Big Data e Computação em Nuvem

Aula 10

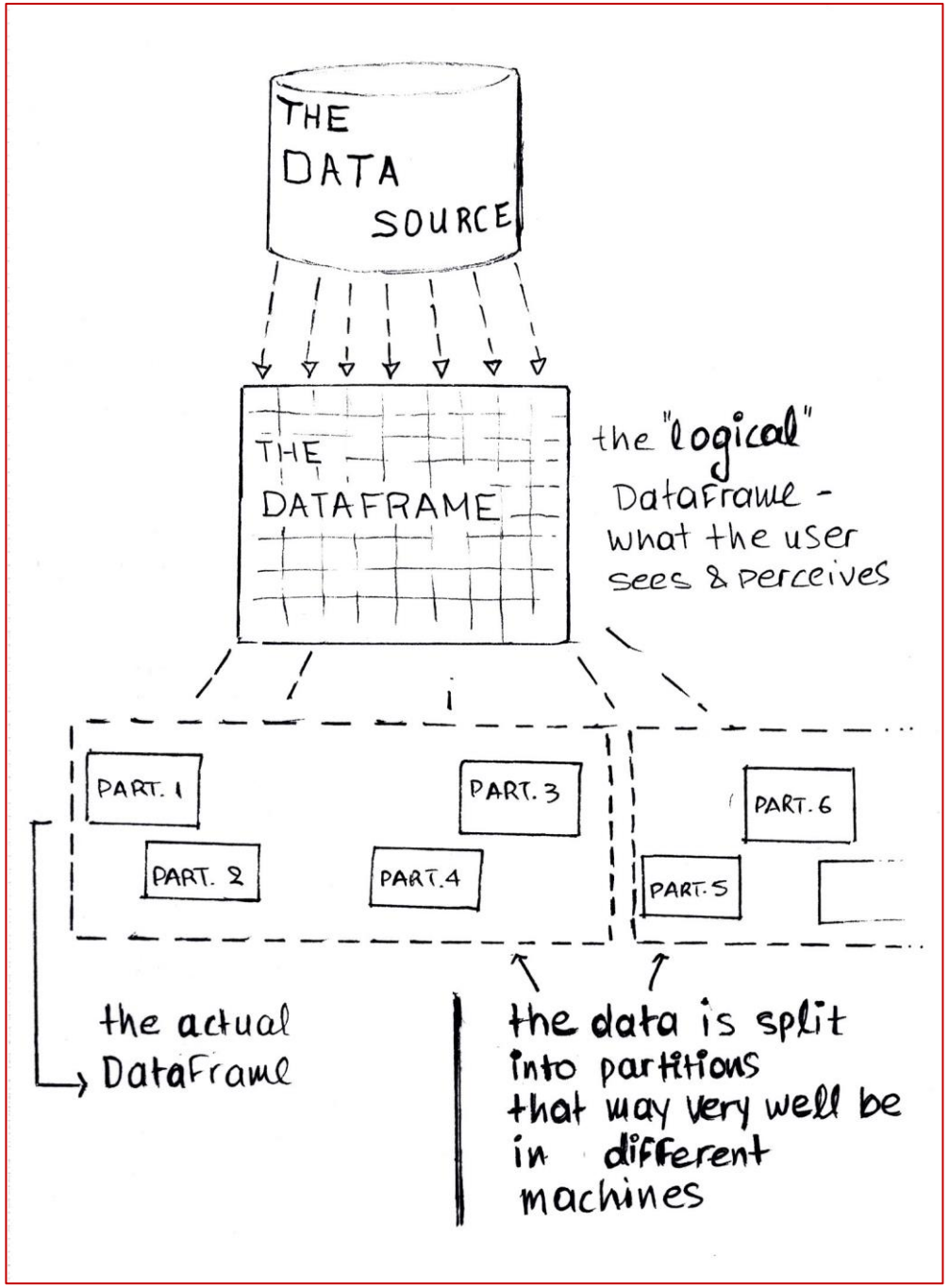
Spark DataFrames - Cache

Prof. Michel Fornaciali, PhD.
Prof. Thanuci Silva, PhD.

