PUE ACADEMY Day
8 de mayo - Madrid 2019

#PUEAcademyDay19

Cloudera Academic

13:05 a 14:05h

Cloudera Workshop: Ingesta y Análisis de Datos en Tiempo Real mediante Kafka y Spark Streaming

Ramon de la Rosa | Big Data and Cloud Specialist at PUE

cloudera ACADEMIC PARTNER

Agenda

- Qué es Big Data
- Cloudera Hadoop
- Aplicaciones Big Data
- Demo caso de uso: PUEAcademyDay Distributed fail to ban
- Cloudera Academy Program CAP

¿Qué es Big Data?

¿Qué es Big Data?

Big Data nació con el objetivo de cubrir unas necesidades no satisfechas por las tecnologías existentes, como es el almacenamiento y tratamiento de grandes volúmenes de datos que poseen unas características muy concretas definidas como las tres **V's** (en la actualidad puede haber más).

Cloudera Apache Hadoop

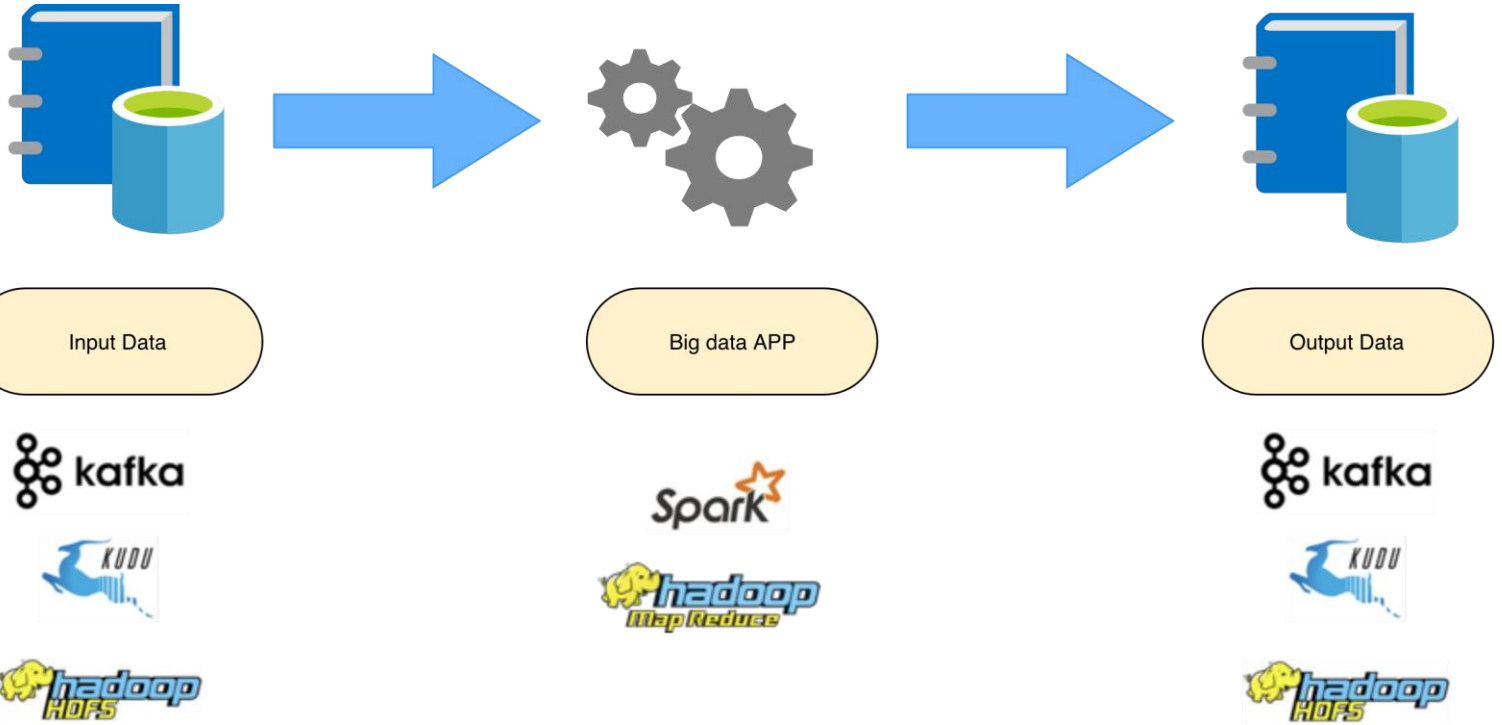
Developer(s)	Apache Software Foundation
Development status	Active
Written in	Java
Operating system	Cross-platform
Type	Distributed file system
License	Apache License 2.0
Website	hadoop.apache.org

