
ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA
DANIEL FABRICIO SALAZAR SANCHEZ
ROBERT ANDRES TORRES FLORES
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION
PROYECTO DE TESIS
JUNIO DEL 2015
CONSTRUCCIÓN DE UN REPOSITORIO POR MEDIO DE LA EXTRACCIÓN Y PROCESAMIENTO DE LA INFORMACIÓN PROVENIENTE DE CDRS GENERADOS POR TELCOS PARA ANALIZAR EL CONSUMO DE LOS DATOS OBTENIDOS, DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM
CONCLUSIONES
RECOMENDACIONES
COMPONENTES HADOOP
HADOOP
NoSQL
Big Data son Todos los Datos
CDR (CALL DETAIL RECORD)
Prezi

DANIEL FABRICIO SALAZAR SANCHEZ
ROBERT ANDRES TORRES FLORES.....

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION

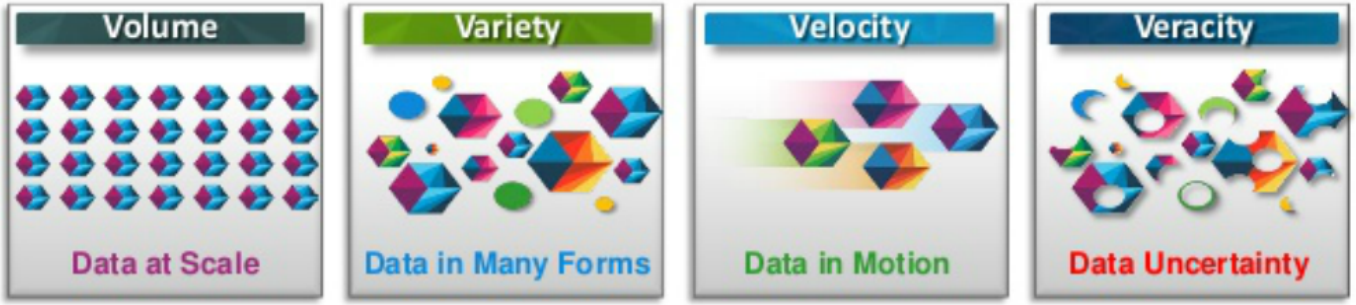
PROYECTO DE TESIS

JUNIO DEL 2015

CONSTRUCCIÓN DE UN REPOSITORIO POR MEDIO DE LA EXTRACCIÓN Y PROCESAMIENTO DE LA INFORMACIÓN PROVENIENTE DE CDRS GENERADOS POR TELCOS PARA ANALIZAR EL CONSUMO DE LOS DATOS OBTENIDOS, DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM


CONCLUSIONES
RECOMENDACIONES
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA
DANIEL FABRICIO SALAZAR SANCHEZ
ROBERT ANDRES TORRES FLORES
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION
PROYECTO DE TESIS
JUNIO DEL 2015
CONSTRUCCIÓN DE UN REPOSITORIO POR MEDIO DE LA EXTRACCIÓN Y PROCESAMIENTO DE LA INFORMACIÓN PROVENIENTE DE CDRS GENERADOS POR TELCOS PARA ANALIZAR EL CONSUMO DE LOS DATOS OBTENIDOS, DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM
COMPONENTES HADOOP
HADOOP
NOSQL
CDR (CALL DETAIL RECORD)
Prezi

BIG DATA SON TODOS LOS DATOS



- Transparencia
- Experimentación
- Segmentación
- Toma de decisiones
- Innovación

→ NUEVA FORMA de TOMAR DECISIONES

enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis

Tipos de Big Data

Web y Social Media
- Clickstream Data
- Twitter Feeds
- Facebook Postings
- Web Content

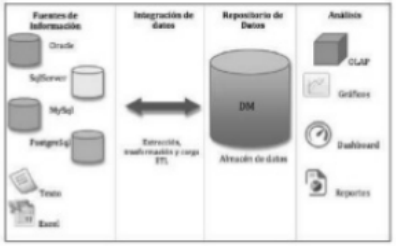
Machine To Machine
- Utility Smart Meter Readings
- RFID Readings
- Oil Rig Sensor Readings
- GPS Signals

Biometrics
- Facial Recognition
- Genetics

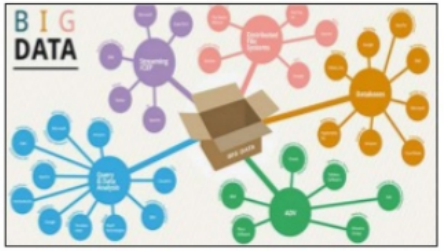
Big Transaction Data
- Healthcare Claims
- Telecommunications Call Detail Records
- Utility Billing Records

Human Generated
- Call Center Voice Recordings
- Email
- Electronic Medical Records

ARQUITECTURA - Diferencias



Vs.



Arq. tradicional

- Centralidad → mainframe, cpd
- BBDD relacionales
- Datos estructurados
- Alm. Convencional:
  - -Silos de información
  - -Datawarehouse

Arq. Big data

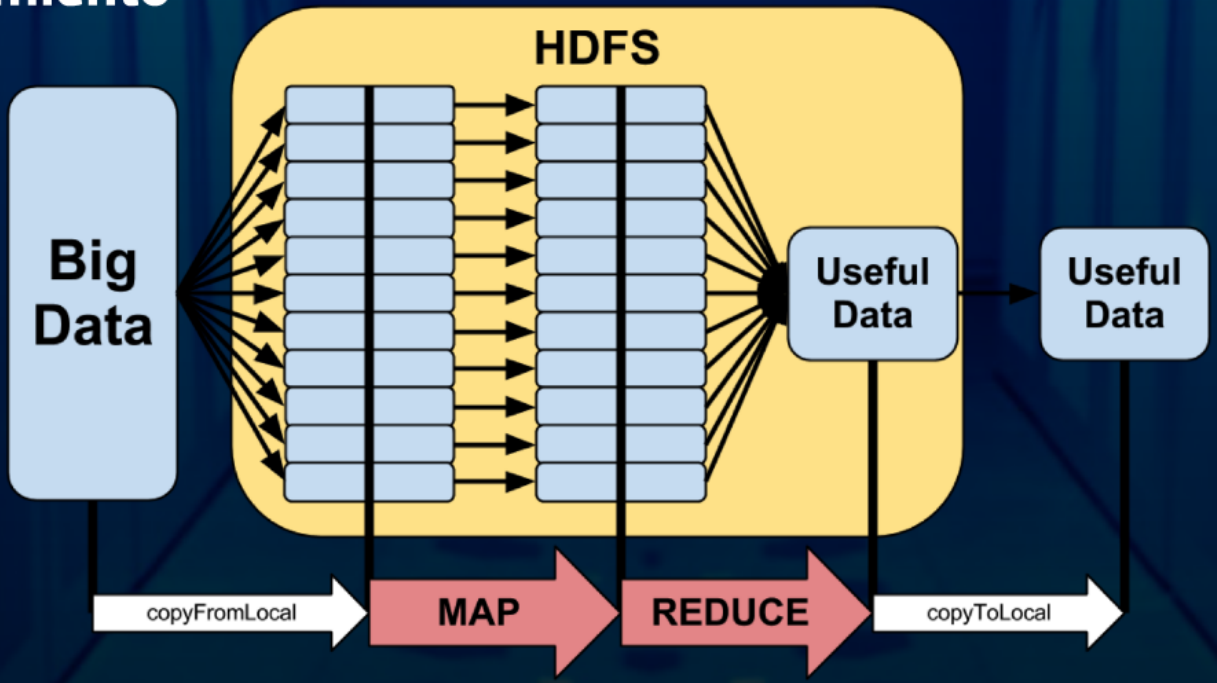
- Alta escalabilidad (Scale-Out)
- Procesamiento paralelo
- Mismo espacio almacen. y procesado→Baja latencia
- Datos no estructurados y est.
- NoSQL
- By-pass de datos (no silos)

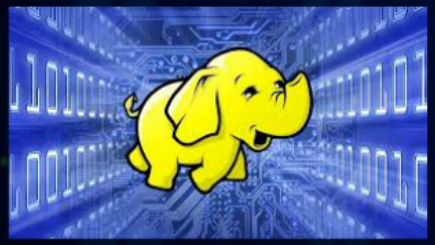
NO Not Only SQL

key	value
firstName	Bugs
lastName	Bunny
location	Earth

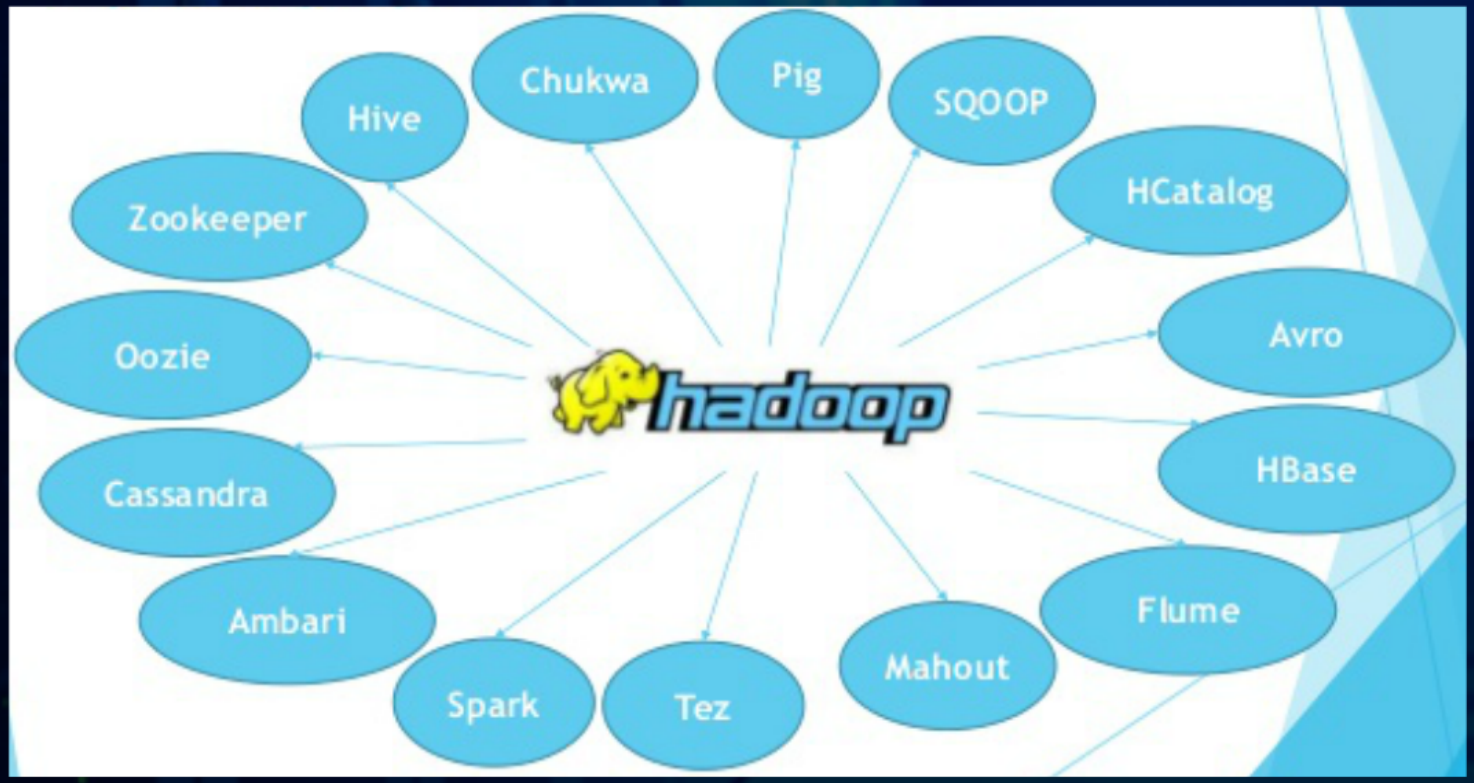
HADOOP

consiste en dividir en dos tareas (mapper – reducer) para manipular los datos distribuidos a nodos de un clúster logrando un alto paralelismo en el procesamiento

