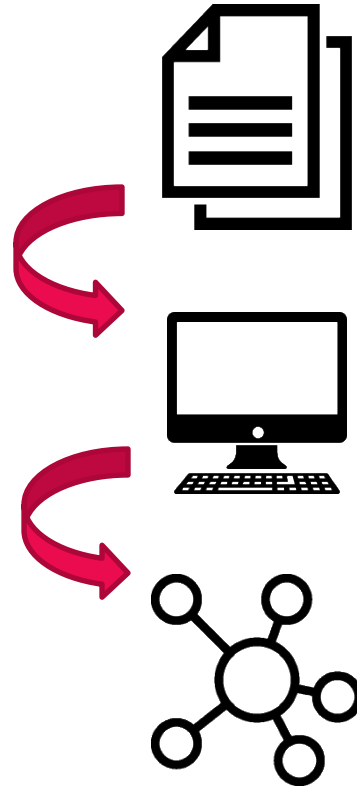
Dica: Primeramente examine cada um dos pais individualmente, então faça uma nova análise considerando todo o casal (A+A, A+B, B+A, etc.)

CN – Ciências da Natureza  
CH – Ciências Humanas  
LC – Linguagens e Códigos  
MT – Matemática  
REDAÇÃO



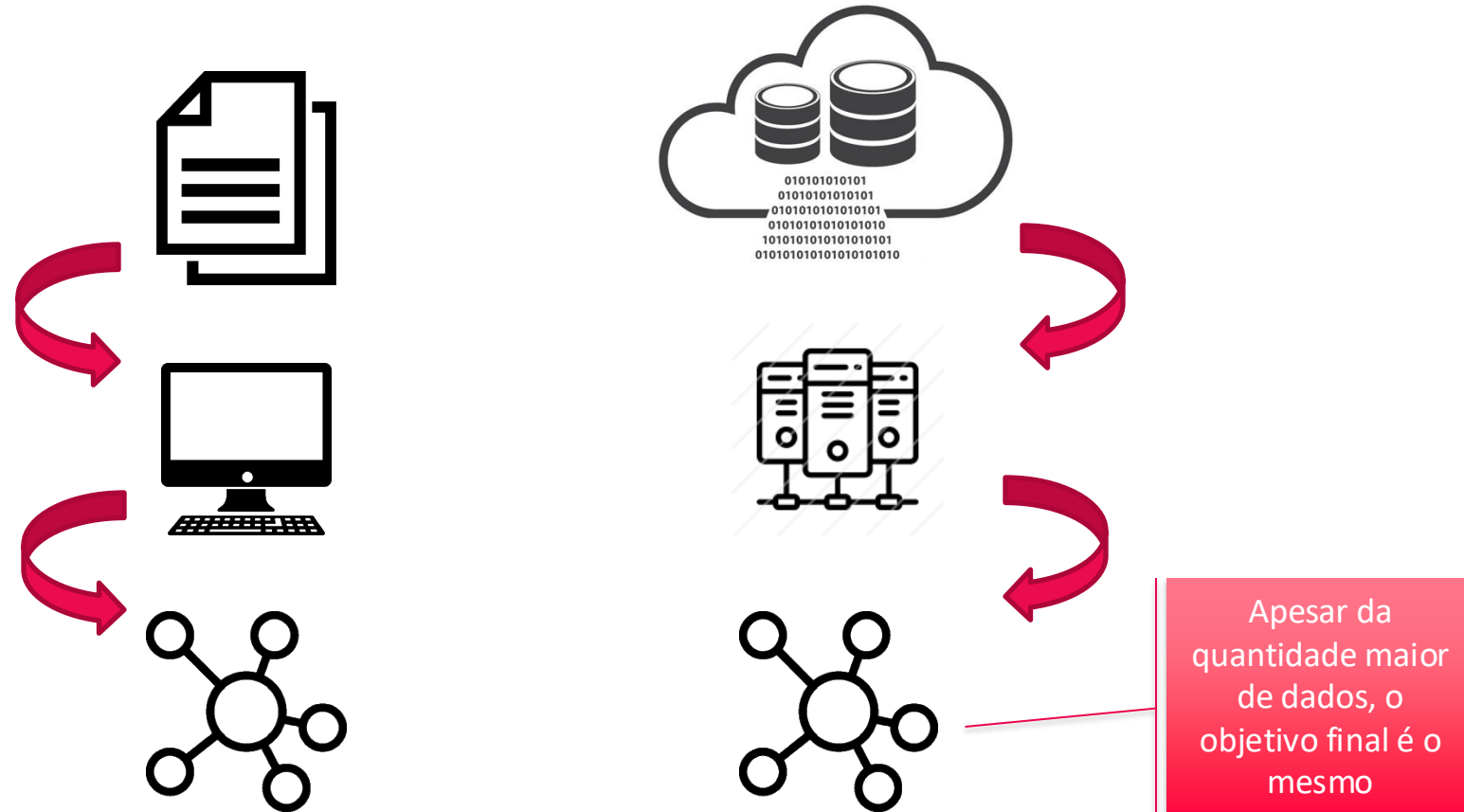
# Machine Learning on Big Data

Como fazer modelos de Machine Learning  
em um conjunto de dados gigante?  
(abordagem usando fundamentos de nuvem e processamento distribuído)



