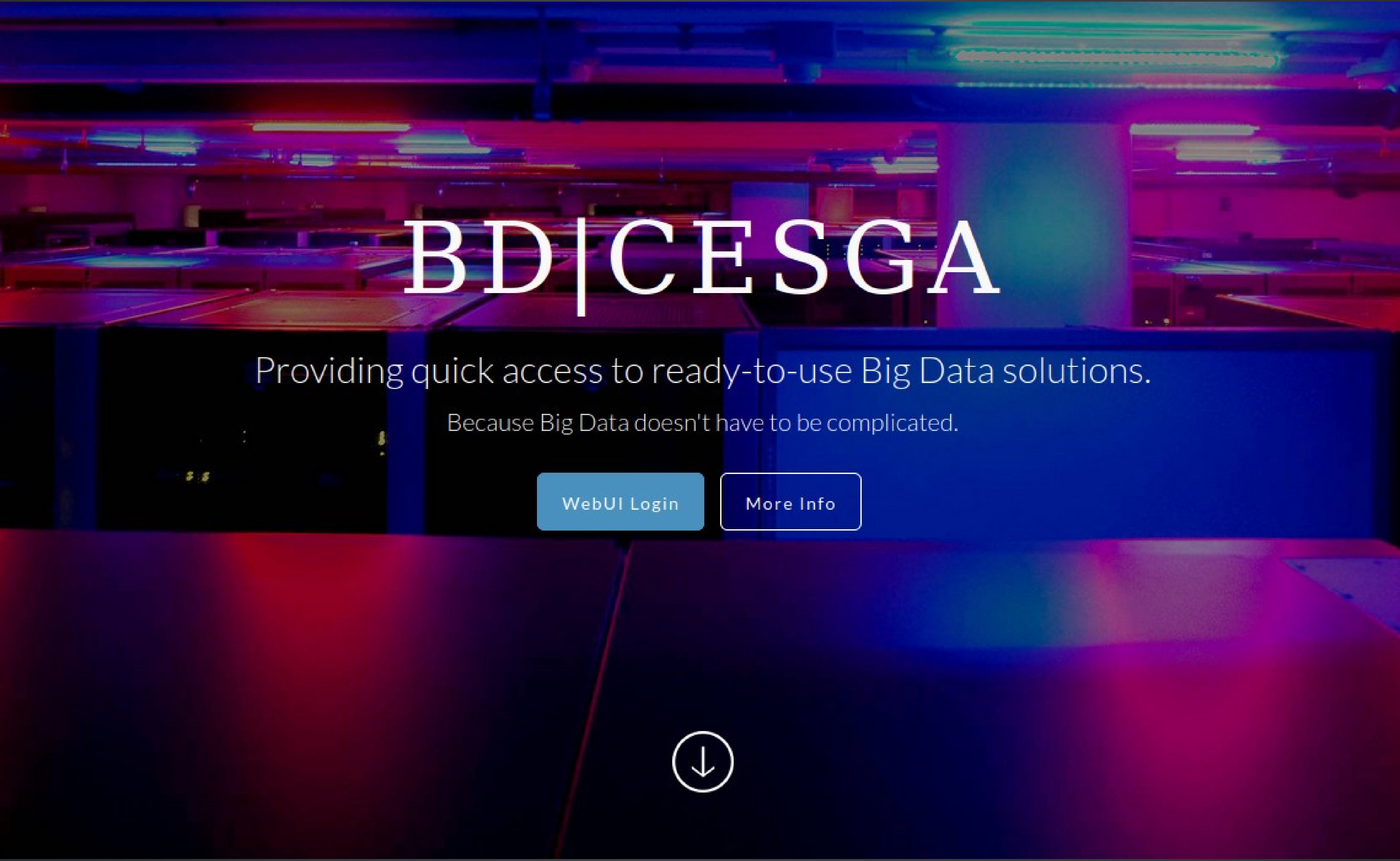
Plataforma BigData

Javier Cacheiro

CESGA

Centro de Supercomputación de Galicia



BD|CESGA

Providing quick access to ready-to-use Big Data solutions.

Because Big Data doesn't have to be complicated.

