



ESCUELA
NACIONAL
DE ESTUDIOS
SUPERIORES
m
UNIDAD MORELIA



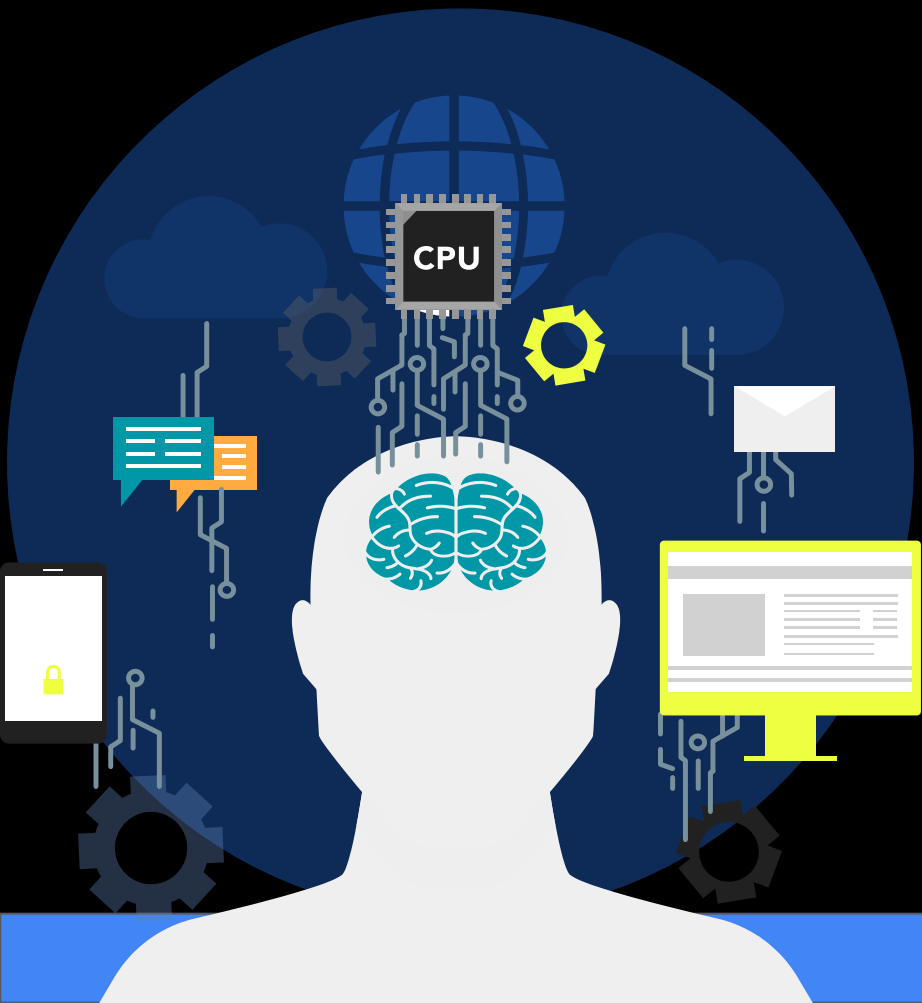
Diplomado en ML y DL aplicado a grandes volúmenes de datos

PAPIME PE103124

Módulo I: ML (Marzo 2024)

Introducción a Dask

Sergio Rogelio Tinoco Martínez
Heberto Ferreira Medina
José Luis Cendejas Valdez



Contenido

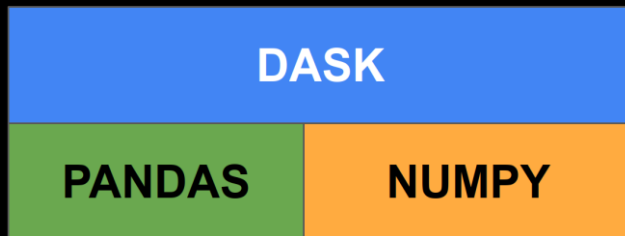
- ¿Qué es Dask?
- ¿Cómo funciona?
- Dask-ML
- Ventajas
- Ejemplos



¿Qué es Dask?

Dask es una herramienta para el procesamiento de grandes volúmenes de datos. Está basada en Pandas y Numpy, por lo que maneja dataframes, matrices y arreglos.

Puede ejecutarse localmente o ampliarse a fin de ejecutarse en un clúster.



