



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Tecnologías para el Big Data Apache Spark

Agenda



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

^ Introducción

^ Apache Spark como procesamiento

^ Ejercicios Prácticos

Compartamos



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Coméntanos sobre el artículo que leíste

Introducción

Objetivos

- Recordar el nuevo paradigma del Big Data.
- Comprender las necesidades de Sistemas Batch



CTIC UNI

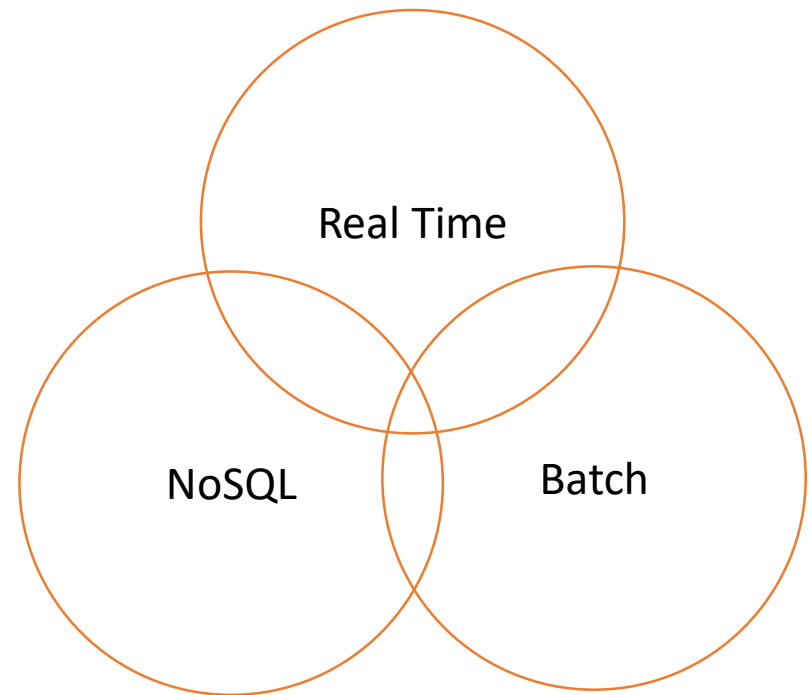
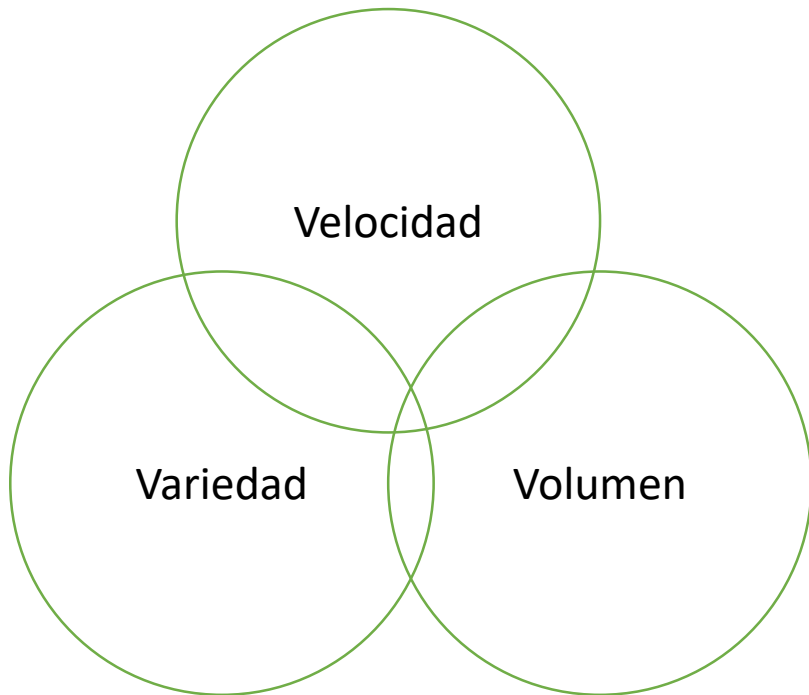
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Introducción