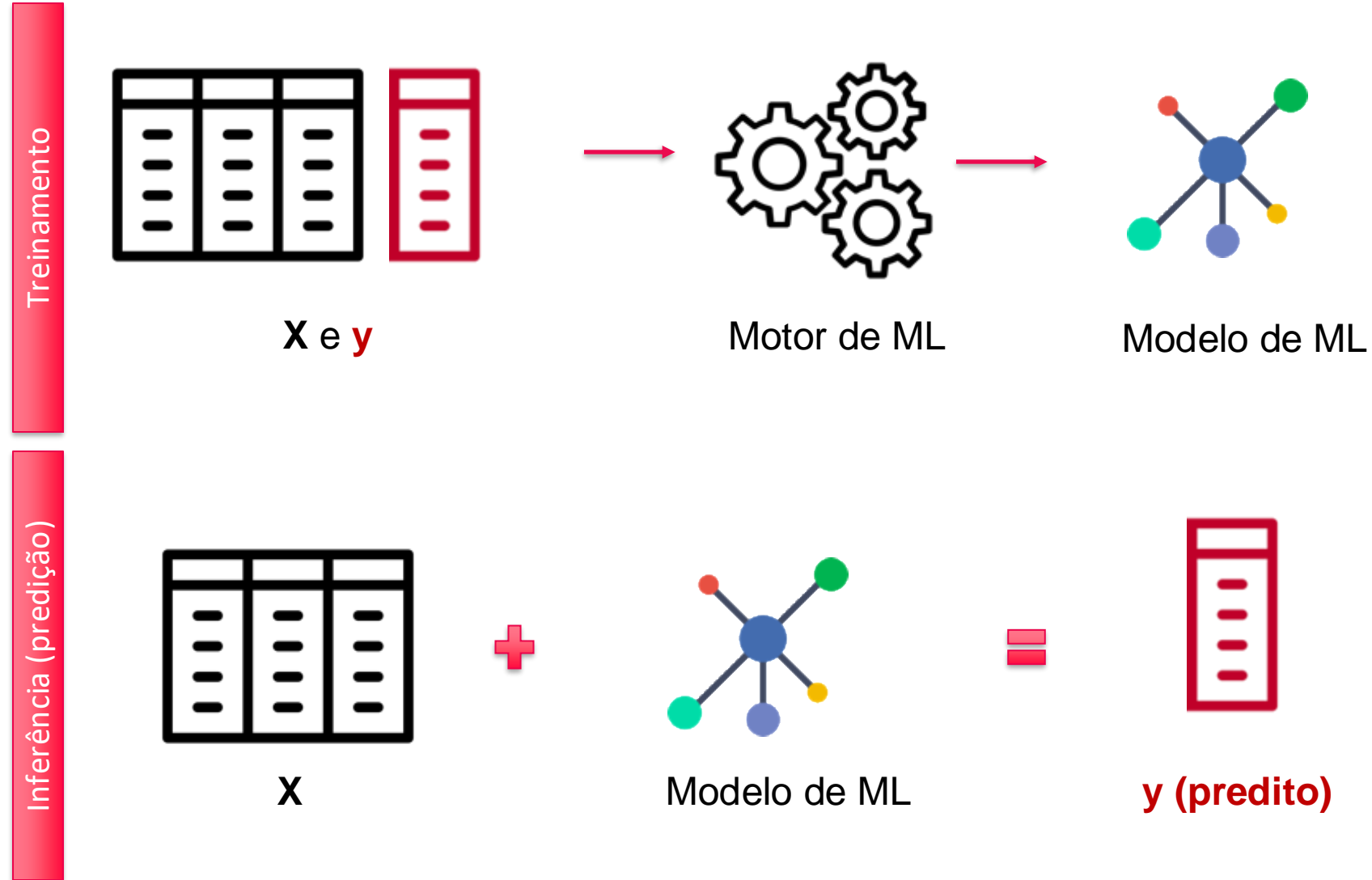
- Vocês já devem ter criado vários modelos de Machine Learning no seu computador, com um conjunto de dados (dataset) controlado.

