

Big Data HPC

Handson utilizando PySpark

SENAI CIMATEC

Tecnologia, Inovação
e Educação para a Indústria

Sistema FIEB



PELO FUTURO DA INOVAÇÃO



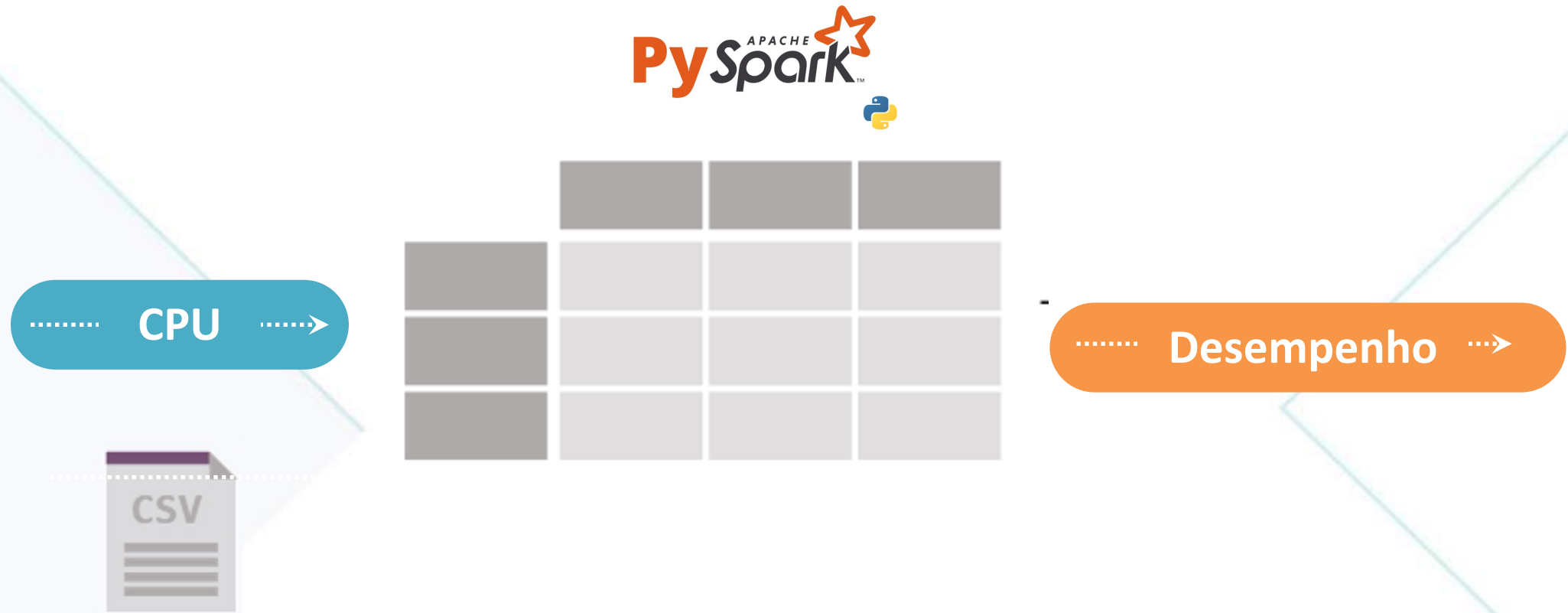
QMS Certification Services

Big Data HPC

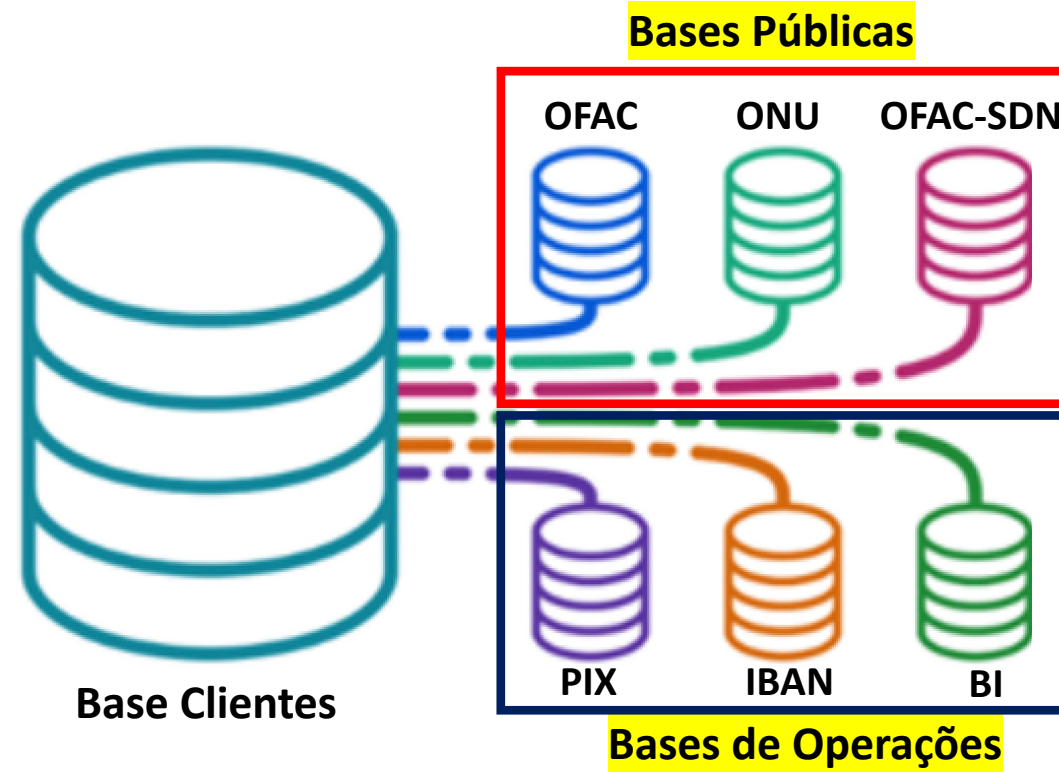