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería



Grandes problemas, grandes soluciones

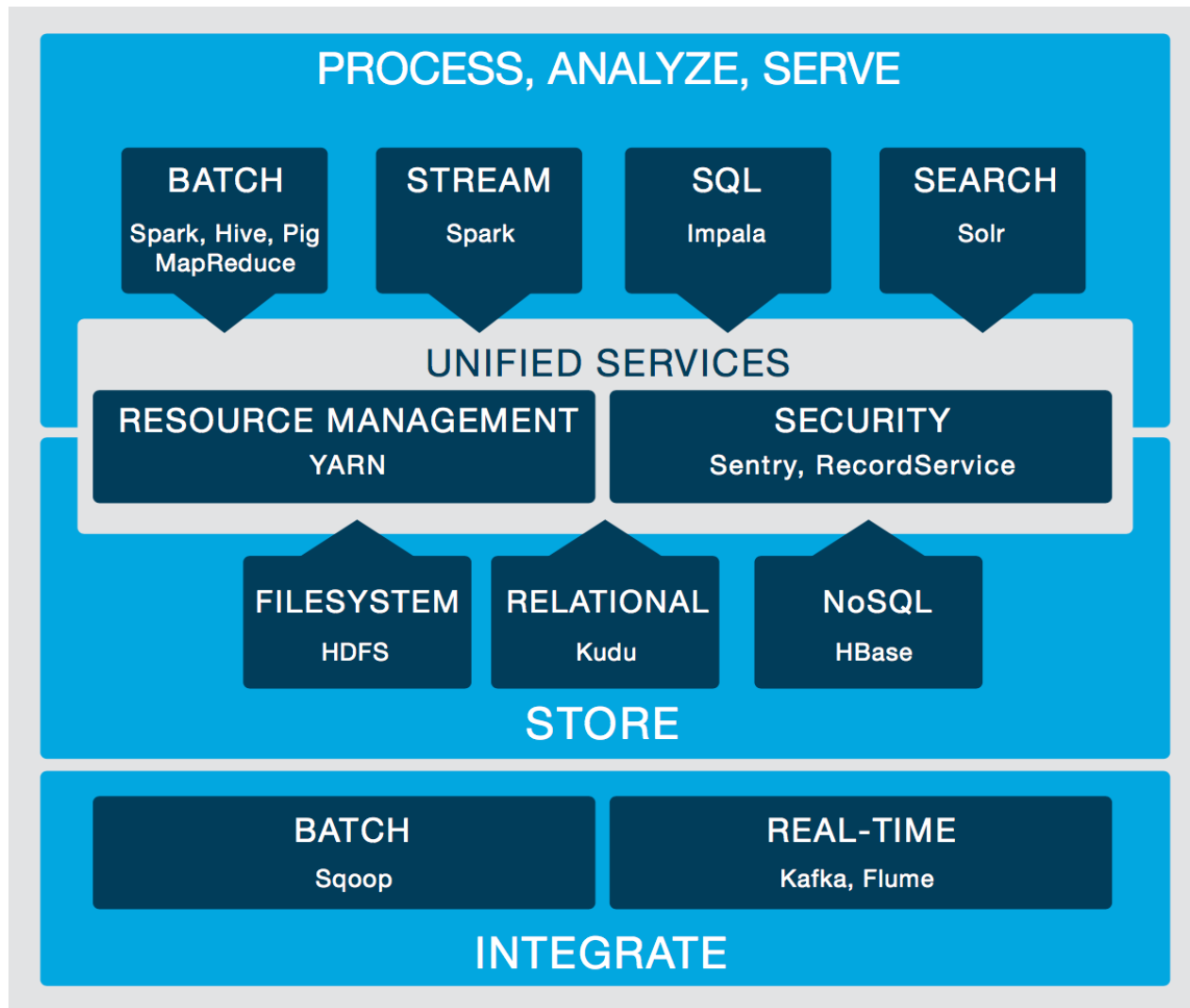
Tecnologías batch procesamiento



Objetivos

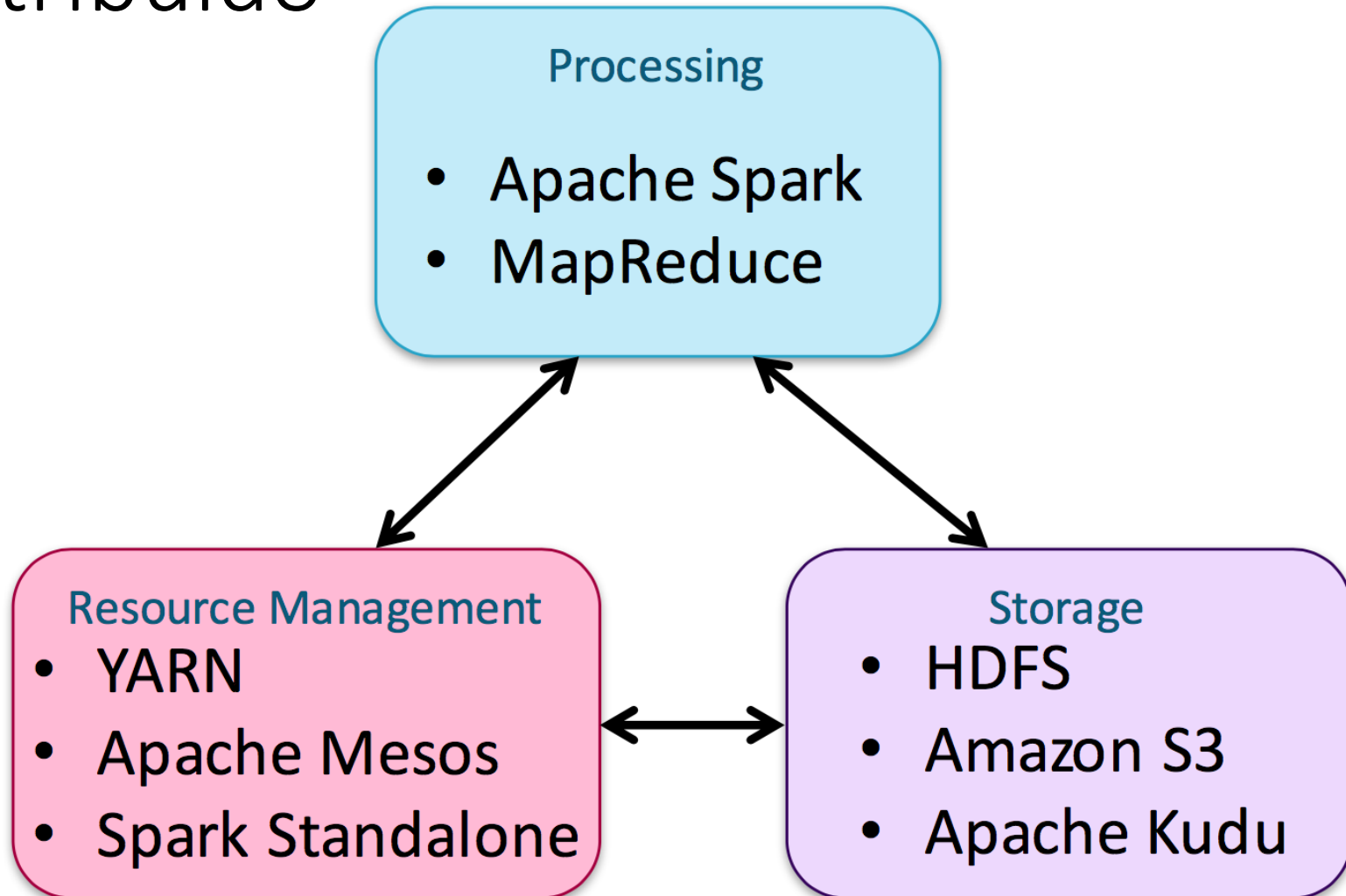
- Entender las necesidades de procesamiento en batch en el ecosistema de big data
- Comprender las necesidades que originaron Spark
- Comprender los componentes de Spark
- Comprender el flujo batch de procesamiento en Spark