Apache Hadoop is an open-source software framework for distributed storage and distributed processing of very large data

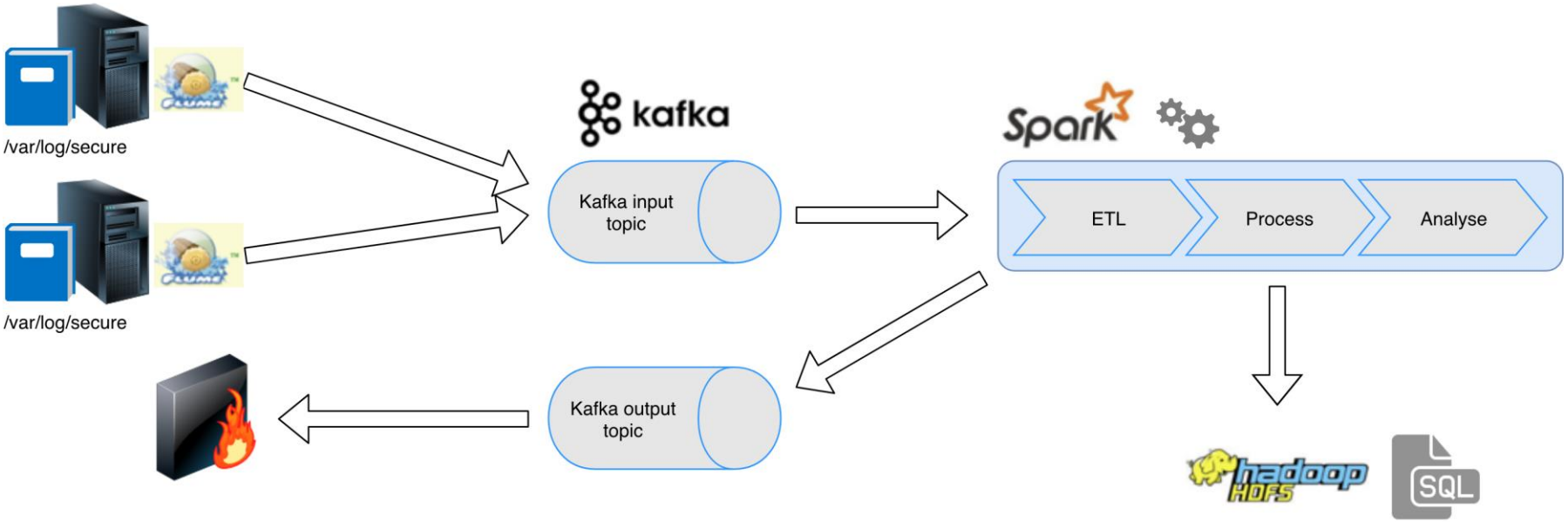
Cloudera was the first commercial software vendor to release a Hadoop Distribution with enterprise features security and governance

Packages included are:
Apache Pig, Apache Hive, Apache HBase, Apache Spark, Apache ZooKeeper, Cloudera Impala, Apache Flume, Apache Sqoop, Apache Oozie, Solr

