

# Frameworks de Big Data

## Projeto

**Eduardo Viegas**



ESCOLA  
**POLITÉCNICA**

# Projeto (70% da nota)

- Desenvolvimento no Hadoop de uma solução distribuída para aprendizagem de máquina

- Extração de características
- Treinamento e teste do modelo
- Deploy do modelo

Hadoop  
(Apache Spark ou MapReduce)

Qualquer ambiente de ML (skLearn, Apache MLib, Weka, ...)

# Projeto

- Desenvolvimento de uma solução de aprendizagem de máquina capaz de prever **alertas** em sensores



Situação de risco que pode acarretar em danos ao sensor

- Sensores emitem periodicamente dados relacionados a temperatura e posicionamento
- Medição é instantânea, não possui informação relacionada a seu histórico/contexto
- Base possui 206 milhões de medições instantâneas de 86920 dispositivos
- Cada dispositivo possui um identificador único no formato MD5

# Projeto

CSV delimitado por ‘;’

Campos
ID do sensor
Hora
Minuto
Ano
Mês
Dia
Temperatura
Latitude
Longitude



~206 milhões de medições  
86920 sensores

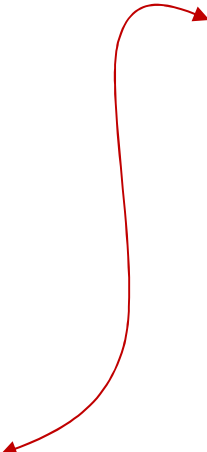
# Extração de Características

- Desenvolvimento de uma solução no Hadoop para extração de características
  - Utilizando o arquivo da base com ~206 milhões de medições
  - Gerar os datasets de treinamento/teste/validação
  - Extrator de características deve ser desenvolvido no MapReduce ou Apache Spark
  - Os datasets gerados devem sumarizar os atributos de acordo com o ID dos medidores
    - Treinamento: Dispositivos com o último caractere do ID entre 0 a 7
    - Teste: Dispositivos com o último caractere do ID entre 8 a A
    - Validação: Dispositivos com o último caractere do ID entre B a F

