Curso Data Engineer: Creando un pipeline de datos
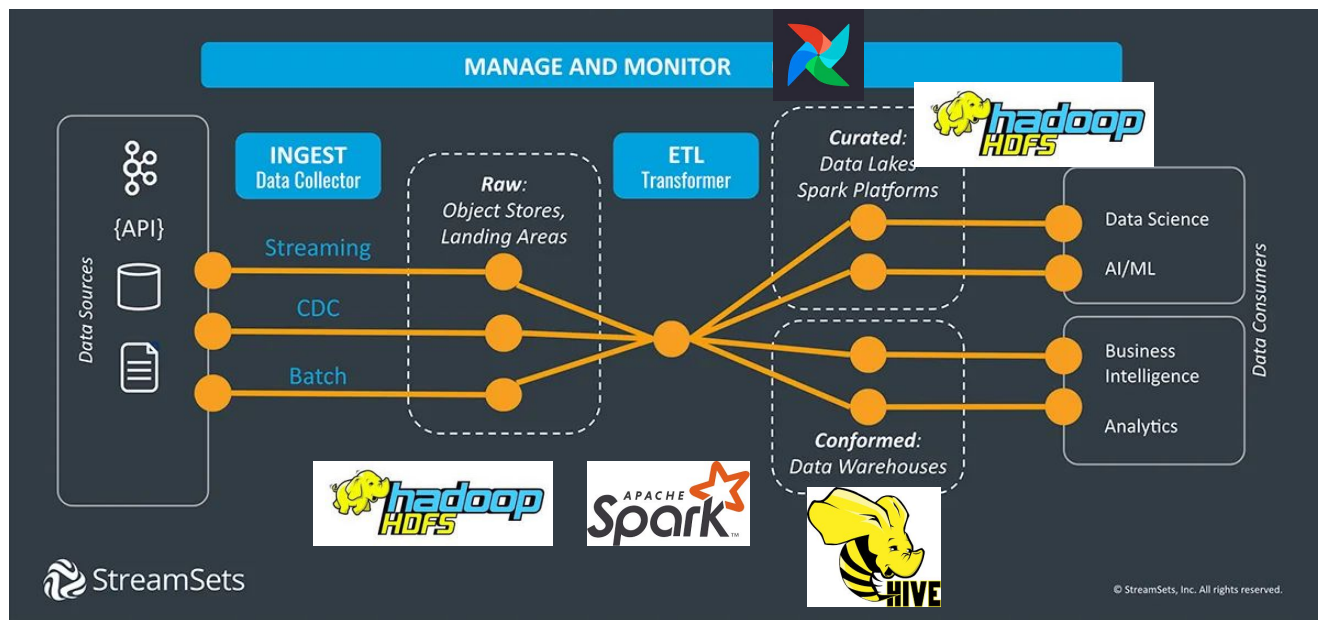
Ingest