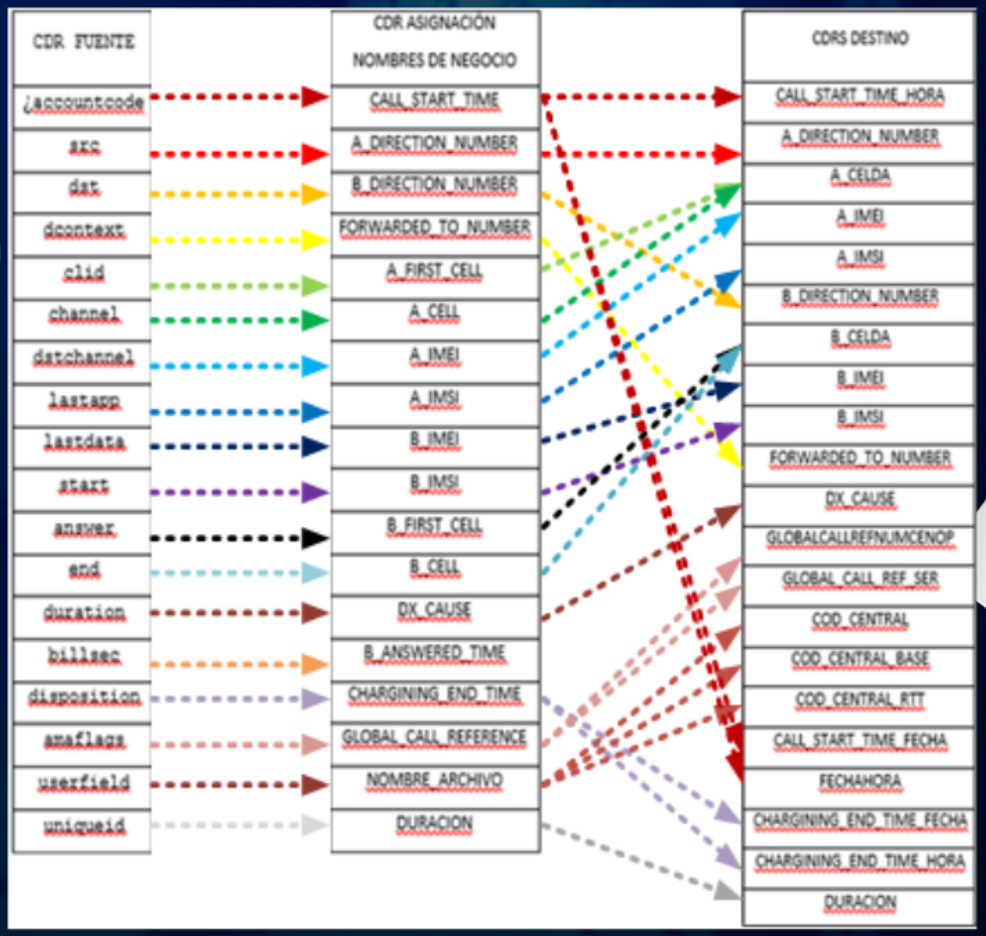




COMPONENTES HADOOP



CDR (CALL DETAIL RECORD)



CREACIÓN, CONFIGURACIÓN E INSTALACIÓN DEL CLÚSTER DE IBM INFOSPHERE BIGINSIGHTS.

Especificación de particiones

Directory	Available disk space
/	10 GB
/tmp	50 GB
/$BIGINSIGHTS_HOME The default directory for this variable is /opt/ibm.	15 GB
/$BIGINSIGHTS_VAR The default directory for this variable is /var/ibm.	5 GB
/home/$USER_HOME The default directory for this variable is /home/biadmin.	5 GB

Verificar identificador ùnico de los discos

```
/dev/sda3: UUID="1632fdf8-2283-4771-9fdd-664964ee7fcf" TYPE="ext3"
/dev/sda1: UUID="8ed83d7a-4e5f-44a1-8448-533da7109312" TYPE="ext3"
/dev/sda2: UUID="59f180e3-931f-4b50-aa94-4b3cb0ab2c0a" TYPE="swap"
```

Cree el usuario biadmin y su grupo

En cada nodo del clúster, como usuario root, cree el grupo BiAdmin y luego agregar el usuario BiAdmin a ella.

Agregue el grupo BiAdmin.

```
groupadd -g 123 biadmin
```

Agregue el usuario BiAdmin al grupo BiAdmin

```
useradd -g biadmin -u 123 biadmin
```

Configurar Protocolo NTP

```
vi /etc/ntpd.conf
```

```
server 0.rhel.pool.ntpd.org
server 1.rhel.pool.ntpd.org
server 2 rhel.pool.ntpd.org
```

```
chkconfig --add ntpd
```

Start the NTPD service.

```
service ntpd start
```

Verify that the clocks are synchronized with a time server.

```
ntpstat
```

Configurar Red en los nodos

```
Kernel IP routing table
Destination     Gateway         Genmask   Flags  Metric  Ref   Use  Iface
...
0.0.0.0         192.0.2.21      0.0.0.0   UG     0       0       0  eth0
)
```

Configuración del ambiente en los nodos del clúster

verificación de pre-requisitos

```
[root@quisrvmed4 ~]# rpm -qa | grep expect
expect-5.44.1.15-5.el6_4.x86_64
```

```
[root@quisrvmed4 ~]# rpm -qa | grep nc-1
nc-1.84-22.el6.x86_64
```

```
# This file controls the state of SELinux on the system.
# SELINUX= can take one of these three values:
#     enforcing - SELinux security policy is enforced.
#     permissive - SELinux prints warnings instead of enforcing.
#     disabled - No SELinux policy is loaded.
SELINUX=disabled
# SELINUXTYPE= can take one of these two values:
```

```
[root@quisrvmed4 ~]# ntpstat
synchronised to NTP server (10.5.1.31) at stratum 3
   time correct to within 31 ms
   polling server every 256 s
```

```
NETWORKING=yes
HOSTNAME=quisrvmed4.otecel.com.ec
NETWORKING_IPV6=no
IPV6INIT=no
~
```

```
[root@quisrvmed4 ~]# lsmod | grep ipv6
ipv6                  317340  292
[root@quisrvmed4 ~]#
```

Deshabilitar Firewall

Verificar en el archivo /etc/hosts que se encuentren las ips y los nombres largos y alias de los nodos del clúster incluido si mismo

```
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4
::1     localhost.localdomain   localhost.localdomain   localhost6      localhost6.localdomain6 localhost       quisrvmed4
10.112.152.163 quisrvmed1.otecel.com.ec quisrvmed1
10.112.152.164 quisrvmed2.otecel.com.ec quisrvmed2
10.112.152.165 quisrvmed3.otecel.com.ec quisrvmed3
10.112.152.166 quisrvmed4.otecel.com.ec quisrvmed4
10.112.152.167 quisrvmed5.otecel.com.ec quisrvmed5
10.112.155.51  newcmserver
10.5.1.74      cmserver-mirror
```

Permisos de ejecucion como root visudo

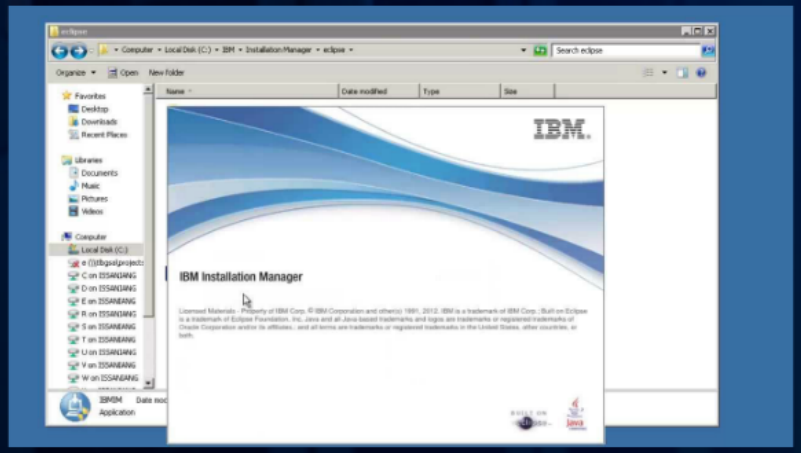
```
## Same thing without a password
# %wheel        ALL=(ALL)       NOPASSWD: ALL
%biadmin ALL=(ALL)  NOPASSWD: ALL
## Allows members of the users group to mount and unmount the
## cdrom as root
# %users  ALL=/sbin/mount /mnt/cdrom, /sbin/umount /mnt/cdrom

## Allows members of the users group to shutdown this system
# %users  localhost=/sbin/shutdown -h now

## Read drop-in files from /etc/sudoers.d (the # here does not mean a comment)
#includedir /etc/sudoers.d
```

Instalación de IBM InfoSphere BigInsights Standard Edition

Comando de instalaciòn Start.sh

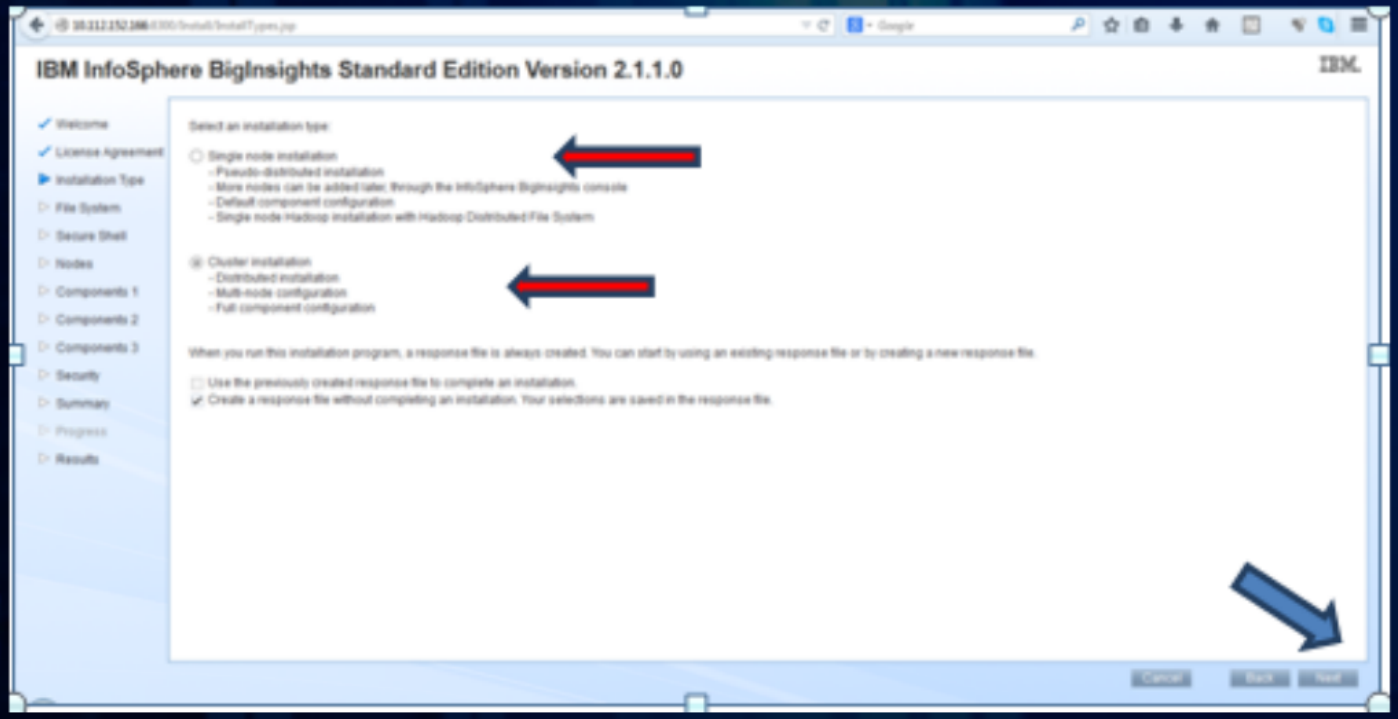
```
[biadmin@quisrvmed4 biginsights-standard-linux64_b20140123_1205]$ ./start.sh
artifacts/ibm-java-sdk-6.0-12.0-linux-x86_64.tgz
Extracting Java ....
Java extraction complete, using JAVA_HOME=/home/biadmin/biginsights-standard-linux64_b20140123_1205/_jvm/ibm-java-x86_64-60
Verifying port 8300 availability
port 8300 available
Starting BigInsights Installer .....
Application server is up and running...


BigInsights Installer started, please use a browser to access:
   http://10.112.152.166:8300/Install

After you are finished, run the following command to stop the installer web server:
   start.sh shutdown
[biadmin@quisrvmed4 biginsights-standard-linux64_b20140123_1205]$
```