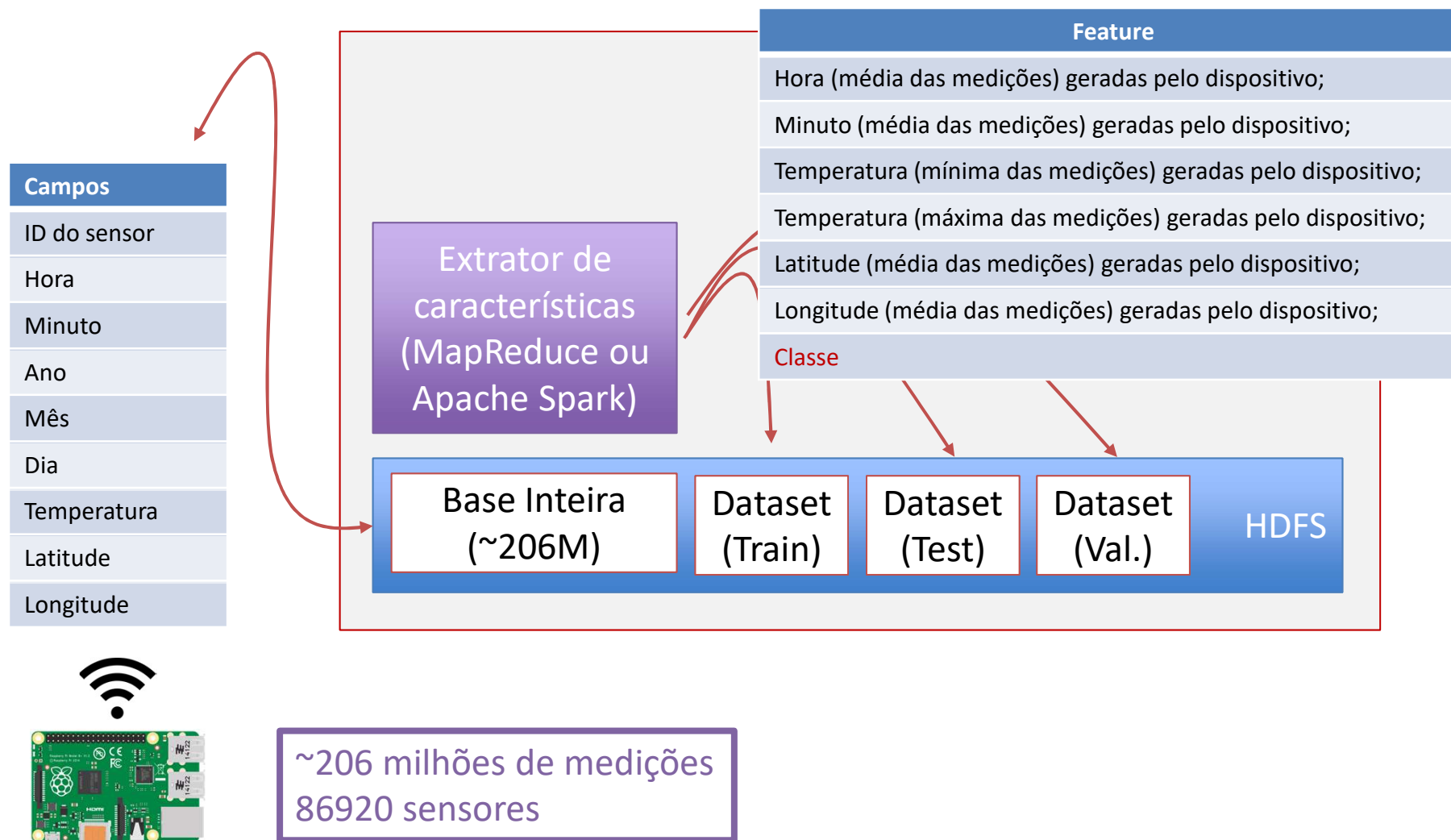
# Extração de Características

- Desenvolvimento de uma solução no Hadoop para extração de características
  - Features a serem extraídas para cada medidor

Feature
Hora (média das medições) geradas pelo dispositivo;
Minuto (média das medições) geradas pelo dispositivo;
Temperatura (mínima das medições) geradas pelo dispositivo;
Temperatura (máxima das medições) geradas pelo dispositivo;
Latitude (média das medições) geradas pelo dispositivo;
Longitude (média das medições) geradas pelo dispositivo;
Classe

- 
- “Frio”: Caso a temperatura (média das medições) geradas pelo dispositivo seja menor que 10
  - “Moderado”: Caso a temperatura (média das medições) geradas pelo dispositivo não seja Frio, e seja menor que 20
  - “Quente”: Caso a temperatura (média das medições) geradas pelo dispositivo não seja Moderado, e seja menor que 25
  - “Alerta”: em caso contrário

