**MANUAL DE INSTALACIÓN Y MONTAJE DE BASES DE DATOS NOSQL (COUCHDB Y ELASTICSEARCH) DE LA DATA PARA LOS APLICATIVOS DEL CLÚSTER DE SOLUCIONES NUVEM CNCFLORA BRASIL**

VERSION 1.0 – 01-12-2016

Elaborado Por: Jaime Alberto Gutiérrez Mejía

# ÍNDICE

[ÍNDICE 2](#_Toc472954621)

[1. OBJETO DEL DOCUMENTO 3](#_Toc472954622)

[2. PARTICIPANTES 3](#_Toc472954623)

[3. OBJETIVOS 3](#_Toc472954624)

[4. PREPARACIÓN DE LOS DATOS PARA EL PROCESO MIGRATORIO 4](#_Toc472954625)

[5. EJECUCIÓN DEL CLIENTE DE GENERACIÓN DE ARCHIVOS DE CARGUE MASIVO EN COUCHDB Y ELASTICSEARCH 9](#_Toc472954626)

[6. REALIZACIÓN DEL CARGUE AL AMBIENTE DE BASES DE DATOS NO RELACIONALES 23](#_Toc472954627)

# OBJETO DEL DOCUMENTO

Este documento tiene como objeto ilustrar cómo debe realizarse el proceso de cargue de bases de datos noSQL (COUCHDB y ELASTICSEARCH) del conjunto de aplicaciones del clúster NUVEM CNCFLORA BRASIL, con el fin de poder registrar de forma automatizada para el core de soluciones de backoffice de la solución, los datos de las correspondientes bases de categorías (familias) de registros de especies biológicas.

# PARTICIPANTES

En este proceso tiene principal participación el Administrador Técnico del Sistema, el jefe de gestión de aplicaciones del Instituto Alexander von Humboldt y los participantes del proceso de despliegue y paso a producción.