# MOTIVAÇÃO



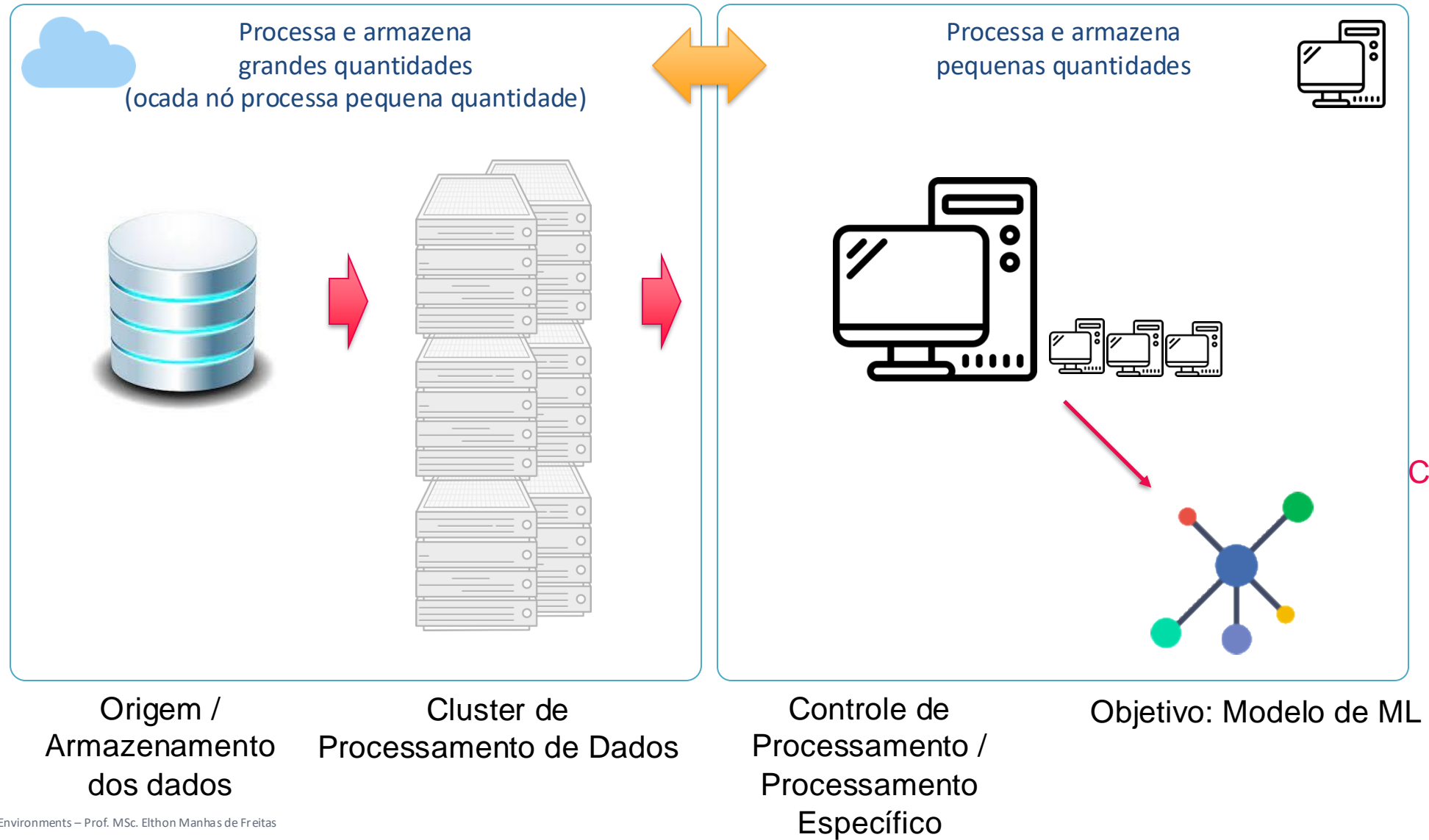
- Mas a realidade do dia a dia nos dá muito mais dados do que conseguimos processar em um só computador.
- Quando buscamos modelos preditivos profissionais que precisam lidar com grandes volumes de dados, precisamos de soluções robustas de processamento.

# Modelos de Machine Learning Visão geral



# Modelos de Machine Learning com Big Data (Visão geral) FIAP MBA+

Existem diversas fases desde a coleta dos dados, até a geração do modelo propriamente dito e temos dois locais de processamento. Em cenários distintos é preciso escolher quais processamentos ocorrem dentro do cluster e quais ocorrem fora do cluster.