Asistente de instalaciòn



Selecciòn tipo de Instalaciòn

- “Custer Installation”, para que pueda crear el clúster de Big Data - Hadoop,
- .Dar clic en “Create a response file without completing an installation”, esto generará un archive de log de instalación para reiniciar en pasos anteriores a la instalación final si existiese necesidad.
- Presionar en “Next”.



- Para el clúster seleccionar “Install Hadoop Distributed File System (HDFS)
- Dar clic en “Overwrite existing files and directories”.
- Presionar en la pestaña “MapReduce general settings”
- Verificar que estén llenos los campos

- Para el clúster seleccionar “Install Hadoop Distributed File System (HDFS)
- Dar clic en “Overwrite existing files and directories”.
- Presionar en la pestaña “MapReduce general settings”
- Verificar que estén llenos los campos
- Presionar en “Next”.

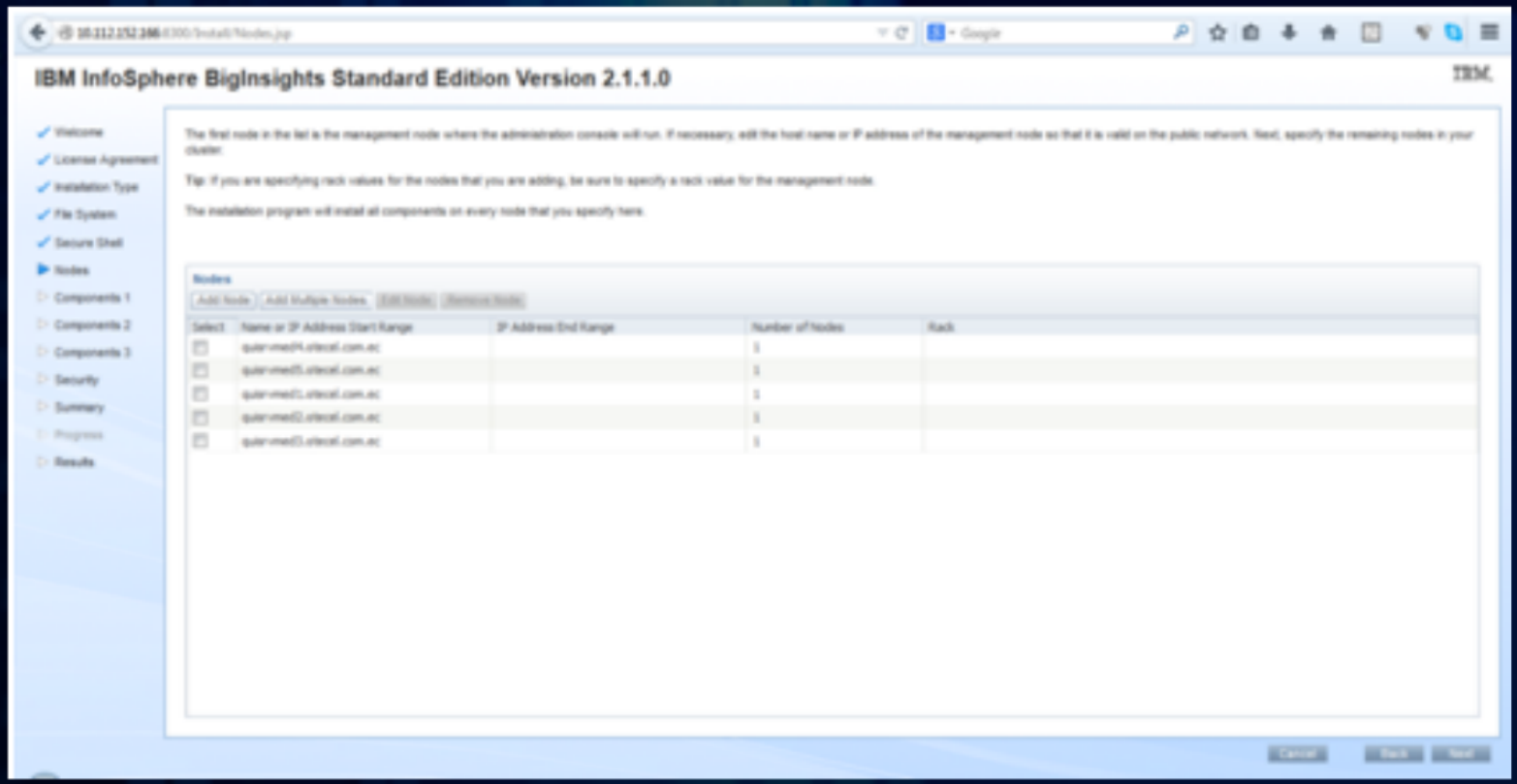
Los nodos del clúster deben tener acceso directo entre ellos tanto por el usuario “root” y para el usuario “biadmin”

- Seleccionar “Use the root user”. Select this option if the following statement is true.
- Especificar el password del usuario root.
- Especificar el usuario de administración de BigIngishts “biadmin”.
- Ingresar la contraseña y confirmamos la misma.
- Ingresar el grupo de administración del grupo de Biginsights “168”.
- Presionar en “Next”.

Ingreso y Configuraciòn de Nodos

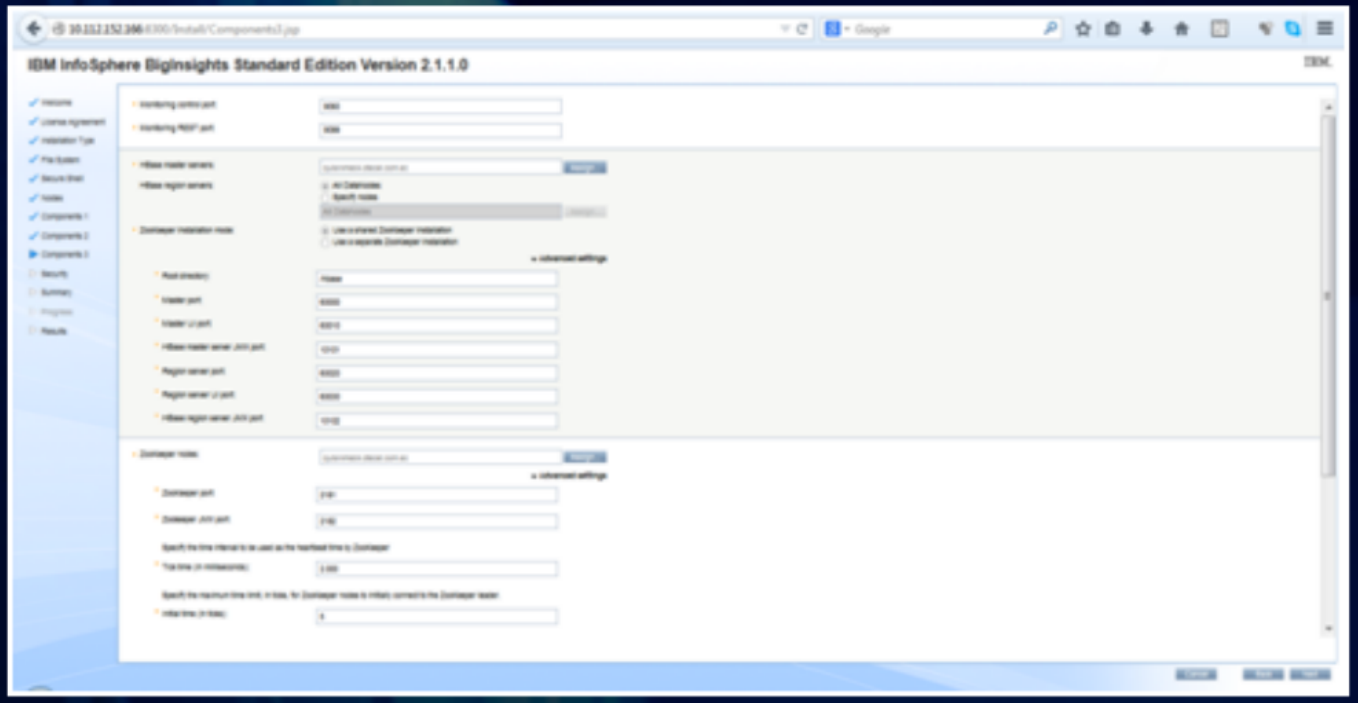
Ingresar los servidores del clúster siguiendo los siguientes pasos:

- Dar clic en Add Node por cada servidor que desea ingresar al clúster:
- Ingresar el nombre del servidor incluido dominio.
- Especificar el password del usuario root del servidor.
- Presionar OK.
- Presionar en "Next



Revisar las configuraciones que se encuentren llenadas, todos los campos que contengan la etiqueta Nodo deben de estar llenado con el servidor "Name Node" principal en este caso "quisrvmed4.TELCOS.com.ec"
Presionar en "Next".

