



Universidad **Ricardo Palma**

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

BIG DATA APLICADO

SESIÓN 6

Expositores:

David Narváez

Eder Pineda

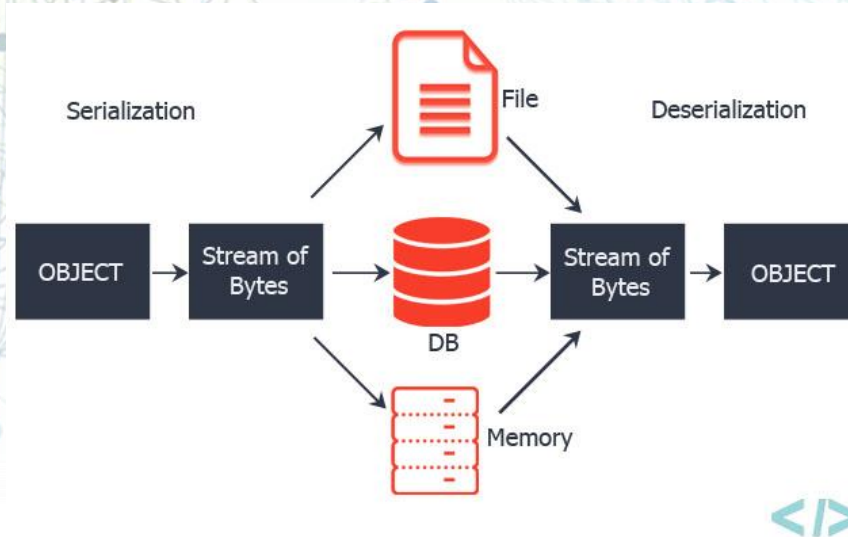
bigdataplicado@gmail.com

AGENDA

- INTRODUCCIÓN:
 - SERIALIZACIÓN
 - TIPOS
 - COMPARACIÓN
 - DISEÑO DE LA ESTRUCTURA DEL HDFS
 - COMANDOS BÁSICO HDFS PARA USUARIOS
 - INGESTA
 - EXPLORANDO HDFS
 - PUT (HDFS COMMAND)
 - SQUOP
 - SPARK

SERIALIZACIÓN

Serialización se refiere al proceso de convertir las estructuras de datos en flujos de bytes ya sea para almacenamiento o transmisión a través de una red.



Por el contrario, deserialización es el proceso de convertir una secuencia de bytes nuevamente en estructuras de datos.

SERIALIZACIÓN

La serialización es fundamental para un sistema de procesamiento distribuido como Hadoop, ya que permite que los datos se conviertan en un formato que pueda ser almacenado eficientemente, como también transferido a través de las conexiones de red.

**Serialización
(Procesamiento de Datos)**

Comunicación entre procesos (RPC)

Almacenamiento de datos

TIPOS DE SERIALIZACIÓN

WRITABLES

- Fue el principal formato de serialización utilizado por Hadoop es **Writables**
- Serialización Nativa de Hadoop
- Compactos y rápidos
- No son fáciles de usar por lenguajes distintos a Java.

Avro

- Especialmente creado para abordar limitaciones de Hadoop Writables
- Provee soporte nativo para MapReduce
- Escrito en JSON y Avro IDL
- Son formatos de almacenamiento en secuencia con un esquema definido

TIPOS DE SERIALIZACIÓN

PARQUET

- *Formato de almacenamiento para hadoop de propósito general*
- *Adecuado para diferentes interfaces de MapReduce y para otros motores como Impala y Spark*
- *Compresión eficiente*
- *Son formatos de almacenamiento columnar con un esquema definido*

ORC

- *Peso liviano*
- *Formato de almacenamiento divisible*
- *Admite Modelo de tipo Hive*
- *Comunmente utilizado en la Distribución de Hadoop de Hortonworks*