Contatos:
MichelSF@insper.edu.br
thanucis@insper.edu.br

Insper

DataFrame



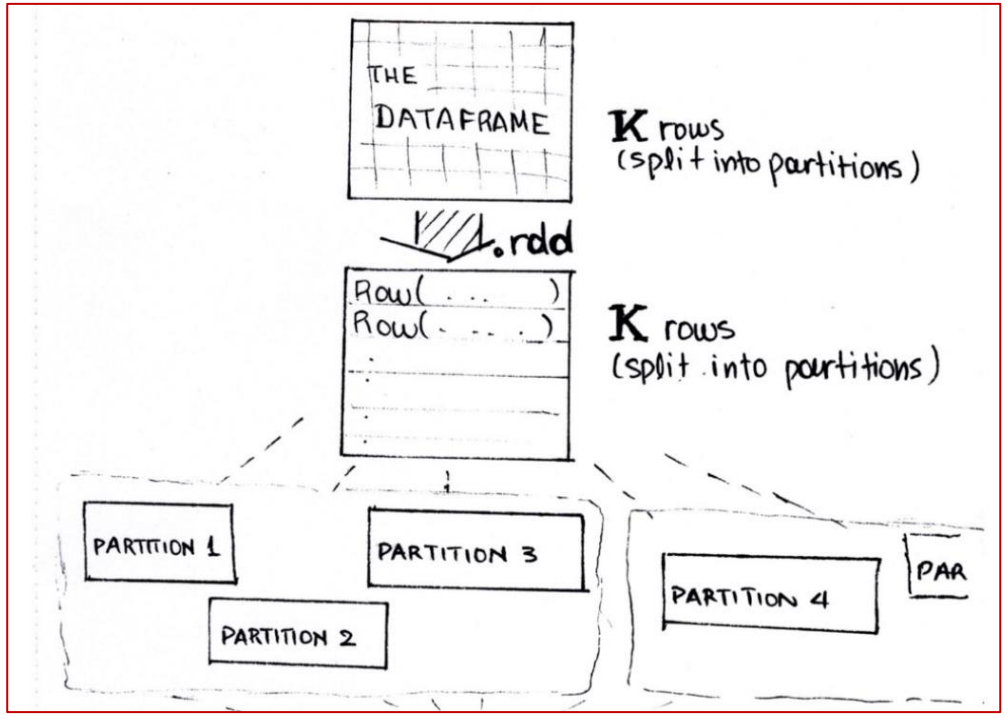


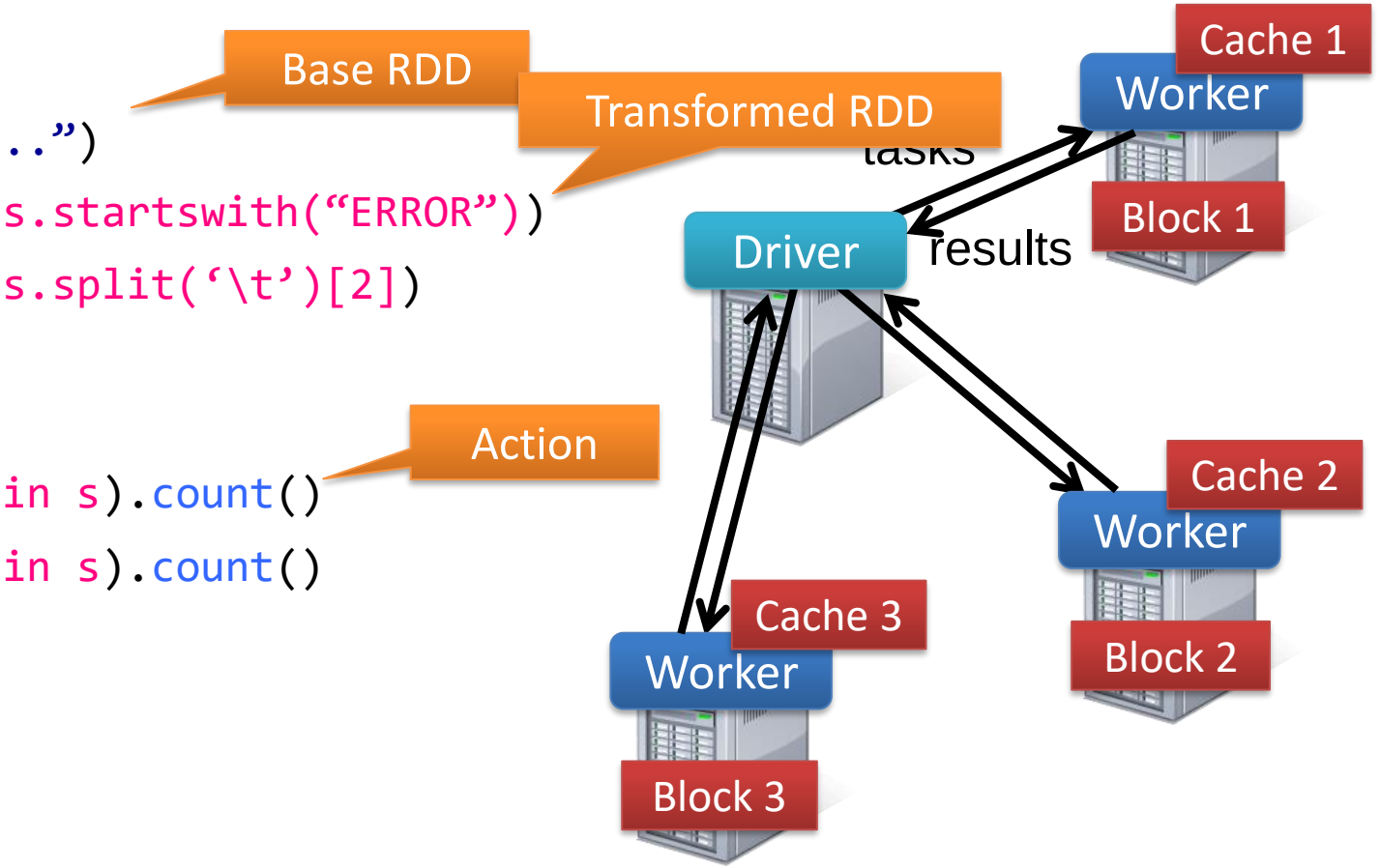
Imagem: https://miro.medium.com/max/1000/1*cXGB03Uf0IJKcew42e_AEw.jpeg