Configuración de componentes Hadoop y finalización de la Instalación

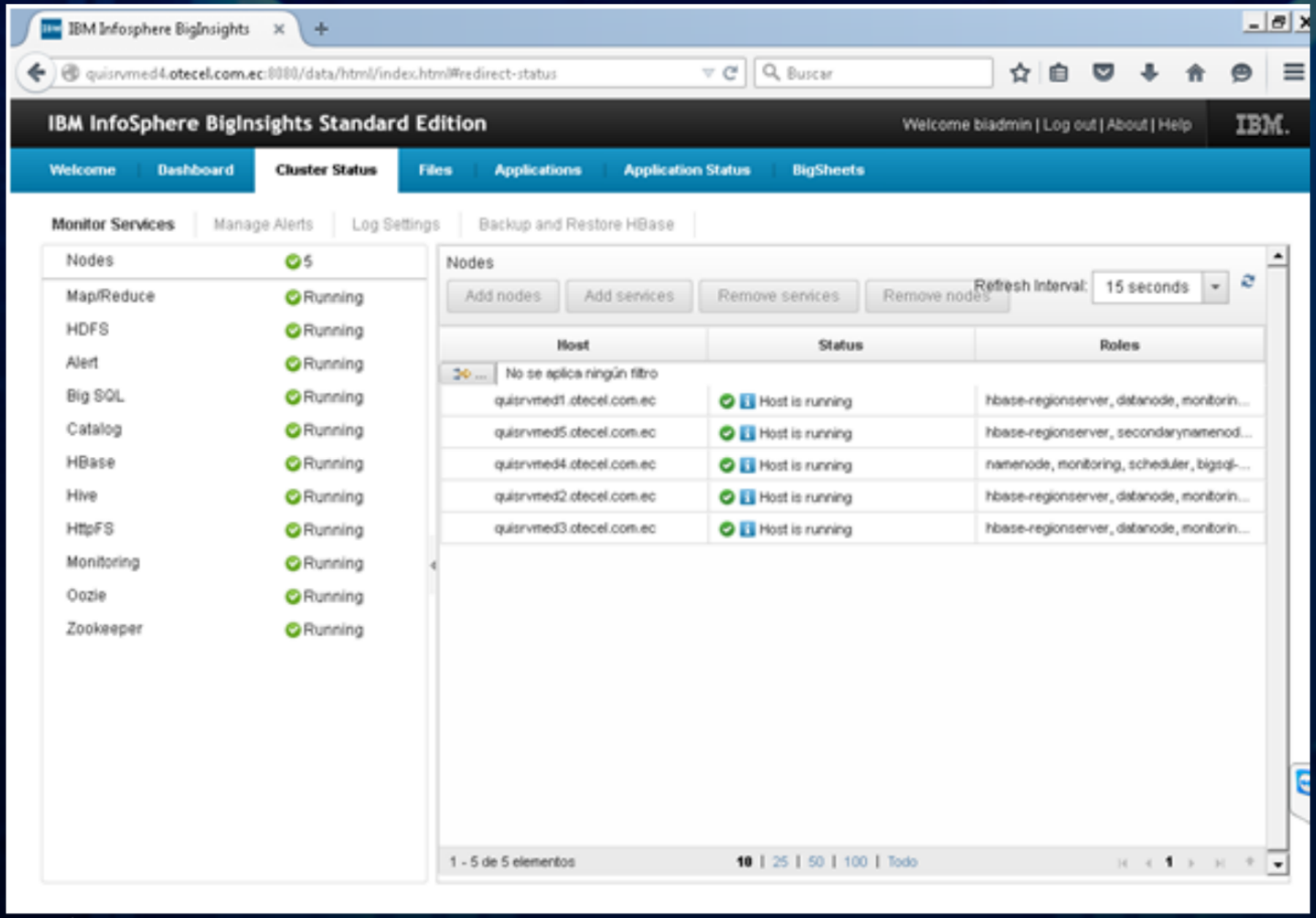


En el directorio donde fue descomprimido el instalador, en este caso "/home/biadmin/biginsights-standard-linux64_b20140123_1205" dirigirse al directorio "/home/biadmin/biginsights-standard-linux64_b20140123_1205/artifacts/security/flatfile".

a. Editar el archivo "biginsights_users.xml"
b. En la etiqueta user name="biadmin" cambiar el password al password que se especificó anteriormente
c. Cerrar el archivo guardando los cambios.
d. Regresar a la consola y presionar "Next".

Por último se revisa las configuraciones y finalizar la instalación

VERIFICACIÒN DEL STATUS DE LOS CLUSTERS



ANÁLISIS, PROCESAMIENTO Y CARGA DE LA INFORMACIÓN DEL CDR

- Búsqueda de información relevante y concerniente a los usuarios

FIELD_NAME	EXAMPLE	DESCRIPTION
Accountcode	12345	An account ID. This field is user-defined and is empty by default.
Src	12565551212	The calling party's caller ID number. It is set automatically and is read-only.
Dst	102	The destination extension for the call. This field is set automatically and is read-only.
Dcontext	PublicExtensions	The destination context for the call. This field is set automatically and is read-only.
Clid	"Big Bird" <12565551212>	The full caller ID, including the name, of the calling party. This field is set automatically and is read-only.
Channel	SIP/0004F2040808-a1bc23ef	The calling party's channel. This field is set automatically and is read-only.
Dstchannel	SIP/0004F2046969-9786b0b0	The called party's channel. This field is set automatically and is read-only.
lastapp	Dial	The last dialplan application that was executed. This field is set automatically and is read-only.
Lastdata	SIP/0004F2046969,30,tT	The arguments passed to the lastapp. This field is set automatically and is read-only.
start	2010-10-26 12:00:00	The start time of the call. This field is set automatically and is read-only.
answer	2010-10-26 12:00:15	The answered time of the call. This field is set automatically and is read-only.
End	2010-10-26 12:03:15	The end time of the call. This field is set automatically and is read-only.
Duration	195	The number of seconds between the start and end times for the call. This field is set automatically and is read-only.
billsec	180	The number of seconds between the answer and end times for the call. This field is set automatically and is read-only.
Disposition	ANSWERED	An indication of what happened to the call. This may be NO ANSWER, FAILED, BUSY, ANSWERED, or UNKNOWN.
Amaflags	DOCUMENTATION	The Automatic Message Accounting (AMA) flag associated with this call. This may be one of the following: OMIT, BILLING, DOCUMENTATION, or Unknown.
Userfield	PerMinuteCharge:0.02	A general-purpose user field. This field is empty by default and can be set to a user-defined string.
Uniqueid	1288112400.1	The unique ID for the src channel. This field is set automatically and is read-only.

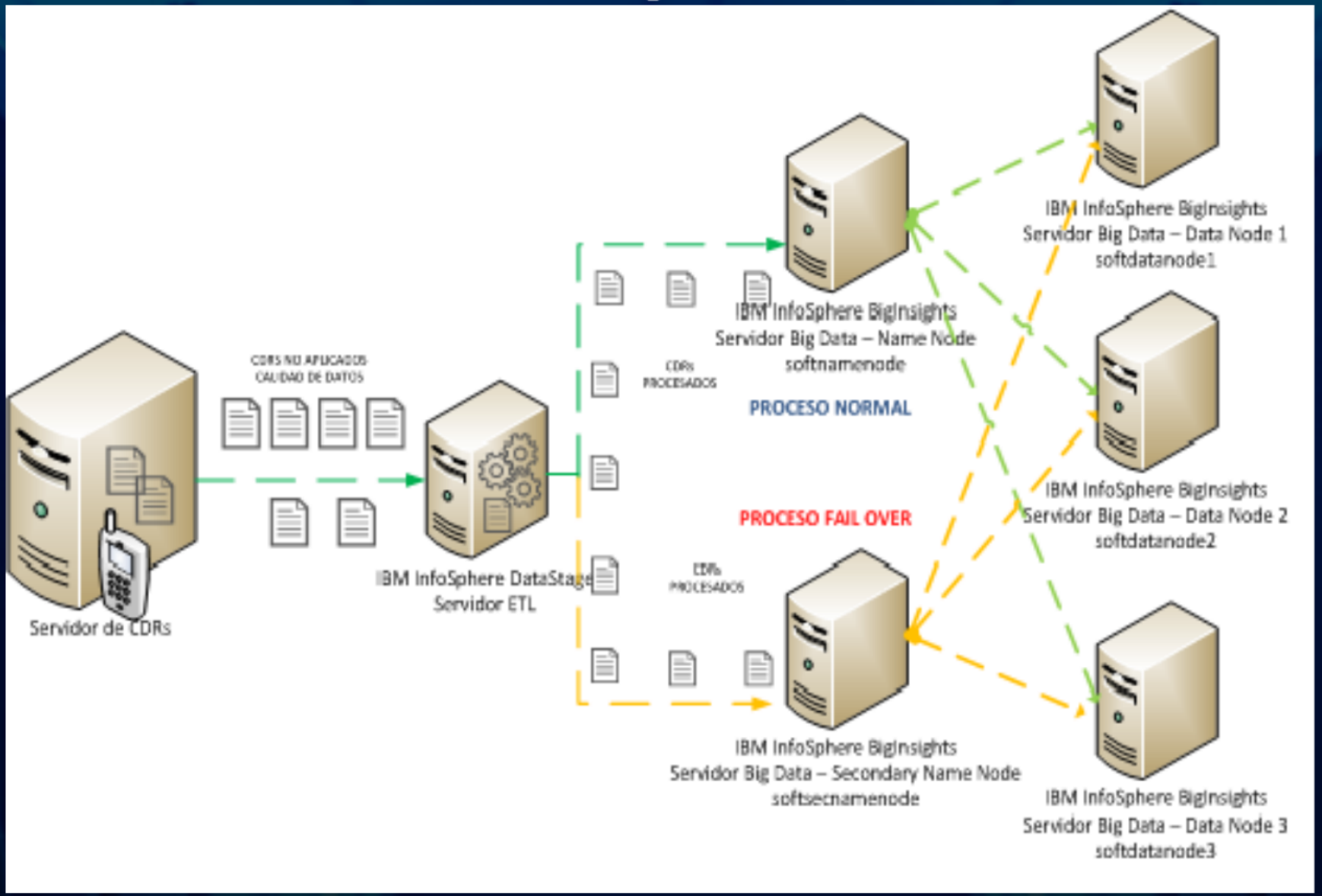
Patrones de procesamiento de la información con calidad de datos

Para la especificación de los patrones de calidad de datos en la información se plantean las siguientes consideraciones:

1. Para los campos de información de números celulares:
    a. Limpieza de caracteres especiales: +,- y ..
    b. Limpieza del número de ubicación regional nacional: 593
2. Para los campos de información de IMEI:
    a. Limpieza de caracteres especiales: +,- y ..
3. Para los campos de información de IMSI:
    a. Limpieza de caracteres especiales: +,- y ..
4. Para los campos de información de Celdas:
    a. Limpieza de caracteres especiales: +,- y ..
5. Para los campos de información de Fechas y Horas:
    a. Limpieza de caracteres especiales: - y :
6. Para el campo de información del CDR:

a. Reasignación de códigos para reducción de tamaño del campo en 3 campos: NOMBRE_ARCHIVO en: COD_CENTRAL, COD_CENTRAL_BASE, COD_CENTRAL_RTT