## Treinamento dentro do Cluster de Processamento



- Requer acesso à configuração das máquinas do Cluster de Processamento
  - Requer contratação das máquinas
  - Complexidade na configuração
- Principais Frameworks
  - Spark MLlib (modelos simples de ML)
  - TensorFlow (já existem distros para Spark)
  - Sparkling Water (para modelos com H2O)

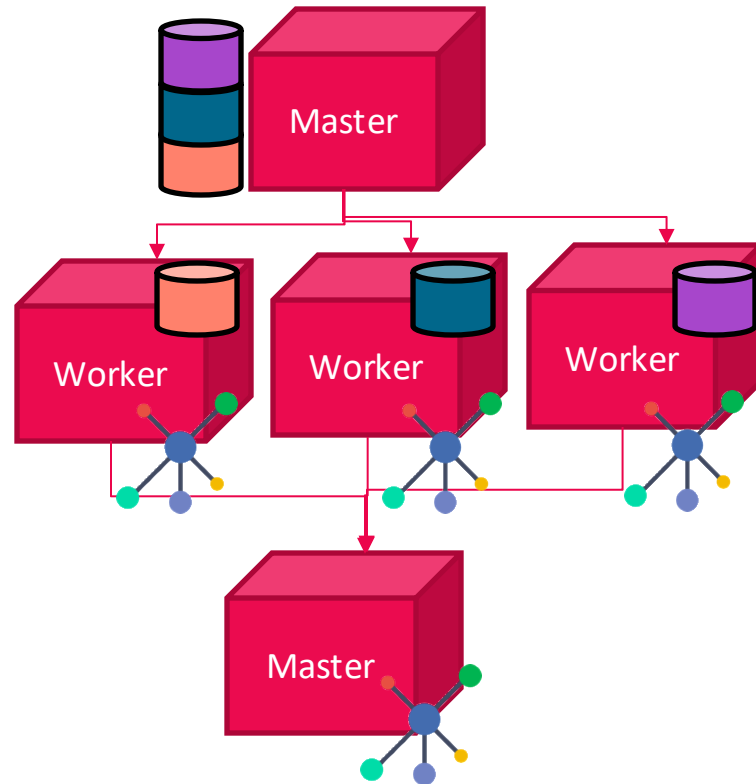
## Treinamento fora do Cluster de Processamento



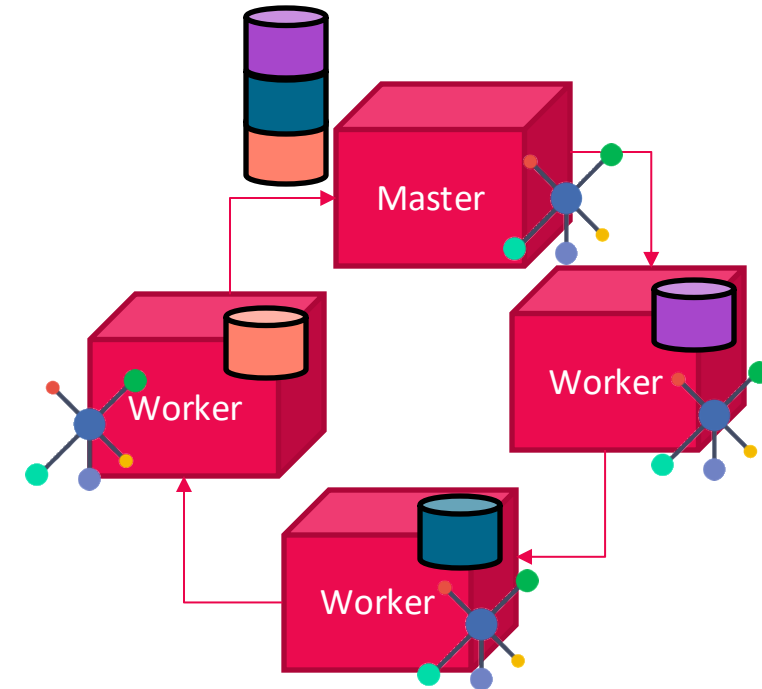
- Requer controle do processamento
  - Necessário saber como dividir a carga
  - Necessário conhecimento de processamento distribuído
    - (checkpoints, joins de modelos, etc.)
- Flexível por algoritmo e Framework
  - Requer maior domínio do Framework utilizado
- Solução independente do Cluster
  - Fácil migração entre soluções de BigData

