Proyecto PAPIME: PE103124

Intervención educativa IIES, ENES-UNAM

UNAM

iies

INSTITUTO DE INVESTIGACIONES
EN ECOSISTEMAS Y SUSTENTABILIDAD

ENES
MORELIA

UTM

Universidad Tecnológica de Morelia

TECNOLÓGICO
NACIONAL DE MÉXICO
en Morelia

Dr. Sergio R.Tinoco M., Dr. Heberto Ferreira M.,
Dr. José L. Cendejas V., MTI. Froylan Hernández R.
MTI. Alberto Valencia García
Est. Jessica Torres, Alejandro Ponce, Gustavo Zárate

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES
UNIDAD MORELIA

iies UNAM

UTM
Universidad Tecnológica de Morelia

DIPLOMADO EN ML Y DL APLICADO AL BIG DATA

MÓDULO II: DEEP LEARNING

DASK

I. S. C. Jéssica Torres :: Dr. José Luis Cendejas Valdez[2]

Dr. Heberto Ferreira Medina[1] :: Dr. Sergio Rogelio Tinoco MartÍnez[3]

Universidad Nacional Autónoma de México[1,3] :: Universidad Tecnológica de Morelia[2]

Instituto de Investigaciones en Ecosistemas y Sustentabilidad[1]

Escuela Nacional de Estudios Superiores Unidad Morelia[3]

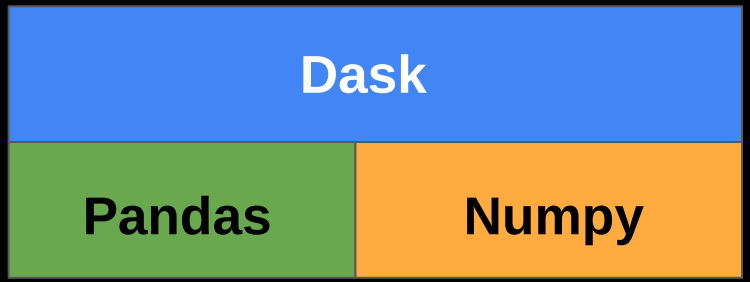
Contenido

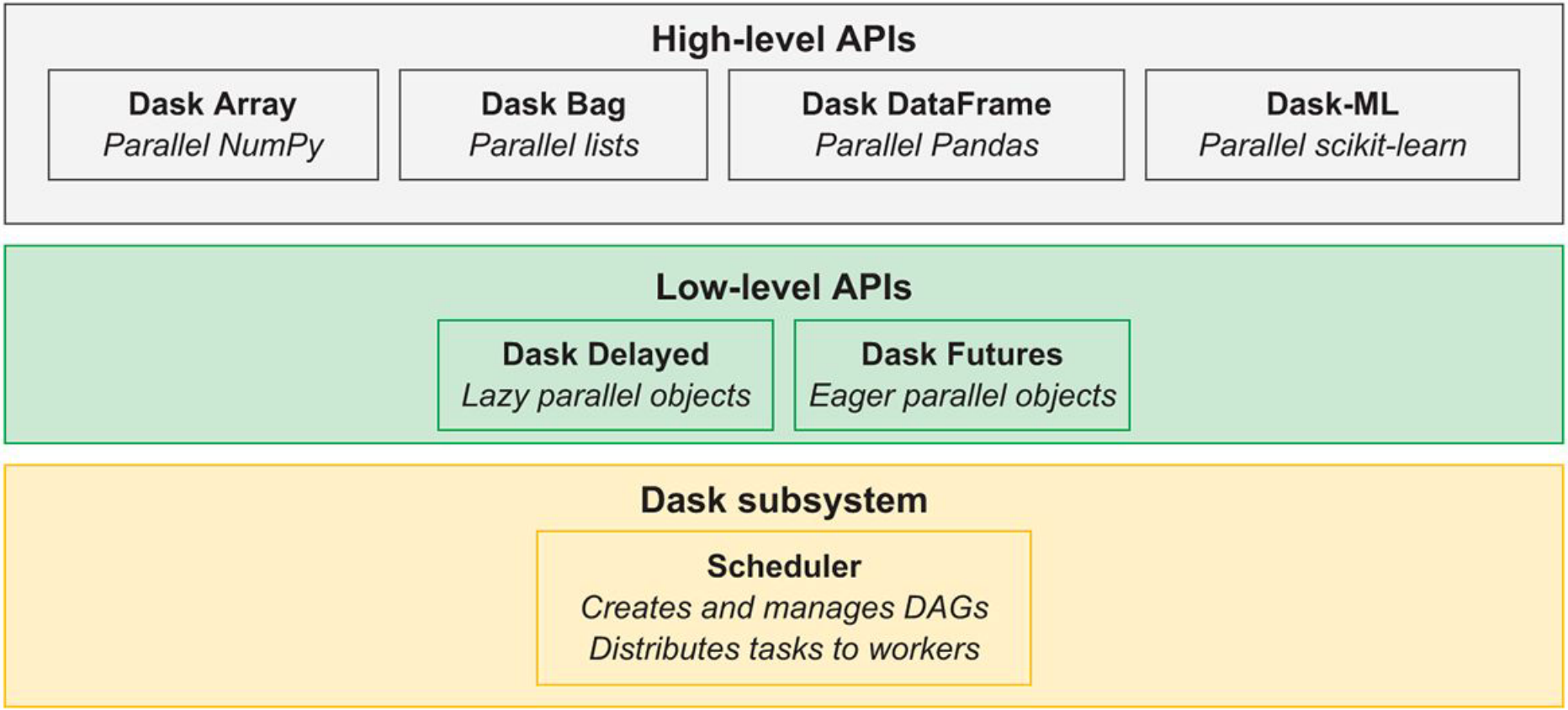
- ¿Qué es Dask?
- ¿Cómo funciona?
- Dask-ML
- Ventajas
- Ejemplos