# Projeto

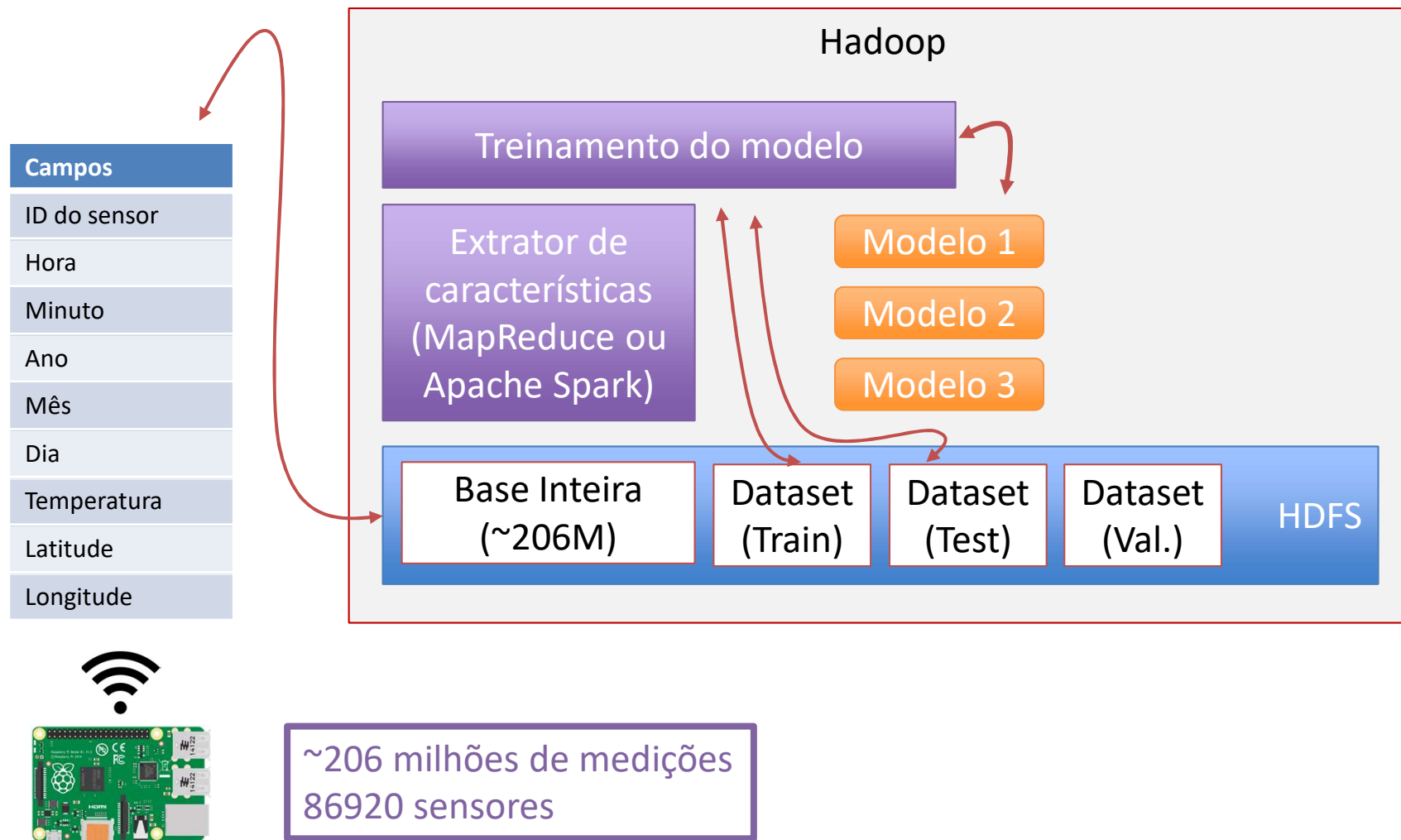


# Treinamento e Teste do Modelo

- Desenvolver um modelo de aprendizagem de máquina para prever a classe com valor “**Alerta**” no dataset gerado
  - Efetue testes com vários classificadores
  - Selecione os 3 melhores classificadores individuais
  - Utilize o dataset de treinamento para gerar o modelo
  - Utilize o dataset de testes para medir a acurácia



