

CONCLUSIONES

- El desarrollo de la parte del Sistema de seguridad, se realizó al cumplir con los requisitos de confiabilidad y el cumplimiento de los requisitos de seguridad, así como a la implementación de los requisitos de los datos que se manejan para la parte de seguridad.
- Big Data es una tecnología que permite almacenar y analizar grandes volúmenes de datos, tanto estructurados como no estructurados, en un formato digital. Esto permite almacenar y analizar grandes volúmenes de datos, tanto estructurados como no estructurados, en un formato digital. Esto permite almacenar y analizar grandes volúmenes de datos, tanto estructurados como no estructurados, en un formato digital.

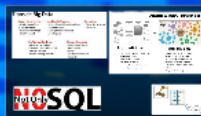
RECOMENDACIONES

- Se recomienda que los parámetros que implementan una solución de Big Data, tengan en cuenta los siguientes aspectos: volumen de datos, tipos de datos e infraestructura para tener una solución adecuada.
- Para que al tener información de Big Data, se pueda tener un análisis de los datos, se recomienda tener un sistema de consultas que permita tener un análisis de los datos, tanto estructurados como no estructurados, en un formato digital. Esto permite almacenar y analizar grandes volúmenes de datos, tanto estructurados como no estructurados, en un formato digital.



TRABAJO DE GRADUACIÓN PARA OBTENER EL TÍTULO DE INGENIERO EN SISTEMAS DE COMPUTACIÓN
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
PROYECTO DE TESIS
JUNIO DEL 2015

CONSTRUCCIÓN DE UN REPOSICIONADOR PARA AYUDAR DE LA EXTRACCIÓN Y PROCESAMIENTO DE LA INFORMACIÓN PROVENIENTE DE COM CALIFICADOS POR TELÉFONO PARA ANALIZAR EL COMPORTAMIENTO DE LOS DATOS DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM





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

- El desarrollo de la parte del Sistema de seguridad, se realizó al cumplir con los requisitos de confiabilidad y el cumplimiento de los requisitos de seguridad, así como a la realización de los tests que se encuentran en el presente documento.
- Big Data es una tecnología que permite almacenar y analizar grandes cantidades de datos que se generan en un momento determinado, pero que no se pueden almacenar en un sistema tradicional de bases de datos.
- Big Data es una tecnología que permite almacenar y analizar grandes cantidades de datos que se generan en un momento determinado, pero que no se pueden almacenar en un sistema tradicional de bases de datos.
- Big Data es una tecnología que permite almacenar y analizar grandes cantidades de datos que se generan en un momento determinado, pero que no se pueden almacenar en un sistema tradicional de bases de datos.

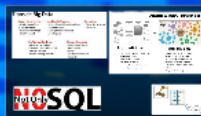
RECOMENDACIONES

- Se recomienda que los parámetros que implementan una solución de Big Data, tengan un consumo de hardware, software y servicios, lo más bajo posible para evitar una solución obsoleta.
- Para el desarrollo de la información de Big Data, se recomienda tener como mínimo un equipo de desarrollo de datos, especialmente en consultas complejas de gran volumen. Si se realiza un trabajo de desarrollo y no tiene un equipo de desarrollo, se recomienda tener un equipo de desarrollo y no tener un equipo de desarrollo.
- Para el desarrollo de la información de Big Data, se recomienda tener como mínimo un equipo de desarrollo de datos, especialmente en consultas complejas de gran volumen. Si se realiza un trabajo de desarrollo y no tiene un equipo de desarrollo, se recomienda tener un equipo de desarrollo y no tener un equipo de desarrollo.
- Big Data no es una tecnología que permita almacenar y analizar grandes cantidades de datos que se generan en un momento determinado, pero que no se pueden almacenar en un sistema tradicional de bases de datos.



ANÁLISIS PARA LA OBTENCIÓN DE DATOS EN TIEMPO REAL
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION
PROYECTO DE TESIS
JUNIO DEL 2015

CONSTRUCCIÓN DE UN REPOSITARIO PARA ANÁLISIS DE LA EXTRACCIÓN Y PROCESAMIENTO DE LA INFORMACIÓN PROVENIENTE DE COMERCIALIZADOS POR TELÉFONO PARA ANALIZAR EL CONSUMO DE LOS DATOS DENTRO DE LA EMPRESA. SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM



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

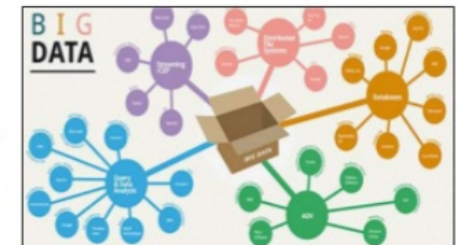
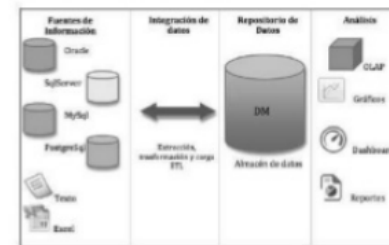
Big Transaction Data

- Healthcare Claims
- Telecommunications Call Detail Records
- Utility Billing Records

Human Generated

- Call Center Voice Recordings
- Email
- Electronic Medical Records

ARQUITECTURA - Diferencias



vs.

Arq. tradicional

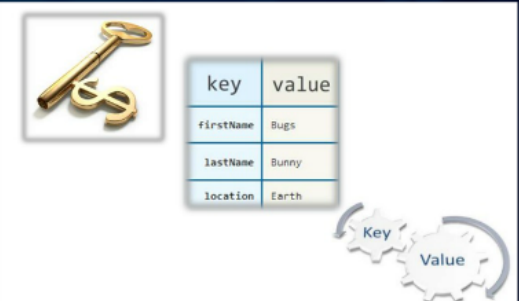
- Centralidad → mainframe, cpd
- BBDD relacionales
- Datos estructurados
- Alm. Convencional:
 - Silos de información
 - Datawarehouse

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)

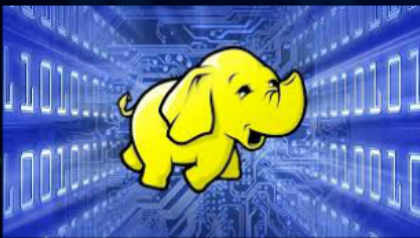
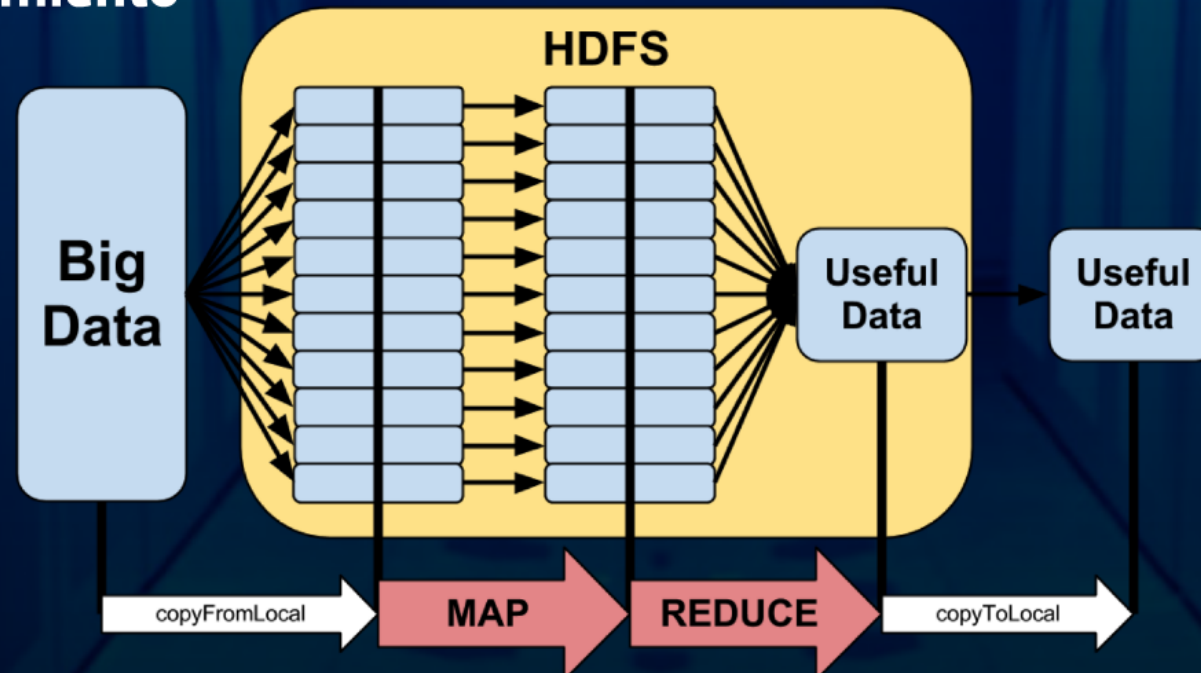
NO SQL