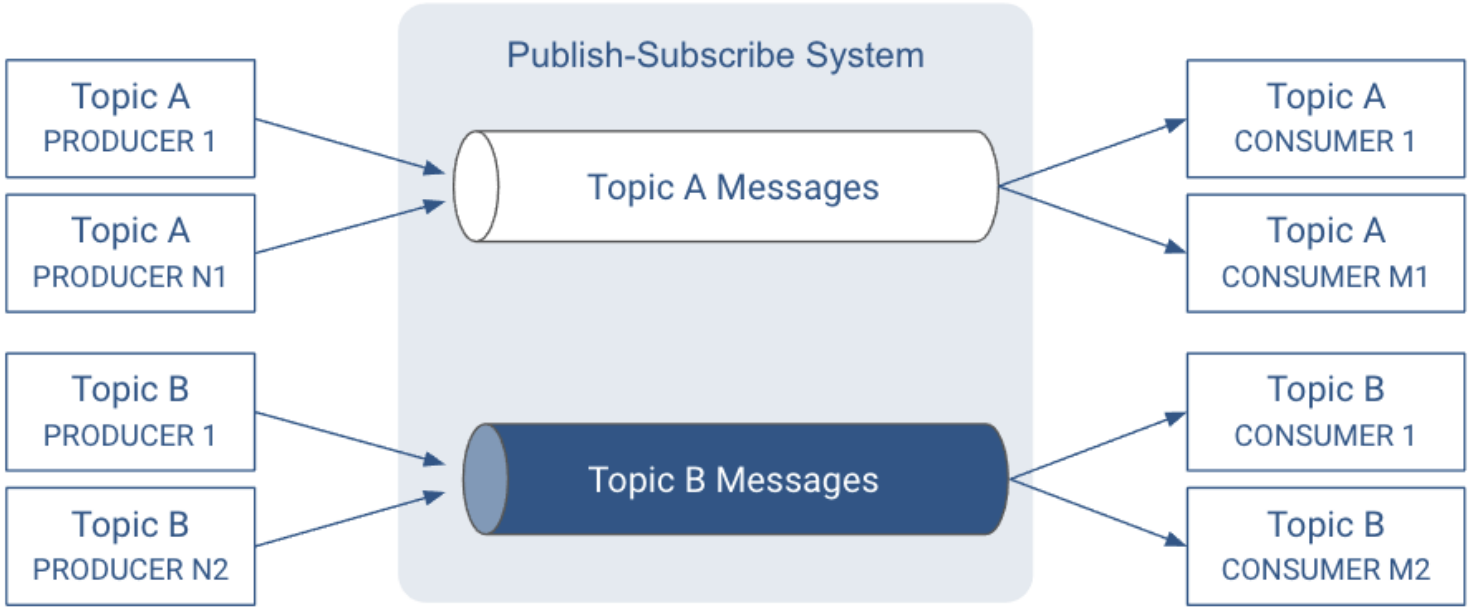
Aplicación Big Data



Demo: Kafka Spark Stream



kafka

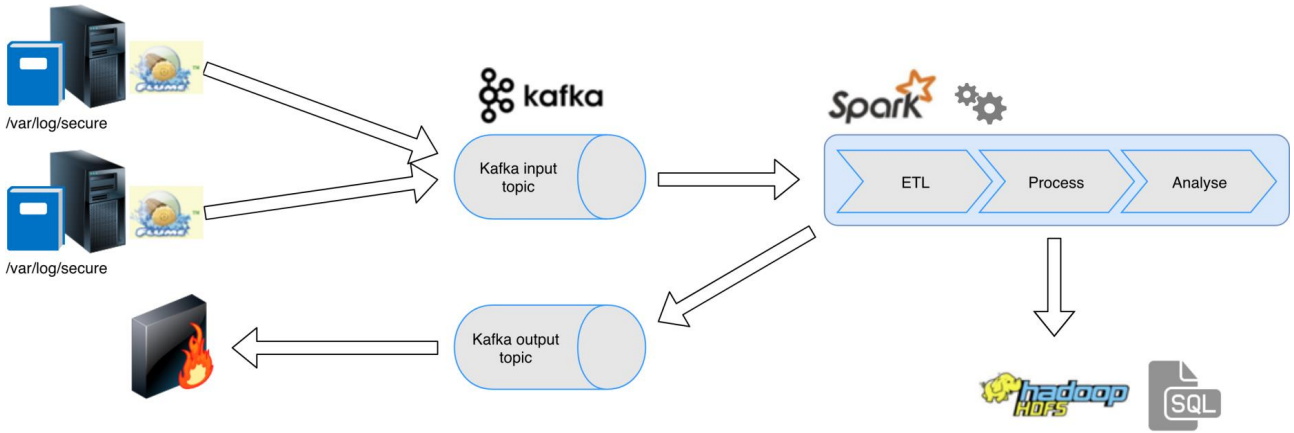

Publish-Subscribe System
Topic A
PRODUCER 1
Topic A
PRODUCER N1
Topic A Messages
Topic A
CONSUMER 1
Topic A
CONSUMER M1
Topic B
PRODUCER 1
Topic B
PRODUCER N2
Topic B Messages
Topic B
CONSUMER 1
Topic B
CONSUMER M2