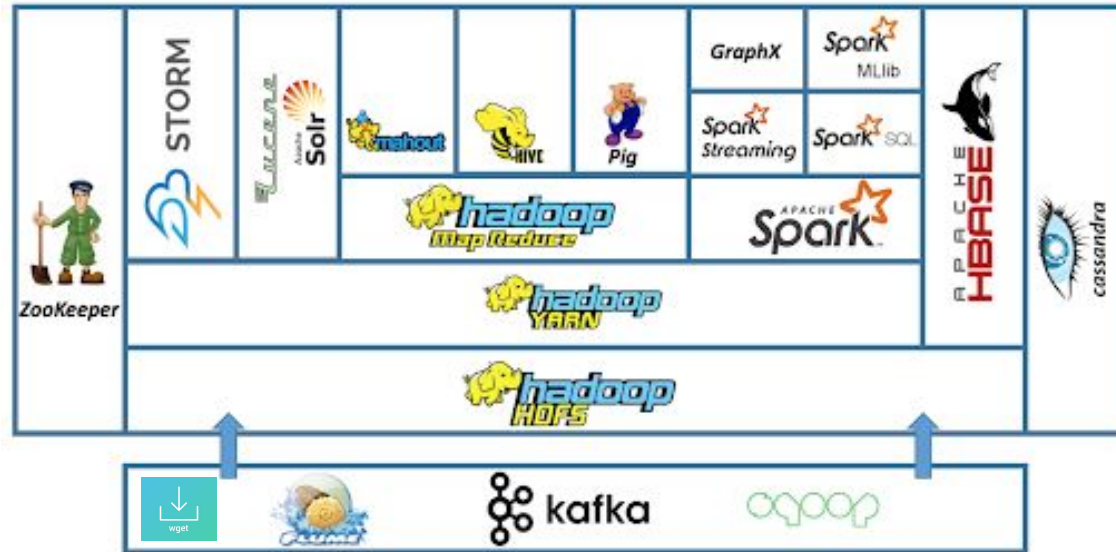# Arquitectura Big Data

Ecosistema Hadoop

Ingest mediante scripts

Ingest con WGET

Blob Storage

Ingest con WGET

Blob Storage

Ingest mediante scripts

Podemos utilizar algunos comandos de linux para hacer ingest de archivos.

Obtenemos los archivos con WGET:

- **wget -P /home/hadoop/landing**

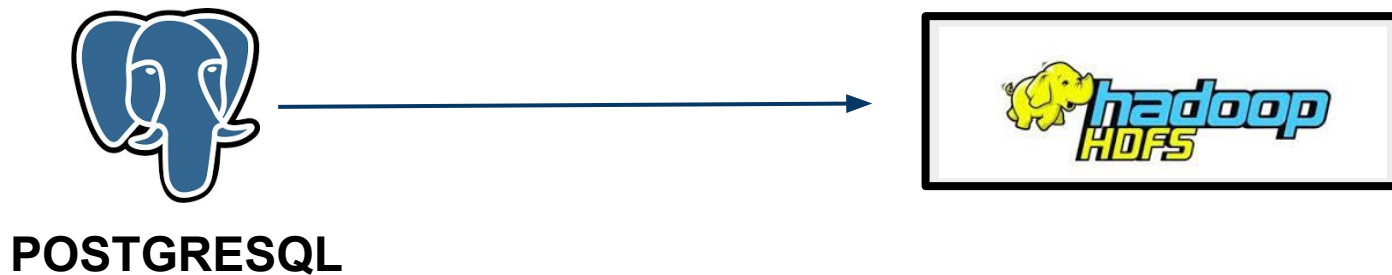 https://dataengineerpublic.blob.core.windows.net/data-engineer/yellow_tripdata_2021-01.csv

Movemos los archivos a HDFS:

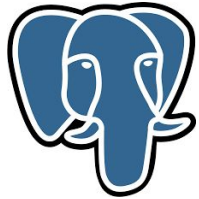- **hdfs dfs -put /home/hadoop/landing/yellow_tripdata_2021-01.csv /ingest**

Ingest
mediante sqoop

Ingest con SQOOP

POSTGRESQL

Ingest con SQOOP

POSTGRESQL

Ingest mediante sqoop

Verificar funcionamiento y versión:

- **sqoop-version**



```
hadoop@5dc251dd43fb:~$ sqoop-version
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2023-03-16 19:02:28,767 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Sqoop 1.4.7
git commit id 2328971411f57f0cb683dfb79d19d4d19d185dd8
Compiled by maugli on Thu Dec 21 15:59:58 STD 2017
hadoop@5dc251dd43fb:~$
```

Ingest mediante sqoop

Listar databases:

**sqoop list-databases **
**—connect jdbc:postgresql://172.17.0.3:5432/northwind **
—username postgres -P
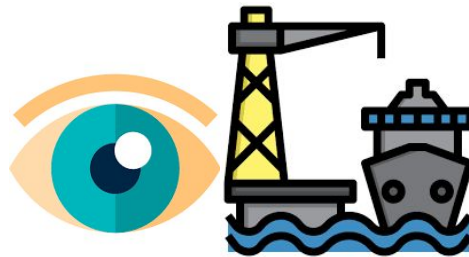

```
2023-03-16 20:36:39,489 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
2023-03-16 20:36:42,458 INFO manager.SqlManager: Using default fetchSize of 1000
postgres
northwind
template1
template0
hadoop@5dc251dd43fb:/$
```