Procedimiento de carga de Información



SCRIPT DE LA TABLA default.telco_cdrs

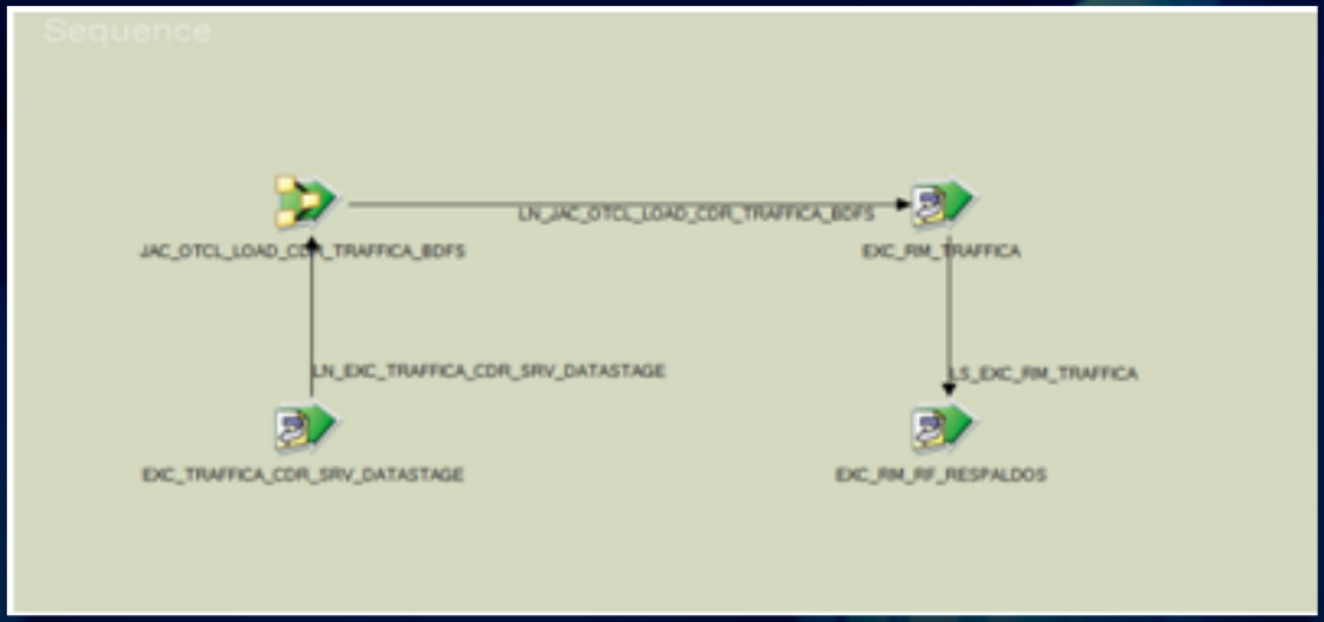
```
CREATE HADOOP TABLE"DEFAULT"."TELCO_CDRS" (
        "CALL_START_TIME_HORA"INTEGER,
        "A_DIRECTION_NUMBER"VARCHAR(50),
        "A_CELDA"INTEGER,
        "A_IMEI"BIGINT,
        "A_IMSI"VARCHAR(50),
        "B_DIRECTION_NUMBER"VARCHAR(50),
        "B_CELDA"INTEGER,
        "B_IMEI"BIGINT,
        "B_IMSI"VARCHAR(50),
        "FORWARDED_TO_NUMBER"VARCHAR(50),
        "DX_CAUSE"SMALLINT,
        "GLOBALCALLREFNUMCENOP"BIGINT,
        "GLOBAL_CALL_REF_SER"VARCHAR(11),
        "COD_CENTRAL"INTEGER,
        "COD_CENTRAL_BASE"INTEGER,
        "COD_CENTRAL_RTT"INTEGER,
        "FECHAHORA"BIGINT,
    "CHARGINIG_END_TIME_FECHA"INTEGER,
      "CHARGINIG_END_TIME_HORA"INTEGER,
      "DURACION"INTEGER

      ) partitionedby (CALL_START_TIME_FECHA INT)
    clusteredby (A_DIRECTION_NUMBER,B_DIRECTION_NUMBER)
    into 48 bucketsstoredasrcfile ALIAS HIVE
METADATA"DEFAULT"."OTC_T_CDRSF2";
```

SCRIPT DE LA TABLA default.telco_cdrstxt

```
CREATE HADOOP TABLE"DEFAULT"."TELCO_CDRSTXT" (
        "CALL_START_TIME_HORA"INTEGER,
        "A_DIRECTION_NUMBER"VARCHAR(50),
        "A_CELDA"INTEGER,
        "A_IMEI"BIGINT,
        "A_IMSI"VARCHAR(50),
        "B_DIRECTION_NUMBER"VARCHAR(50),
        "B_CELDA"INTEGER,
        "B_IMEI"BIGINT,
        "B_IMSI"VARCHAR(50),
        "FORWARDED_TO_NUMBER"VARCHAR(50),
        "DX_CAUSE"SMALLINT,
        "GLOBALCALLREFNUMCENOP"BIGINT,
        "GLOBAL_CALL_REF_SER"VARCHAR(11),
        "COD_CENTRAL"INTEGER,
        "COD_CENTRAL_BASE"INTEGER,
        "COD_CENTRAL_RTT"INTEGER,
        "CALL_START_TIME_FECHA"INTEGER,
        "FECHAHORA"BIGINT,
          "CHARGINIG_END_TIME_FECHA"INTEGER,
      "CHARGINIG_END_TIME_HORA"INTEGER,
      "DURACION"INTEGER

      )    ROWFORMATDELIMITEDFIELDSTERMINATEDBY','
      LINES TERMINATEDBY'\n'
      STOREDASTEXTFILE ALIAS HIVE
METADATA"DEFAULT"."OTC_T_CDRSTXT_F2";
```

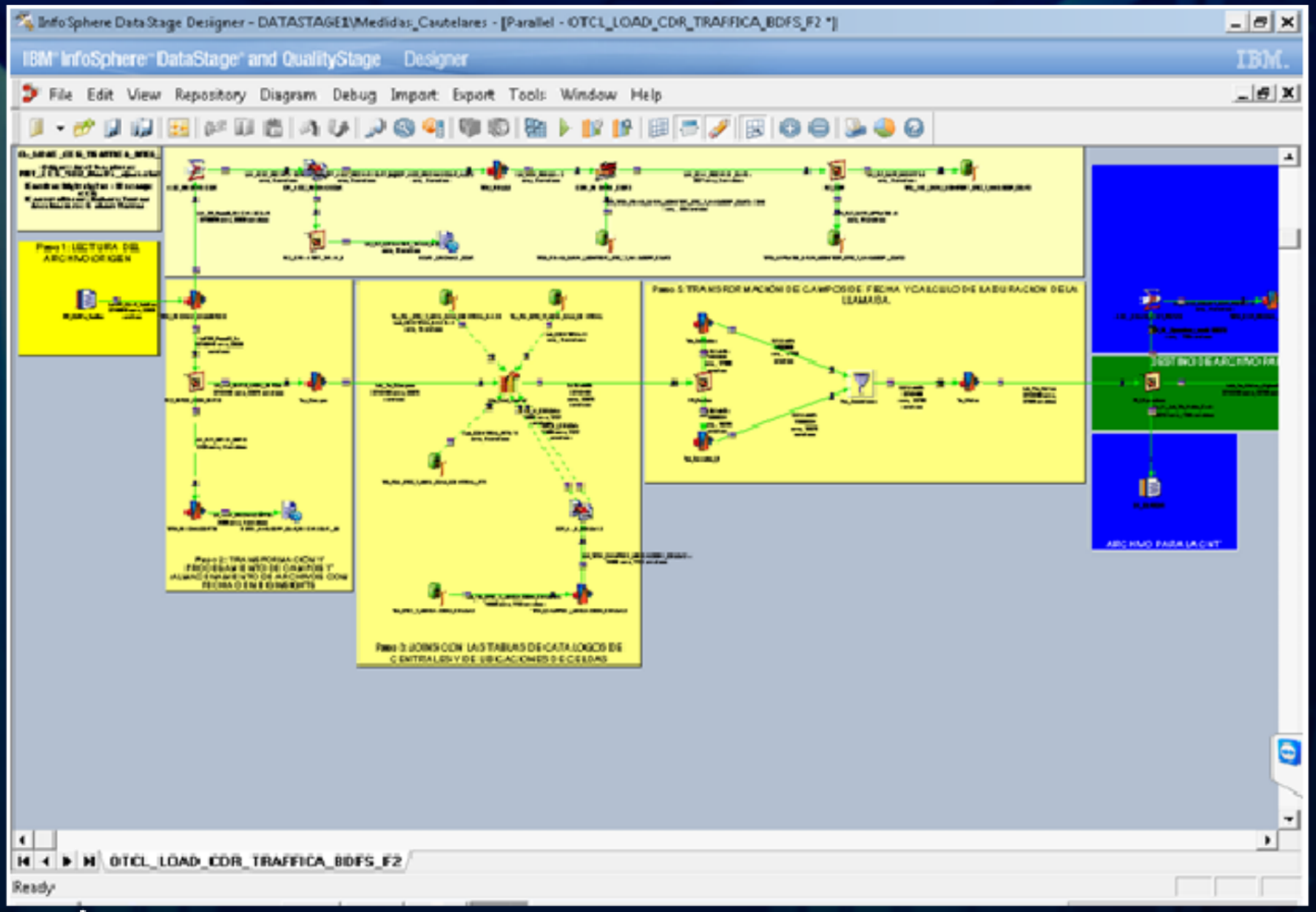
CAPTURAS DEL DESARROLLO DE LOS ETLS

- ETL_SECUENCE_JOB_BIG DATA_TELCO

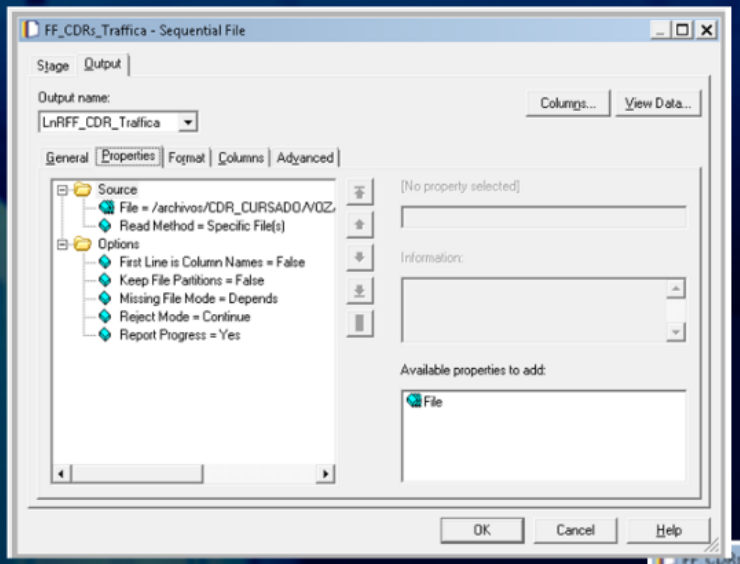


- SEGURIDADES
- Autenticación.
- Nivel de datos
- Información protegida al estar en formato RCFile que se encuentra en binario.
- Nivel de aplicativo
- Restricciones de acceso al sistema por consola Web.
- Restricción de modificar la estructura o integridad del archivo si no se posee la credencial adecuada