Not Only

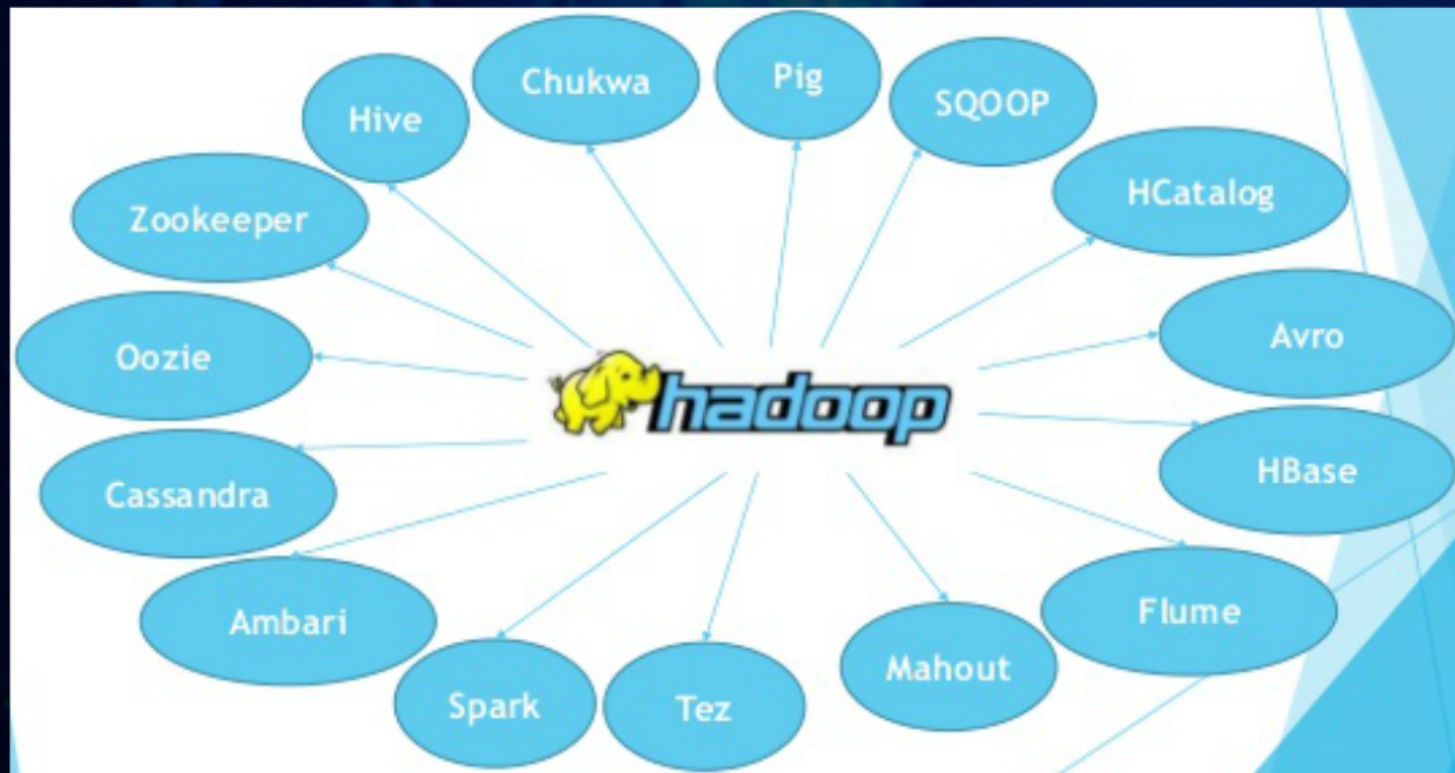


HADOOP

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)

Rec Calls

Account Code	Name	Date	Time
+91 9234823451	Venka	22/01/2013	03:13 PM
+91 9000123551	Vijay S	22/01/2013	01:14 PM
91 9000003451	Sang	26/01/2013	10:20 AM
923456789	Sang	20/01/2013	01:14 PM
923456789	Sang	20/01/2013	10:20 AM
923456789	Sang	20/01/2013	10:05 AM
923456789	Sang	20/01/2013	09:45 AM
923456789	Sang	20/01/2013	09:30 AM

CDR FUENTE	CDR ASIGNACIÓN	CDR DESTINO
	NOMBRES DE NEGOCIO	
<u>accountcode</u>	<u>CALL START TIME</u>	<u>CALL START TIME HORA</u>
<u>sec</u>	<u>A.DIRECTION NUMBER</u>	<u>A.DIRECTION NUMBER</u>
<u>dst</u>	<u>B.DIRECTION NUMBER</u>	<u>A.CELDA</u>
<u>dcontext</u>	<u>FORWARDED TO NUMBER</u>	<u>A.IMEI</u>
<u>clid</u>	<u>A.FIRST CELL</u>	<u>A.IMSI</u>
<u>channel</u>	<u>A.CELL</u>	<u>B.DIRECTION NUMBER</u>
<u>dstchannel</u>	<u>A.IMEI</u>	<u>B.CELDA</u>
<u>lastapp</u>	<u>A.IMSI</u>	<u>B.IMEI</u>
<u>lastdata</u>	<u>B.IMEI</u>	<u>B.IMSI</u>
<u>start</u>	<u>B.IMSI</u>	<u>FORWARDED TO NUMBER</u>
<u>answex</u>	<u>B.FIRST CELL</u>	<u>DX.CAUSE</u>
<u>end</u>	<u>B.CELL</u>	<u>GLOBALCALLREFNUMCENOP</u>
<u>duration</u>	<u>DX.CAUSE</u>	<u>GLOBAL CALL REF SER</u>
<u>billsec</u>	<u>B.ANSWERED TIME</u>	<u>COO.CENTRAL</u>
<u>disposition</u>	<u>CHARGING END TIME</u>	<u>COO.CENTRAL BASE</u>
<u>amaflags</u>	<u>GLOBAL CALL REFERENCE</u>	<u>COO.CENTRAL RTT</u>
<u>userfield</u>	<u>NOMBRE ARCHIVO</u>	<u>CALL START TIME FECHA</u>
<u>uniqueid</u>	<u>DURACION</u>	<u>FECHAHORA</u>
		<u>CHARGING END TIME FECHA</u>
		<u>CHARGING END TIME HORA</u>
		<u>DURACION</u>



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

Especificación de particiones

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

Configurar Protocolo NTP

```
vi /etc/ntp.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.

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

```
service ntpd start
```

Verify that the clocks are synchronized with a time server.