Ecosistema Big Data



Cloudera

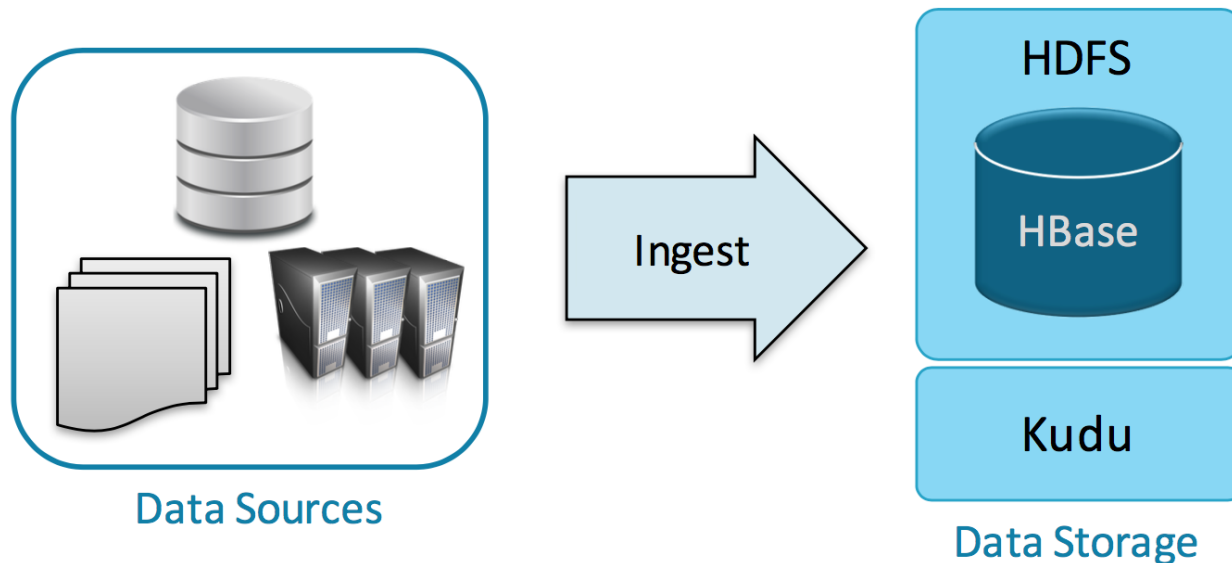
Procesamiento distribuido



Ingesta de data y almacenamiento



- Hadoop ingesta data de muchas fuentes y muchos formatos
- Tradicionales como DB
- Logs, event data o archivos importados





Apache Spark: Un motor para procesamiento de datos a gran escala

- Propósito general
- Corre en Hadoop y procesa data en HDFS
- Soporta un amplio rango de flujos de trabajo
 - Machine Learning
 - BI
 - Streaming
 - Batch processing
 - Querying structured data

Hadoop MapReduce



- El motor procesamiento original de Hadoop
- Basado en Java
- Algunas herramientas existentes actualmente usan código de MapReduce
- La principal forma de procesamiento antes de la introducción de Spark

Recuérdame



CTIC UNI

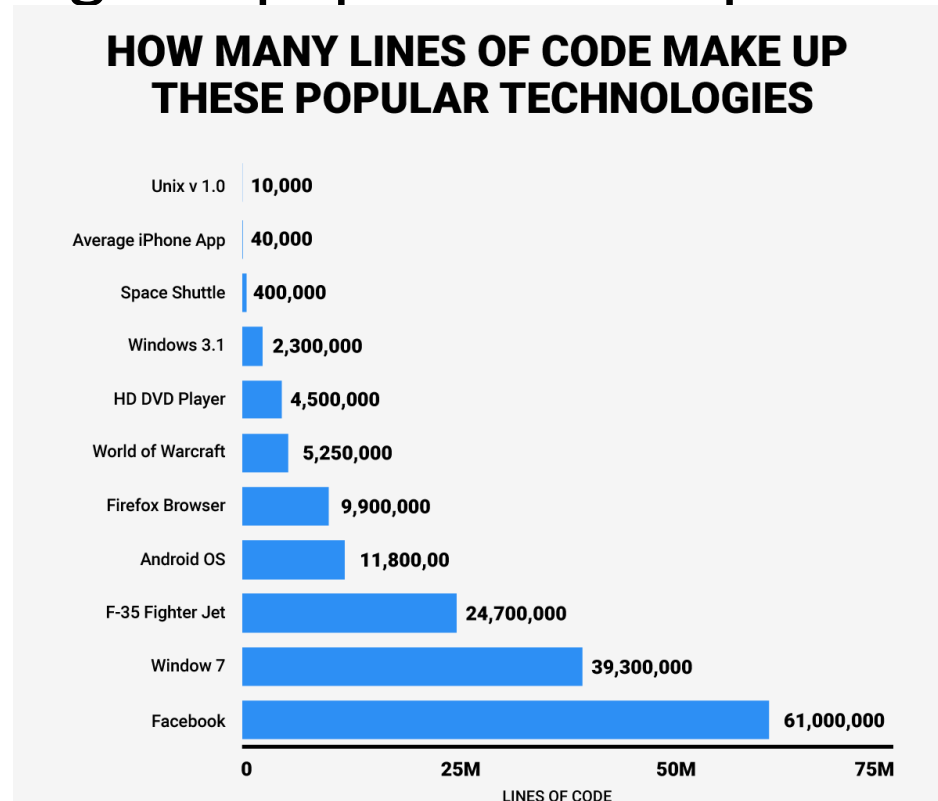
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Comparte lo que recuerdas de las últimas
clases ¿Hadoop? ¿HDFS? ¿Hive? ¿Hue?
¿Azure? ¿Amazon? ¿Impala? ...