High-Level APIs

Dask Array

Parallel NumPy

Dask Bag

Parallel Lists

Dask DataFrame

Parallel Pandas

Dask ML

Parallel Scikit-Learn

Low-Level APIs

Dask Delayed

Lazy Parallel Objects

Dask Futures

Eager Parallel Objects

Task Schedulers

Local
Scheduler

Distributed
Scheduler

Mesos

YARN

Kubernetes

Dask Built-In Schedulers

External Resource Managers

¿Qué es Dask? – 2 –

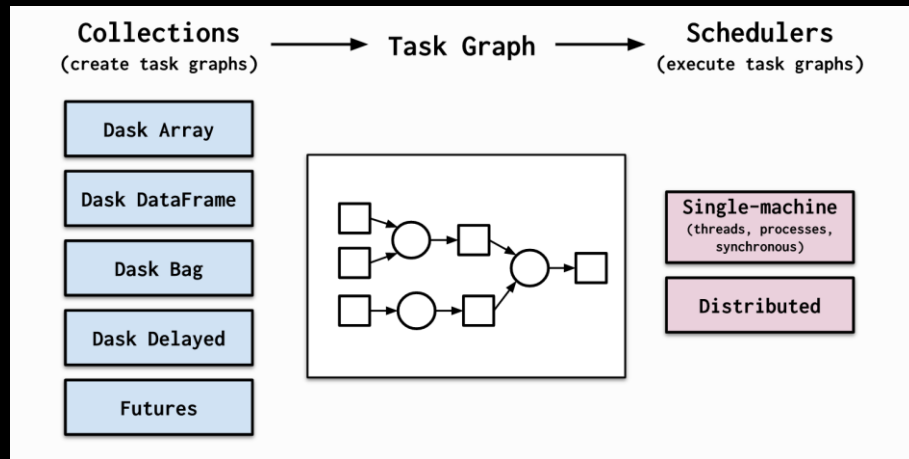
Entre las principales ventajas de Dask es su similitud con pandas.

```
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()
```

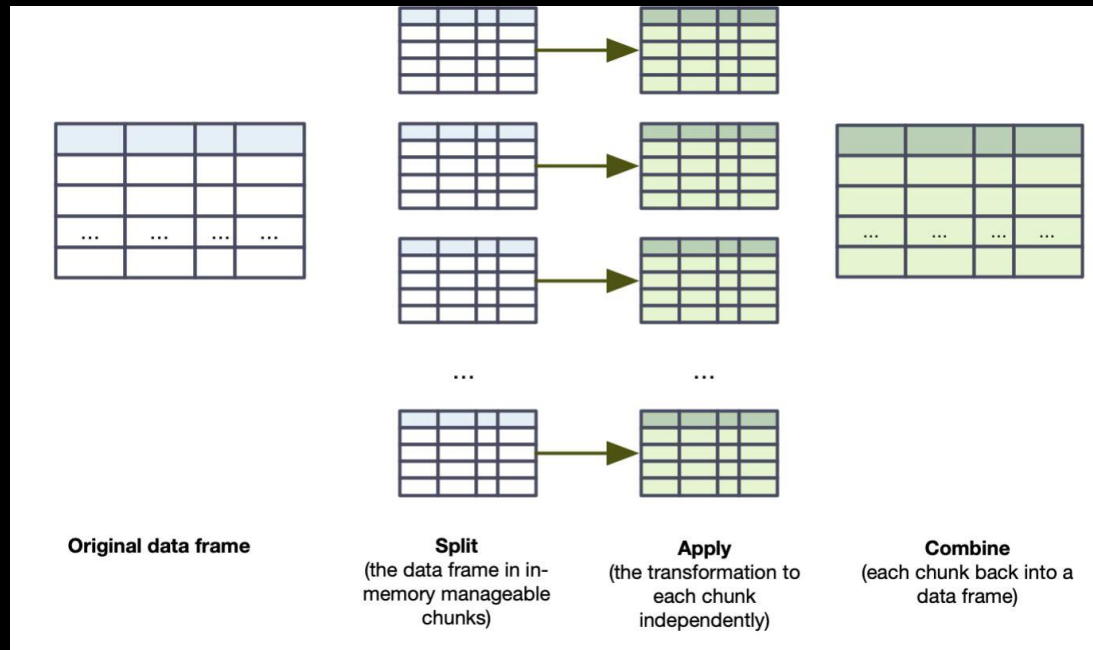
¿Qué es Dask? – 3 –

Además de usarse de forma local, Dask puede usarse de forma distribuida, ya que puede dividir los datos en diferentes fragmentos tanto en memoria como en disco.



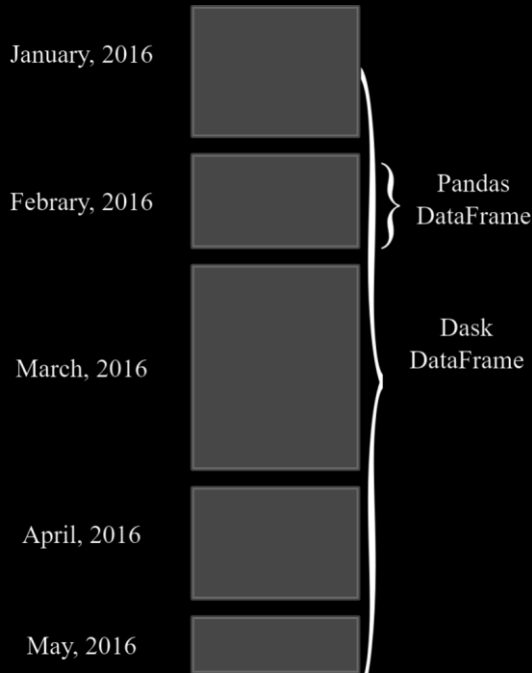
Dask se basa en “chunks”

Los “chunks” (“cachos”) son particiones definidas de los datos. Cuando se trabaja con chunks, se repiten todas las operaciones sobre los datos en cada partición.