```
ntpstat
```

Verificar identificador único de los discos

```
/dev/sda3: UUID="1632fd8-2283-4771-9fdd-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 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 newcmsserver
10.5.1.74 cmsserver-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 instalación 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]#
```

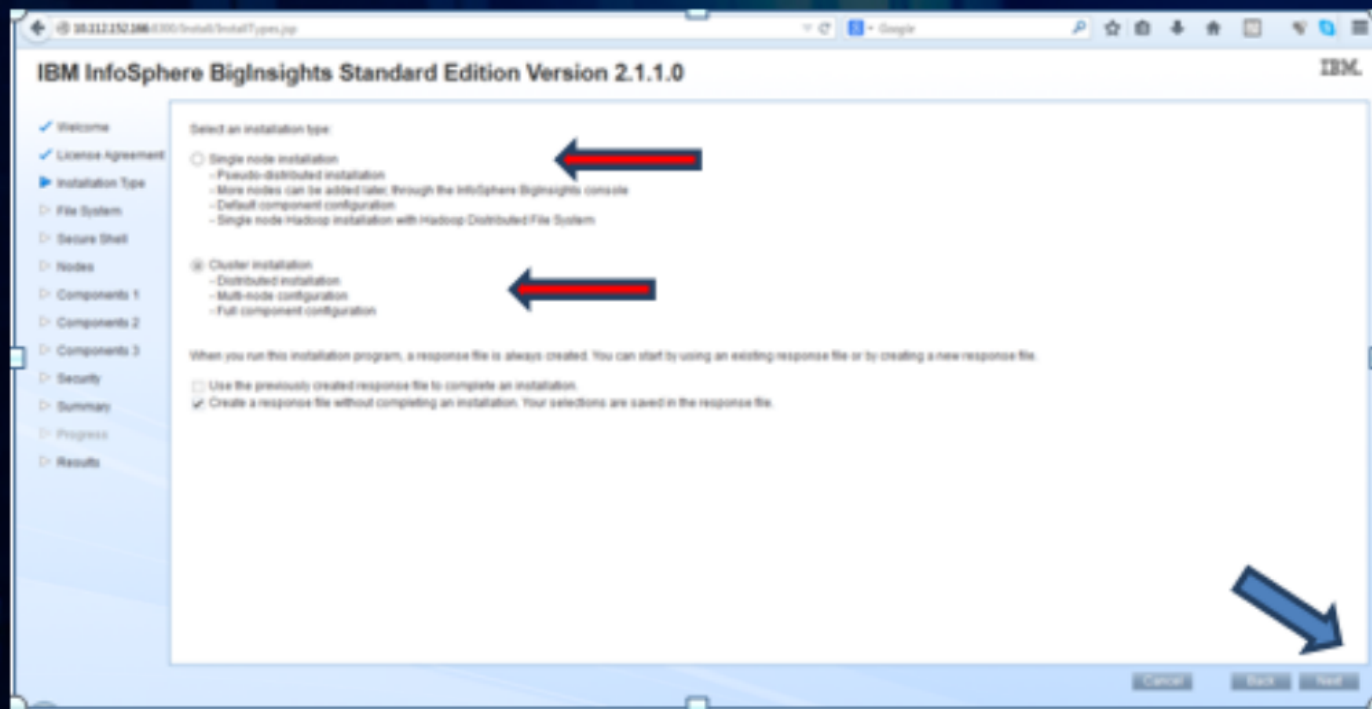


Asistente de instalación



Selección tipo de Instalación

- “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"

IBM InfoSphere BigInsights Standard Edition Version 2.1.1.0

To begin installing a InfoSphere BigInsights cluster, you specify settings for the type of file system to use in your cluster, the installation directories, and the MapReduce directories.

Install Hadoop Distributed File System (HDFS)
 Install General Parallel File System (GPFS)
 Use an existing General Parallel File System (GPFS)
 Use an existing shared directory space

Specify the root, installation and data directories for InfoSphere BigInsights.

- The installation directory contains all of the files required for each of the InfoSphere BigInsights components to run correctly.
- The data directory contains all of the log files and associated data for each component.
- Other local file system directories that are specified in the installation program will be appended to the root directory path.

If the installation directories already exist, then select **Overwrite existing files and directories** to overwrite the existing directories.

Overwrite existing files and directories

The directories that you specify must be accessible by the current user. If you do not start your directory names with a forward slash (/), then your directory is appended to the root directory. The root directory must start with a forward slash (/).

• Root directory: /
 • Installation directory: opt/ibm/biginsights
 • Data directory: var/ibm/biginsights
 If specified as relative path (without a leading /), this directory will be appended to /

MapReduce general settings

Cancel Back 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 BigInsights "biadmin".
- Ingresar la contraseña y confirmamos la misma.
- Ingresar el grupo de administración del grupo de Biginsights "168".
- Presionar en "Next".

IBM InfoSphere BigInsights Standard Edition Version 2.1.1.0

Passwordless secure shell (SSH) is required for communication between nodes in a cluster. The installation program can configure passwordless SSH for the BigInsights administrator if it is not already configured.

The installation program uses the root user, or a user with root security privileges, to complete the following configuration tasks:

- Create an InfoSphere BigInsights administrator user if that user does not already exist
- Distribute the passwordless SSH configuration
- Create and configure all of the required InfoSphere BigInsights directories if the BigInsights administrator user ID does not have the privileges

Select the option that you want to install the product with:

Nodes
 Components 1
 Components 2
 Components 3

Use the root user. Select this option if all the following statements are true:
 - The root user is configured for passwordless SSH access from the current node to all nodes, including to itself
 - The InfoSphere BigInsights administrator user ID exists on the current node