# Estratégias de Treinamento Distribuído

## Paralelismo de Dados e Seleção (K-Fold)



## Sequencial com dados parciais



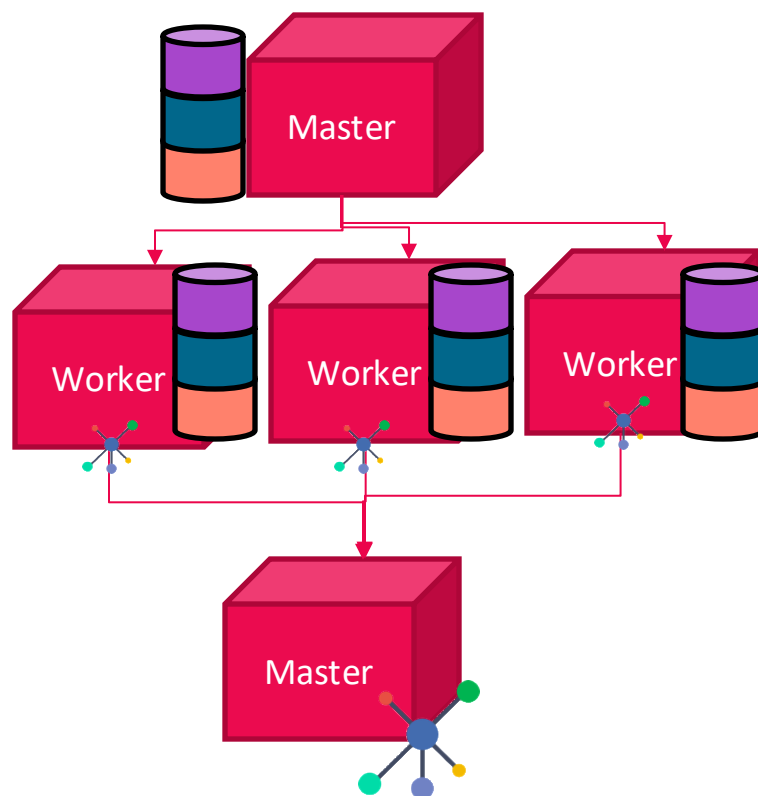
Cada Worker tem acesso a uma parte dos dados, e criam um modelo final  
Requer um processamento final para escolha do melhor modelo

Cada Worker tem acesso a uma parte dos dados, e criam um modelo único final



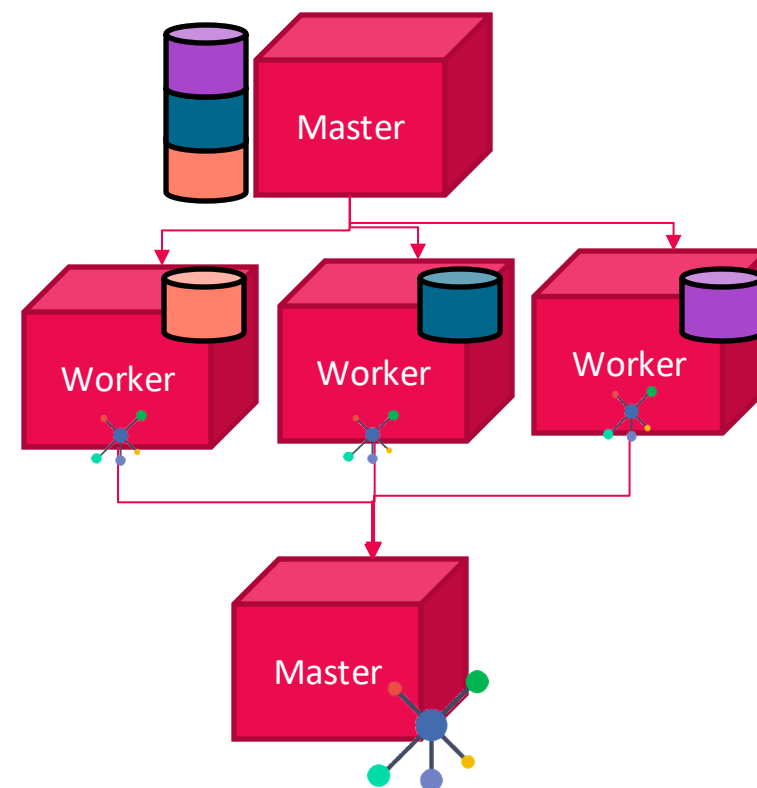
# Estratégias de Treinamento Distribuído

## Paralelismo de Processamento



Todos os Workers tem acesso ao dado completo, e criam partes do modelo  
Requer um processamento final para unificar as partes do modelo  
Internamente, no worker, está havendo vários treinamentos sequenciais