- ETL_PARALLEL_JOB_BIG DATA_TELCOCDRS

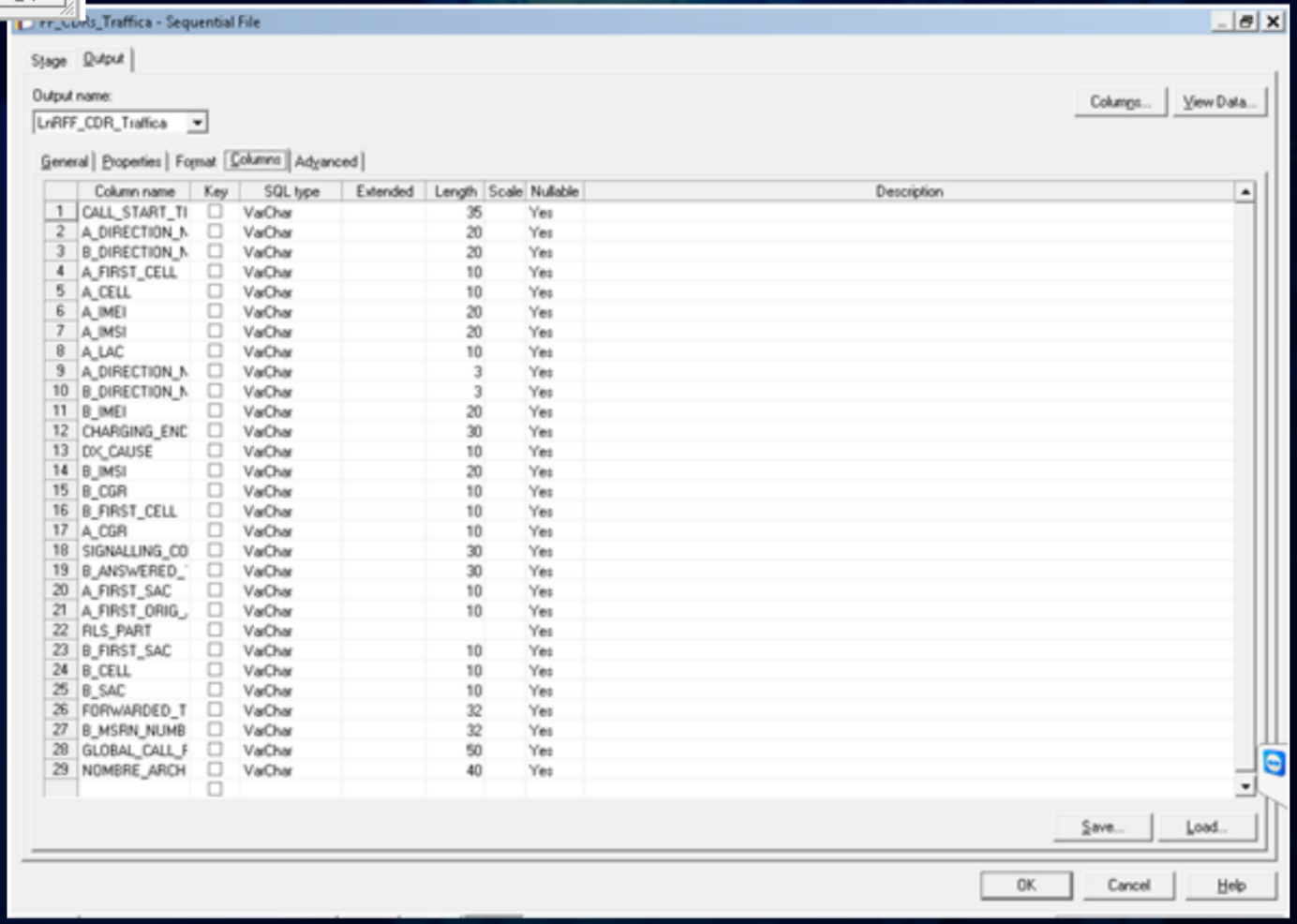


EXTRACCIÒN

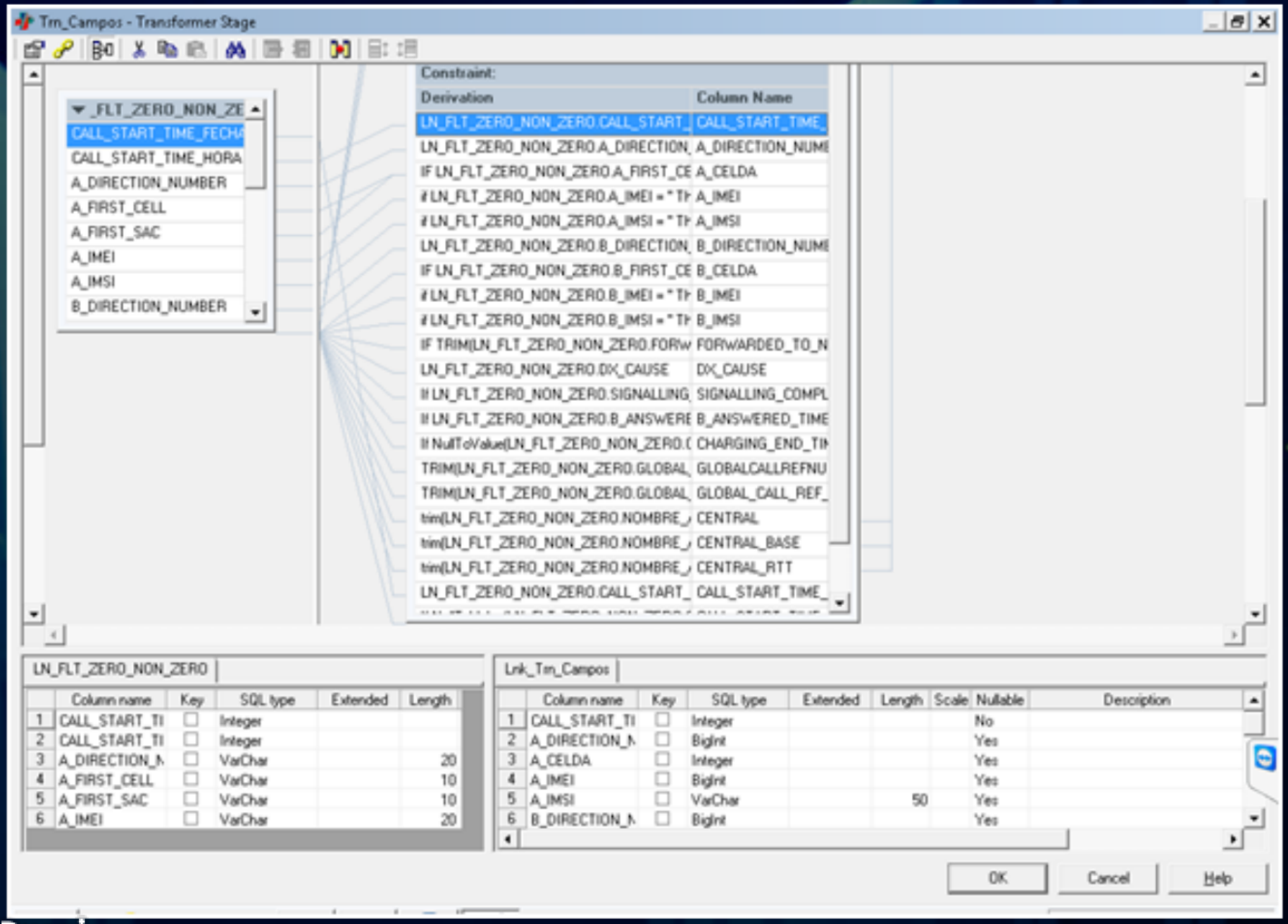


PROPIEDADES

COLUMNAS



TRANSFORMACIÒN



CARGA

biadmin@quisrvmed4:~/Archivos/Scripts - Xshell 4

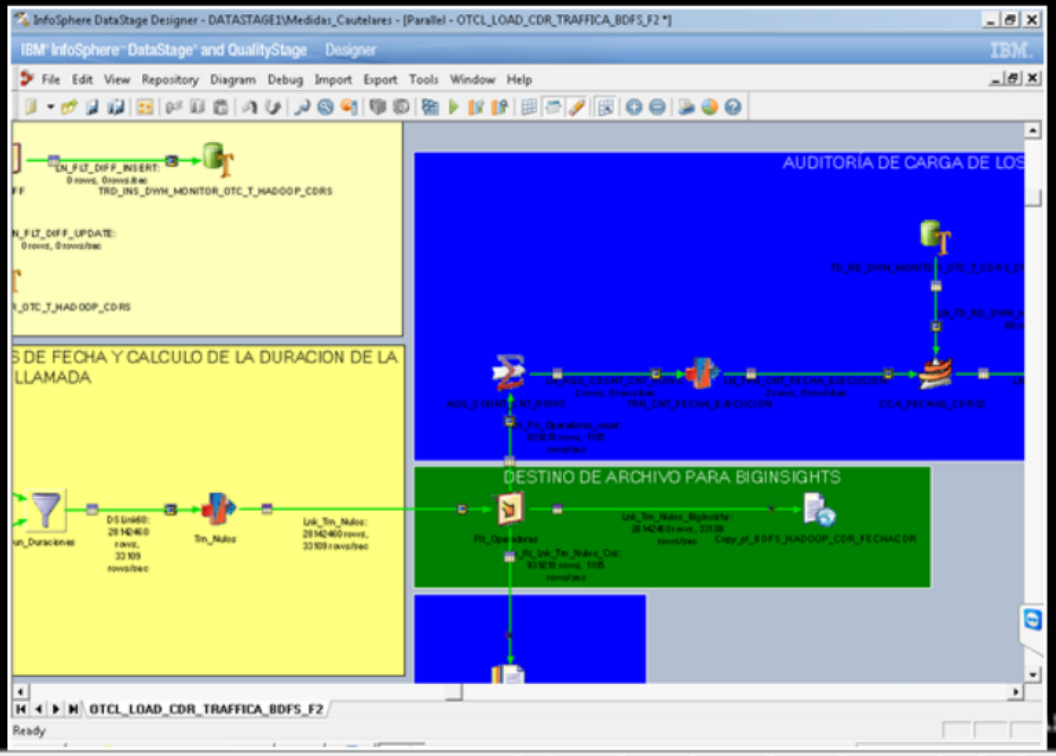
File Edit View Tools Window Help

New Reconnect

1 otecel-quisrvmed4

```
#!/bin/bash
PARAM=$1
/opt/ibm/biginsights/hive/bin/hive -e "LOAD DATA INPATH '/user/biadmin/Archivos/Trafica/CDRs/CDR$1F2.out' INTO TABLE otc_t_cdrsf2_txt"
hadoop fs -cat /user/biadmin/Archivos/Trafica/CDRs/FechasXcdr/CDR_FECHAS$PARAM.out >> stage$PARAM.out
while read line
do
/opt/ibm/biginsights/hive/bin/hive -e "INSERT INTO TABLE DEFAULT.OTC_T_CDRSF2 PARTITION (CALL_START_TIME_FECHA=$line) SELECT CALL_START_TIME_H
ORA,A_DIRECTION_NUMBER,A_CELDA,A_IMEI,A_IMSI,B_DIRECTION_NUMBER,B_CELDA,B_IMEI,B_IMSI,FORWARDED_TO_NUMBER,DX_CAUSE,GLOBALCALLREFNUMCENOP,GLOBA
L_CALL_REF_SER,COD_CENTRAL,COD_CENTRAL_BASE,COD_CENTRAL_RTT,FECHAHORA,CHARGINIG_END_TIME_FECHA,CHARGINIG_END_TIME_HORA,DURACION FROM DEFAULT.O
TC_T_CDRSF2_TXT where CALL_START_TIME_FECHA=$line"
done < stage$PARAM.out
hadoop fs -rm /biginsights/hive/warehouse/otc_t_cdrsf2_txt/CDR$1F2.out
```