 - The InfoSphere BigInsights administrator user ID has passwordless sudo privileges on the current node

Use the root user. Select this option if the following statement is true:
 - You can provide the password for the root user

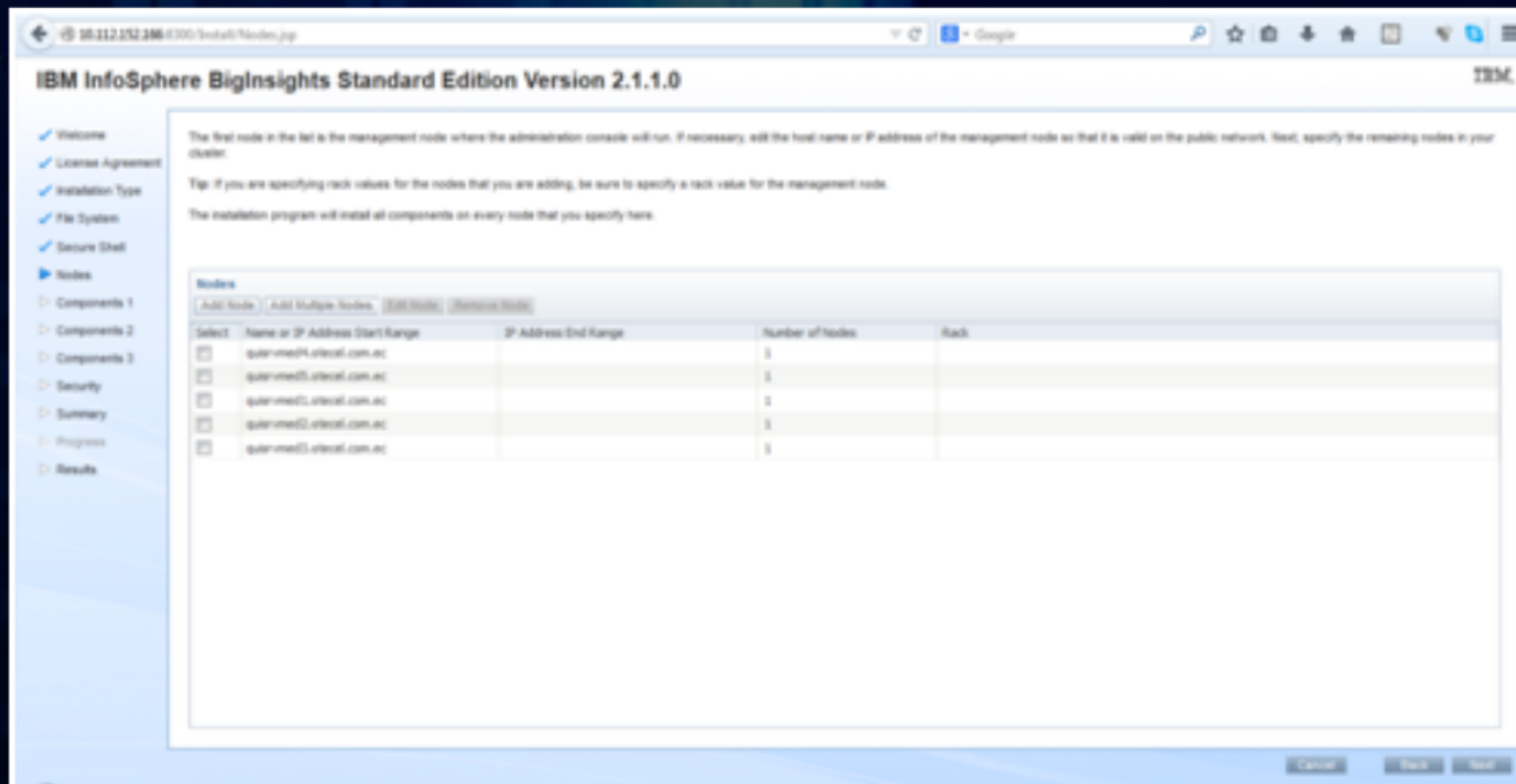
• Root password: *****
 Do not specify root as the BigInsights administrator user ID or group ID. This user will not be given sudo privileges.
 • InfoSphere BigInsights administrator user ID: biadmin
 • InfoSphere BigInsights administrator password: *****
 • Confirm administrator password: *****
 The InfoSphere BigInsights administrator group is used for InfoSphere BigInsights security. Ensure that no other users will be added to this group.
 • InfoSphere BigInsights administrator group ID: 168

Cancel Back Next

Ingreso y Configuración 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".

VERIFICACIÓN DEL STATUS DE LOS CLUSTERS

IBM InfoSphere BigInsights Standard Edition

Welcome bladmin | Log out | About | Help

Welcome | Dashboard | **Cluster Status** | Files | Applications | Application Status | BigSheets

Monitor Services | Manage Alerts | Log Settings | Backup and Restore HBase

Nodes 5

MapReduce Running

HDFS Running

Alert Running

Big SQL Running

Catalog Running

HBase Running

Hive Running

HttpFS Running

Monitoring Running

Oozie Running

Zookeeper Running

Nodes

Add nodes | Add services | Remove services | Remove nodes | Refresh Interval: 15 seconds

Host	Status	Roles
No se aplica ningún filtro		
quitrvmcd1.otecel.com.ec	Host is running	hbase-regionserver, datanode, monitorin...
quitrvmcd5.otecel.com.ec	Host is running	hbase-regionserver, secondarynamenod...
quitrvmcd4.otecel.com.ec	Host is running	namenode, monitoring, scheduler, bigsq...
quitrvmcd2.otecel.com.ec	Host is running	hbase-regionserver, datanode, monitorin...
quitrvmcd3.otecel.com.ec	Host is running	hbase-regionserver, datanode, monitorin...

1 - 5 de 5 elementos | 10 | 25 | 50 | 100 | Todo

ANÁLISIS, PROCESAMIENTO Y CARGA DE LA INFORMACIÓN DEL CDR

- Búsqueda de información relevante y concerniente a los usuarios

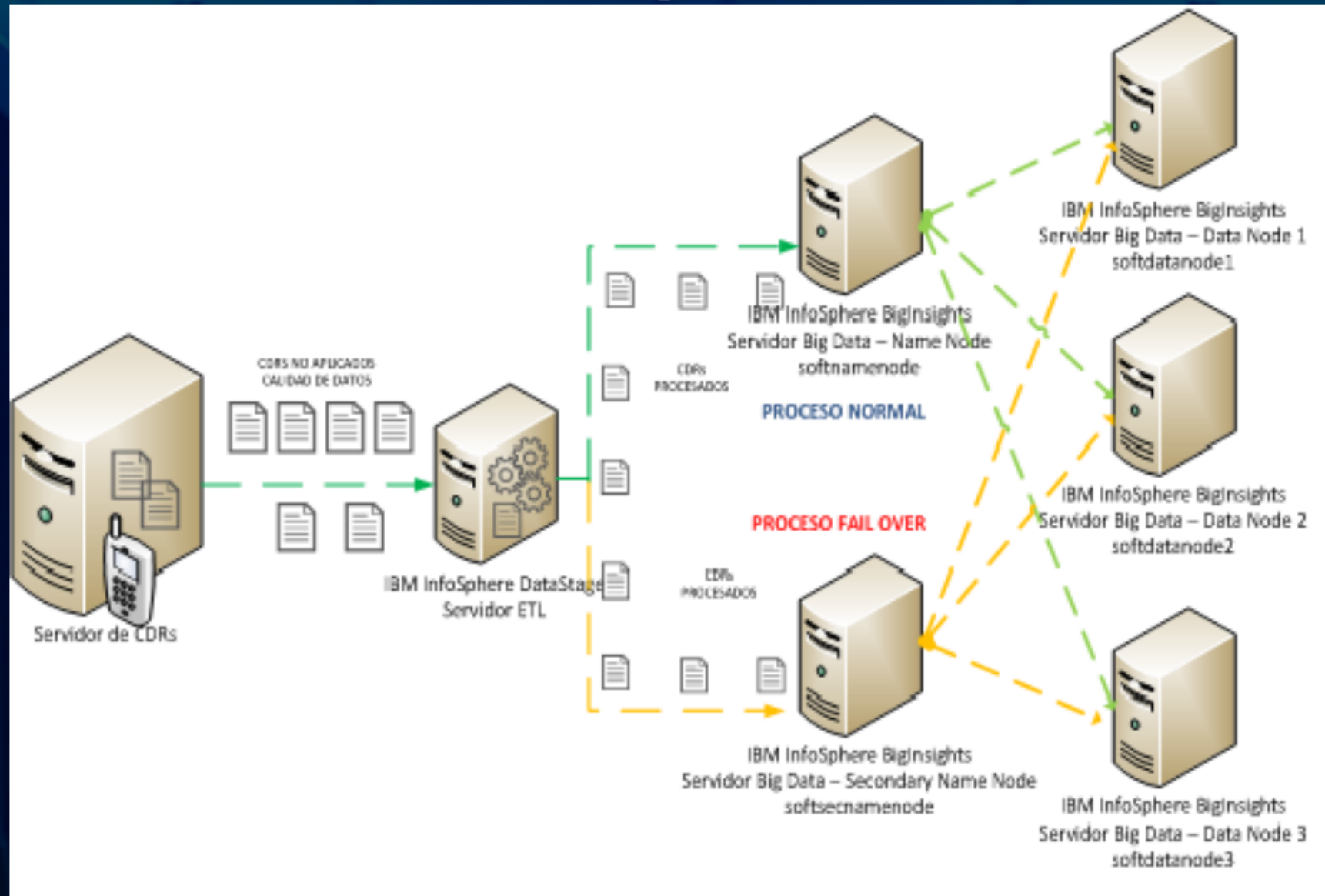
FIELD_NAME	EXAMPLE	DESCRIPTION
Accountcode	12345	An account ID. This field is user-defined and is empty by default.
Src	12565551212	The calling party's caller ID number. It is set automatically and is read-only.
Dst	102	The destination extension for the call. This field is set automatically and is read-only.
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/0004F20 40808-a1bc23ef	The calling party's channel. This field is set automatically and is read-only.
Dstchannel	SIP/0004F20 40969-9786b0b0	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/0004F20 40969,30,IT	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.
Amflags	DOCUMENT ATION	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. ¹⁴
Uniqueid	1288112400. 1	The unique ID for the <i>src</i> 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

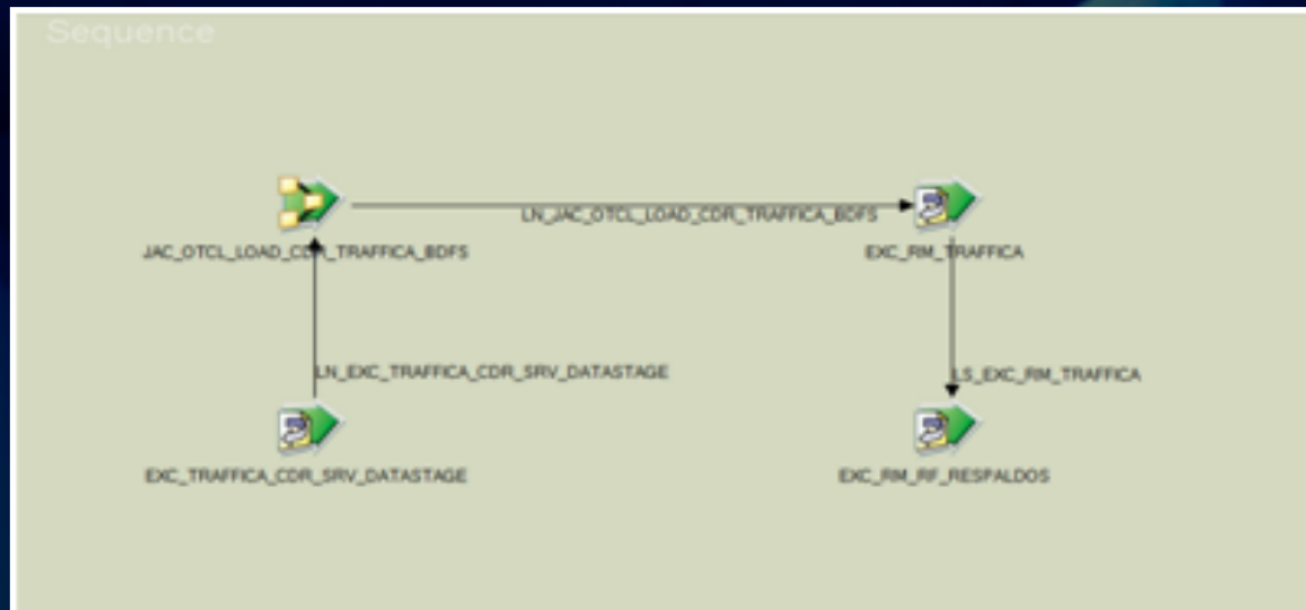
```
CREATE HADOOP TABLE "DEFAULT"."TELCO_CDRS" (  
  "CALL_START_TIME_HORA" INTEGER,  
  "A_DIRECTION_NUMBER" VARCHAR(50),  
  "A_CELDA" INTEGER,  
  "A_IMEI" BIGINT,  
  "A_IMSI" VARCHAR(50),  
  "B_DIRECTION_NUMBER" VARCHAR(50),  
  "B_CELDA" INTEGER,  
  "B_IMEI" BIGINT,  
  "B_IMSI" VARCHAR(50),  
  "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  
  