## Paralelismo de Dados

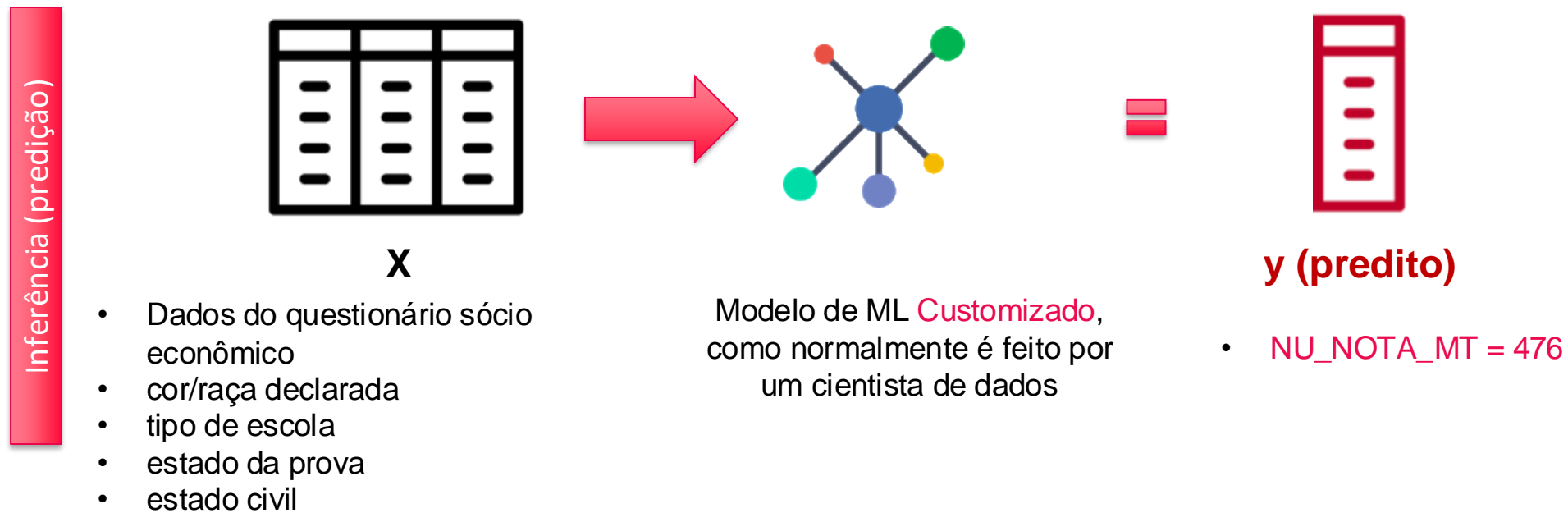


Cada Worker tem acesso a uma parte dos dados, e criam partes do modelo  
Requer um processamento final para unificar as partes do modelo