Ingest mediante sqoop

Importar tablas:

```
sqoop import \
    --connect jdbc:postgresql://172.17.0.3:5432/northwind \
    --username postgres\
    --table region\
    --m 1 \
    --P \
    --target-dir /sqoop/ingest \
    --as-parquetfile \
    --delete-target-dir
```
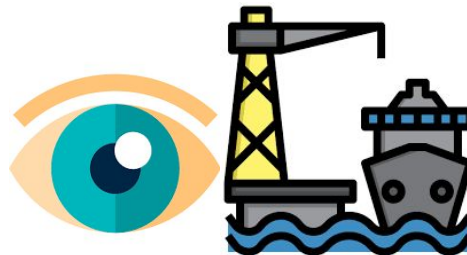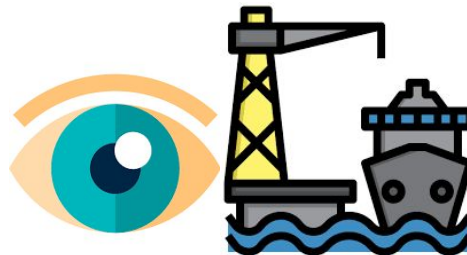


```
        Total time spent by all maps in occupied slots (ms)=8675
        Total time spent by all reduces in occupied slots (ms)=0
        Total time spent by all map tasks (ms)=8675
        Total vcore-milliseconds taken by all map tasks=8675
        Total megabyte-milliseconds taken by all map tasks=13324800
    Map-Reduce Framework
        Map input records=4
        Map output records=4
        Input split bytes=87
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=66
        CPU time spent (ms)=4570
        Physical memory (bytes) snapshot=275968000
        Virtual memory (bytes) snapshot=2981875712
        Total committed heap usage (bytes)=180355072
        Peak Map Physical memory (bytes)=275968000
        Peak Map Virtual memory (bytes)=2981875712
    File Input Format Counters
        Bytes Read=0
    File Output Format Counters
        Bytes Written=0
2023-03-16 20:06:32,380 INFO mapreduce.ImportJobBase: Transferred 1.8496 KB in 38.8773 seconds (48.7174 bytes/sec)
2023-03-16 20:06:32,391 INFO mapreduce.ImportJobBase: Retrieved 4 records.
```

Ingest mediante sqoop

Importar tablas con filtro:

```
sqoop import \
    –connect jdbc:postgresql://172.17.0.3:5432/northwind \
    –username postgres\
    –table region\
    –m 1 \
    –P \
    –target-dir /sqoop/ingest/southern \
    –as-parquetfile \
    –where "region_description = 'Southern' " \
    –delete-target-dir
```



```
            HDFS: Number of large read operations=0
            HDFS: Number of write operations=10
            HDFS: Number of bytes read erasure-coded=0
    Job Counters
            Launched map tasks=1
            Other local map tasks=1
            Total time spent by all maps in occupied slots (ms)=8319
            Total time spent by all reduces in occupied slots (ms)=0
            Total time spent by all map tasks (ms)=8319
            Total vcore-milliseconds taken by all map tasks=8319
            Total megabyte-milliseconds taken by all map tasks=12777984
    Map-Reduce Framework
            Map input records=1
            Map output records=1
            Input split bytes=87
            Spilled Records=0
            Failed Shuffles=0
            Merged Map outputs=0
            GC time elapsed (ms)=74
            CPU time spent (ms)=4060
            Physical memory (bytes) snapshot=254566400
            Virtual memory (bytes) snapshot=2971721728
            Total committed heap usage (bytes)=181403648
            Peak Map Physical memory (bytes)=254566400
            Peak Map Virtual memory (bytes)=2971721728
    File Input Format Counters
            Bytes Read=0
    File Output Format Counters
            Bytes Written=0
2023-03-16 20:20:21,436 INFO mapreduce.ImportJobBase: Transferred 1.8115 KB in 30.653 seconds (60.5161 bytes/sec)
2023-03-16 20:20:21,447 INFO mapreduce.ImportJobBase: Retrieved 1 records.
```
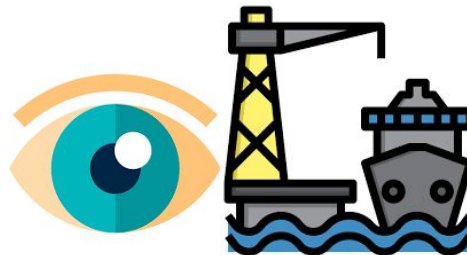
Ingest mediante sqoop