TIPOS DE SERIALIZACIÓN

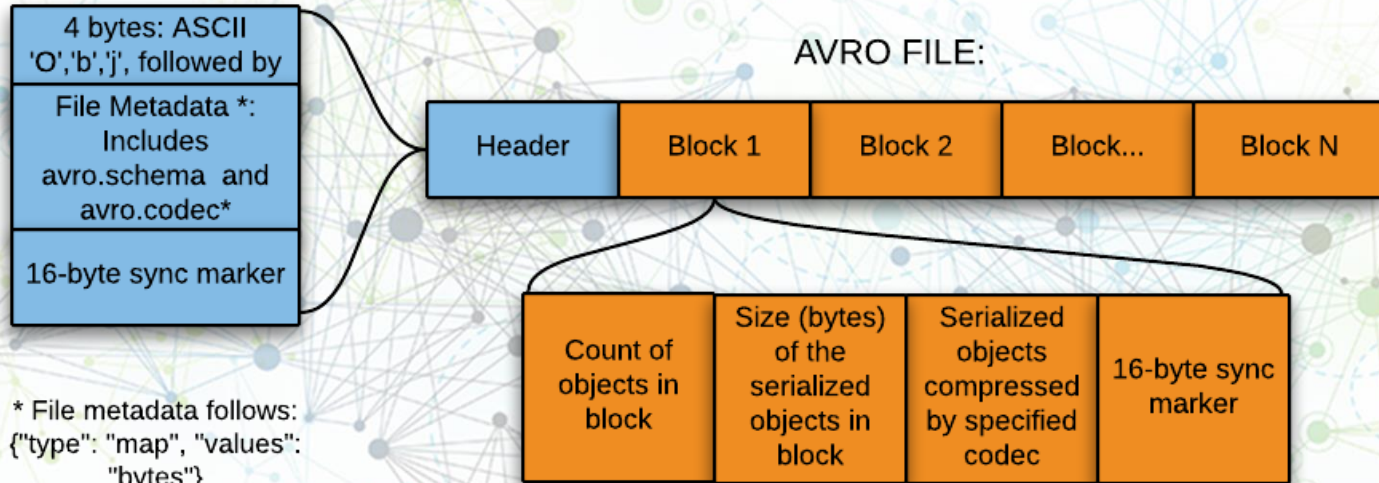
THRIFT

- *Creado por Facebook*
- *Usado como framework para implementar interfaces cross-language*
- *No provee soporte nativo para Hadoop*

Protocol Buffers

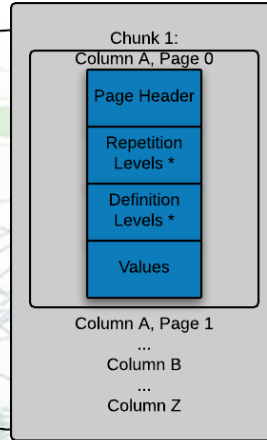
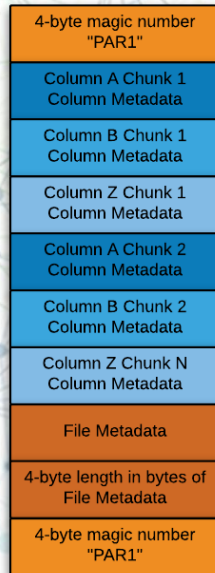
- *Creados por Google*
- *Creados para realizar transferencia de datos entre servicios escritos en diferentes lenguajes*
- *No provee soporte nativo para Hadoop*