# Aprendizado através de Prática

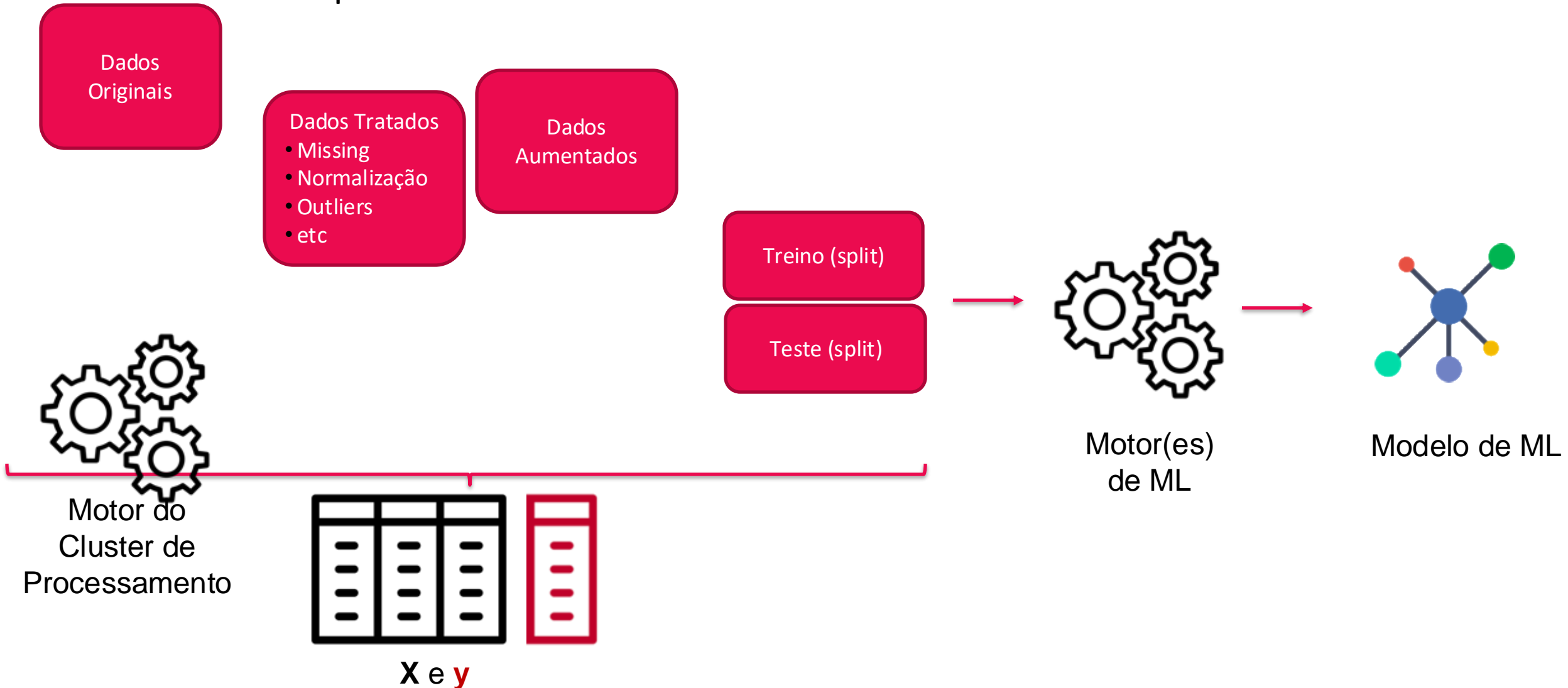
## Conhecendo nossa atividade

- Queremos criar um modelo customizado de Advanced Analytics em que, dado informações conhecidas do candidato, possamos prever qual a expectativa de nota na **prova de matemática** deste candidato.



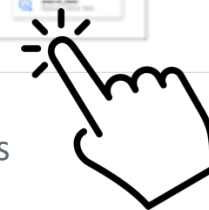
# Pré-Processamento dos dados

- Devido volume, costumamos maximizar a execução do pré-processamento de dados no cluster de processamento de dados



# Atividade – ETL em Big Data para Feature Engineering

- Nosso objetivo é criar um modelo de Machine Learning Simples utilizando Scikit-Learning que tenta “adivinhar” a nota dos candidatos em matemática (**NU\_NOTA\_MT**).
  - Como variável de entrada, utilizaremos as colunas do questionário sócio econômico, acrescidas de cor/raça, tipo de escola, estado da prova e estado civil. Ou seja, utilize as colunas:
    - NU\_ANO, Q001, Q002, Q003, Q004, Q005, Q006, Q007, Q008, Q009, Q010, Q011, Q012, Q013, Q014, Q015, Q016, Q017, Q018, Q019, Q020, Q021, Q022, Q023, Q024, Q025, Q026, Q027, TP\_ESTADO\_CIVIL, TP\_COR\_RACA, TP\_ESCOLA, CO\_UF\_PROVA.
  - Random Forest só irá aceitar entradas numéricas.
    - Como existem colunas textuais, precisamos transformá-las em categóricas numéricas \*
  - Devido diferença de colunas, vamos usar o ano como uma coluna de entrada \*
  - Considere apenas alunos que fizeram a prova de matemática (**TP\_PRESENCA\_MT=1**)
  - Missing Values (alternativas?):
    - Caso o campo não esteja preenchido, preencher com zero.
    - Alternativa: Como temos uma grande quantidade de dados, podemos considerar apenas os alunos que possuem todos os campos necessários preenchidos (Not Null)
  - Treino e Teste (devem ser determinísticos)
    - Colocar 80% dos dados em uma tabela chamada `enem.ml_treino` em teu dataset
    - Colocar 20% dos dados em uma tabela chamada `enem.ml_teste` em teu dataset
  - Pesquisar comandos úteis como:
    - CASE .. WHEN .. END, MOD, CONCAT, ML.LABEL\_ENCODER, RAND, farm\_fingerprint, GENERATE\_UUID