¿Qué es Dask?

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

Puede ejecutarse localmente o ampliarse para 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
Dask subsystem
Scheduler
Creates and manages DAGs
Distributes tasks to workers

¿Qué es Dask? -2-

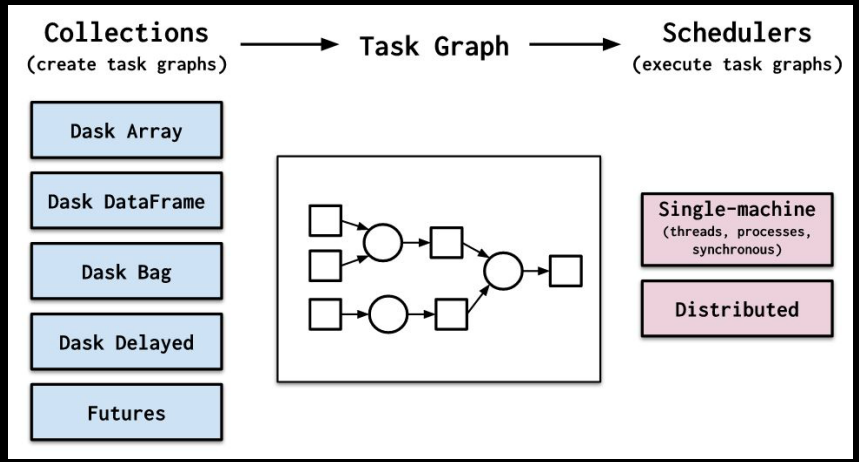
Una de las principales ventajas que tiene Dask, es su similitud con Pandas:

```
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()
```

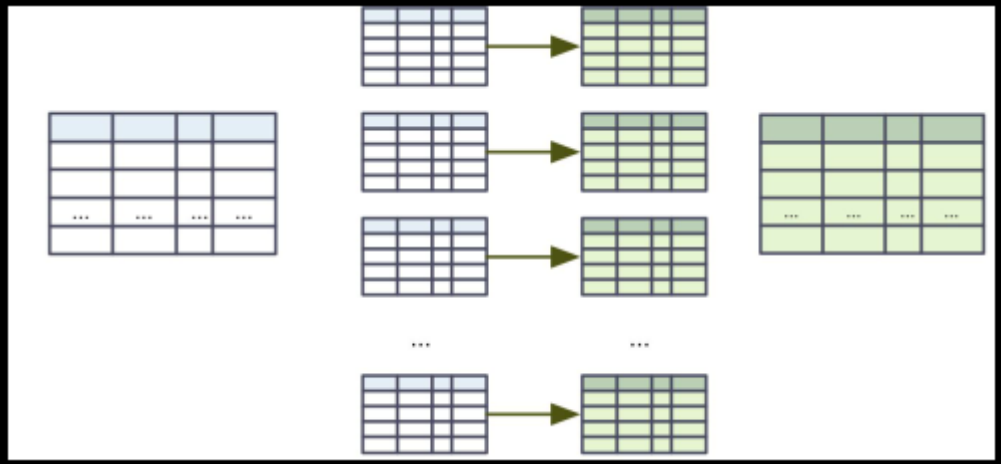
¿Qué es Dask? -3-

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



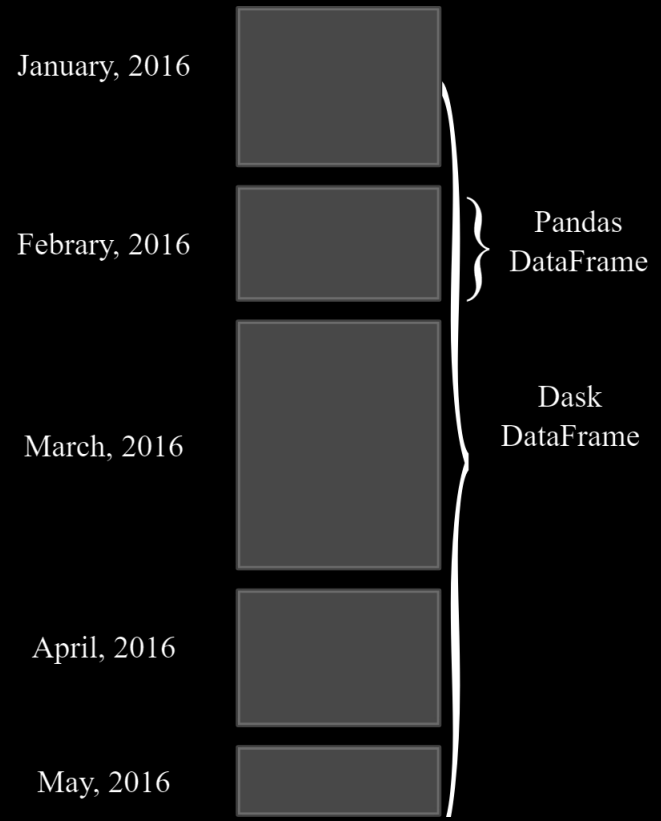
Dask y los “*chunks*”