Historia Spark



- **2009** Inició como un proyecto de clase en la universidad de Berkeley con la idea de construir un framework para la administración de clusters ¡2000 líneas de código! Equipo liderado por Matei Zahara.



Historia Spark



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

- **2010** Publica el paper de Spark
- **2012** Publica el paper de RDD (Resilient Distributed Datasets)
- **2014** Pasa a ser un proyecto incubado por Apache



cloudera



HORTONWORKS®



databricks™

NETFLIX



Apache Spark



- Escrito en Scala
 - Lenguaje de programación funcional que corre en el JVM
- Spark shell
 - Interactivo para aprendizaje o exploración de data
 - Python, Scala o R
- Spark applications
 - Para procesamiento de grandes volúmenes de datos
 - Java, Scala o Python

¿Qué es Spark?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Es una manera sencilla y rápida de trabajar grandes volúmenes de data de forma distribuida.

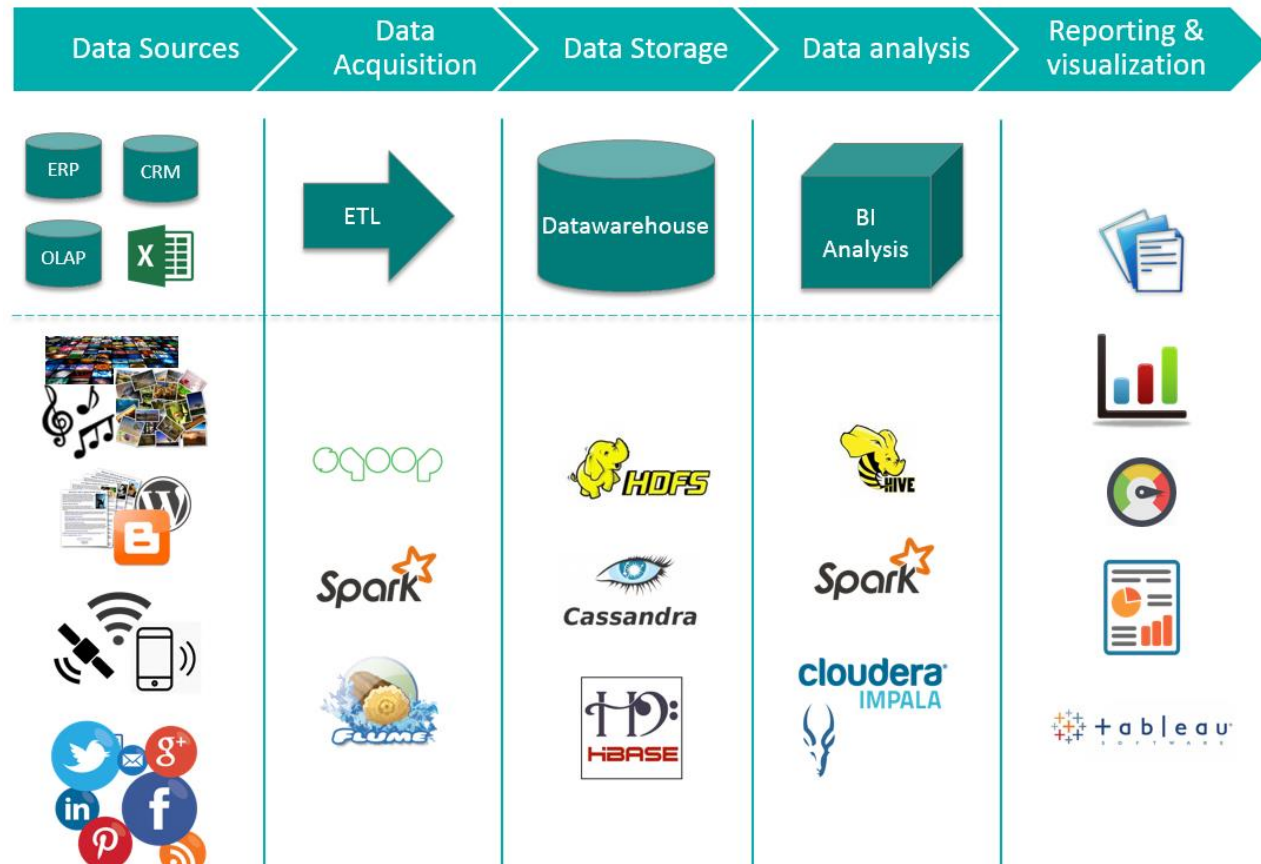


Divide y vencerás

¿Qué es Spark?



Es un motor de procesamiento de código libre construido en Scala para ser rápido, fácil de entender y proporcionar analítica sofisticada.



Comunicándonos



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

En equipos coméntanos cómo es el flujo de un proyecto en el que estés o hayas trabajado, ¿Se parece a las 5 fases de un proyecto Big Data?

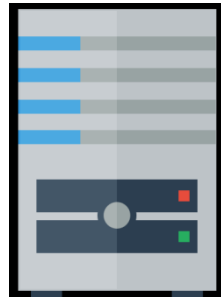
¿Qué hace Spark?



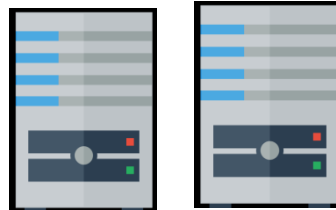
CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Spark permite manejar hasta varios petabytes de datos a la vez distribuido a través de miles de servidores virtuales o físicos.