Spark

Original author(s)	Matei Zaharia
Developer(s)	Apache Software Foundation, UC Berkeley AMPLab, Databricks
Initial release	May 26, 2014; 4 years ago
Stable release	v2.4.2 / April 23, 2019; 1 day ago
Repository	https://github.com/apache/spark
Written in	Scala, Java, Python, R[1]
Operating system	Microsoft Windows, macOS, Linux
Available in	Scala, Java, SQL, Python, R
Type	Data analytics, machine learning algorithms
License	Apache License 2.0
Website	spark.apache.org

- Apache Spark es un framework open souce de computación distribuida
- Componentes
  - Spark Core (RDD)
  - Spark SQL (Data Frames)
  - Spark Streamming
  - MLLib
  - GrapX
- Se puede programar en
  - Scala (*)
  - Python (*)
  - Java
  - R

Demo: Kafka Spark Stream

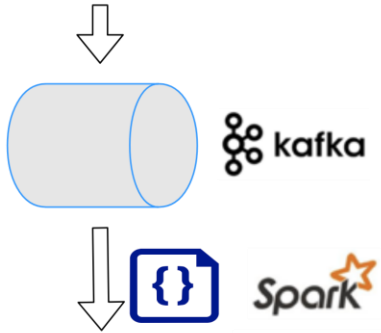


GitHub

https://github.com/rdelaros/cap-puedacademyday19

Lectura de topic

```
May  3 12:04:10 s2 sshd[323856]: input_userauth_request: invalid user elpa [preauth]
May  3 12:04:10 s2 sshd[323856]: Received disconnect from 167.86.88.236 port 55366:11: Normal Shutdown, Thank you for playing [preauth]
May  3 12:04:10 s2 sshd[323856]: Disconnected from 167.86.88.236 port 55366 [preauth]
May  3 12:04:41 s2 sshd[324056]: Invalid user dantoy23 from 167.86.88.236 port 50776
```




```
+----+--------------------+--------+---------+------+--------------------+-------------+
| key|               value|   topic|partition|offset|           timestamp|timestampType|
+----+--------------------+--------+---------+------+--------------------+-------------+
|null|[41 70 72 20 32 3...|new_logs|        0|     0|2019-05-03 12:03:...|            0|
|null|[41 70 72 20 32 3...|new_logs|        0|     1|2019-05-03 12:03:...|            0|
+----+--------------------+--------+---------+------+--------------------+-------------+
```

	timestamp	topic	value
0	2019-05-03 12:28:27.525	new_logs	Apr 29 05:10:14 s2 sudo: pam_unix(sudo:session...
1	2019-05-03 12:28:27.525	new_logs	Apr 29 05:11:14 s2 sudo: pam_unix(sudo:session...
2	2019-05-03 12:28:27.525	new_logs	Apr 29 05:11:14 s2 sudo: icinga : TTY=unknown...
3	2019-05-03 12:28:27.525	new_logs	Apr 29 05:11:14 s2 sudo: pam_unix(sudo:session...

Tabla de logs

```
+----+--------------------+--------+---------+------+--------------------+-------------+
| key|               value|   topic|partition|offset|           timestamp|timestampType|
+----+--------------------+--------+---------+------+--------------------+-------------+
|null|[41 70 72 20 32 3...|new_logs|        0|     0|2019-05-03 12:03:...|            0|
|null|[41 70 72 20 32 3...|new_logs|        0|     1|2019-05-03 12:03:...|            0|
+----+--------------------+--------+---------+------+--------------------+-------------+
```

```
May  3 12:04:10   s2   sshd[323856]: input_userauth_request: invalid user elpa [preauth]
```