Importar tablas desde una query:

```
sqoop import \
    –connect jdbc:postgresql://172.17.0.3:5432/northwind \
    –username postgres\
    –query "select * from region where region_id = 3 AND \$CONDITIONS"\
    –m 1 \
    –P \
    –target-dir /sqoop/ingest \
    –as-parquetfile \
    –delete-target-dir
```



```
            HDFS: Number of large read operations=0
            HDFS: Number of write operations=10
            HDFS: Number of bytes read erasure-coded=0
    Job Counters
            Launched map tasks=1
            Other local map tasks=1
            Total time spent by all maps in occupied slots (ms)=8319
            Total time spent by all reduces in occupied slots (ms)=0
            Total time spent by all map tasks (ms)=8319
            Total vcore-milliseconds taken by all map tasks=8319
            Total megabyte-milliseconds taken by all map tasks=12777984
    Map-Reduce Framework
            Map input records=1
            Map output records=1
            Input split bytes=87
            Spilled Records=0
            Failed Shuffles=0
            Merged Map outputs=0
            GC time elapsed (ms)=74
            CPU time spent (ms)=4060
            Physical memory (bytes) snapshot=254566400
            Virtual memory (bytes) snapshot=2971721728
            Total committed heap usage (bytes)=181403648
            Peak Map Physical memory (bytes)=254566400
            Peak Map Virtual memory (bytes)=2971721728
    File Input Format Counters
            Bytes Read=0
    File Output Format Counters
            Bytes Written=0
2023-03-16 20:20:21,436 INFO mapreduce.ImportJobBase: Transferred 1.8115 KB in 30.653 seconds (60.5161 bytes/sec)
2023-03-16 20:20:21,447 INFO mapreduce.ImportJobBase: Retrieved 1 records.
```