TIPOS DE SERIALIZACIÓN

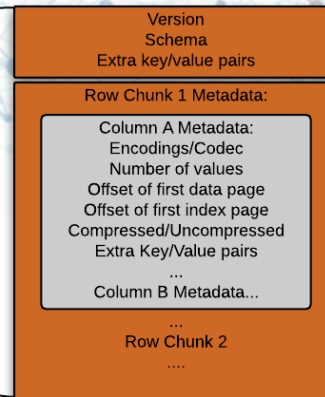


TIPOS DE SERIALIZACIÓN

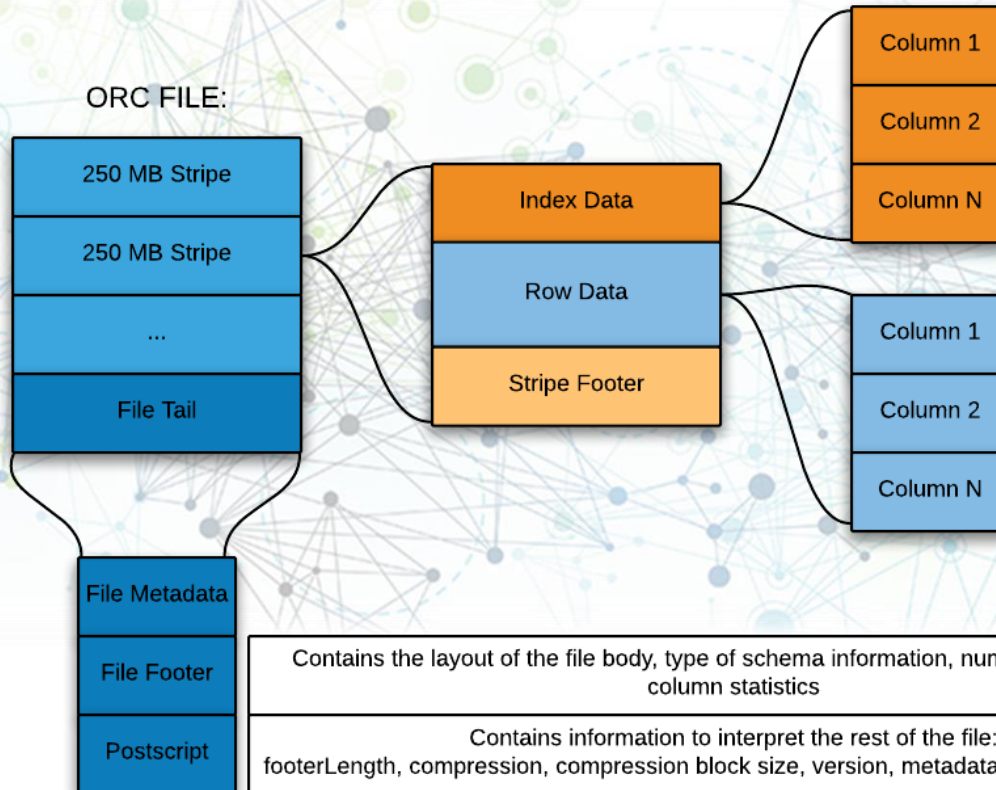
PARQUET FILE:



* Repetition levels and
definition levels allow for
nested schema structures



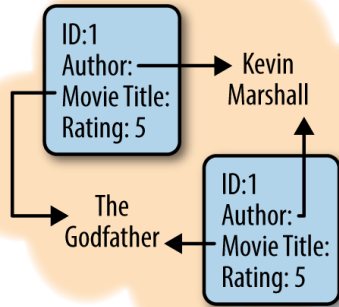
TIPOS DE SERIALIZACIÓN



SERIALIZACIÓN - COMPARACIÓN

<https://svds.com/dataformats/>

Computer Memory



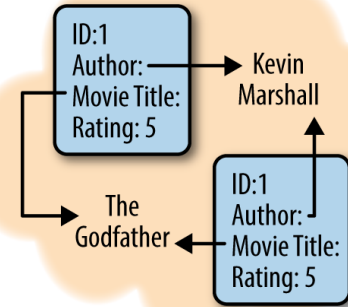
Serialization

Text File

1, Kevin, The Godfather, 5
2, Marshall, The Godfather, 3

De-Serialization

Computer Memory



serialized



Any questions?

deserialized



DISEÑO DE LA ESTRUCTURA DEL HDFS

/User/<username>

Data, aplicaciones y archivos de configuración que pertenecen solo a un usuario en específico. No es parte de un proceso de negocio

/etl

Data en varias etapas o capas que ha sido procesada por un flujo ETL. Este directorio debe tener permisos de lectura y escritura para los procesos ETL.