Ampliación capacidade

Discos **2TB → 18TB**

Fase 1: 211 discos (completada)

Fase 2: 165 discos (1T 2024)

Características: puntos fortes

Capacidade: **2.5PB** (CDH)

I/O agregada **30GB/s**

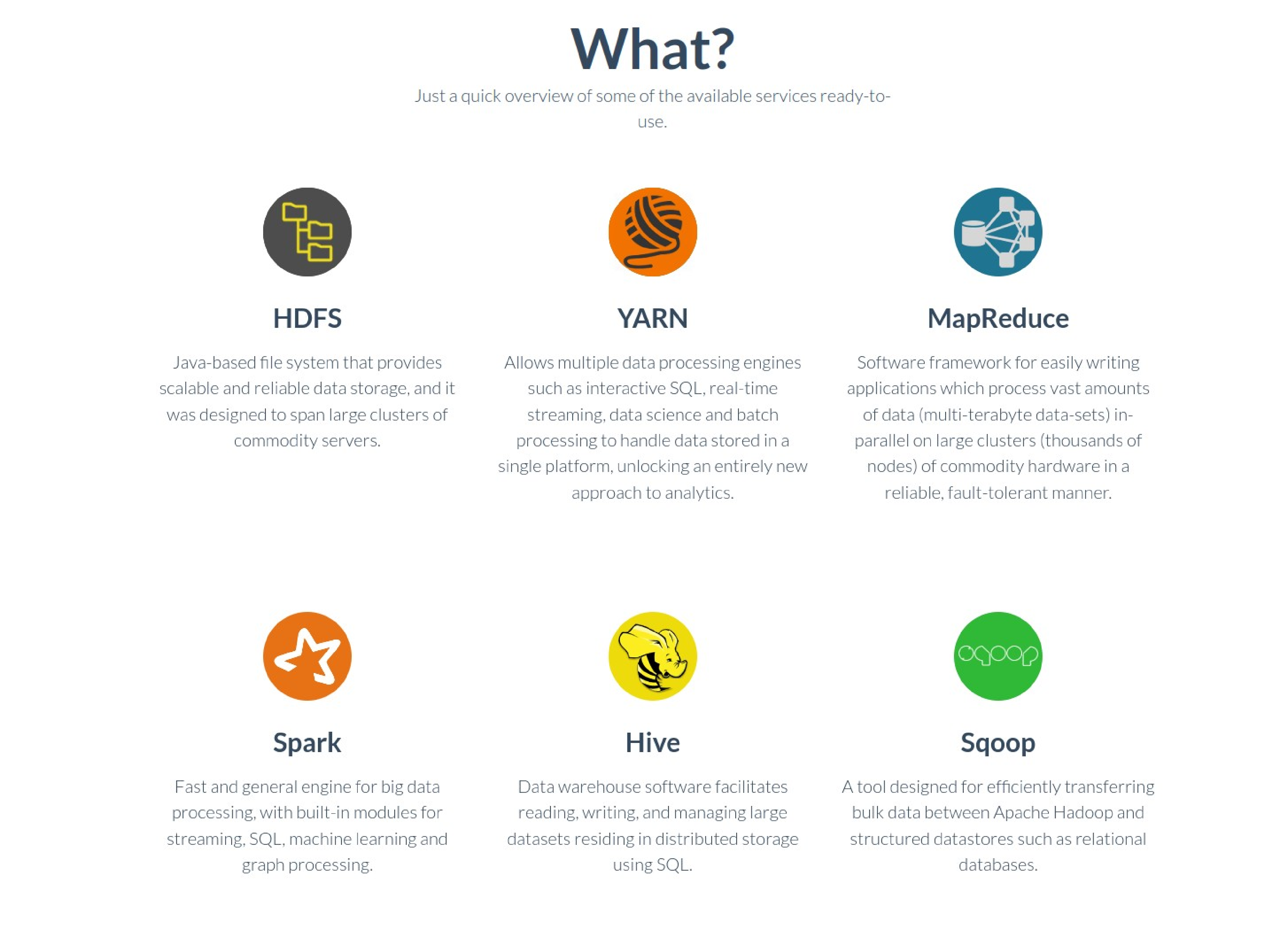
10GbE conectividade entre nodos

Características: puntos débiles

64GB RAM por nodo

12 cores por nodo

Software disponĩble



What?

Just a quick overview of some of the available services ready-to-use.

HDFS

Java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers.

YARN

Allows multiple data processing engines such as interactive SQL, real-time streaming, data science and batch processing to handle data stored in a single platform, unlocking an entirely new approach to analytics.

MapReduce

Software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

Spark

Fast and general engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing.

Hive

Data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.