Driver



Workers



Workers



Workers



¿Qué hace Spark?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Tiene una serie de librerías y APIs que soportan lenguajes como Scala, Java, Python y R

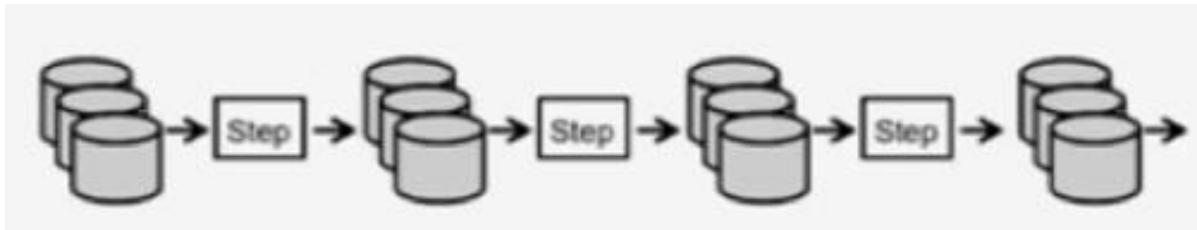
Tiene buena integración con HDFS, pero también puede leer de Hbase, Cassandra, MongoDB o Amazon S3 como Data Storage



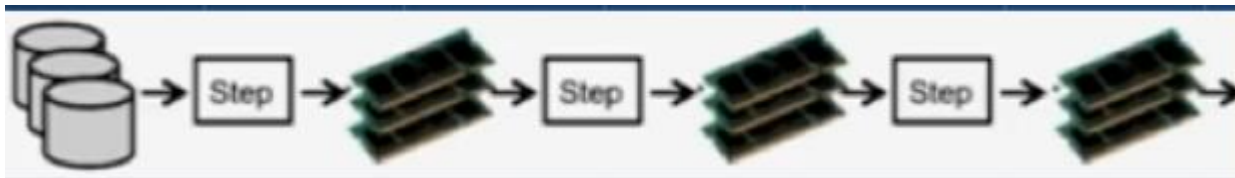
¿Por qué es tan rápido?



La mayoría de herramientas lee y escribe de disco
(Hadoop con Map Reduce)



Spark permite persistir la data en memoria RAM para futuras interacciones (tengo toda la RAM del cluster)