"HadoopHiveCDRF2.sh_bkcup" 9L, 848C

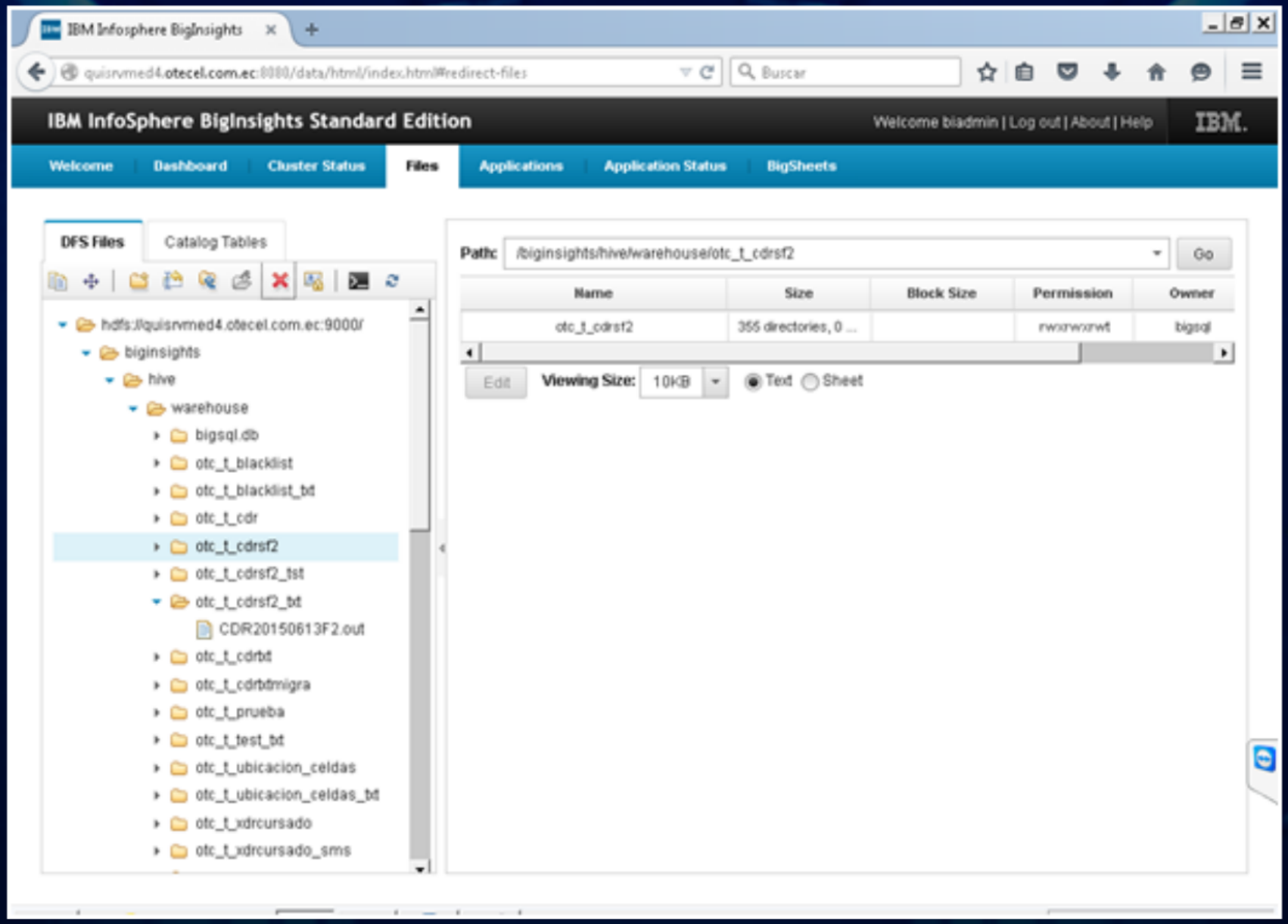


Connected to 10.112.152.166:22.