/tmp

Data temporal generada por herramientas o compartida entre usuarios

/data

Set de datos que han sido ingestados y/o procesados y son compartidos para toda la organización . Este directorio es de acceso de solo lectura por los usuarios. Almacena la RAW data

/app

Este directorio almacena todas las aplicaciones y artefactos como JAR files, Hive HQL files, Python files, etc.



DISEÑO DE LA ESTRUCTURA DEL HDFS

HOW DO DATA LAKES WORK?

The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.

STRUCTURED DATA

1. Information in rows and columns
2. Easily ordered and processed with data mining tools

1

The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.

2

The reservoir of water is a dataset, where you run analytics on all the data.

3

The outflow of water is the analyzed data.

4

Through this process, you are able to "sift" through all the data quickly to gain key business insights.

UNSTRUCTURED DATA

1. Raw, unorganized data
2. Emails
3. PDF files
4. Images, video and audio
5. Social media tools



COMANDOS BÁSICOS HDFS PARA USUARIOS

Comando *dfs*:

Ejecuta un comando de tipo filesystem en el sistema de archivos soportado por Hadoop.

El siguiente link muestra los **COMMAND_OPTIONS** de *hadoop dfs*. → [File System Shell Guide](#).

- appendToFile
- cat
- checksum
- chgrp
- chmod
- chown
- copyFromLocal
- copyToLocal
- count
- cp
- createSnapshot
- deleteSnapshot
- df
- du
- dus

- expunge
- find
- get
- getfacl
- getfattr
- getmerge
- help
- ls
- lsr
- mkdir
- moveFromLocal
- moveToLocal
- mv
- put
- renameSnapshot

- rm
- rmdir
- rmr
- setfacl
- setfattr
- setrep
- stat
- tail
- test
- text
- touchz
- truncate
- usage



COMANDOS BÁSICOS HDFS PARA USUARIOS

Comando ls:

Para un archivo devuelve el Stat; mientras que para un directorio devuelve la lista de su contenido.



```
[root@ip-10-0-0-74 /]# hdfs dfs -ls /
Found 5 items
drwxr-xr-x   - root   supergroup          0 2018-04-16 11:38 /data
drwxr-xr-x   - hbase  hbase              0 2018-04-17 20:57 /hbase
drwxrwxr-x   - solr   solr               0 2018-03-11 16:46 /solr
drwxrwxr-t   - root   root              0 2018-04-11 00:39 /tmp
drwxr-xr-x   - hdfs   supergroup         0 2018-03-18 14:35 /user
[root@ip-10-0-0-74 /]# hdfs dfs -ls /user
Found 9 items
drwxr-xr-x   - admin  admin              0 2018-03-12 00:23 /user/admin
drwxrwxrwx   - mapred hadoop          0 2018-03-11 16:47 /user/history
drwxrwxr-t   - hive   hive              0 2018-03-11 16:47 /user/hive
drwxrwxr-x   - hue    hue               0 2018-03-11 16:48 /user/hue
drwxrwxr-x   - impala impala            0 2018-03-11 16:46 /user/impala
drwxrwxr-x   - oozie  oozie             0 2018-03-11 16:46 /user/oozie
drwxr-xr-x   - root   supergroup        0 2018-03-19 10:27 /user/root
drwxr-x--x   - spark  spark             0 2018-03-11 19:12 /user/spark
drwxrwxr-x   - sqoop2 sqoop             0 2018-03-18 22:08 /user/sqoop2
```


COMANDOS BÁSICOS HDFS PARA USUARIOS

Comando `ls -R`:

Versión recursiva de `ls`.



```
[ec2-user@ip-10-0-0-74 ~]$ hdfs dfs -ls -R /data
-rw-r--r--  3 root supergroup  42322948 2018-03-11 23:18 /data/Ventas_Piloto2.csv
-rw-r--r--  3 root supergroup   477907 2018-04-16 11:38 /data/clasespark
drwxr-xr-x  - root supergroup         0 2018-03-19 10:44 /data/ing_sqoop
-rw-r--r--  3 root supergroup         0 2018-03-19 10:44 /data/ing_sqoop/_SUCCESS
-rw-r--r--  3 root supergroup    326 2018-03-19 10:44 /data/ing_sqoop/part-m-00000
```

COMANDOS BÁSICOS HDFS PARA USUARIOS

Comando mkdir:

Crea directorios en hdfs.



```
[root@ip-10-0-0-74 /]# hdfs dfs -mkdir /data/ingesta
[root@ip-10-0-0-74 /]# hdfs dfs -ls /data/
Found 4 items
-rw-r--r--   3 root supergroup   42322948 2018-03-11 23:18 /data/Ventas_Piloto2.csv
-rwxrwxr-x   3 root supergroup    477907 2018-04-16 11:38 /data/clasespark
drwxr-xr-x   - root supergroup      0 2018-03-19 10:44 /data/ing_sqoop
drwxr-xr-x   - root supergroup      0 2018-04-18 01:01 /data/ingesta
[root@ip-10-0-0-74 /]# █
```

INGESTA - EXPLORANDO HDFS



INGESTA - EXPLORANDO HDFS

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities ▾

Browse the file system

Logs

Overview 'ip-10-0-0-74.ec2.internal:8020' (active)

Started:	Tue Apr 17 19:56:10 -0500 2018
Version:	2.6.0-cdh5.14.0, r9b197d35839383c798c618ba917ccaa196a17699
Compiled:	Sat Jan 06 16:38:00 -0500 2018 by jenkins from Unknown
Cluster ID:	cluster4
Block Pool ID:	BP-958870086-10.0.0.74-1520801153671

INGESTA - EXPLORANDO HDFS

Hadoop Overview Datanodes Snapshot Startup Progress Utilities ▾

Browse Directory

/data/ingesta

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	7.87 KB	Wed Apr 18 00:22:55 -0500 2018	3	128 MB	Untitled.ipynb
-rw-r--r--	root	supergroup	71.88 KB	Wed Apr 18 00:16:55 -0500 2018	3	128 MB	venta_diaria.java

INGESTA - EXPLORANDO HDFS

The screenshot displays the Hadoop web interface with a modal window open for file information. The background shows the 'Browse Directory' page for '/data/ingesta' with a table of files and a 'Go!' button. The modal window, titled 'File information - venta_diaria.java', contains a 'Download' link and a 'Block information' dropdown menu. The dropdown is open, showing 'Block 0' selected. Below the dropdown, the following details are listed: Block ID: 1073756150, Block Pool ID: BP-958870086-10.0.0.74-1520801153671, Generation Stamp: 15402, Size: 73602, and Availability: a list of three IP addresses. A 'Close' button is at the bottom right of the modal.

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse Directory

/data/ingesta

Permission	Owner	Group
-rw-r--r--	root	supergro
-rw-r--r--	root	supergro

Hadoop, 2017.

File information - venta_diaria.java

Download

Block information — Block 0 ▼

Block 0

Block ID: 1073756150
Block Pool ID: BP-958870086-10.0.0.74-1520801153671
Generation Stamp: 15402
Size: 73602
Availability:

- ip-10-0-0-74.ec2.internal
- ip-10-0-0-212.ec2.internal
- ip-10-0-0-43.ec2.internal

Close

Block Size	Name
128 MB	Untitled.ipynb
128 MB	venta_diaria.java

Go!