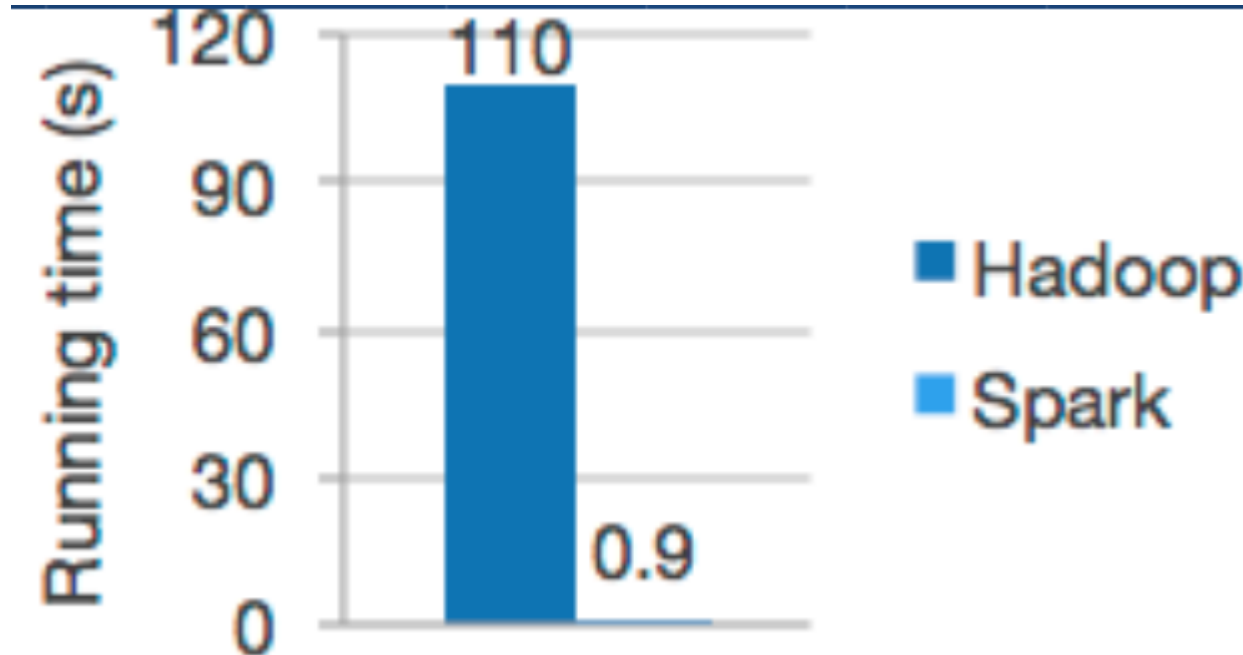


¿Qué tan rápido es?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería



Regresión logística en Spark y Hadoop

What is Apache Spark?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

<https://www.youtube.com/watch?v=SxAxAhn-BDU>

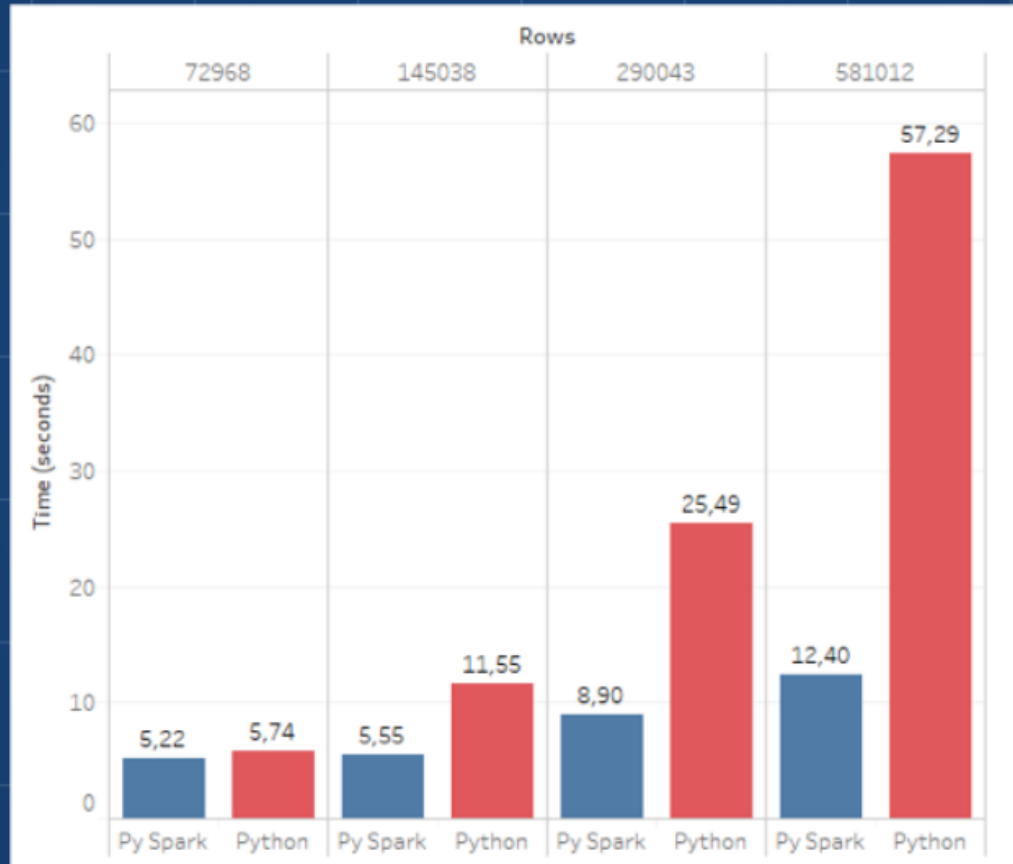
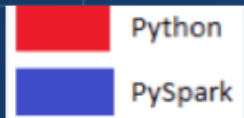
De Cloudera

¿Qué tan rápido es?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería



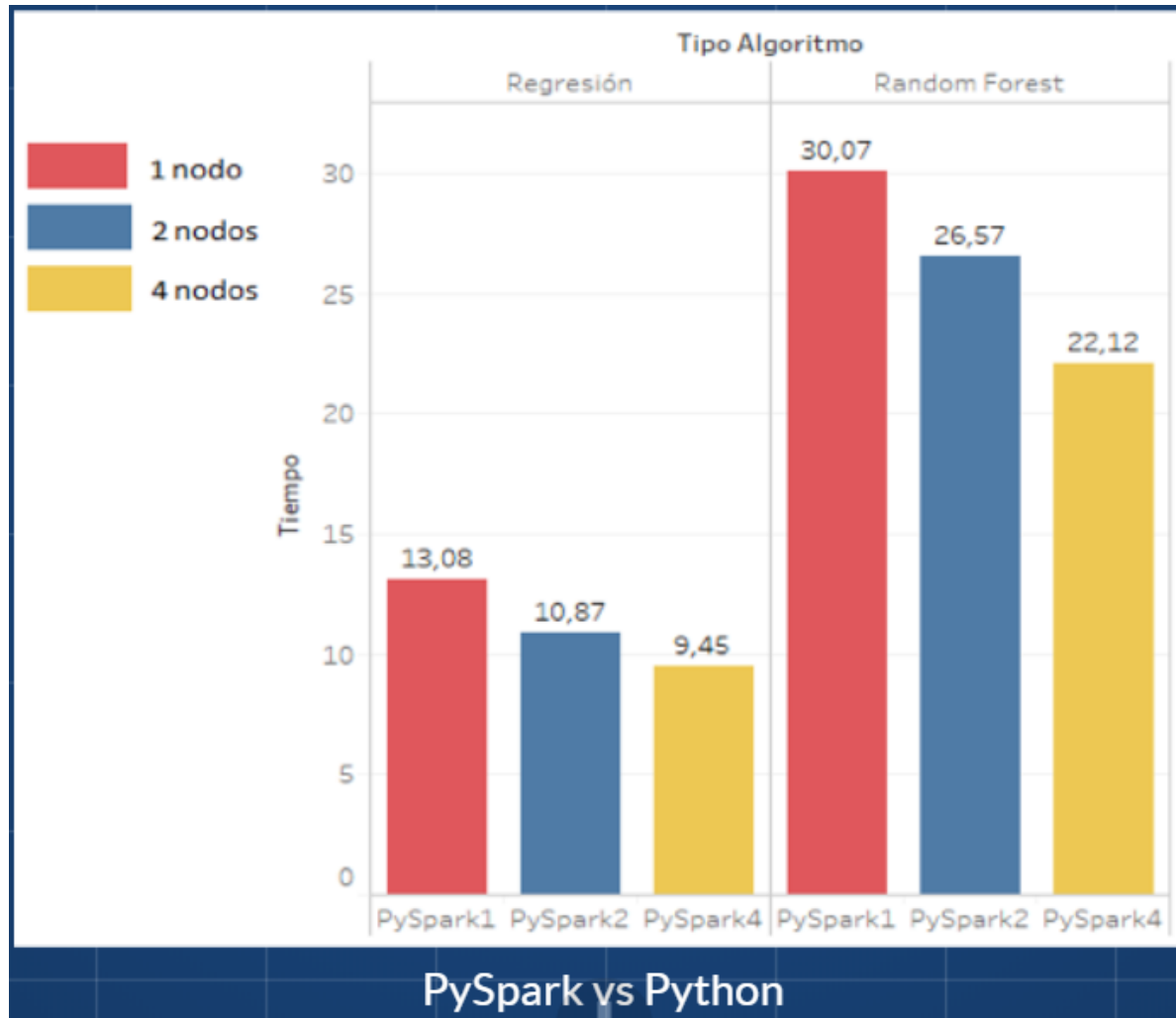
Regresión Logística en PySpark vs Python

¿Qué tan rápido es?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería



Mitos en Spark



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

¿Cómo está relacionado Apache Spark con Hadoop?

Spark puede correr en un cluster de Hadoop y gestionarlo por YARN. Puede procesar data en HDFS o Hive, pero también puede correr sola (standalone).