timestamp | server | service | log_entry

Spark

```
+------------------+------+-------------+--------------------+
|         timestamp|server|      service|           log_entry|
+------------------+------+-------------+--------------------+
|2019-04-28 17:06:01|    s3|        sudo:|  icinga : TTY=un...|
|2019-04-28 17:05:45|    s3|sshd[218391]:| Invalid user ora...|
|2019-04-28 17:05:45|    s3|sshd[218391]:| input_userauth_r...|
|2019-04-28 17:05:45|    s3|sshd[218391]:| Disconnected fro...|
|2019-04-28 17:05:45|    s3|sshd[218391]:| Received disconn...|
|2019-04-28 17:05:21|    s3|sshd[218365]:| Disconnected fro...|
|2019-04-28 17:05:21|    s3|sshd[218365]:| Received disconn...|
|2019-04-28 17:05:20|    s3|sshd[218365]:| input_userauth_r...|
|2019-04-28 17:05:20|    s3|sshd[218365]:| Invalid user ora...|
|2019-04-28 17:05:01|    s3|        sudo:|  icinga : TTY=un...|
+------------------+------+-------------+--------------------+
```

SSH Invalid users

```
Invalid user dantoy23 from 167.86.88.236 port 55406
```

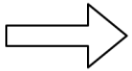
user ip

```
+-------------------+------+-------------+--------------------+
|          timestamp|server|      service|           log_entry|
+-------------------+------+-------------+--------------------+
|2019-04-28 17:06:01|    s3|        sudo:|  icinga : TTY=un...|
|2019-04-28 17:05:45|    s3|sshd[218391]:| Invalid user ora...|
|2019-04-28 17:05:45|    s3|sshd[218391]:| input_userauth_r...|
|2019-04-28 17:05:45|    s3|sshd[218391]:| Disconnected fro...|
|2019-04-28 17:05:45|    s3|sshd[218391]:| Received disconn...|
|2019-04-28 17:05:21|    s3|sshd[218365]:| Disconnected fro...|
|2019-04-28 17:05:21|    s3|sshd[218365]:| Received disconn...|
|2019-04-28 17:05:20|    s3|sshd[218365]:| input_userauth_r...|
|2019-04-28 17:05:20|    s3|sshd[218365]:| Invalid user ora...|
|2019-04-28 17:05:01|    s3|        sudo:|  icinga : TTY=un...|
+-------------------+------+-------------+--------------------+
```

```
+-------------------+------+--------+---------------+
|          timestamp|server|    user|             ip|
+-------------------+------+--------+---------------+
|2019-05-03 12:04:41|    s5|dantoy23|  167.86.88.236|
|2019-05-03 12:04:24|    s5|      pi|180.139.114.144|
|2019-05-03 12:04:14|    s5|      pi|180.139.114.144|
|2019-05-03 12:04:10|    s5|    elpa|  167.86.88.236|
|2019-05-03 12:03:40|    s5|  chello|  167.86.88.236|
|2019-05-03 12:03:11|    s5| abcd123|  167.86.88.236|
|2019-05-03 12:02:41|    s5|   kulot|  167.86.88.236|
|2019-05-03 12:02:10|    s5| jen1414|  167.86.88.236|
|2019-05-03 12:01:39|    s5|    dio2|  167.86.88.236|
|2019-05-03 12:01:09|    s5|    dio1|  167.86.88.236|
+-------------------+------+--------+---------------+
```