  ) partitioned by (CALL_START_TIME_FECHA INT)  
  clustered by (A_DIRECTION_NUMBER,B_DIRECTION_NUMBER)  
  into 48 buckets stored as rcfile 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  
  
  ) ROWFORMAT DELIMITED FIELDSTERMINATED BY ','  
  LINES TERMINATED BY '\n'  
  STORED AS TEXTFILE ALIAS HIVE  
  METADATA "DEFAULT"."OTC_T_CDRSTXT_F2";
```

CAPTURAS DEL DESARROLLO DE LOS ETLs

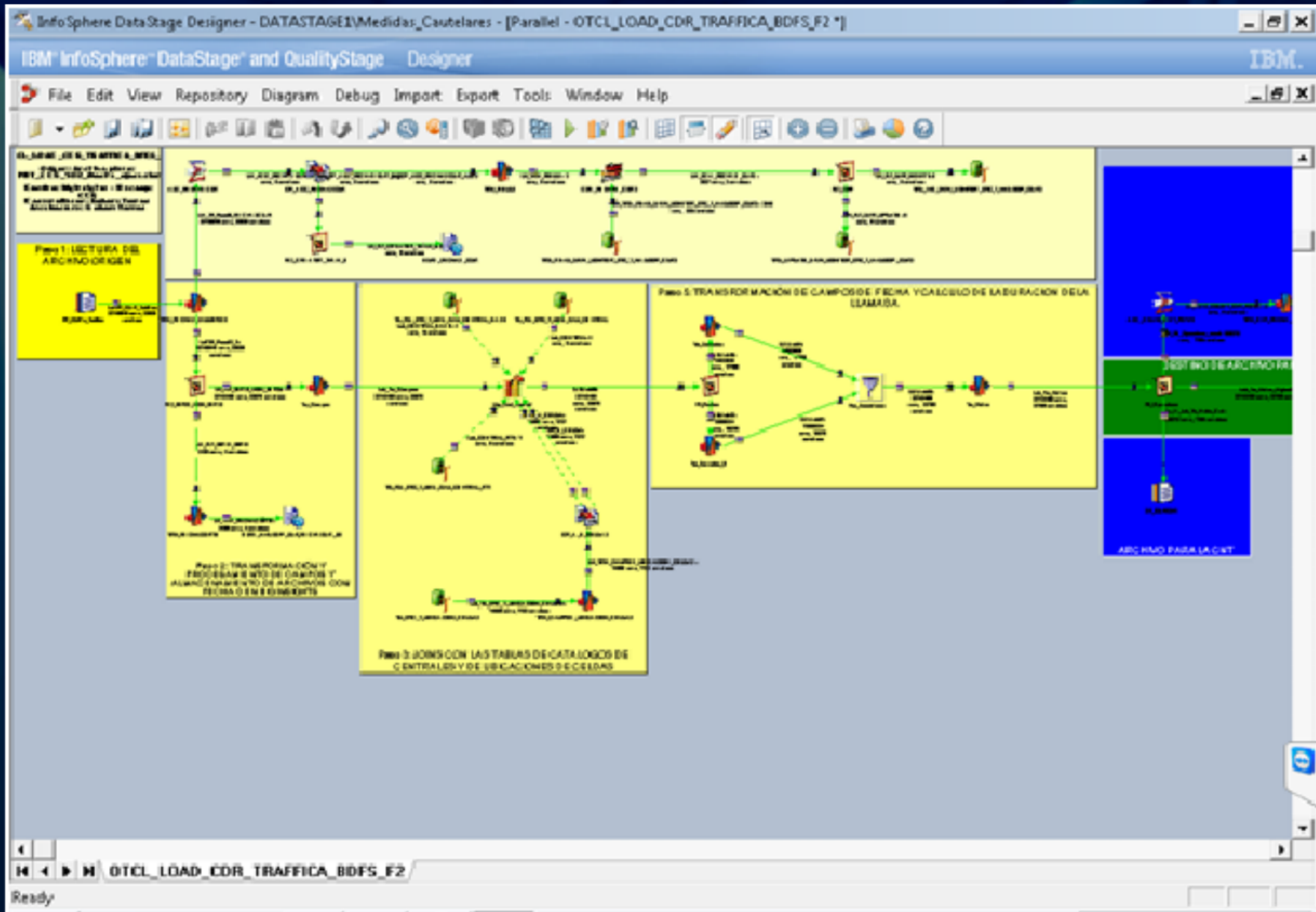
▪ ETL_SECUENCE_JOB_BIG DATA_TELCO



- 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



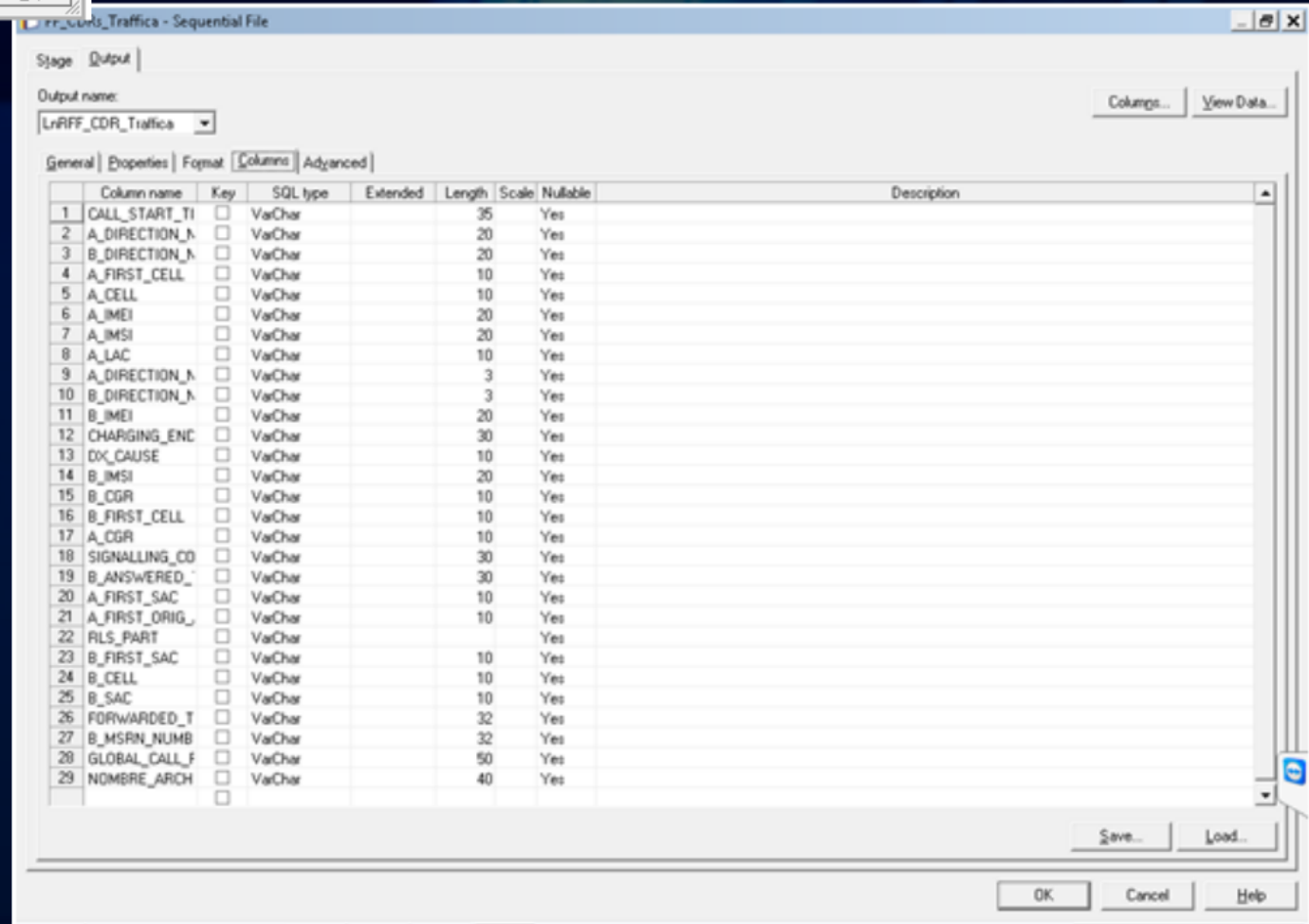
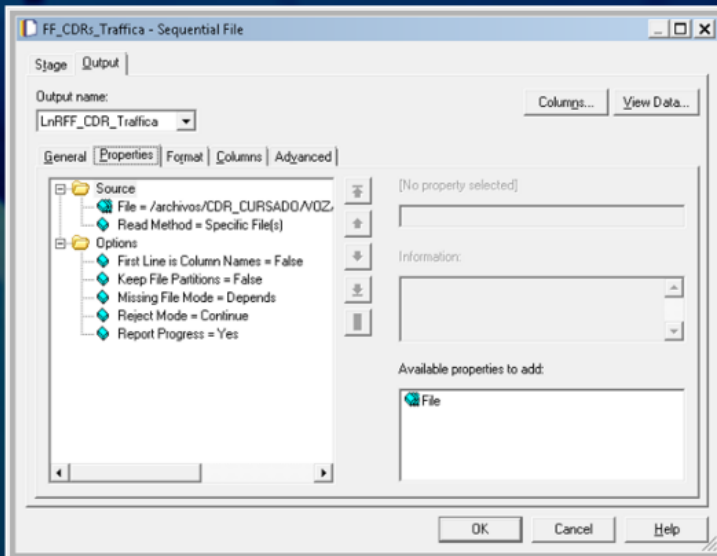
- ETL_PARALLEL_JOB_BIG DATA_TELCOCDRS



EXTRACCIÓN

COLUMNAS

PROPIEDADES



TRANSFORMACIÓN

Trn_Campos - Transformer Stage

Constraint:

Derivation	Column Name
LN_FLT_ZERO_NON_ZERO.CALL_START_	CALL_START_TIME_
LN_FLT_ZERO_NON_ZERO.A_DIRECTION_	A_DIRECTION_NUME
IF LN_FLT_ZERO_NON_ZERO.A_FIRST_CE	A_CELDA
#LN_FLT_ZERO_NON_ZERO.A_IMEI = * T	A_IMEI
#LN_FLT_ZERO_NON_ZERO.A_IMSI = * T	A_IMSI
LN_FLT_ZERO_NON_ZERO.B_DIRECTION_	B_DIRECTION_NUME
IF LN_FLT_ZERO_NON_ZERO.B_FIRST_CE	B_CELDA
#LN_FLT_ZERO_NON_ZERO.B_IMEI = * T	B_IMEI
#LN_FLT_ZERO_NON_ZERO.B_IMSI = * T	B_IMSI
IF TRIM(LN_FLT_ZERO_NON_ZERO.FORw	FORWARDED_TO_N
LN_FLT_ZERO_NON_ZERO.DX_CAUSE	DX_CAUSE
#LN_FLT_ZERO_NON_ZERO.SIGNALLING	SIGNALLING_COMPL
#LN_FLT_ZERO_NON_ZERO.B_ANSWERE	B_ANSWERED_TIME
#NullOfValue(LN_FLT_ZERO_NON_ZERO.C	CHARGING_END_TI
TRIM(LN_FLT_ZERO_NON_ZERO.GLOBAL	GLOBALCALLREFNU
