Working with Large Dataset

Working with Large Dataset

- Pandas, NumPy and Scikit-Learn can’t work efficiently

Working with Large Dataset

- Pandas, NumPy and Scikit-Learn can’t work efficiently
- Time consuming and Computationally Expensive

Analytics Vidhya
Learn everything about analytics

Working with Large Dataset

- Use system with higher specs

Working with Large Dataset

- Use system with higher specs
- Alternative platforms (AWS, Colab etc)

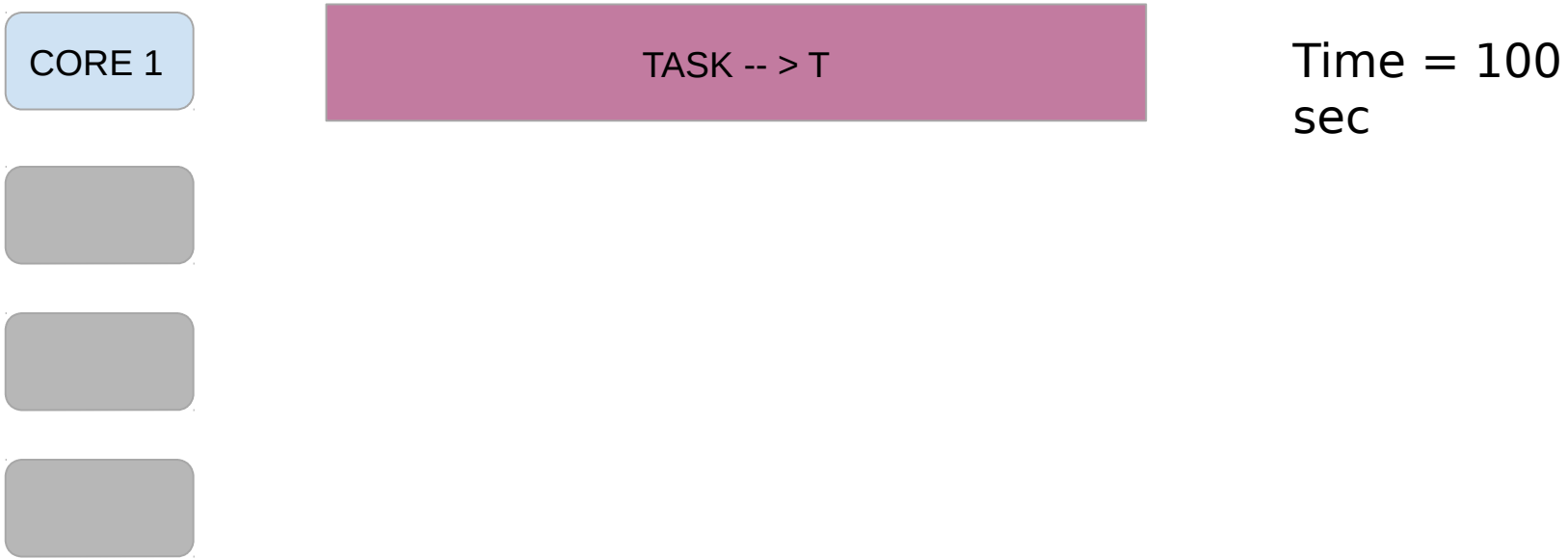
Working with Large Dataset

- Use system with higher specs
- Alternative platforms (AWS, Colab etc)
- Open source library

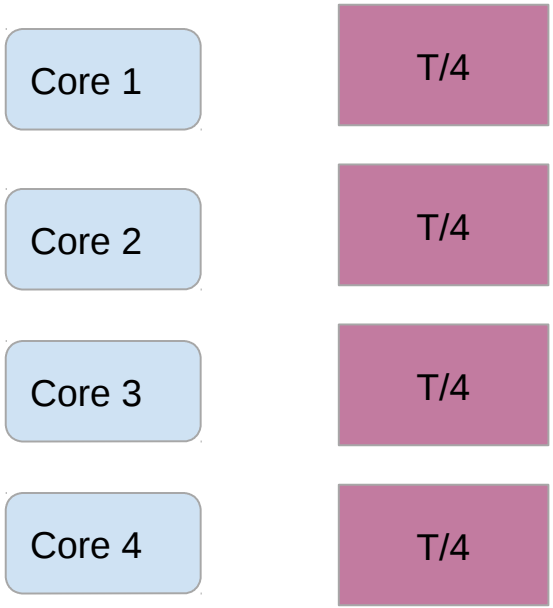
Why Use DASK?

- Dask is a parallel computing python library

Why Use DASK?



Why Use DASK?



Time = 25 sec

Analytics Vidhya
Learn everything about analytics

Why Use DASK?

- Dask is a parallel computing python library
- Similar API as python libraries (pandas and NumPy)

```
import pandas as pd
df = pd.read_csv('2015-01-01.csv')
df.groupby(df.user_id).value.mean()
```

```
import dask.dataframe as dd
df = dd.read_csv('2015-*-*.csv')
df.groupby(df.user_id).value.mean().compute()
```

Analytics Vidhya
Learn everything about analytics

Why Use DASK?

- Dask is a parallel computing python library
- Similar API as python libraries (pandas and NumPy)
- Easy to write codes

Setting up your System

pip install “dask[complete]”