Insper

Spark – Mecanismo de Cache

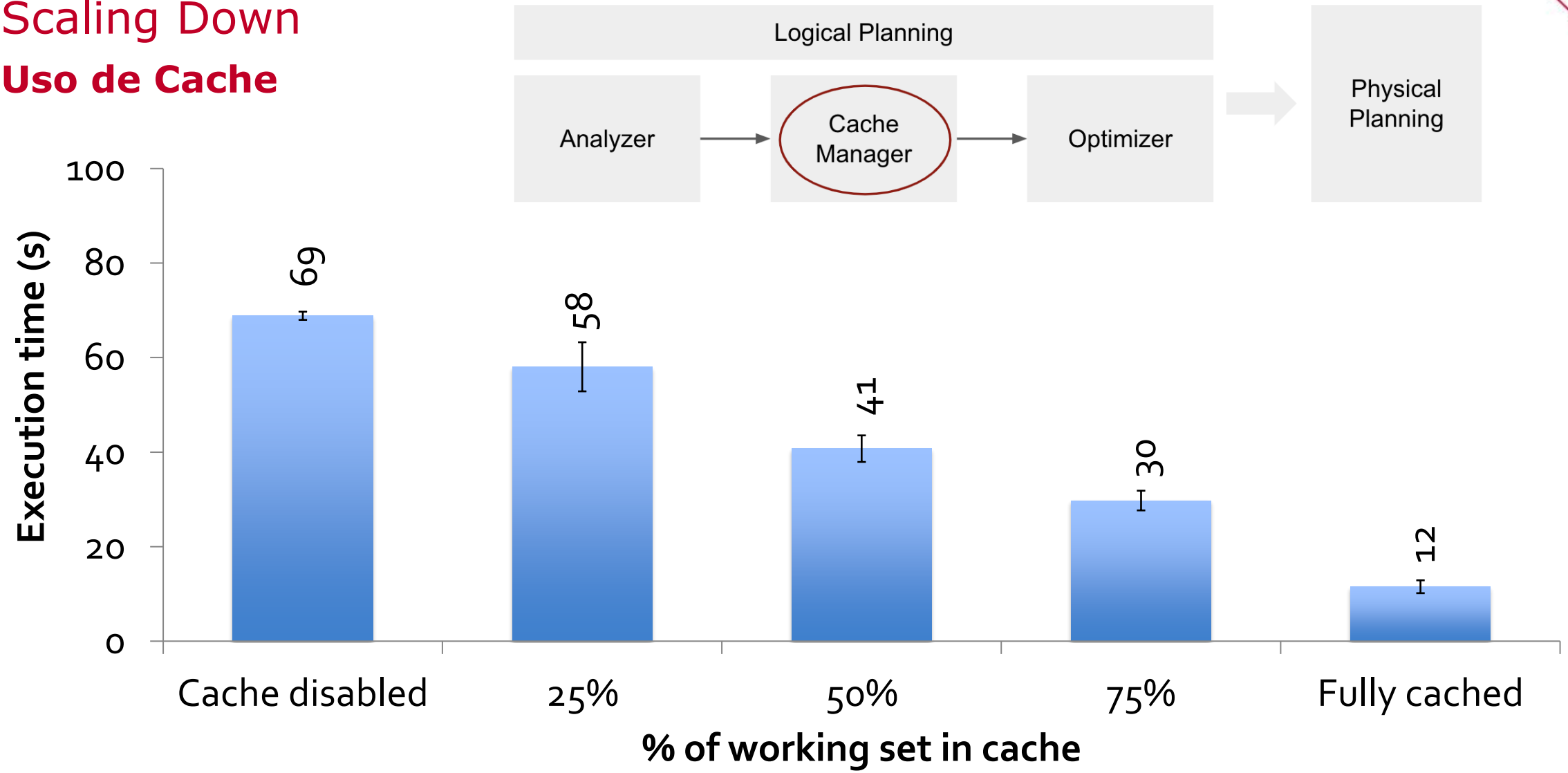
```
lines = spark.textFile(“hdfs://...”)
errors = lines.filter(lambda s: s.startswith(“ERROR”))
messages = errors.map(lambda s: s.split(‘\t’)[2])
messages.cache()

messages.filter(lambda s: “foo” in s).count()
messages.filter(lambda s: “bar” in s).count()
. . .
```



Scaling Down

Uso de Cache



Insper

Uso do cache

```
inputRDD = sc.textFile("words.txt").cache()

print("Number of words: ",inputRDD.count())
print("Number of distinct words: ", inputRDD.distinct().count())
```

Insper