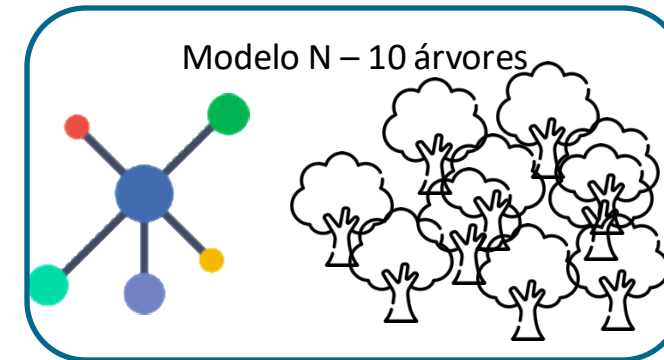
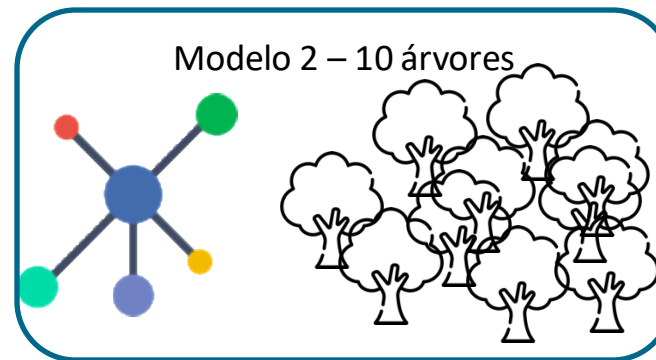
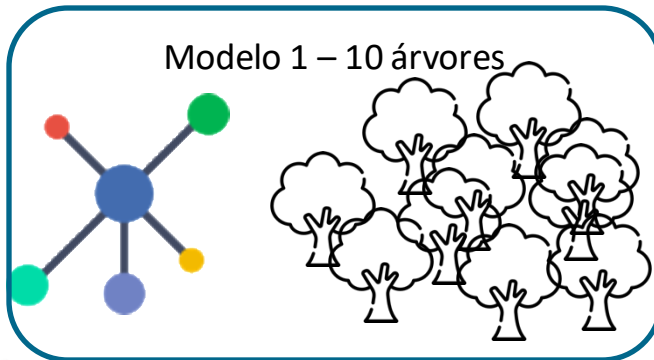
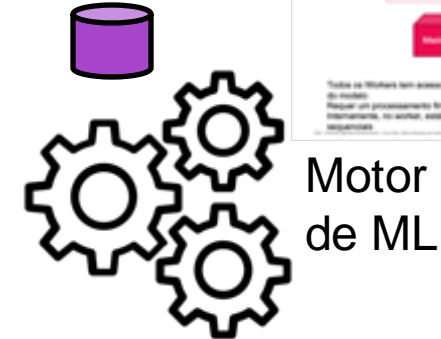
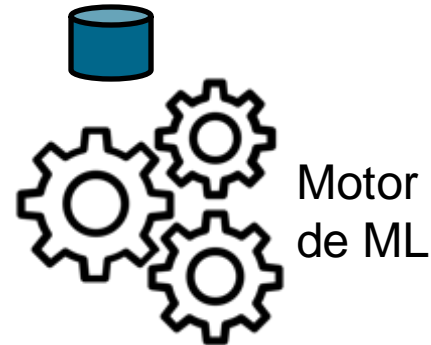
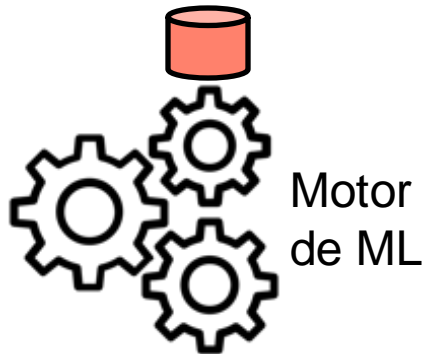


Obs.: Pode-se explorar pela interface gráfica do BigQuery, mas deve submeter o Notebook pelo Portal, template no Link:

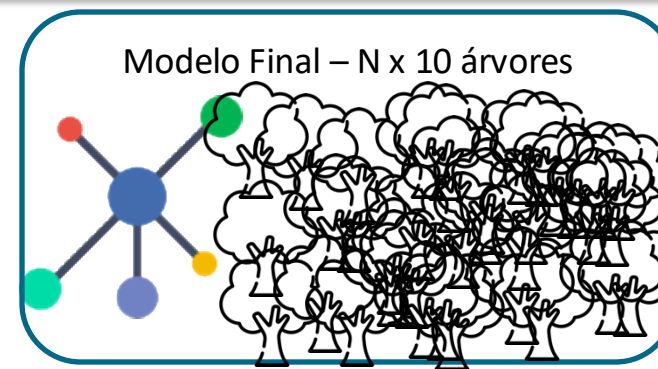
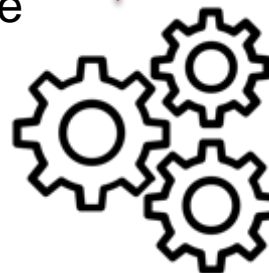
[https://colab.research.google.com/drive/1hdwvcM73gQRzn1hw\\_IlliuD66xDQr0cE](https://colab.research.google.com/drive/1hdwvcM73gQRzn1hw_IlliuD66xDQr0cE)

# Modelo de ML com BigData – Paralelismo de Dados

Para nossa abordagem, vamos criar N modelos de ML com Random Forest, e gerar uma Floresta final com todas as árvores



Processo de Unificação



# Atividade: Modelo de ML com BigData – Paralelismo de Dados

- Gere 12 florestas (modelos do tipo Random Forest Regressor) para um décimo dos dados disponíveis na base de treino
  - Cada floresta só pode possuir 10 árvores;
  - Cada floresta só pode usar a parte dos dados disponíveis que lhe é atribuída;
  - Discuta com a turma alternativas para essa divisão;
- Gere uma nova floresta com as árvores das florestas anteriores
  - Você precisa pesquisar o código para fazer esse *merge*
- Obtenha uma fração dos dados de teste e:
  - Faça uma predição usando a floresta resultando com as 120 árvores;
  - Faça uma predição usando uma floresta qualquer que você gerou;
  - Calcule o Root Mean Square Error das duas florestas e avalie essa métrica nos dois casos
- Ao final, vamos explorar possíveis soluções

Roteiro:

<https://colab.research.google.com/drive/1yOvqUkVMxFmspUFjtH5J0hUzUVmWCYRy>

FIAP