Handson utilizando PySpark

Murilo Boratto

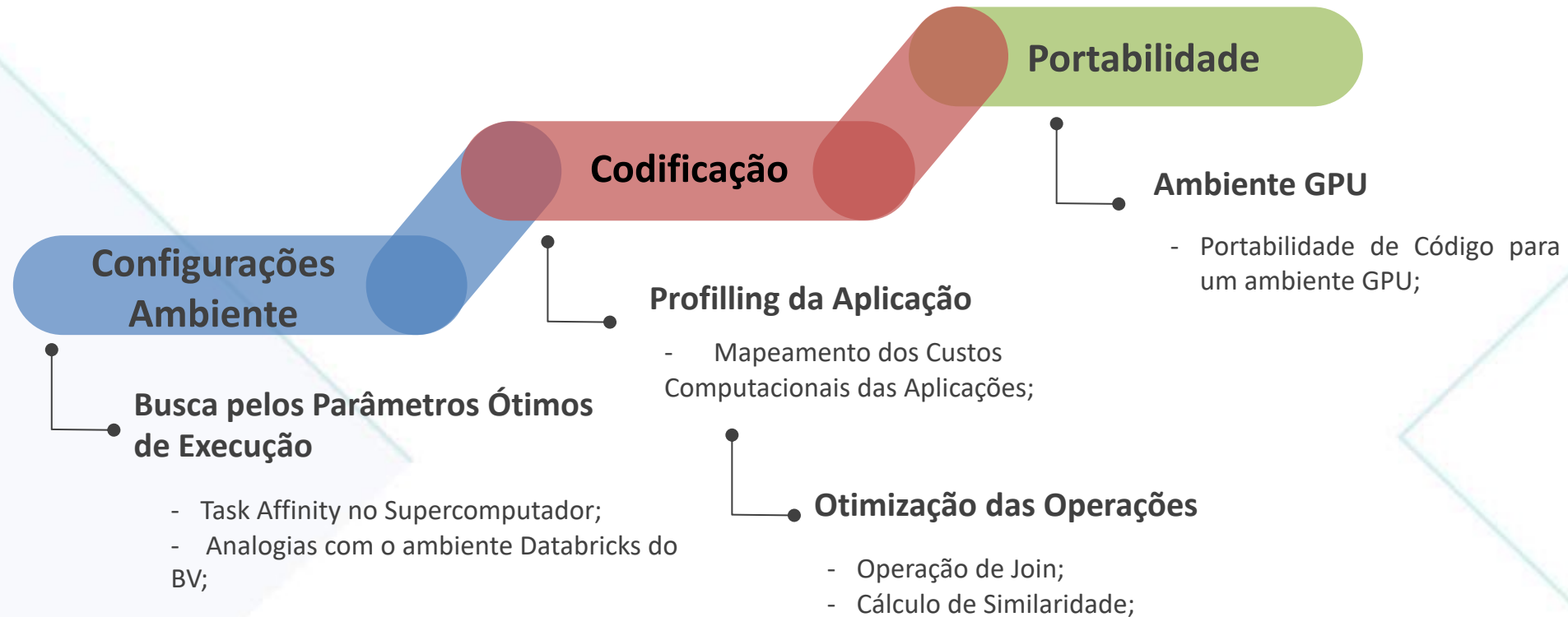
Estrutura do Estudo de Caso



Estudos de Caso