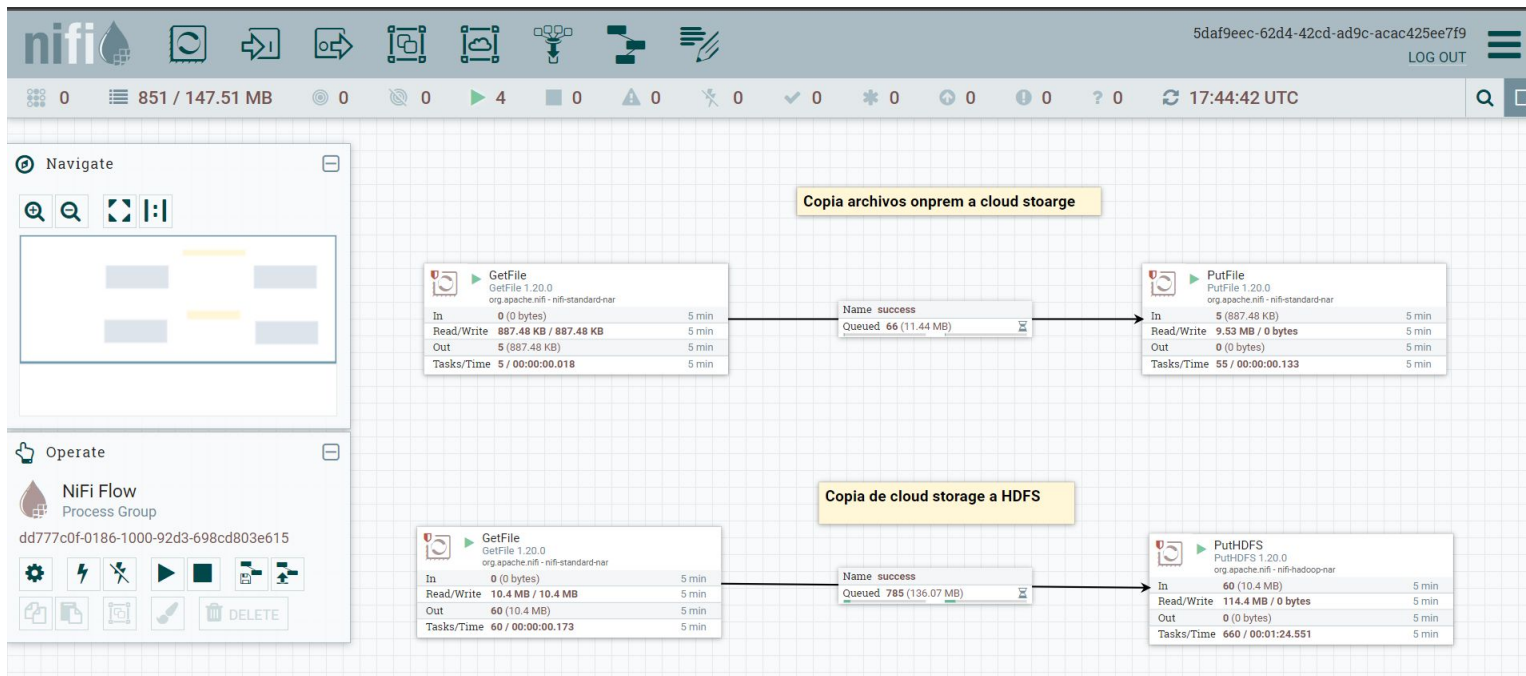# Ingest
mediante nifi

APACHE nifi

ON-PREMISE → CLOUD → hadoop HDFS

Ingest con APACHE nifi

APACHE nifi

APACHE nifi

APACHE nifi

GetFile

Processor Details

▶ Running   ⚙ STOP & CONFIGURE

SETTINGS	SCHEDULING	PROPERTIES	RELATIONSHIPS	COMMENTS

Required field

Property		Value
Input Directory	❓	/home/nifi/ingest
File Filter	❓	starwars.csv
Path Filter	❓	No value set
Batch Size	❓	10
Keep Source File	❓	true
Recurse Subdirectories	❓	true
Polling Interval	❓	0 sec
Ignore Hidden Files	❓	true
Minimum File Age	❓	0 sec
Maximum File Age	❓	No value set
Minimum File Size	❓	0 B
Maximum File Size	❓	No value set

APACHE nifi

PutFile

APACHE nifi

PutHDFS

APACHE nifi

- **Instalación**:
  - Instalado en la VM
  - instalar desde docker (docker pull apache/nifi)
- **Usr y contraseña:**
  - Usr: d30eb1a2-3bfe-4c85-9ea4-9562915a70e6
  - Pass: NvxFSKesWIiU1K4XL1AQJwovv9z7TW4h
  - /opt/nifi/nifi-current/bin nifi.sh set-single-user-credentials nifi <password>
  - En caso que lo instalen desde docker buscar el usr y pass en docker logs nifi
- **Archivos de configuración Hadoop:**
  - core-site.xml:
    https://github.com/fpineyro/homework-0/blob/2767f00cf9c16774dbb10fc2d7b8d17f11114750/core-site.xml
  - hdfs-site.xml:
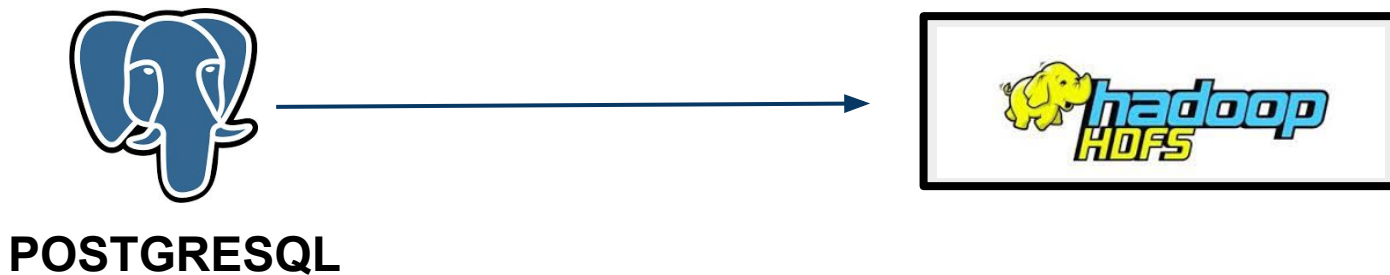   https://github.com/fpineyro/homework-0/blob/2767f00cf9c16774dbb10fc2d7b8d17f11114750/hdfs-site.xml

Ejercicio

Ejercicios

- Ingest

  - SQOOP

  - NIFI

Ingest con SQOOP

POSTGRESQL

Ingest con APACHE nifi

APACHE nifi