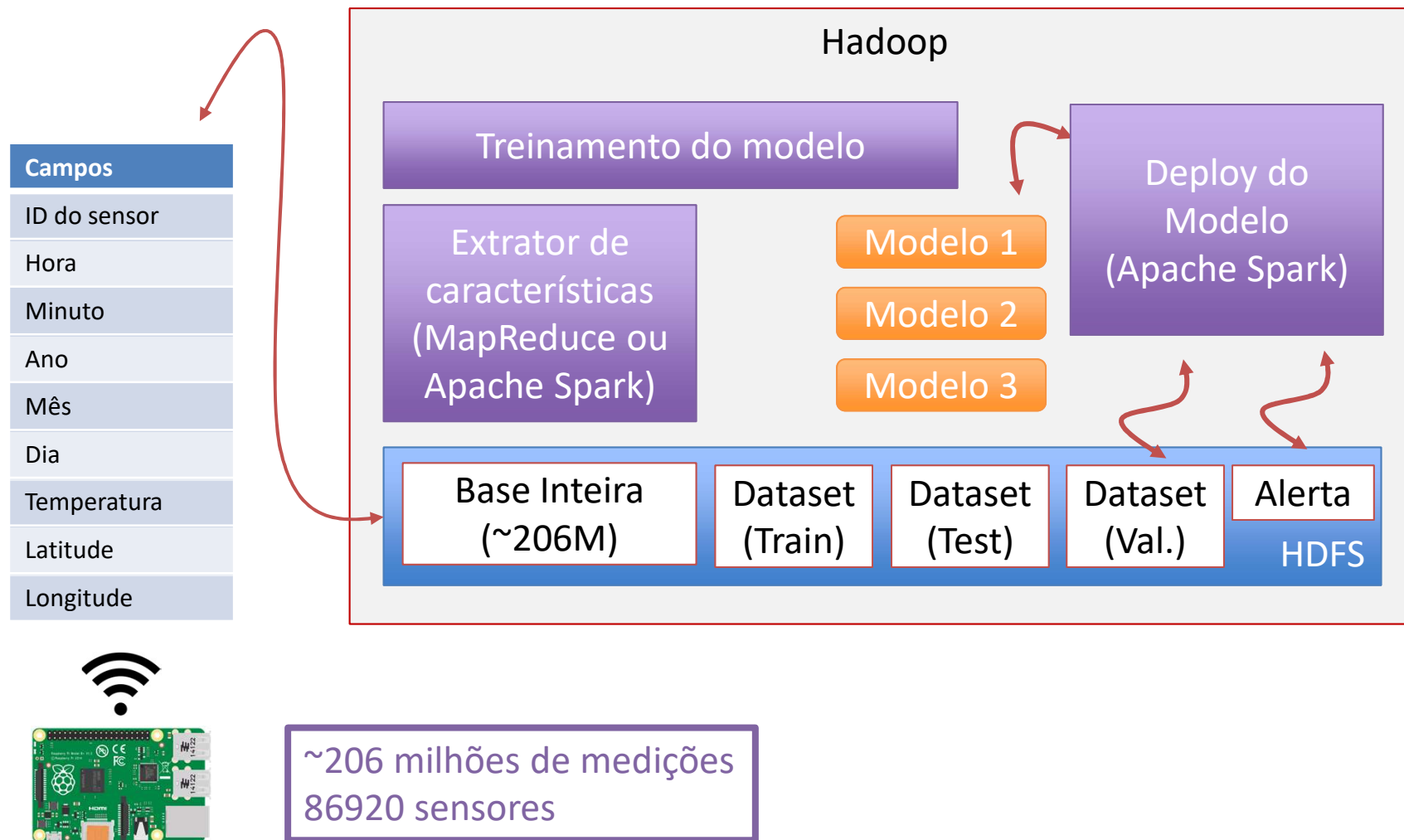
# Projeto



# Deploy do Modelo

- Desenvolver um módulo utilizando o **Apache Spark** que:
  - Carrega do disco os três modelos gerados
  - Classifica os eventos do dataset de validação armazenado no HDFS
  - Determina a classe através do voto majoritário
  - Gera um arquivo no HDFS com os IDs dos sensores que foram classificados como Alerta
  - Exibe a matriz de confusão final
  - O módulo que efetua a classificação deve ser executado nos 3 workers!
    - Utilize a função **repartition(3)** no RDD de classificação

# Projeto



# Entregáveis

- Cada equipe (até 3 alunos) deve entregar um ZIP com três pastas
  - Extrator de características:
    - código fonte
  - Treinamento do modelo:
    - Código fonte, com a gravação do modelo em disco e avaliação de pelo menos 5 técnicas de classificação
    - Tela exibindo as acurácias dos modelos avaliados
  - Deploy:
    - código fonte
    - tela exibindo a matriz de confusão

# Dicas =)

- Recomendo o desenvolvimento do extrator no MapReduce
- Para geração dos modelos no sklearn
  - Gravem os modelos em disco
  - Copiem os modelos para o diretório local dos spark-workers



ESCOLA  
**POLITÉCNICA**