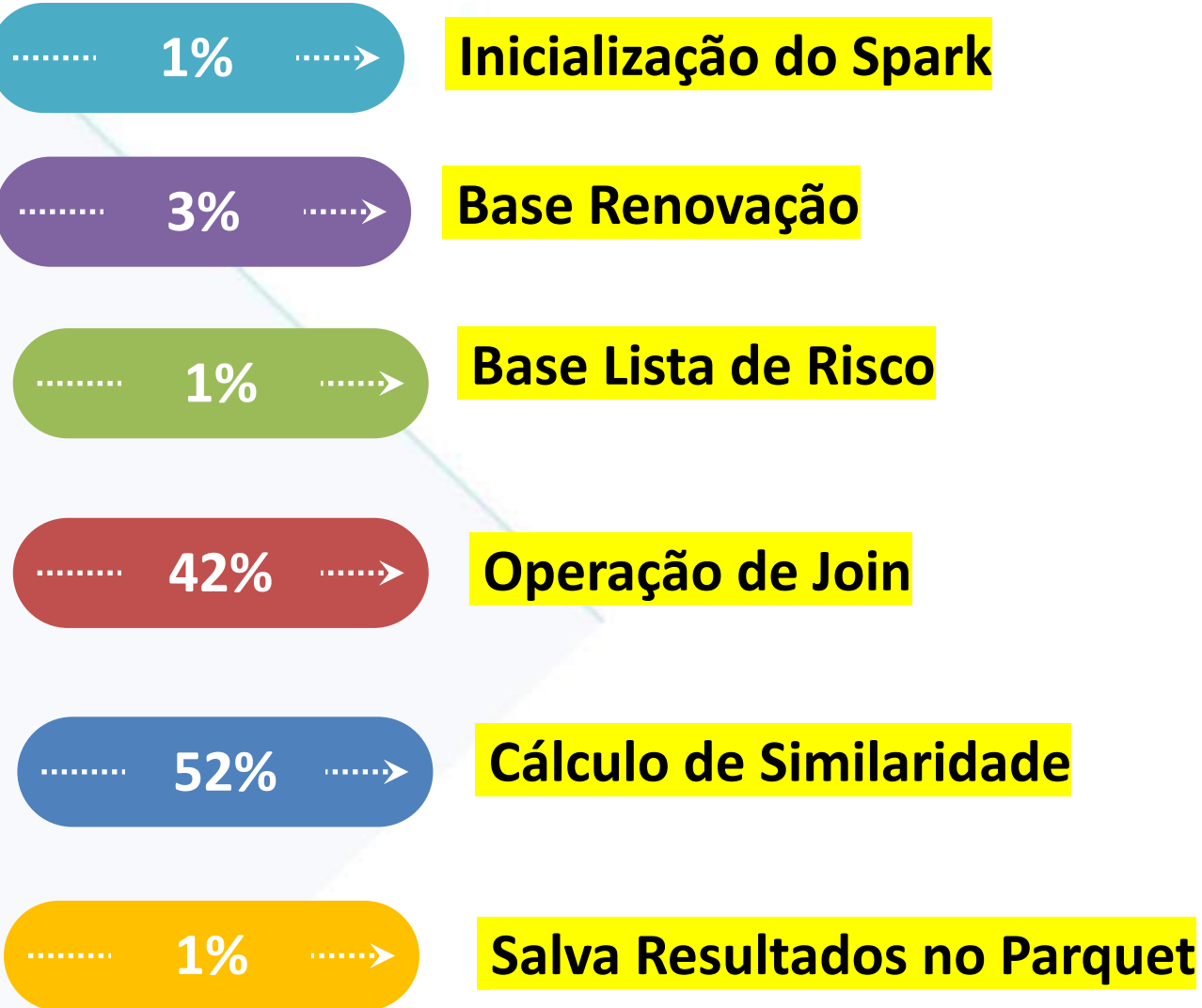
Níveis de Otimização



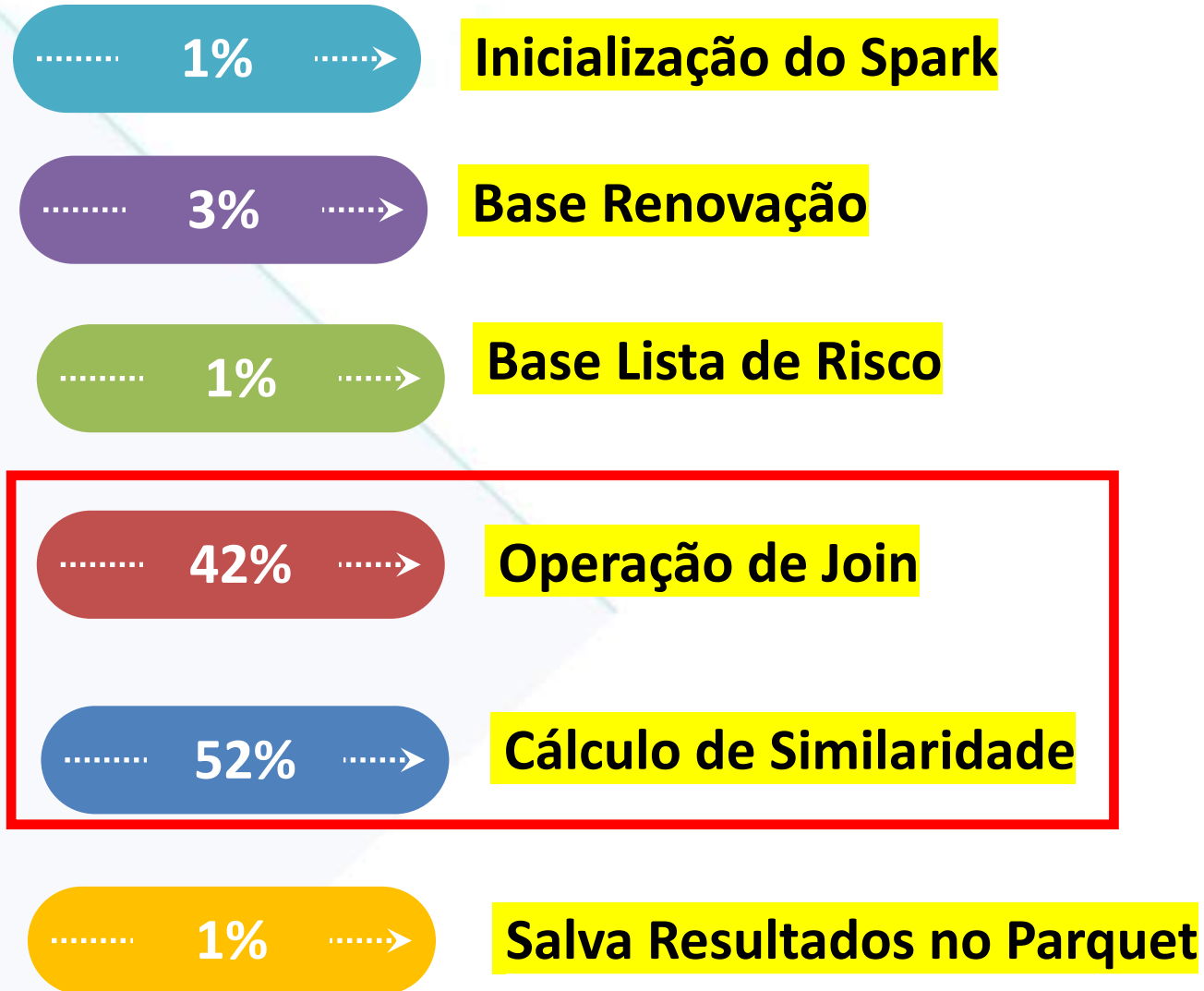
Base Pública OFAC-SDN – Profiling – Original