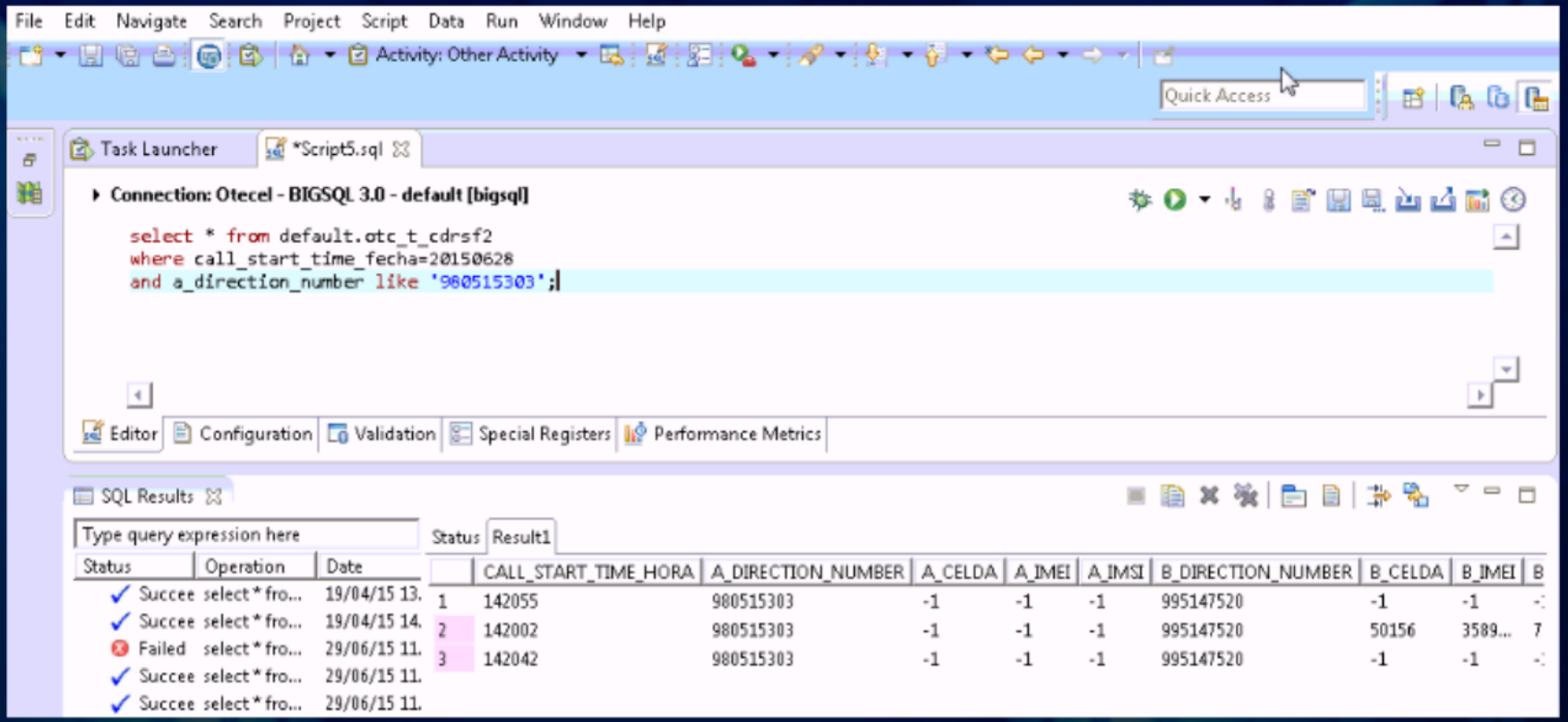
SSH2 xterm 142x39 1,1 1 session CAP NUM

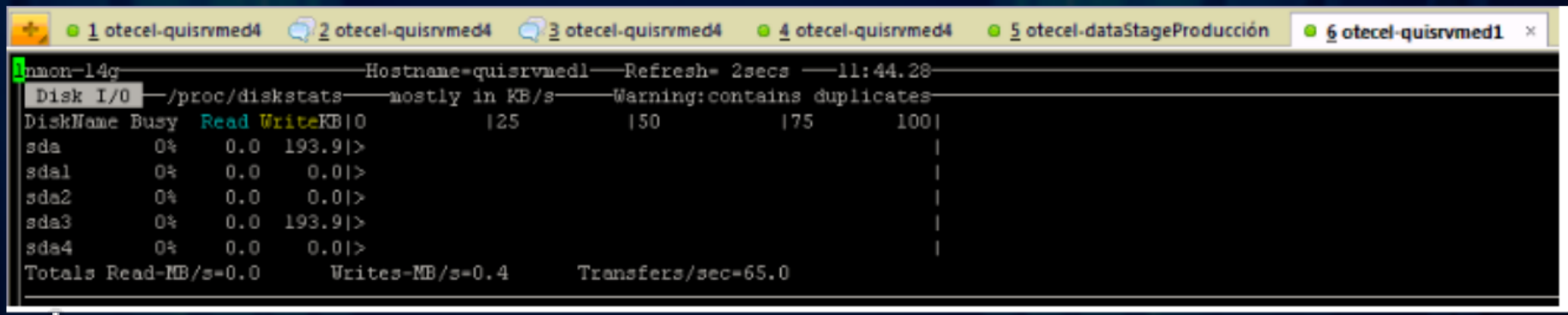
Prezi

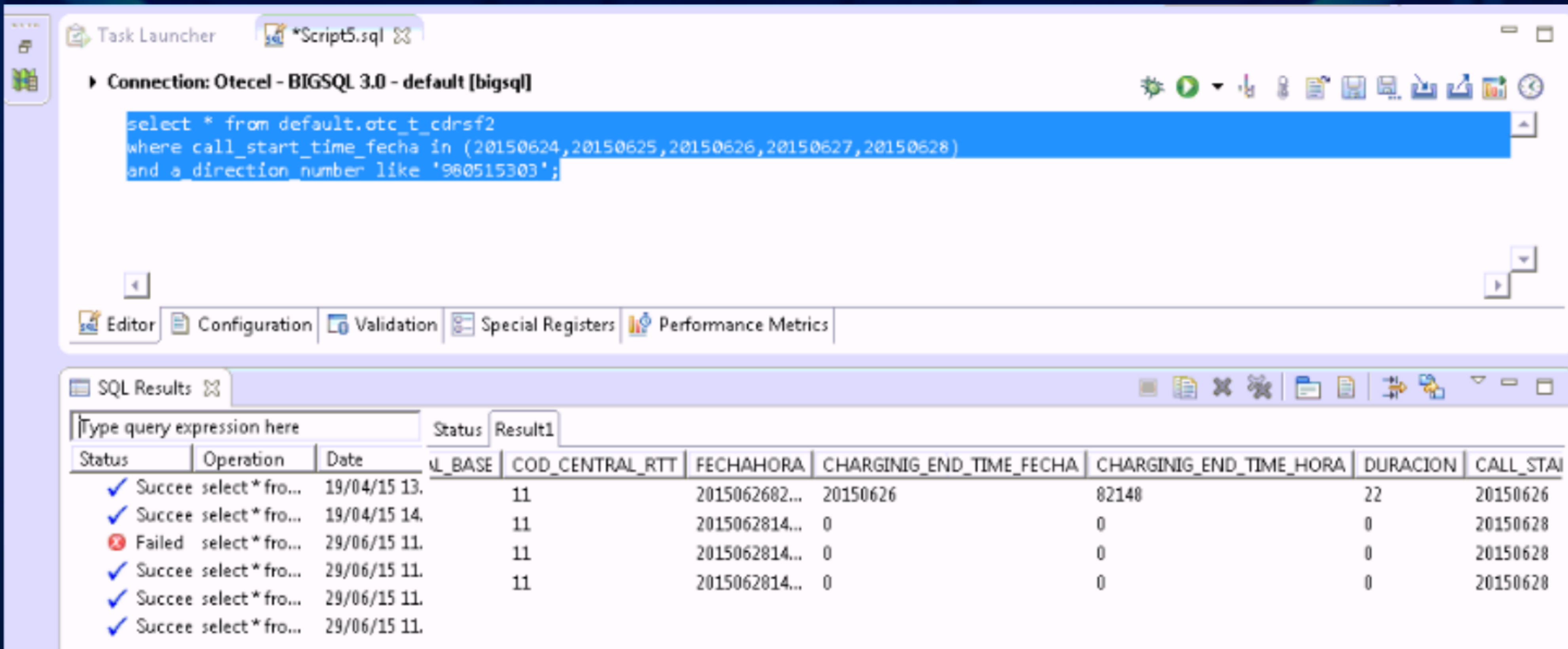
EXPLORACIÒN DE ARCHIVOS PLANOS Y RCFILES

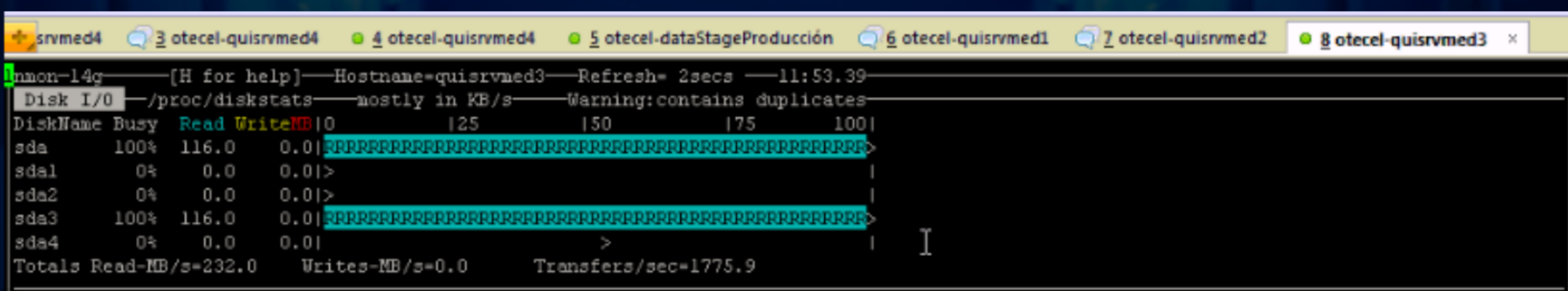


CONSULTA DE DATOS EXTRAIDOS EN BIGSQL

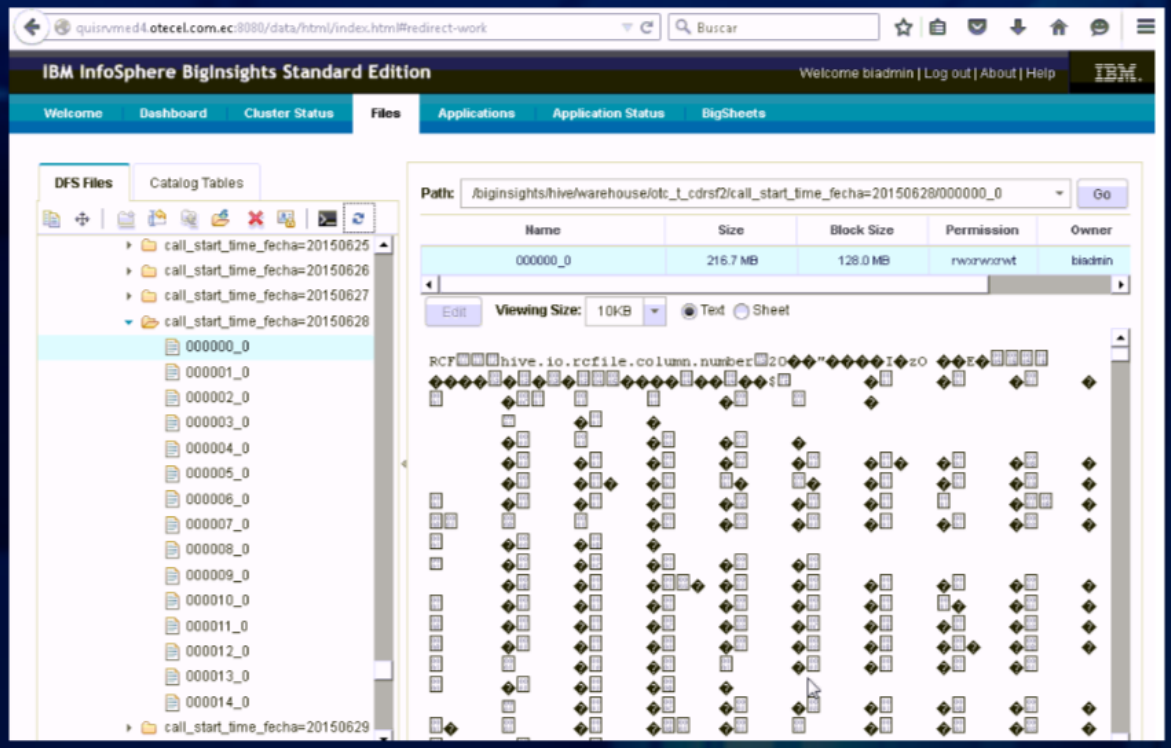


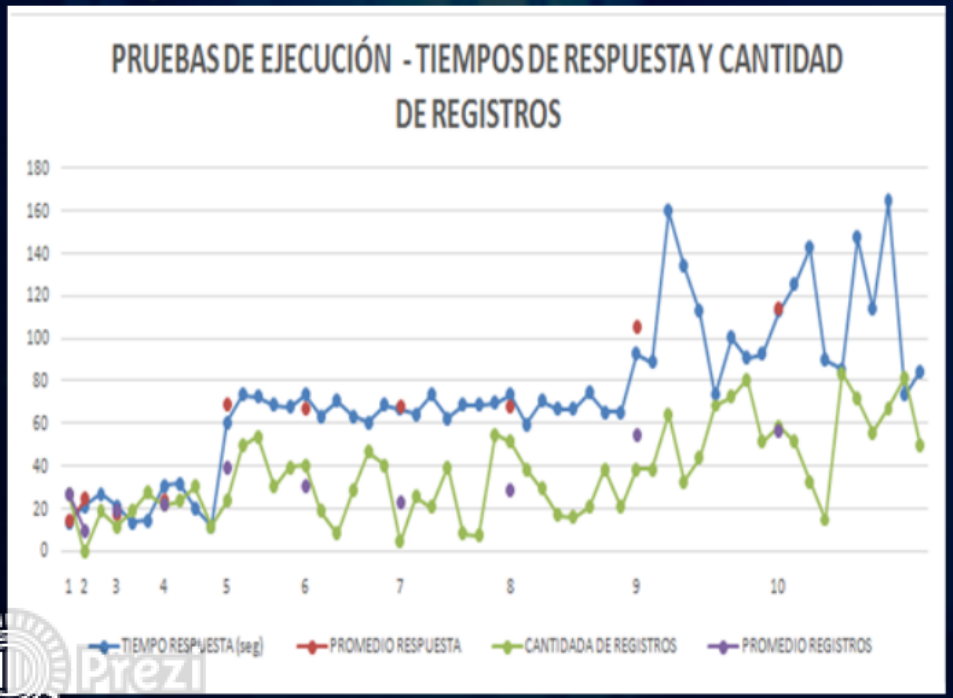



Task Launcher
*Script5.sql
Connection: Otecel - BIGSQL 3.0 - default [bigsql]
select * from default.otc_t_cdrsf2
where call_start_time_fecha in (20150624,20150625,20150626,20150627,20150628)
and a_direction_number like '980515303';
Editor
Configuration
Validation
Special Registers
Performance Metrics
SQL Results
Type query expression here
Status
Result1
Status Operation Date
Succee select * fro... 19/04/15 13.
Succee select * fro... 19/04/15 14.
Failed select * fro... 29/06/15 11.
Succee select * fro... 29/06/15 11.
Succee select * fro... 29/06/15 11.
Succee select * fro... 29/06/15 11.
L_BASE COD_CENTRAL_RTT FECHAHORA CHARGINIG_END_TIME_FECHA CHARGINIG_END_TIME_HORA DURACION CALL_STA
11 2015062682... 20150626 82148 22 20150626
11 2015062814... 0 0 0 20150628
11 2015062814... 0 0 0 20150628
11 2015062814... 0 0 0 20150628


srvmed4
3 otecel-quisrvmed4
4 otecel-quisrvmed4
5 otecel-dataStageProducción
6 otecel-quisrvmed1
7 otecel-quisrvmed2
8 otecel-quisrvmed3
nmon-14g [H for help] Hostname=quisrvmed3 Refresh= 2secs 11:53.39
Disk I/O /proc/diskstats mostly in KB/s Warning:contains duplicates
DiskName Busy Read WriteMB|0 |25 |50 |75 100|
sda 100% 116.0 0.0|RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR>
sda1 0% 0.0 0.0|>
sda2 0% 0.0 0.0|>
sda3 100% 116.0 0.0|RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR>
sda4 0% 0.0 0.0| >
Totals Read-MB/s=232.0 Writes-MB/s=0.0 Transfers/sec=1775.9

SEGURIDADES





PRUEBAS DE EJECUCIÒN

CONCLUSIONES

- La tendencia de la curva del tiempo de respuesta, se estabiliza cuando se incrementa la cantidad de consultas y el número de registros retornados, esto se debe a la naturaleza de la estructura de la tabla que se encuentra particion y con bucketing.
- En Big Data es importante tener una arquitectura heterogénea de componentes, considerando que son difíciles de acoplar hacia otros sistemas que no sean empresariales o que no tengan conexiones nativas por el "Paralelismo".
- En arquitecturas de Big Data se puede notar como ventaja, que al tener réplica de datos en 3 en los datanodes, no s tiene pérdida de servicio, considerando que la probabilidad de fallo es de 1 a 10000 y que se diseña para que los no de datos tengan un balanceo de carga y colas de procesos tras falla de uno o más nodos de datos, además, envia ale tras cada fallo de uno o más componentes.
- Utilizando las herramientas de IBM se comprueba la alta compatibilidad, comunicación y manejo de información en productos de IBM y de otros sistemas ya que éstas se integran con la mayoría de fuentes de datos, sistemas y recur tecnológicos, sean open source o de carácter privativo.

RECOMENDACIONES

- Se recomienda que las personas que implementen una solución de Big Data, tengan conocimientos de Networking, sistemas operativos, bases de datos e infraestructura para crear una solución adecuada.

- Para optimizar más aún la información en Big Data, se necesita tener conocimientos de minería de datos y optimización de datos, especialmente en consultas masivas de gran retorno. Si es mal estructurada la información y no tiene un adecuado manejo, no se evidencia ningún beneficio de poseer una arquitectura de esta naturaleza.

- No es necesario invertir en Hardware especializado para procesamiento y almacenamiento de grandes volúmenes de datos, debido a que se diseña para alto desempeño y escalabilidad a todo nivel tanto horizontal como vertical.

- Big Data no solo puede albergar información en texto plano, también se podría utilizar para analizar las llamadas efectuadas y extraer la información de voz a texto para realizar análisis de patrones delictivos, sean de índole criminalística o judicial y de esta manera brindar un servicio de protección, monitoreo y acción ante eventos malignos.


CONCLUSIONES
RECOMENDACIONES
ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA
DANIEL FABRICIO SALAZAR SANCHEZ
ROBERT ANDRES TORRES FLORES
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION
PROYECTO DE TESIS
JUNIO DEL 2015
CONSTRUCCIÓN DE UN REPOSITORIO POR MEDIO DE LA EXTRACCIÓN Y PROCESAMIENTO DE LA INFORMACIÓN PROVENIENTE DE CDRS GENERADOS POR TELCOS PARA ANALIZAR EL CONSUMO DE LOS DATOS OBTENIDOS, DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM
COMPONENTES HADOOP
HADOOP
NOSQL
Big Data son Todos los Datos
CDR (CALL DETAIL RECORD)
Prezi