INGESTA - EXPLORANDO HDFS

The image displays the Hue web interface for exploring HDFS. The top navigation bar includes the Hue logo, a 'Consulta' dropdown menu, and a search bar labeled 'Search data and saved documents...'. The left sidebar shows a file tree with 'HDFS' selected, and a red box highlights the 'HDFS' icon. The main panel shows the 'Impala' query editor with a sample query: 'Ejemplo: SELECT * FROM nombre de tabla o pulse CTRL + espacio'. The bottom panel shows a detailed view of the 'data' directory, listing files like 'Ventas_Piloto2.csv', 'clasespark', 'ing_sqoop', and 'ingesta'.

Hue Interface Components:

- Top Bar:** Hue logo, 'Consulta' dropdown, search bar.
- Left Sidebar:** File tree showing 'HDFS' selected. A red box highlights the 'HDFS' icon.
- Main Panel:** 'Impala' query editor. Sample query: 'Ejemplo: SELECT * FROM nombre de tabla o pulse CTRL + espacio'.
- Bottom Panel:** Detailed view of the 'data' directory. Files listed: 'Ventas_Piloto2.csv', 'clasespark', 'ing_sqoop', 'ingesta'.

INGESTA - EXPLORANDO HDFS



Consulta

Search data and saved documents...

Jobs

admin

Ingesta

Filtrar...

Untitled.ipynb

venta_diaria.java

Ver como binario

Editar archivo

Descargar

Ver ubicación de archivo

Actualizar

Última modificación
18/04/2018 5:16

Usuario
root

Inicio

/ data / ingesta / **venta_diaria.java**

```
// ORM class for table 'venta_diaria'  
// WARNING: This class is AUTO-GENERATED. Modify at your own risk.  
//  
// Debug information:  
// Generated date: Mon Mar 19 10:27:43 EDT 2018  
// For connector: org.apache.sqoop.manager.MySQLManager  
import org.apache.hadoop.io.BytesWritable;  
import org.apache.hadoop.io.Text;  
import org.apache.hadoop.io.Writable;  
import org.apache.hadoop.mapred.lib.db.DBWritable;  
import com.cloudera.sqoop.lib.JdbcWritableBridge;  
import com.cloudera.sqoop.lib.DelimiterSet;  
import com.cloudera.sqoop.lib.FieldFormatter;  
import com.cloudera.sqoop.lib.RecordParser;
```


INGESTA - EXPLORANDO HDFS

Hue Consulta Search data and saved documents... Jobs admin

Explorador de archivos

ingesta

Filtrar...

- Untitled.ipynb
- venta_diaria.java

venta_diaria.java

Tamaño	71,9 KB
Usuario	root
Grupo	supergroup
Permisos	-rw-r--r--
Fecha	April 17, 2018 10:16 PM

Página 1 a 18 de 18

ingesta / venta_diaria.java

```
// Generated date: Mon Mar 19 10:27:43 EDT 2018
// For connector: org.apache.sqoop.manager.MySQLManager
import org.apache.hadoop.io.BytesWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.Writable;
import org.apache.hadoop.mapred.lib.db.DBWritable;
import com.cloudera.sqoop.lib.JdbcWritableBridge;
import com.cloudera.sqoop.lib.DelimiterSet;
import com.cloudera.sqoop.lib.FieldFormatter;
import com.cloudera.sqoop.lib.RecordParser;
```

INGESTA – PUT (HDFS COMMANDS)

Comando put:

Copia una o multiples fuentes del filesystem local a un file system destino.



- p : Conserva fecha de acceso y modificación, propietario y permisos.*
- f : Sobreescribe el archivo si ya existe.*
- l : Permite al DataNode persistir el archivo en disco, fuerza un factor de replicación en 1. **Usar con cuidado.***
- d : Salta la creación de archivo temporal con el sufijo **._COPYING_**.*