Base Pública OFAC-SDN – Profiling – Original



Base Pública OFAC-SDN – Profiling – Original



Base Pública OFAC-SDN – Profiling – Original

Aplicação Original

```
df_match = df_renov.join(df_lista_risco,  
                        ((df_renov['TpPessoa'] == df_lista_risco['TIPO']) &  
                         (df_renov['m'] == df_lista_risco['m'])), how='inner')
```

Aplicação Otimizada

```
df_joined = df_renov.join(F.broadcast(df_lista_risco), ["TIPO_RELACAO"])
```

Base Pública OFAC-SDN – Profiling – Original

Aplicação Original

```
def lev(x, y):
    return fuzz.ratio(x,y)

similarity = udf(lev, IntegerType())

df = df.withColumn("SIMILARITY",
similarity(df.NOME,df.NOME_RISCO))

fil = df.filter(df.SIMILARITY) > 90)
```

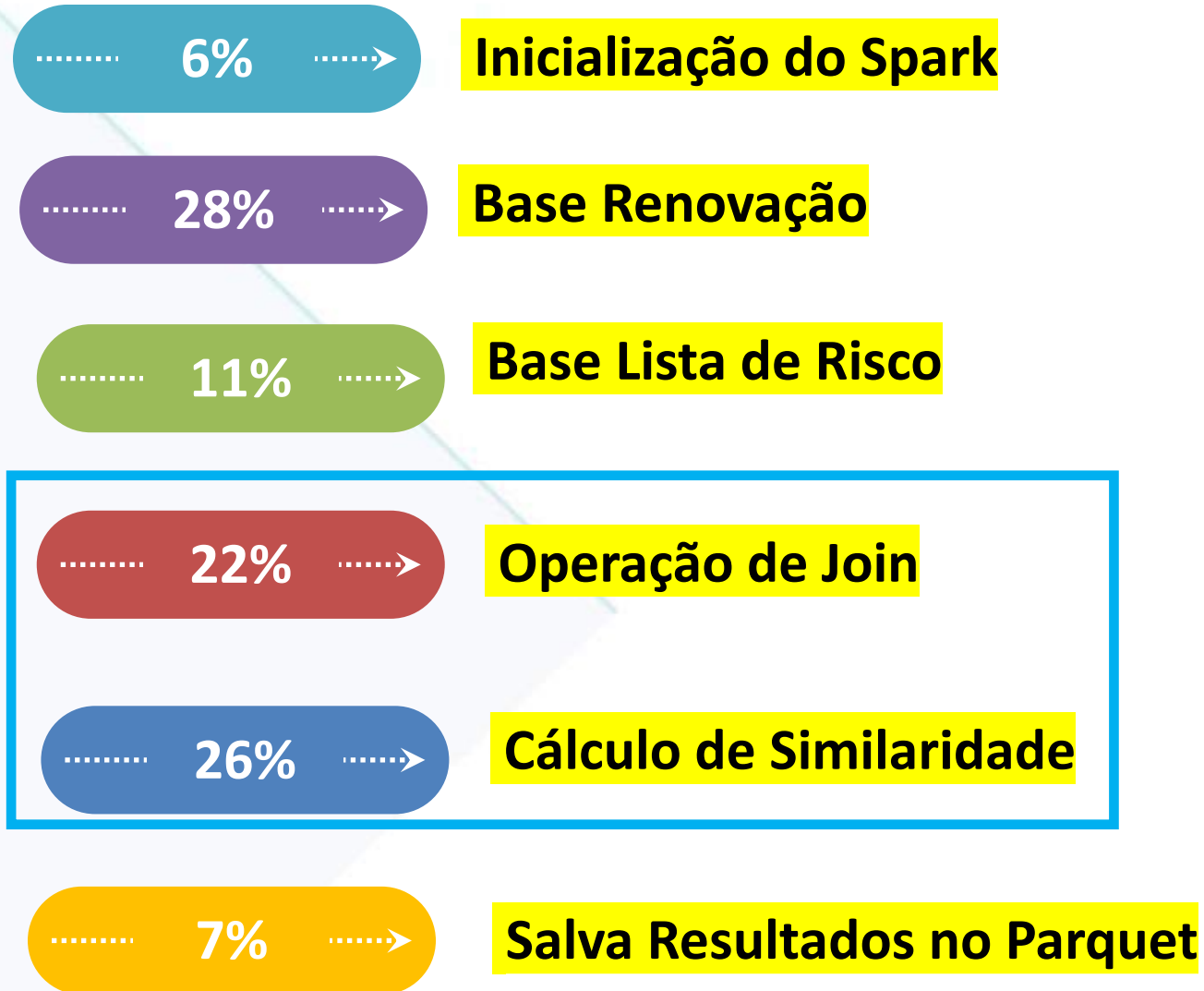
Aplicação Otimizada

```
lev      = F.levenshtein
str1     = F.col("NOME")
str2     = F.col("NOME_RISCO")
len_s1   = F.length(str1)
len_s2   = F.length(str2)

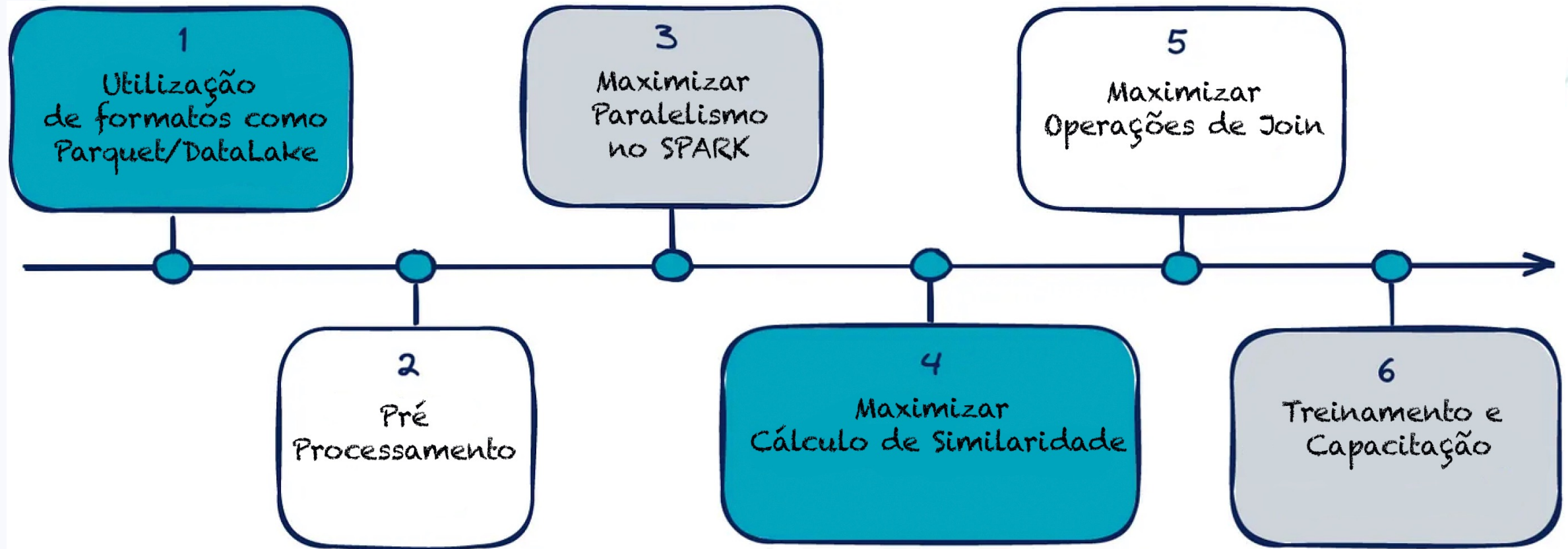
df_lev = df_joined.withColumn(
    "SIMILARITY", (100 * (1 - (lev(str1, str2) /
        (len_s1 + len_s2))))).cast("int"))

fil = df_lev.filter(F.col("SIMILARITY") > 90)
```