Dask se basa en “chunks” – 2 –

Con Dask, los cálculos se realizan de manera similar a Pandas, sin la necesidad de realizar métodos iterativos para cada partición.



Dask-ML

Dask-ML es una herramienta de aprendizaje automático escalable que utiliza Dask junto con otras bibliotecas de aprendizaje automático, tales como Scikit-Learn, XGBoost, entre otras.



Dask-ML – 2 –

Al igual que Scikit-Learn, Dask-ML cuenta con modelos de aprendizaje automático escalables a grandes volúmenes de datos.

```
from dask_ml.linear_model import LinearRegression
from dask_glm.datasets import make_regression

X, y = make_regression()
lr = LinearRegression()
lr.fit(X, y)
lr.predict(X)
lr.score(X, y)
```

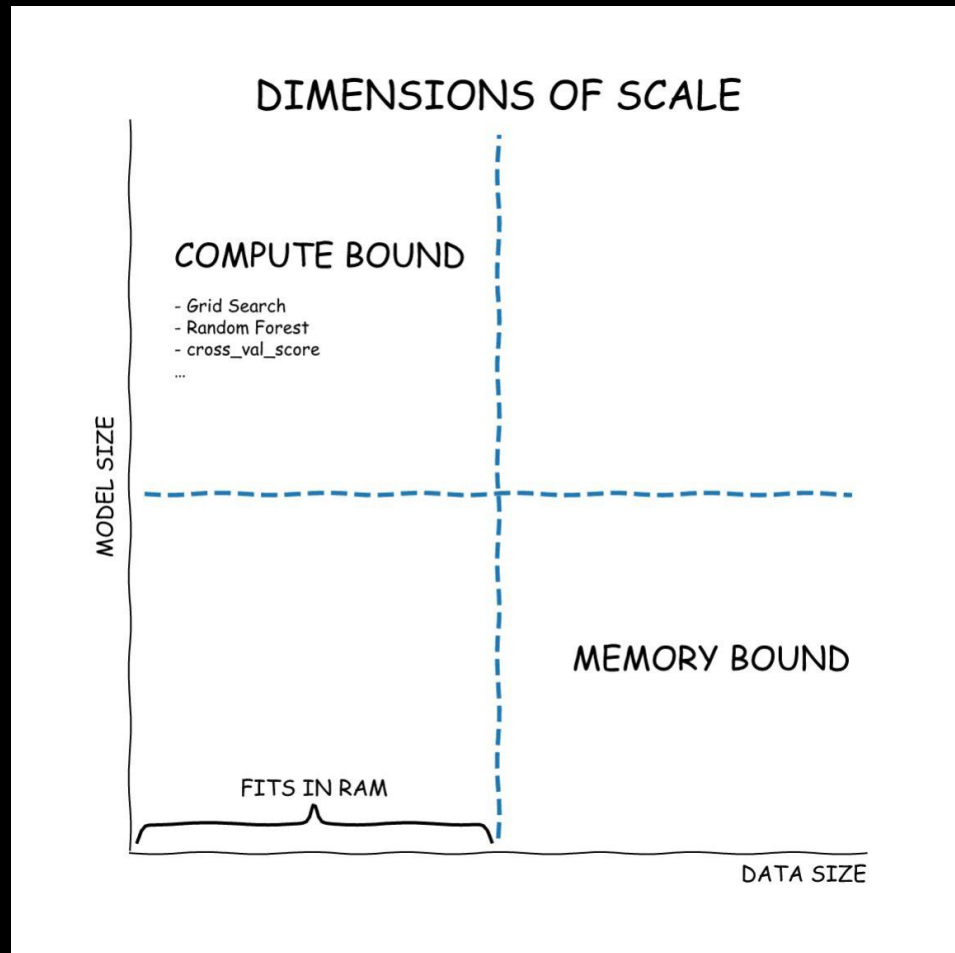
Dask-ML – 3 –

También cuenta con métodos de preprocesamiento de datos, semejantes a los de Scikit-Learn.

```
from dask_ml.preprocessing import OneHotEncoder
import dask.array as da
import numpy as np

enc = OneHotEncoder(sparse=True)
X = da.from_array(np.array([[ 'A' ], [ 'B' ], [ 'A' ], [ 'C' ]]), chunks=2)
enc = enc.fit(X)
result = enc.transform(X)
result
```

- Clústeres de Dask
 - Joblib para Scikit-Learn.
 - Optimizador de hiperparámetros.
- Colecciones de alto nivel
 - Arrays.
 - Bags.
 - Dataframes.
- dask-ml
 - .model_selection
 - .linear_model
 - .clustering
 - .preprocessing



Ventajas

- ✓ Pandas en paralelo.
- ✓ Dask demora y llega a evitar errores de falta de memoria que podrían suceder utilizando Pandas.
- ✓ Tiene un mejor control de los recursos (en comparación a Pandas).
- ✓ Sigue las APIs de Pandas, Numpy y Scikit-Learn.
- ✓ Más fácil de configurar que otras plataformas (como Hadoop y Spark).