¿Qué tan grande puede ser un cluster?

El cluster más grande que se tiene registro tiene 8000 nodos. En términos de data ha sido usado para ordenar 100TB de datos 3X más rápido que Hadoop.

Mitos en Spark



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

¿Toda mi data tiene que ocupar exactamente la RAM?

No, Spark manda a disco la data que no alcance en RAM, lo que permite trabajar con datasets enormes.

¿Necesito una versión especial de Scala o Python?

No, Spark no necesita ningun componente para Scala o Python.

¿Por qué Scala?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Why is Apache Spark implemented in Scala?

 Answer

Request ▼

Follow 85 Comment Downvote



6 Answers





Matei Zaharia, CTO @ Databricks

Answered Dec 2, 2014 · Upvoted by Ashesh Ambasta, Lead backend engineer at CentralApp writing Scala services



When we started Spark, we wanted it to have a concise API for users, which Scala did well. At the same time, we wanted it to be fast (to work on large datasets), so many scripting languages didn't fit the bill. Scala can be quite fast because it's statically typed and it compiles in a known way to the JVM. Finally, running on the JVM also let us call into other Java-based big data systems, such as Cassandra, HDFS and HBase.

Since we started, we've also added APIs in Java (which [became much nicer with Java 8](#) ) and [Python](#) .

¿Por qué Scala?



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Why is Apache Spark implemented in Scala?

 Answer

Request ▼

Follow 85 Comment Downvote



6 Answers





Matei Zaharia, CTO @ Databricks

Answered Dec 2, 2014 · Upvoted by Ashesh Ambasta, [Lead backend engineer at CentralApp writing Scala services](#)



When we started Spark, we wanted it to have a concise API for users, which Scala did well. At the same time, we wanted it to be fast (to work on large datasets), so many scripting languages didn't fit the bill. Scala can be quite fast because it's statically typed and it compiles in a known way to the JVM. Finally, running on the JVM also let us call into other Java-based big data systems, such as Cassandra, HDFS and HBase.

Since we started, we've also added APIs in Java (which [became much nicer with Java 8](#) ) and [Python](#) .

Ecosistema Spark



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Spark SQL +
DataFrames

Streaming

MLlib
Machine Learning

GraphX
*Graph
Computation*

Spark Core API

R

SQL

Python

Scala

Java

Spark Core



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Es la base de todo proyecto, es el motor de toda la plataforma para todos los demás componentes. Provee capacidad computacional en memoria.

Soporta una variedad de lenguajes como Scala, Java, Python y R.

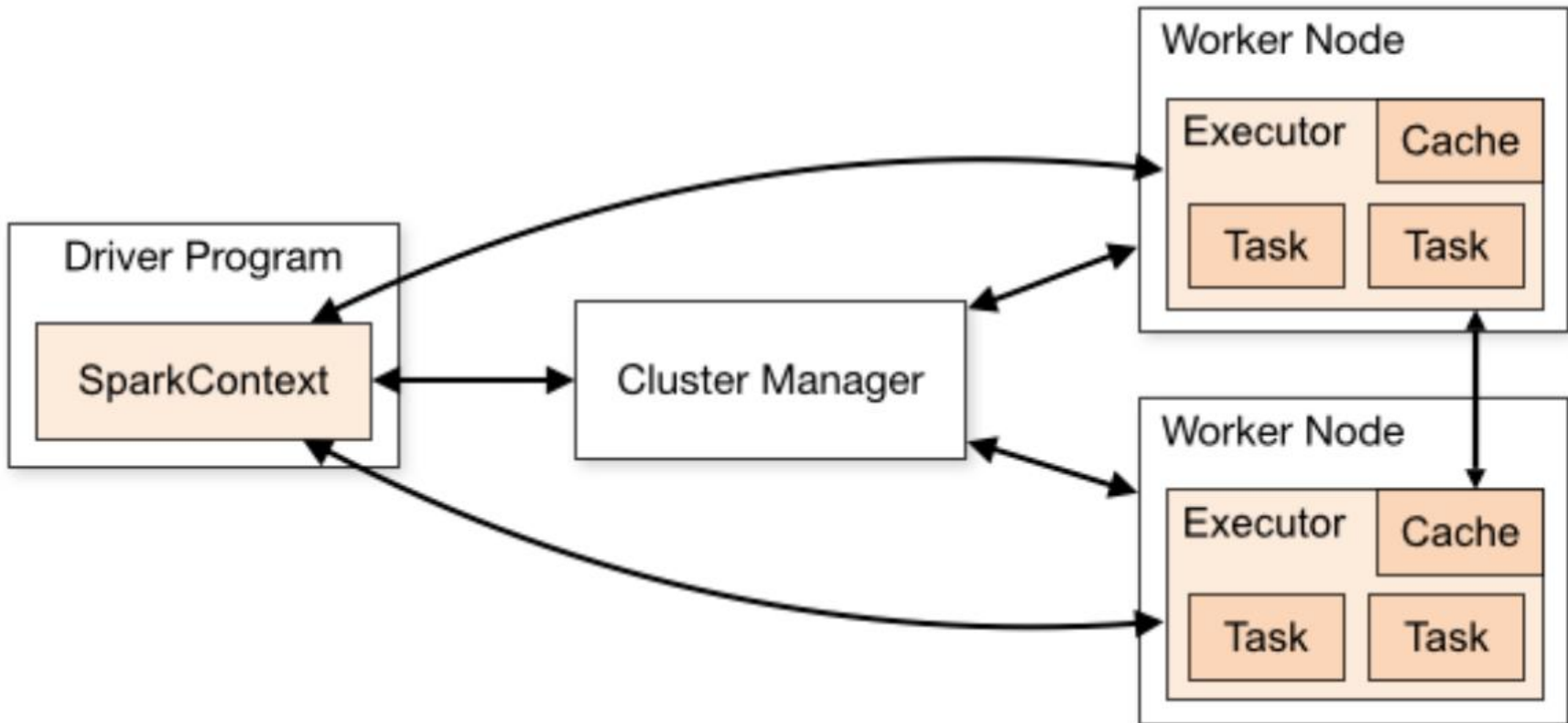


Arquitectura



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería



Spark SQL



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Es el módulo para trabajar con datos estructurados, usando SQL.

```
context = HiveContext(sc)
results = context.sql(
  "SELECT * FROM people")
names = results.map(lambda p:
  p.name)
```