Los *chunks* son “particiones” definidas de los datos. Cuando se trabaja con chunks, todas las operaciones se deben realizar para cada partición



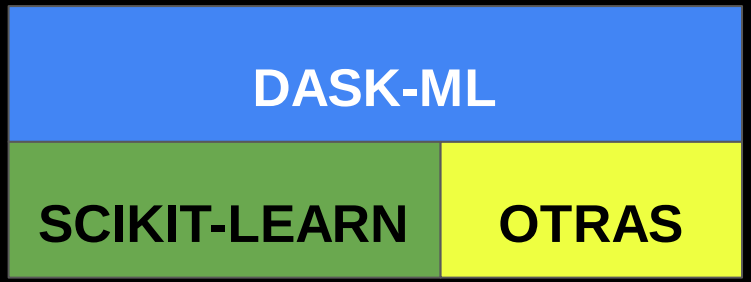
Dask y los “*chunks*” -2-

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 implementada en Dask, junto con bibliotecas de aprendizaje automático como Scikit-Learn, XGBoost y otras



Dask ML

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

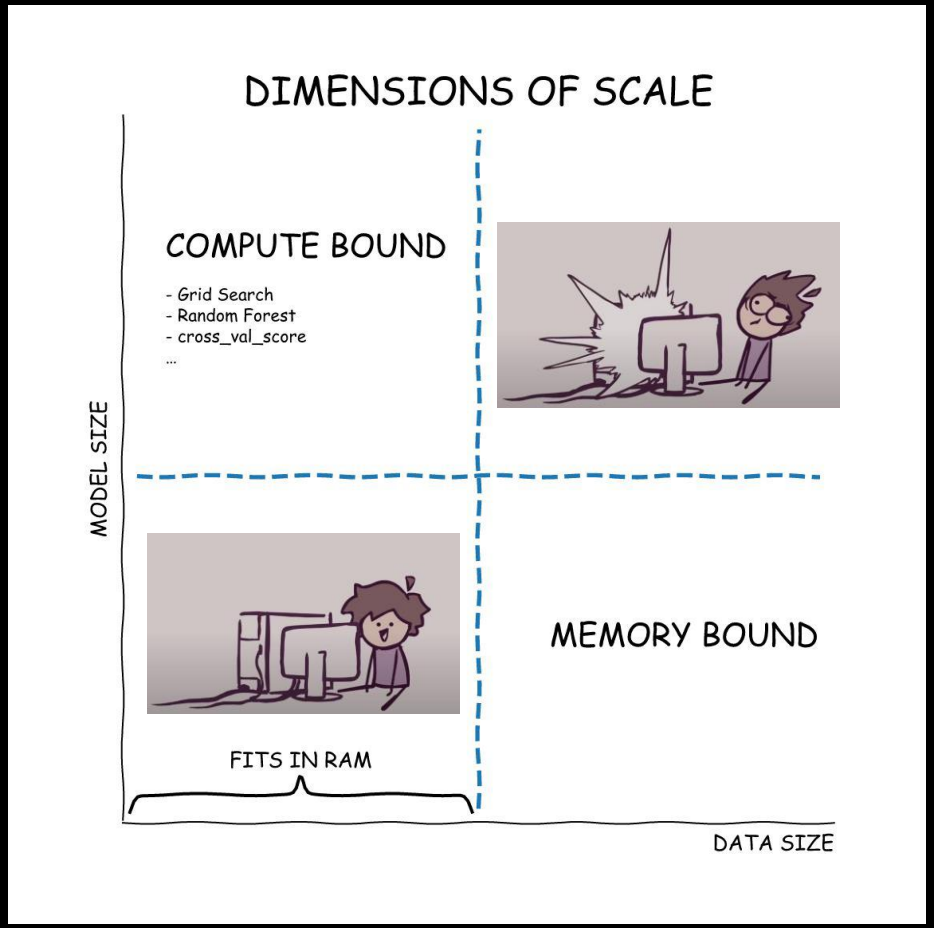
También cuenta con métodos de preprocesamiento de datos, iguales a Scikit-learn

```
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
```

Clústers 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 los errores de falta de memoria que se pueden tener con Pandas
- Tiene un mejor control de los recursos (comparado con Pandas)
- Sigue las API de Pandas, Numpy y Scikit-Learn
- Más fácil de configurar que otros servicios (como Hadoop y Spark)