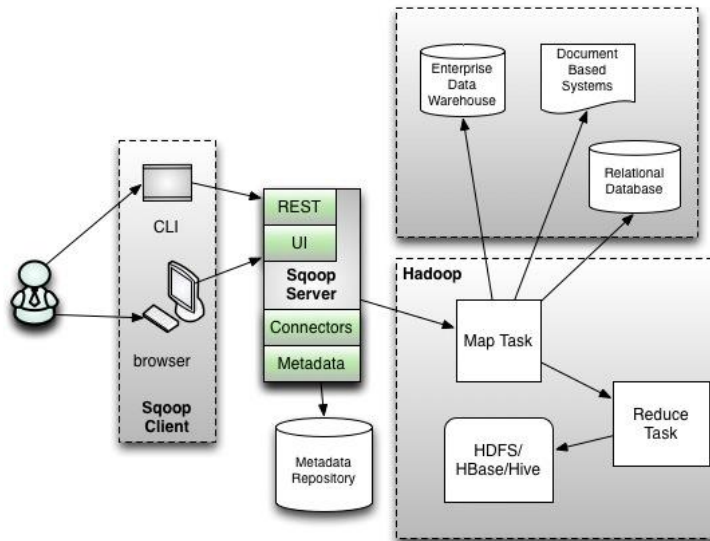
```
[root@ip-10-0-0-74 /]# hdfs dfs -put venta_diaria.java /data/ingesta
[root@ip-10-0-0-74 /]# hdfs dfs -ls /data/ingesta
Found 1 items
-rw-r--r--   3 root supergroup      73602 2018-04-18 01:16 /data/ingesta/venta_d
iaria.java
[root@ip-10-0-0-74 /]#
```

```
[root@ip-10-0-0-74 /]# hdfs dfs -put //34.200.224.183/usr/jupyter_site/Untitled.ipynb /data/ingesta
[root@ip-10-0-0-74 /]# hdfs dfs -ls /data/ingesta
Found 2 items
-rw-r--r--   3 root supergroup      8059 2018-04-18 01:22 /data/ingesta/Untitled.ipynb
-rw-r--r--   3 root supergroup      73602 2018-04-18 01:16 /data/ingesta/venta_diaria.java
[root@ip-10-0-0-74 /]#
```



INGESTA - SQOOP

Apache Sqoop(TM) es una herramienta diseñada para realizar una eficiente transferencia de gran volumen de data entre Apache Hadoop y almacenes de datos estructurados como las bases de datos relacionales.





INGESTA - SQOOP

Consultar Bases de Datos

```
# sqoop list-databases --connect  
jdbc:mysql://dbdata01.ccpetoqzkfsy.us-east-  
1.rds.amazonaws.com:3306 --username admin --password  
admin123
```

Consultar Tablas

```
# sqoop list-tables --connect  
jdbc:mysql://dbdata01.ccpetoqzkfsy.us-east-  
1.rds.amazonaws.com:3306/DBRIO --username admin --  
password admin123
```





INGESTA - SQOOP

Importar una tabla

```
# sqoop import --connect jdbc:mysql://dbdata01.ccpetoqzkfsy.us-east-1.rds.amazonaws.com:3306/DBRIO --username admin --password admin123 --table venta_diaria --target-dir /data/ingesta/sqoop/venta_diaria1 --m 1
```

Importar una tabla en formato Avro

```
# sqoop import --connect jdbc:mysql://dbdata01.ccpetoqzkfsy.us-east-1.rds.amazonaws.com:3306/DBRIO --username admin --password admin123 --table venta_diaria --target-dir /data/ingesta/sqoop/venta_diaria2 --m 1 --as-avrodatafile
```





INGESTA - SQOOP

Importar una tabla en formato Parquet

```
# sqoop import --connect jdbc:mysql://dbdata01.ccpetoqzkfsy.us-east-1.rds.amazonaws.com:3306/DBRIO --username admin --password admin123 --table venta_diaria --target-dir /data/ingesta/sqoop/venta_diaria4 --m 1 --as-parquetfile
```

Importar un query

```
# sqoop import --connect jdbc:mysql://dbdata01.ccpetoqzkfsy.us-east-1.rds.amazonaws.com:3306/DBRIO --username admin --password admin123 --query 'SELECT periodo, margen from venta_diaria where $CONDITIONS' --target-dir /data/ingesta/sqoop/venta_diaria5 --m 1 --as-parquetfile
```