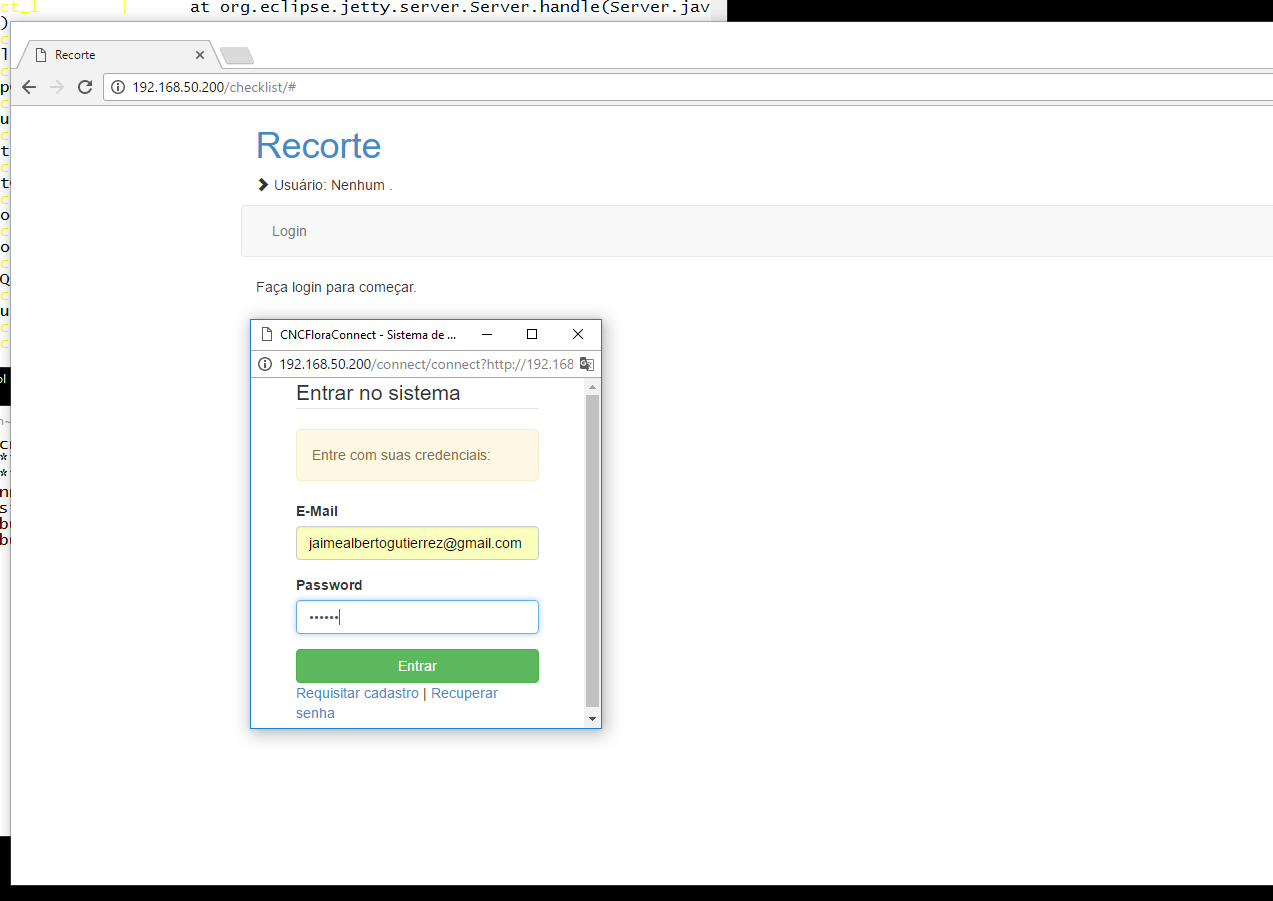
# OBJETIVOS

1. Realizar la preparación de los datos para el cargue al Sistema
2. Ejecutar el cliente de construcción de los archivos de generación de cargue masiva de registros para COUCHDB y la indización de los mismos con Elasticsearch.
3. Desarrollar el proceso de cargue de los datos desde la consola del Sistema operativo del ambiente de ejecución
4. Validar la data cargada en los aplicativos del clúster NUVEM

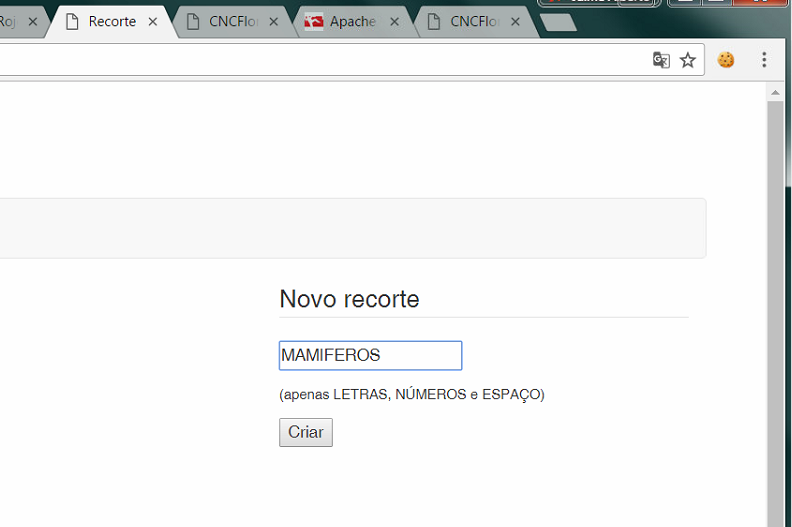
# PREPARACIÓN DE LOS DATOS PARA EL PROCESO MIGRATORIO

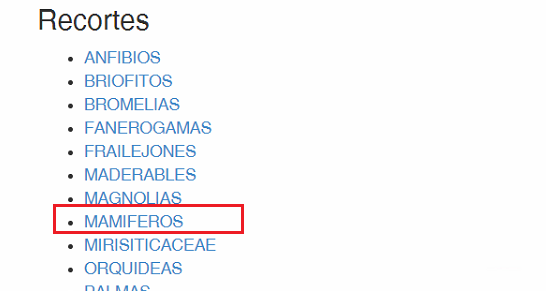
Con el objeto de desarrollar el proceso de migración de datos a la solución de NUVEM CNCFLORA, el usuario deberá preparar la data de las diferentes bases de datos, previa creación de las bases concretas (denominadas en el ambiente de producción RECORTES). Para ello, se deberá ingresar al Sistema por alguna de las dos siguientes vías:

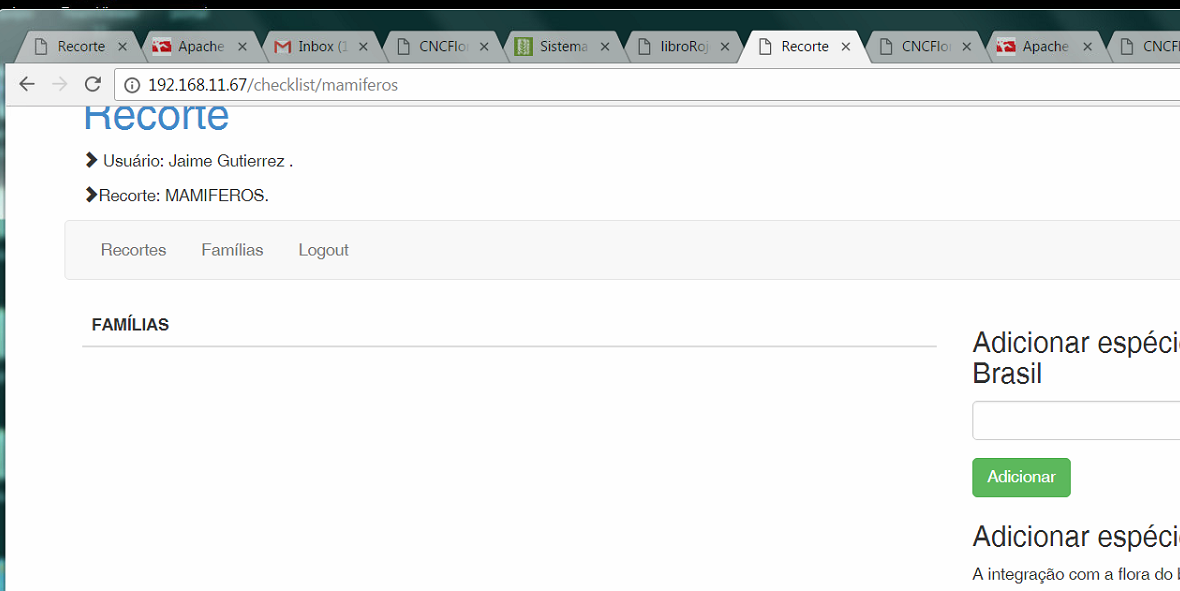
1. **VÍA ADMINISTRACIÓN CHECKLIST:** Para construir las bases requeridas, el usuario deberá ingresar por el aplicativo Web de LISTA DE CHEQUEO (Checklist) <http://192.168.11.67/checklist>, autenticarse con el usuario previamente creado desde el aplicativo de gestión de SSO (ENTRYPOINT: <http://192.168.11.67/entrypoint/>), tal como se muestra a continuación:

****

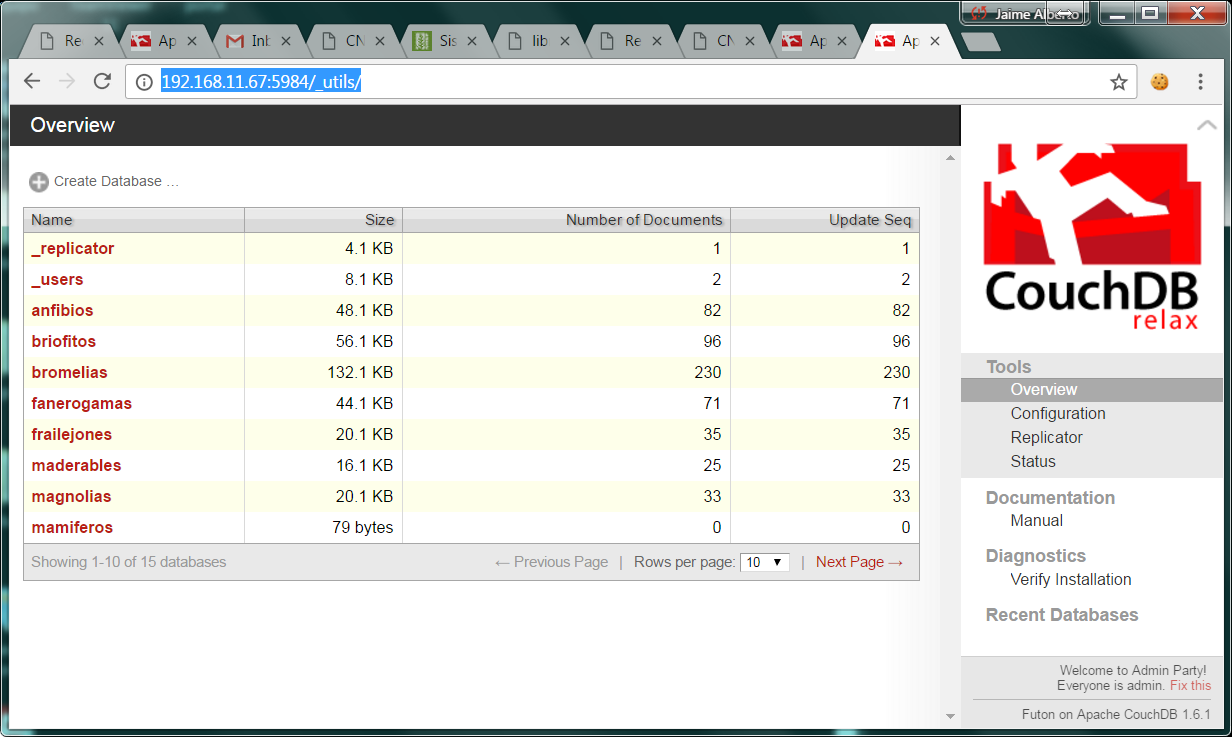
Una vez el usuario se autentique se puede proceder a crear la base, escribiendo el nombre de la misma en el formulario Nuevo Recorte:







1. **VÍA ADMINISTRACIÓN FUTON:** Si fuera a realizarse el proceso directamente desde el gestor de base de datos no relacional de COUCHDB, se deberá realizar inicio de sesión en la consola del administrador de la base (FUTON: <http://192.168.11.67:5984/_utils/>), para poder realizar la creación de las bases de agrupación de especies (familias).

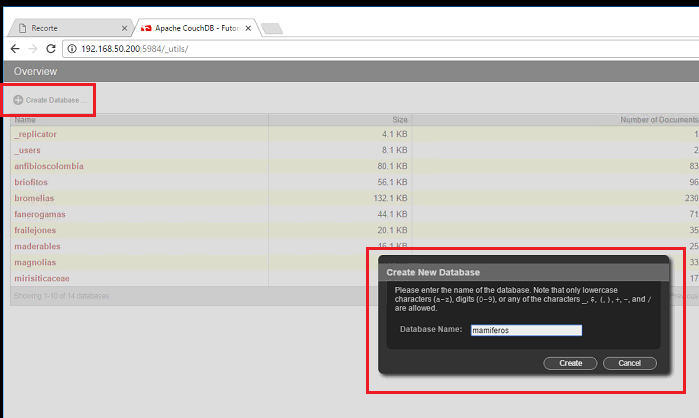
****

Proceder a iniciar sesión en la parte inferior con las credenciales creadas para el administrador. En el caso de este proyecto:

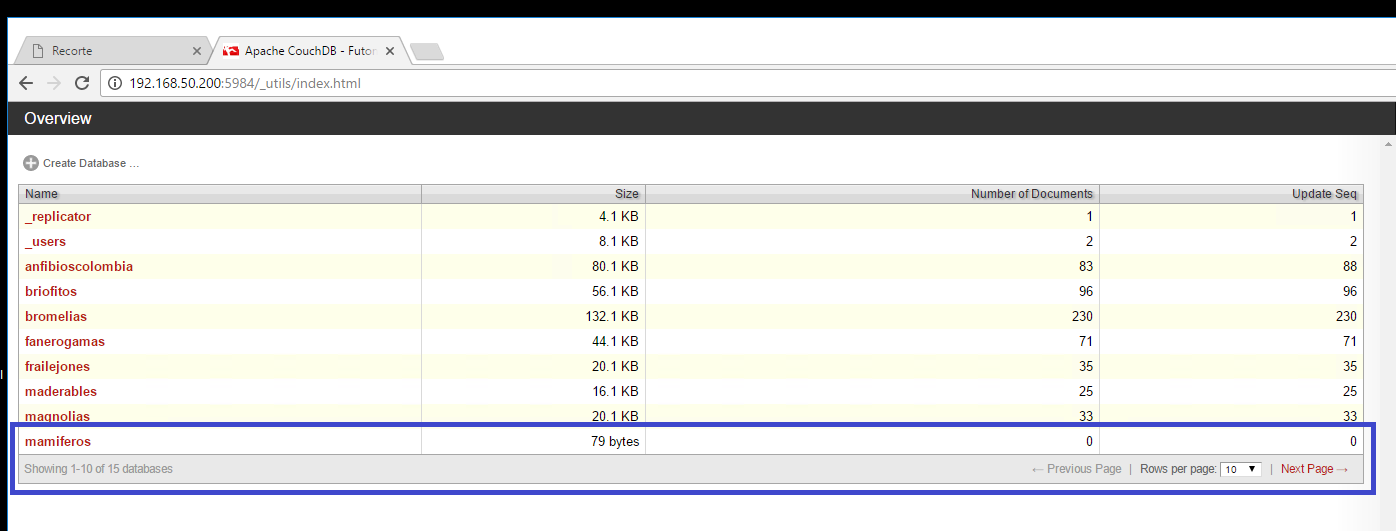
**Username: admin**

**Password: ncc&123x**

Una vez se haya iniciado sesión en el administrador de FUTON, se procede a escoger la opción CREATE DATABASE y se ingresa el nombre del nuevo recorte o base para CouchdDB.



Automáticamente la base quedará creada y lista para la inserción de nuevos registros de especies biológicas.