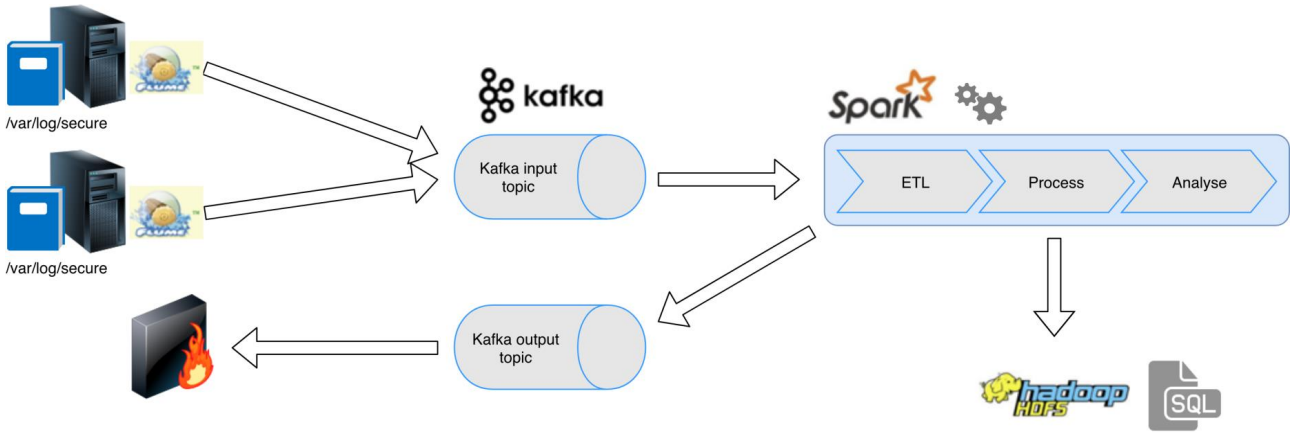
Windowing

timestamp	server	user	ip
2019-05-03 12:04:41	s5	dantoy23	167.86.88.236
2019-05-03 12:04:24	s5	pi	180.139.114.144
2019-05-03 12:04:14	s5	pi	180.139.114.144
2019-05-03 12:04:10	s5	elpa	167.86.88.236
2019-05-03 12:03:40	s5	chello	167.86.88.236
2019-05-03 12:03:11	s5	abcd123	167.86.88.236
2019-05-03 12:02:41	s5	kulot	167.86.88.236
2019-05-03 12:02:10	s5	jen1414	167.86.88.236
2019-05-03 12:01:39	s5	dio2	167.86.88.236
2019-05-03 12:01:09	s5	dio1	167.86.88.236



start	end	ip	count
2019-05-03 12:00:00	2019-05-03 12:05:00	167.86.88.236	10
2019-05-03 12:00:00	2019-05-03 12:05:00	202.22.142.111	1
2019-05-03 12:00:00	2019-05-03 12:05:00	180.139.114.144	2
2019-05-03 11:55:00	2019-05-03 12:00:00	167.86.88.236	10
2019-05-03 11:50:00	2019-05-03 11:55:00	194.156.28.8	1
2019-05-03 11:50:00	2019-05-03 11:55:00	167.86.88.236	10
2019-05-03 11:45:00	2019-05-03 11:50:00	167.86.88.236	10
2019-05-03 11:40:00	2019-05-03 11:45:00	167.86.88.236	10
2019-05-03 11:40:00	2019-05-03 11:45:00	128.199.245.4	1
2019-05-03 11:40:00	2019-05-03 11:45:00	40.76.50.216	1

Demo: Kafka Spark Stream



GitHub

https://github.com/rdelaros/cap-puedacademyday19

Nuevas profesiones en

- Arquitecto Big Data
- Desarrollador Big Data
- Data Analyst
- Científico de datos
- Administrador de Hadoop

SQL | Java | Python

Linux | Scala | Ansible

Kudu | Spark | Hbase

Hive | Impala | Hadoop

Kafka | NiFi

Cloudera Academy Program CAP

- Cursos
  - Introduction to Hadoop and Big Data
  - Developer Training for Spark and Hadoop
- Máquinas virtuales
  - 1 máquina virtual por curso simulando un cluster
  - Cloudera Quick Start Virtual Machine
- Licencia Cloudera Enterprise
- Más información: www.pue.es/cloudera-academy


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
JUSPAXAR
GOZAIMASHITA
EFCHARISTO
KOMAPSUMNIDA
MAAKE
GRAZIE
MEHRBANI
PALDIES
SUKSAMA
EKHMET
YAQHANYELAY
TASHAKKUR ATU
TINGKI
BÏYAN
SHUKRIA
BOLZÏN
MERCI

PUE ACADEMY Day

www.pue.es

¡Gracias!

#PUEAcademyDay19

pueacademy@pue.es

BCN: 93 206 02 49

MAD: 91 162 06 69

pue

PROGRAMAS EDUCATIVOS

ORACLE ACADEMY · cloudera ACADEMIC PARTNER · Microsoft Imagine Academy · vmware IT ACADEMY · Linux Professional Institute · Networking Academy

PROGRAMAS DE CERTIFICACIÓN

EC-Council Associate · Certified Associate PROGRAM PROVIDER · AUTODESK. Certification · Microsoft Office Specialist · Microsoft Technology Associate · unity Certification · pue mobile