INGESTA - SQOOP



Common arguments

Import control arguments

Validation <https://sqoop.apache.org/docs/1.4.6/SqoopUserGuide.html>

Incremental import arguments

Input parsing arguments

Parameters for overriding mapping





INGESTA - SGOOP

Crear un Shell Script

```
# touch ingesta_sgoop.sh
```

Editar el Shell Script con el código Sgoop

```
# vim ingesta_sgoop.sh
```

Presionar esc

Pegar el código Sgoop

Presionar esc

Digitir :w

Digitir :q!

Dar permisos a la shell

```
# chmod 777 ingesta_sgoop.sh
```

Ejecutar la Shell

```
# ./ingesta_sgoop.sh
```





INGESTA - SQOOP

Feature	Sqoop 1	Sqoop 2
Connectors for all major RDBMS	Supported.	Not supported. Workaround: Use the generic JDBC Connector which has been tested on the following databases: Microsoft SQL Server, PostgreSQL, MySQL and Oracle. This connector should work on any other JDBC compliant database. However, performance might not be comparable to that of specialized connectors in Sqoop.
Kerberos Security Integration	Supported.	Supported.
Data transfer from RDBMS to Hive or HBase	Supported.	Not supported. 1.Workaround: Follow this two-step approach.Import data from RDBMS into HDFS 2.Load data into Hive or HBase manually using appropriate tools and commands such as the LOAD DATA statement in Hive
Data transfer from Hive or HBase to RDBMS	1.Not supported. Workaround: Follow this two-step approach.Extract data from Hive or HBase into HDFS (either as a text or Avro file) 2.Use Sqoop to export output of previous step to RDBMS	Not supported. Follow the same workaround as for Sqoop 1.





INGESTA - SQQOOP

Ejercicio 1

*Realizar la consulta de las tablas que existen en la base de datos
db_a05768_ederp, utilizando los siguientes datos:*

Server: MYSQL5009.site4now.net:3306

User: a05768_ederp

Password: eder2018

Ejercicio 2

Ingestar alguna de las tablas encontradas en el ejercicio 1.



INGESTA - SPARK

Ingesta de datos desde una DB:

Ingesta Spark DB

```
In [1]: import findspark
findspark.init()
##findspark.init("/home/ubuntu/spark-2.2.1-bin-hadoop2.7")
import pyspark
```

```
In [2]: from pyspark.sql import SparkSession
```

```
In [3]: spark = SparkSession.builder.appName('ingesta_spark').getOrCreate()
```

```
In [4]: url = "jdbc:mysql://MYSQL5009.site4now.net:3306/db_a05768_ederp"

connectionProperties = {
    'user' : 'a05768_ederp',
    'password' : 'eder2018'
}
```

```
In [5]: df_mysql = spark.read.jdbc(url=url, table='stg_formato_01_hist', properties=connectionProperties)
```

```
In [6]: df_mysql.count()
```

```
Out[6]: 5652
```

```
In [8]: df_mysql.write.parquet('/user/dnarvaez/stg_formato_01_hist_db/', mode="append")
```



INGESTA - SPARK

Ejecución de Aplicaciones Spark

*El script **spark-submit** se usa para iniciar aplicaciones en un clúster.*

```
[root@ip-10-0-0-74 app]# ls
codegen_venta_diaria.java  ingesta_spark_file.py  ingesta_sqoop_1.sh  ingesta_sqoop.sh  QueryResult.java
[root@ip-10-0-0-74 app]# spark2-submit ingesta_spark_file.py
```


A complex network graph with numerous nodes and connecting lines. The nodes are colored in shades of blue, green, and grey, and are of varying sizes. The lines are thin and connect the nodes in a dense, interconnected pattern. The background is a light, textured grey. The text "GRACIAS!" is centered in the middle of the image in a dark blue, serif font. There are blue diagonal bars in the top-left and bottom-right corners.

GRACIAS!