Base Pública OFAC-SDN – Profiling – Original



Recomendações para Otimizar a Aplicação PLD Spark do Banco BV



Resultados Experimentais

PLD	OFAC	ONU	OFAC-SDN
ORIGINAL	135,20	860,92	9476,56
OTIMIZAÇÕES CPU	58,51	299,50	3157,90
SPEEDUP CPU	2,31X	2,87X	3X

Tabela: Tempos de Execução (seg.) comparando o desempenho das Bases Públicas em sistemas CPU.

Resultados Experimentais

PLD	OFAC	ONU	OFAC-SDN
ORIGINAL	135,20	860,92	9476,56
OTIMIZAÇÕES CPU	58,51	299,50	3157,90
OTIMIZAÇÕES GPU	22,10	143,00	1353,79
SPEEDUP CPU	2,31X	2,87X	3X
SPEEDUP GPU	6X	6X	7X

Tabela: Tempos de Execução (seg.) comparando o desempenho das Bases Públicas em sistemas CPU e 4-GPUs.

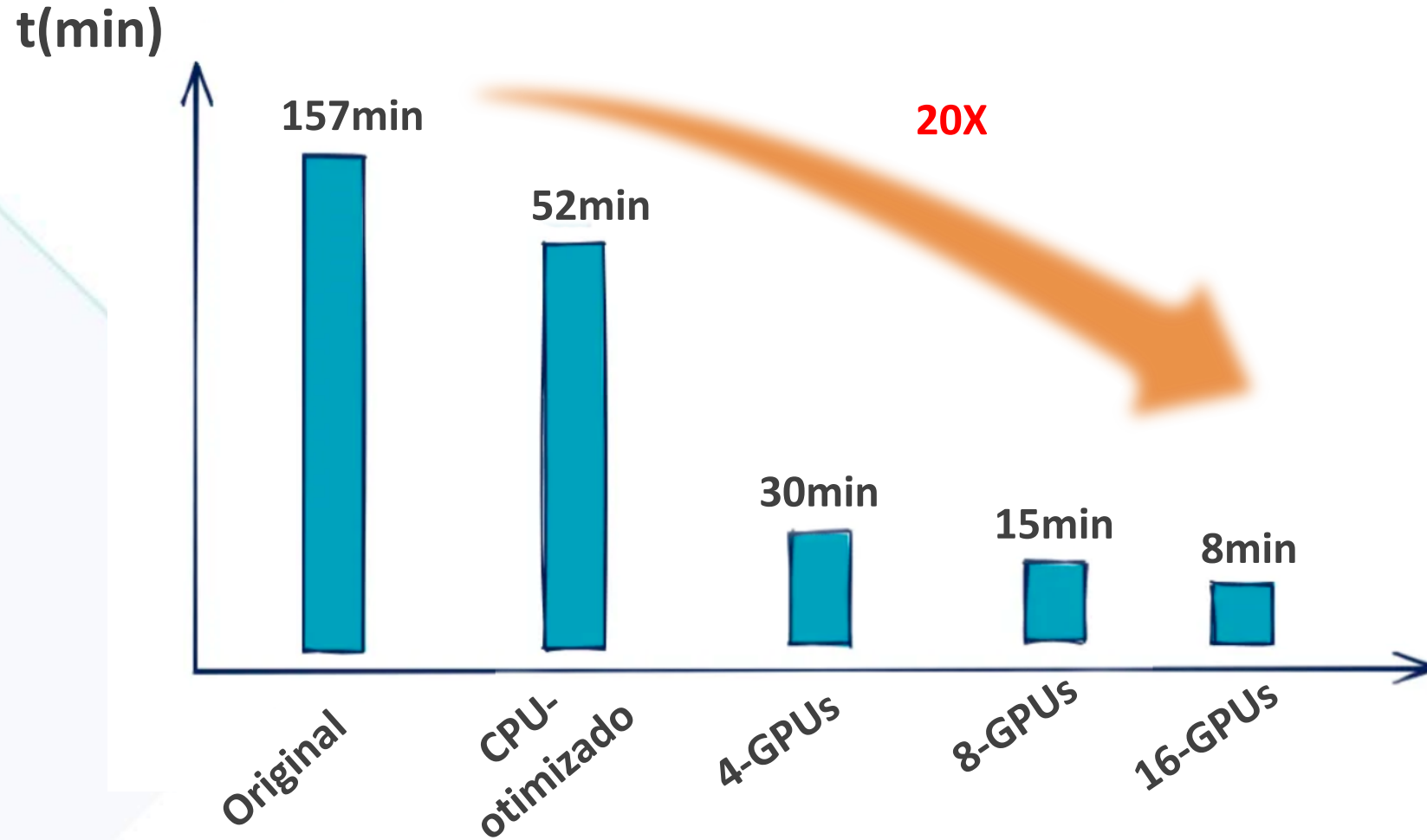
Resultados Experimentais

	4-GPUs	8-GPUs	16-GPUs
PLD	OFAC-SDN	OFAC-SDN	OFAC-SDN
ORIGINAL	9476,56	9476,56	9476,56
OTIMIZAÇÕES CPU	3157,90	3157,90	3157,90
OTIMIZAÇÕES GPU	1353,79	900,00	480,10
SPEEDUP CPU	3X	3X	3X
SPEEDUP GPU	7X	10X	20X

Tabela: Tempos de Execução (seg.) comparando o desempenho das Bases Públicas em sistemas CPU e multi-GPU.

Resultados Experimentais (OFAC-SDN)

SENAI CIMATEC



- **A utilização de Sistemas Multi-GPU torna-se praticamente indispensável para as bases com grande quantidade de dados.**
- **Uma série de otimizações tanto a nível de hardware/software podem ser aplicadas as aplicações melhorando ainda mais o desempenho.**
- **A idéia é simular com bases maiores, afim de comparar o desempenho computacional real obtido.**

Big Data HPC

Handson utilizando PySpark

SENAI CIMATEC

Tecnologia, Inovação e Educação para a Indústria

Sistema FIEB



PELO FUTURO DA INOVAÇÃO



QMS Certification Services