Sqoop

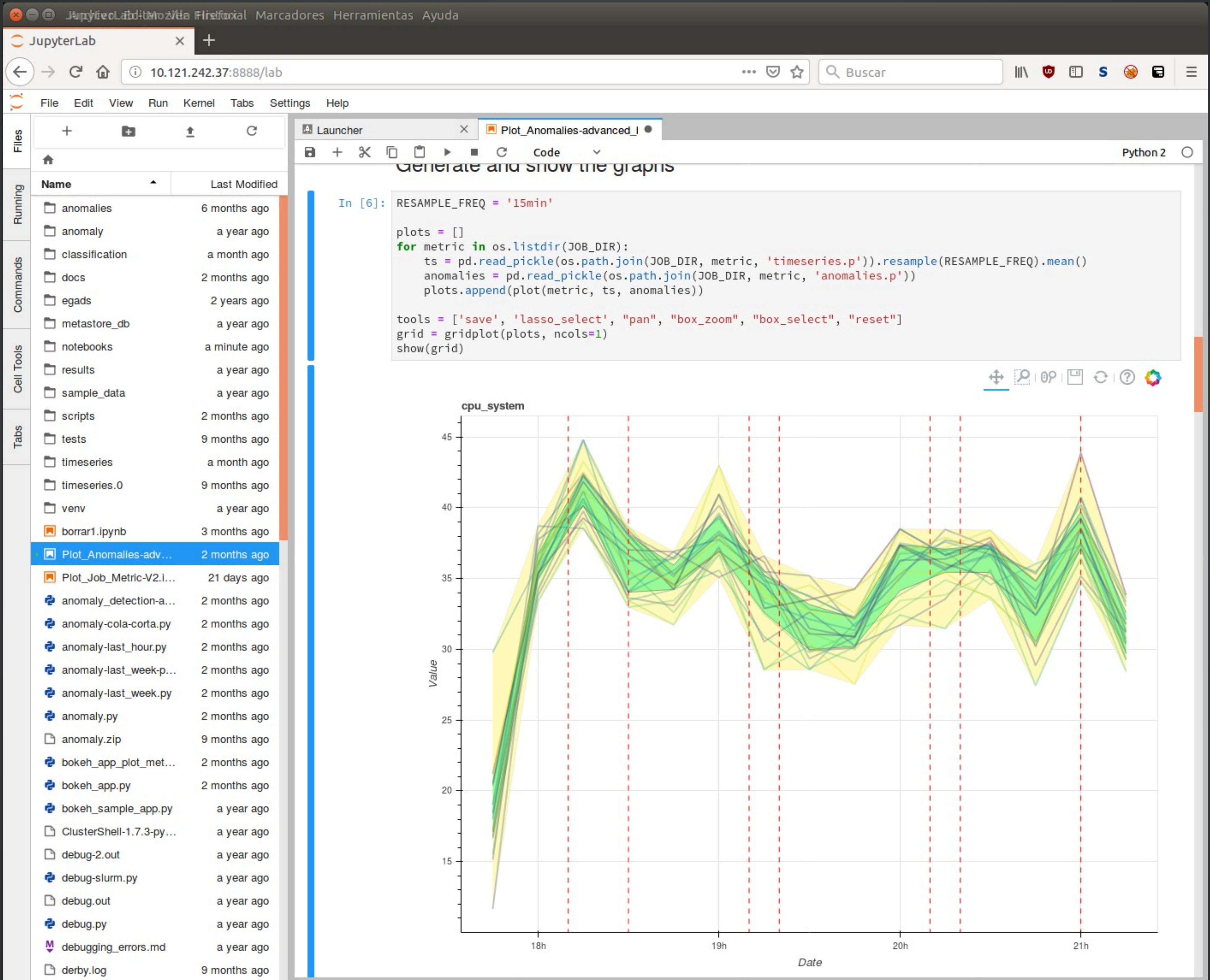
A tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.

Casos de Uso

Aplicacións con necesidade de **procesar grandes volumes de datos** pero con pouca necesidade de cálculo

JUPYTER

INTERACTIVE COMPUTING



Name	Last Modified
anomalies	6 months ago
anomaly	a year ago
classification	a month ago
docs	2 months ago
egads	2 years ago
metastore_db	a year ago
notebooks	a minute ago
results	a year ago
sample_data	a year ago
scripts	2 months ago
tests	9 months ago
timeseries	a month ago
timeseries.0	9 months ago
venv	a year ago
borrar1.ipynb	3 months ago
Plot_Anomalies-adv…	2 months ago
Plot_Job_Metric-V2.i…	21 days ago
anomaly_detection-a…	2 months ago
anomaly-cola-corta.py	2 months ago
anomaly-last_hour.py	2 months ago
anomaly-last_week-p…	2 months ago
anomaly-last_week.py	2 months ago
anomaly.py	2 months ago
anomaly.zip	9 months ago
bokeh_app_plot_met…	2 months ago
bokeh_app.py	2 months ago
bokeh_sample_app.py	a year ago
ClusterShell-1.7.3-py…	a year ago
debug-2.out	a year ago
debug-slurm.py	a year ago
debug.out	a year ago
debug.py	a year ago
debugging_errors.md	a year ago
derby.log	9 months ago