TRIM(LN_FLT_ZERO_NON_ZERO.GLOBAL	GLOBAL_CALL_REF_
tim(LN_FLT_ZERO_NON_ZERO.NOMBRE_	CENTRAL
tim(LN_FLT_ZERO_NON_ZERO.NOMBRE_	CENTRAL_BASE
tim(LN_FLT_ZERO_NON_ZERO.NOMBRE_	CENTRAL_RTT
LN_FLT_ZERO_NON_ZERO.CALL_START_	CALL_START_TIME_

LN_FLT_ZERO_NON_ZERO

Column name	Key	SQL type	Extended	Length
1 CALL_START_TI	<input type="checkbox"/>	Integer		
2 CALL_START_TI	<input type="checkbox"/>	Integer		
3 A_DIRECTION_N	<input type="checkbox"/>	VarChar		20
4 A_FIRST_CELL	<input type="checkbox"/>	VarChar		10
5 A_FIRST_SAC	<input type="checkbox"/>	VarChar		10
6 A_IMEI	<input type="checkbox"/>	VarChar		20

Lnk_Trn_Campos

Column name	Key	SQL type	Extended	Length	Scale	Nullable	Description
1 CALL_START_TI	<input type="checkbox"/>	Integer				No	
2 A_DIRECTION_N	<input type="checkbox"/>	Bigint				Yes	
3 A_CELDA	<input type="checkbox"/>	Integer				Yes	
4 A_IMEI	<input type="checkbox"/>	Bigint				Yes	
5 A_IMSI	<input type="checkbox"/>	VarChar		50		Yes	
6 B_DIRECTION_N	<input type="checkbox"/>	Bigint				Yes	

OK Cancel Help

CARGA

The image displays a terminal window and the IBM InfoSphere DataStage Designer interface. The terminal window shows a shell script for loading data from a Hadoop file into a Hive table. The script uses `LOAD DATA INPATH` to load `CDR#1F2.out` into `otc_t_cdraf2_txt`, and a `while read line` loop to insert each line into the table. The `done` command is followed by `stage?PARAM.out` and a `hadop fs -rm` command to delete the source file.