Aplicar Funciones

```
context.jsonFile("s3n://...")
  .registerTempTable("json")
results = context.sql(
  """SELECT *
  FROM people
  JOIN json ...""")
```

Leer distintas fuentes

Spark Streaming



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Hace sencillo construir aplicaciones en Streaming

```
TwitterUtils.createStream(...)  
    .filter(_.getText.contains("Spark"))  
    .countByWindow(Seconds(5))  
)
```

Contar tweets en un periodo de tiempo

```
stream.join(historicCounts).filter {  
    case (word, (curCount, oldCount)) =>  
        curCount > oldCount  
}
```

Encontrar palabras con una mayor frecuencia que la data histórica

Spark MLlib



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Es la librería que hace escalable aplicar Machine Learning con Spark

```
data = spark.read.format("libsvm")\
    .load("hdfs://...")
model = KMeans(k=10).fit(data)
```

Kmeans PySpark

Ramdon Forest, Regresión Logística, Naive Bayes, árboles de decisión

Spark Graph X



CTIC UNI

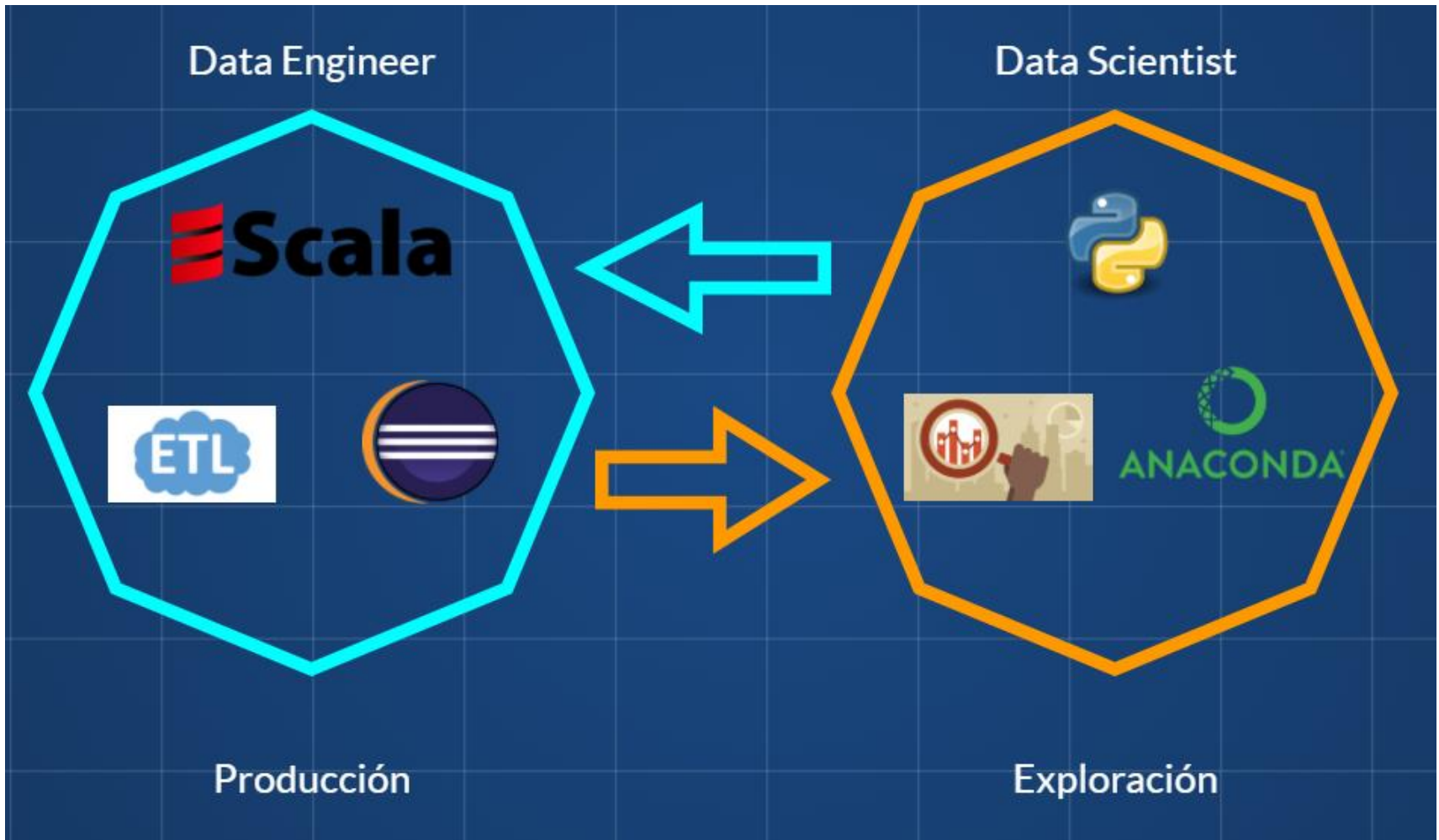
Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Es el API para grafos y computación paralela en grafos

```
graph = Graph(vertices, edges)
messages = spark.textFile("hdfs://...")
graph2 = graph.joinVertices(messages) {
  (id, vertex, msg) => ...
}
```

GraphX en Scala

¿Cómo se trabaja con Spark?





CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Preguntas



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Learning Scala

Bibliografía



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

<http://udacity.com/>

<https://databricks.com/>

<https://cognitiveclass.ai/>



CTIC UNI

Centro de Tecnologías de Información y Comunicaciones
Universidad Nacional de Ingeniería

Tecnologías para el Big Data Apache Spark