Generate and show the graphs

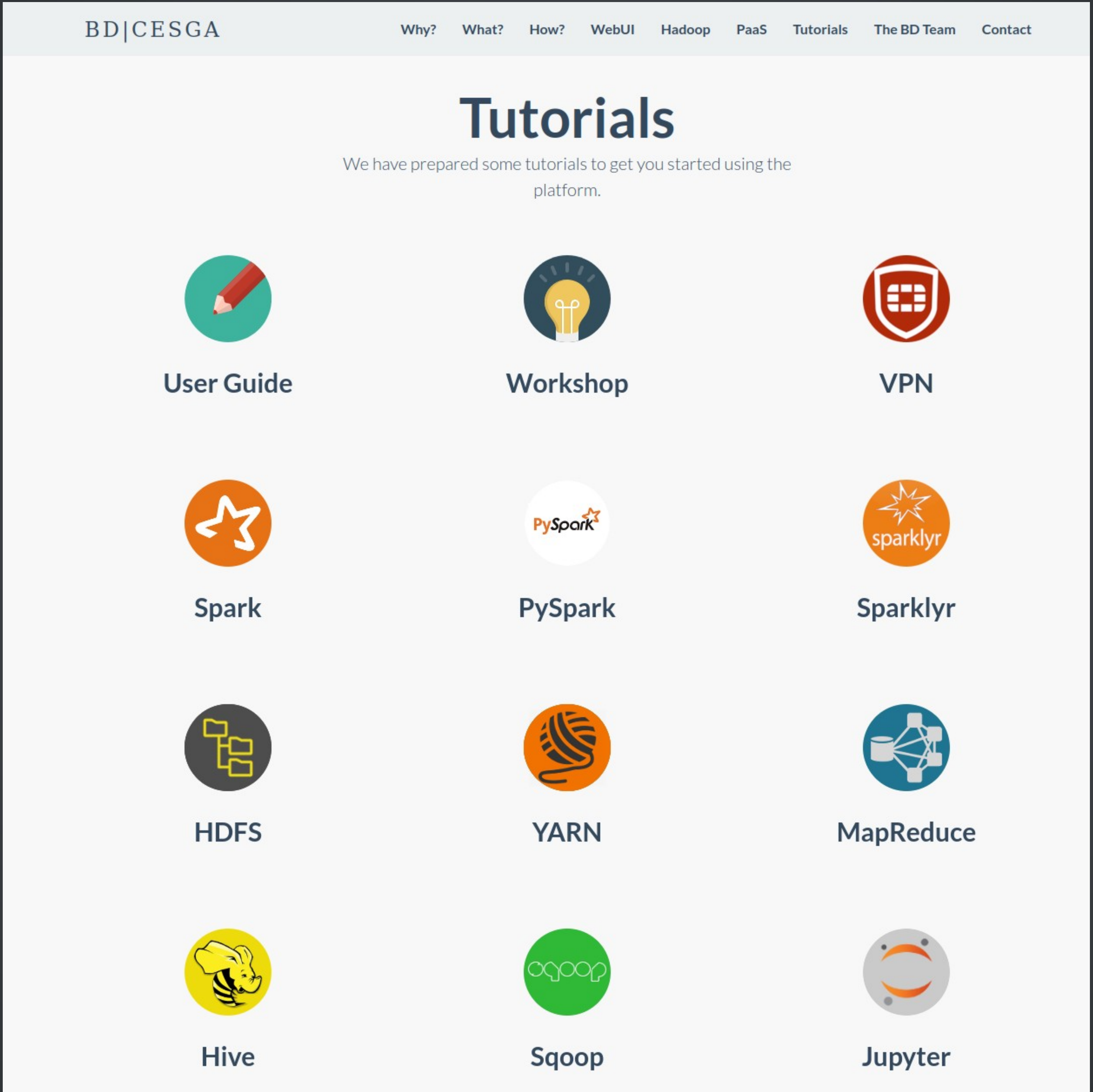
In [6]:

```
RESAMPLE_FREQ = '15min'

plots = []
for metric in os.listdir(JOB_DIR):
    ts = pd.read_pickle(os.path.join(JOB_DIR, metric, 'timeseries.p')).resample(RESAMPLE_FREQ).mean()
    anomalies = pd.read_pickle(os.path.join(JOB_DIR, metric, 'anomalies.p'))
    plots.append(plot(metric, ts, anomalies))

tools = ['save', 'lasso_select', "pan", "box_zoom", "box_select", "reset"]
grid = gridplot(plots, ncols=1)
show(grid)
```

```
module load anaconda3
start_jupyter-lab
```

→ Só para usar con Spark ←

Titoriais



Tutorials

We have prepared some tutorials to get you started using the platform.

User Guide

Workshop

VPN

Spark

PySpark

Sparklyr

HDFS

YARN

MapReduce

Hive

Sqoop

Jupyter

Referencias

- Portal, Titoriais, Guía de uso
  - https://bigdata.cesga.gal
- Curso Spark
  - https://github.com/javicacheiro/pyspark_course