```
biadmin@quisrvmed4:~/Archivos/Scripts - Xshell 4
File Edit View Tools Window Help
New Reconnect
otecel-quisrvmed4
~/bin/bash
PARAM=#1
/opt/iba/biginsights/hive/bin/hive -e "LOAD DATA INPATH '/user/biadmin/Archivos/Trafica/CDRs/CDR#1F2.out' INTO TABLE otc_t_cdraf2_txt"
hadop fs -cat /user/biadmin/Archivos/Trafica/CDRs/Fechas%cdr/CDR_FECHAS%PARAM.out >> stage?PARAM.out
while read line
do
/opt/iba/biginsights/hive/bin/hive -e "INSERT INTO TABLE DEFAULT.OTC_T_CDRAF2 PARTITION (CALL_START_TIME_FECHA=$line) SELECT CALL_START_TIME_M
ORA,A_DIRECTION_NUMBER,A_CELDA,A_IMEI,A_INSI,B_DIRECTION_NUMBER,B_CELDA,B_IMEI,B_INSI,FORWARDED_TO_NUMBER,DX_CAUSE,GLOBALCALLREFNUMCENOP,GLOBA
L_CALL_REF_SEP,COD_CENTRAL,COD_CENTRAL_BASE,COD_CENTRAL_RTT,FECHAMORA,CHARGINIG_END_TIME_FECHA,CHARGINIG_END_TIME_HORA,DURACION FROM DEFAULT.O
TC_T_CDRAF2_TXT where CALL_START_TIME_FECHA=$line"
done < stage?PARAM.out
hadop fs -rm /biginsights/hive/warehouse/otc_t_cdraf2_txt/CDR#1F2.out
```

The IBM InfoSphere DataStage Designer window shows a job design for "OTCL_LOAD_CDR_TRAFFICA_BDFS_F2". The job is titled "AUDITORIA DE CARGA DE LOS" and "DESTINO DE ARCHIVO PARA BIGINSIGHTS". It includes stages for "IN_FUT_OFF_INSERT", "IN_FUT_OFF_UPDATE", "IN_OTC_T_HADOOP_CDRAF2", "ADD_COUNT_COLUMN", "INSERT_DATE_TIME_DURATION", "COPY_OF_BDFS_HADOOP_CDR_FECHACDR", and "COPY_OF_BDFS_HADOOP_CDR_FECHACDR". The job is currently running, as indicated by the "Ready" status at the bottom.

Connected to 10.112.152.166:22.

SSH2 xterm 142x39 1,1 1 session CAP NUM



EXPLORACIÓN DE ARCHIVOS PLANOS Y RCFILES

The screenshot displays the IBM InfoSphere BigInsights Standard Edition web interface. The browser address bar shows the URL: `quisrvmed4.otecel.com.ec:8080/data/html/index.html#redirect-files`. The page title is "IBM InfoSphere BigInsights Standard Edition" and the user is logged in as "biadmin".

The navigation menu includes: Welcome, Dashboard, Cluster Status, **Files**, Applications, Application Status, and BigSheets.

The main content area is divided into two panes:

- DFS Files:** A file explorer showing the directory structure. The path is `hdfs://quisrvmed4.otecel.com.ec:9000r/biginsights/hive/warehouse/otc_t cdrsf2`. The selected directory is `otc_t cdrsf2`.
- File Details:** A table showing the details of the selected directory.

Name	Size	Block Size	Permission	Owner
otc_t cdrsf2	355 directories, 0 ...		rwxrwxrwt	bigsql

Additional controls include a "Viewing Size" dropdown set to 10kB and radio buttons for "Text" (selected) and "Sheet".

CONSULTA DE DATOS EXTRAIDOS EN BIGSQL

The screenshot shows the BigSQL IDE interface. The main window displays a SQL query:

```
select * from default.otc_t_cdrsf2
where call_start_time_fecha=20150628
and a_direction_number like '980515303';
```

Below the query editor, the 'SQL Results' window shows a table with the following data:

Status	Operation	Date	CALL_START_TIME_HORA	A_DIRECTION_NUMBER	A_CELDA	A_IMEI	A_IMSI	B_DIRECTION_NUMBER	B_CELDA	B_IMEI	B
✓	Succee select * fro...	19/04/15 13.	1	142055	980515303	-1	-1	995147520	-1	-1	-:
✓	Succee select * fro...	19/04/15 14.	2	142002	980515303	-1	-1	995147520	50156	3589...	7
✗	Failed select * fro...	29/06/15 11.	3	142042	980515303	-1	-1	995147520	-1	-1	-:
✓	Succee select * fro...	29/06/15 11.									
✓	Succee select * fro...	29/06/15 11.									

The screenshot shows a terminal window displaying disk I/O statistics. The output is as follows:

```
Hostname=quisrvmed1 Refresh= 2secs 11:44.28
Disk I/O /proc/diskstats mostly in KB/s Warning:contains duplicates
DiskName Busy Read WriteKB|0 |25 |150 |175 |100|
sda 0% 0.0 193.9|> |
sda1 0% 0.0 0.0|> |
sda2 0% 0.0 0.0|> |
sda3 0% 0.0 193.9|> |
sda4 0% 0.0 0.0|> |
Totals Read-MB/s=0.0 Writes-MB/s=0.4 Transfers/sec=65.0
```

Task Launcher *Script5.sql

Connection: Otecel - BIGSQL 3.0 - default [bigsql]

```
select * from default.otc_t_cdnsf2
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

Status	Operation	Date	VL_BASE	COD_CENTRAL_RTT	FECHAHORA	CHARGINIG_END_TIME_FECHA	CHARGINIG_END_TIME_HORA	DURACION	CALL_STAI
✓	Succee select * fro...	19/04/15 13.	11		2015062682...	20150626	82148	22	20150626
✓	Succee select * fro...	19/04/15 14.	11		2015062814...	0	0	0	20150628
✗	Failed select * fro...	29/06/15 11.	11		2015062814...	0	0	0	20150628
✓	Succee select * fro...	29/06/15 11.	11		2015062814...	0	0	0	20150628
✓	Succee select * fro...	29/06/15 11.							
✓	Succee select * fro...	29/06/15 11.							

srvmed4 otecel-quisrvmed4 otecel-quisrvmed4 otecel-dataStageProducción otecel-quisrvmed1 otecel-quisrvmed2 otecel-quisrvmed3

```
hmon-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 WriteKB |0 125 150 175 100|
sda 100% 116.0 0.0|#####>
sda1 0% 0.0 0.0|>
sda2 0% 0.0 0.0|>
sda3 100% 116.0 0.0|#####>
sda4 0% 0.0 0.0|>
Totals Read-KB/s=232.0 Writes-KB/s=0.0 Transfers/sec=1775.9
```

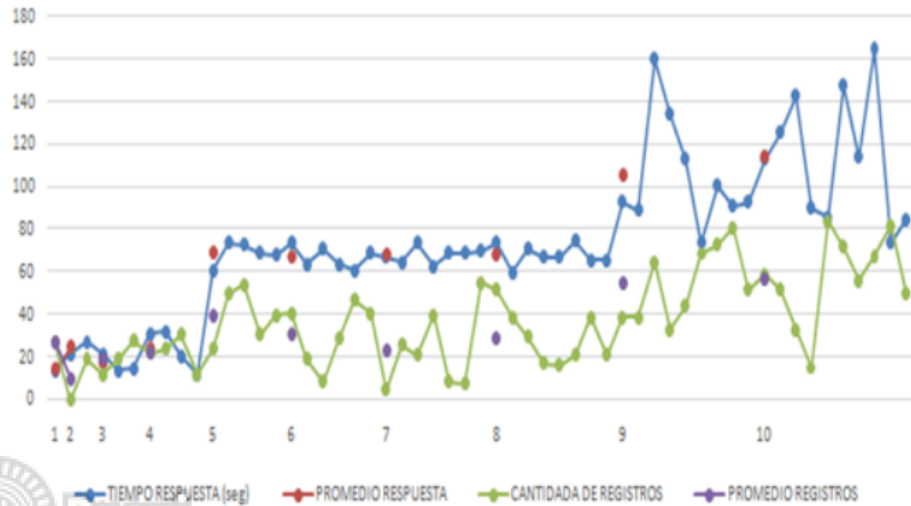

SEGURIDADES

The screenshot shows the IBM InfoSphere BigInsights Standard Edition interface. The top navigation bar includes 'Welcome', 'Dashboard', 'Cluster Status', 'Files', 'Applications', 'Application Status', and 'BigSheets'. The 'Files' tab is active, displaying a file browser for the path '/biginsights/hive/warehouse/otc_t_cdsrcf2/call_start_time_fecha=20150628/000000_0'. A table view shows the following data:

Name	Size	Block Size	Permission	Owner
000000_0	216.7 MB	128.0 MB	rwxrwxrwt	bladrin

Below the table, a preview of the file content is shown, displaying a large number of records with columns for 'hive.io.rcfile.column.number' and 'I'.

PRUEBAS DE EJECUCIÓN - TIEMPOS DE RESPUESTA Y CANTIDAD DE REGISTROS



PRUEBAS DE EJECUCIÓN



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 particionada 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 se tiene pérdida de servicio, considerando que la probabilidad de fallo es de 1 a 10000 y que se diseña para que los nodos de datos tengan un balanceo de carga y colas de procesos tras falla de uno o más nodos de datos, además, envía alertas 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 recursos 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

El desarrollo de la parte del Sistema de seguridad, se realizó de acuerdo con los requisitos de funcionalidad y de seguridad requeridos, así como a la estructura de flujo de datos de la información que se maneja en el sistema.

Big Data es una tecnología que permite el almacenamiento de información, procesamiento y análisis de grandes volúmenes de datos en tiempo real, que se genera a una velocidad que supera a la capacidad de los sistemas tradicionales de almacenamiento de datos.

El uso de Big Data en la industria de seguros, permite a las compañías de seguros analizar los datos de los clientes y mejorar sus servicios, así como a los asegurados obtener mejores condiciones de pólizas y coberturas.

El uso de Big Data en la industria de seguros, permite a las compañías de seguros analizar los datos de los clientes y mejorar sus servicios, así como a los asegurados obtener mejores condiciones de pólizas y coberturas.

RECOMENDACIONES

- Se recomienda que los parámetros que implementan una solución de Big Data, tengan en cuenta el hardware, software, estándares, bases de datos e infraestructura para tener una solución adecuada.
- Para que se realice la información de Big Data, se recomienda tener como mínimo de 100 GB de datos, para poder realizar consultas y análisis de datos, especialmente en consultas complejas de gran volumen. Si se necesita un volumen de datos menor, se puede utilizar un sistema de almacenamiento de datos en la nube, como Amazon S3, que permite almacenar grandes volúmenes de datos de manera segura y accesible.
- Es necesario tener un hardware adecuado para procesar y almacenar grandes volúmenes de datos, especialmente en consultas complejas de gran volumen. Se debe tener en cuenta el costo de adquisición y mantenimiento del hardware, así como el costo de energía y refrigeración.
- Big Data no es una solución única para todos los problemas. Se debe evaluar cuidadosamente los requisitos de cada caso de uso, y seleccionar la solución más adecuada para cada caso.



TRABAJO DE GRADUACIÓN PARA OBTENER EL TÍTULO DE INGENIERO EN SISTEMAS DE COMPUTACIÓN
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
PROYECTO DE TESIS
JUNIO DEL 2015

CONSTRUCCIÓN DE UN REPOSICIONADOR PARA AYUDAR DE LA EXTRACCIÓN Y PROCESAMIENTO DE LA INFORMACIÓN PROVENIENTE DE COMS GARANTADOS POR TELCELS PARA ANALIZAR EL COMPORTAMIENTO DE LOS DATOS DENTRO DE LA EMPRESA SOFTCONSULTING SA UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM