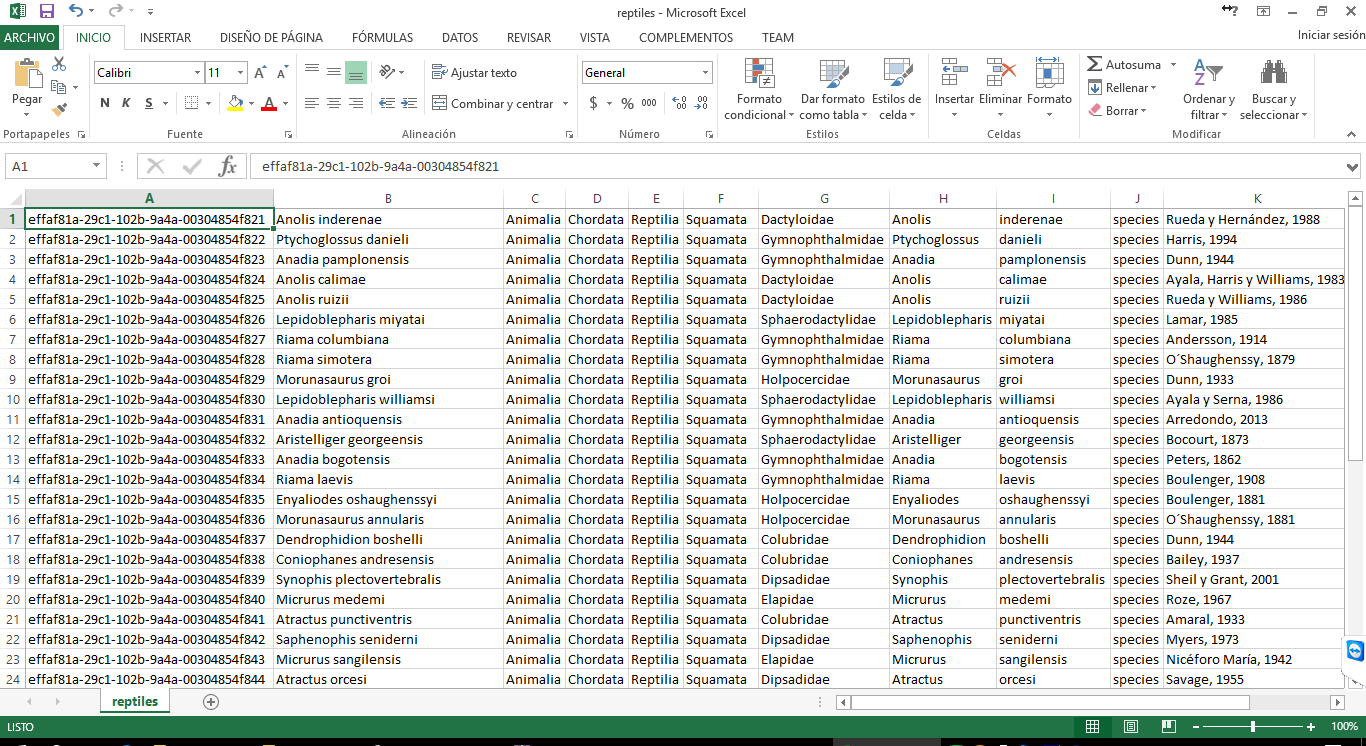


# EJECUCIÓN DEL CLIENTE DE GENERACIÓN DE ARCHIVOS DE CARGUE MASIVO EN COUCHDB Y ELASTICSEARCH

Para el proceso de cargue masivo de registros a las bases de datos del Sistema, se debe preparar la creación de los registros, empleando una plantilla de cargue masiva, la cual incluye los siguientes campos:

|  |  |
| --- | --- |
| **Campo** | **Descripción** |
| id | Identificador único de la Especie |
| nombreCientifico | Nombre científico nominal de la Especie |
| kingdom | Familia taxonómica a la que se asocia la especie |
| phylum | Filum de agrupación de la Especie |
| clase | Clase de categorización |
| orden | Orden de distribución |
| family | Familia de la especie |
| genus | Genus taxonómico |
| specificEpithet | Epíteto de clasificación organizacional |
| rango | Rango de definición (especie, subespecie o derivación) |
| autor | Autor científico de investigación de la especie |
| nombreComun | Nombre común o genérico de la especie |

Al realizar la organización, el documento debe quedar así:



La estructura del ID debe seguir la nomenclatura que se cita a continuación:

**effaf81a-29c1-102b-9a4a-00304854f826**

* Primer bloque: 8 dígitos hexadecimales
* Segundo bloque: 4 dígitos hexadecimales
* Tercer bloque: 4 dígitos hexadecimales
* Cuarto bloque: 4 dígitos hexadecimales
* Quinto bloque: 12 dígitos hexadecimales

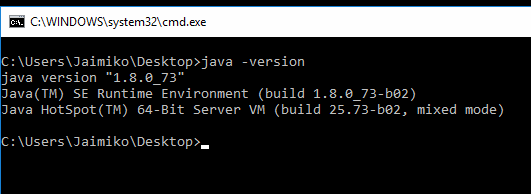
Y cada ID de registro debe ser absolutamente diferente del siguiente, para mantener la integridad de entidad en este modelo.

Una vez el archivo se organice en formato CSV, el paso siguiente, es verificar que no existan los símbolos de tabulación (TAB), punto y coma (;), dado que este último es el separador de campos en el archivo de interface y finalmente las comillas dobles o simples, las cuales deben ser removidas de los registros.

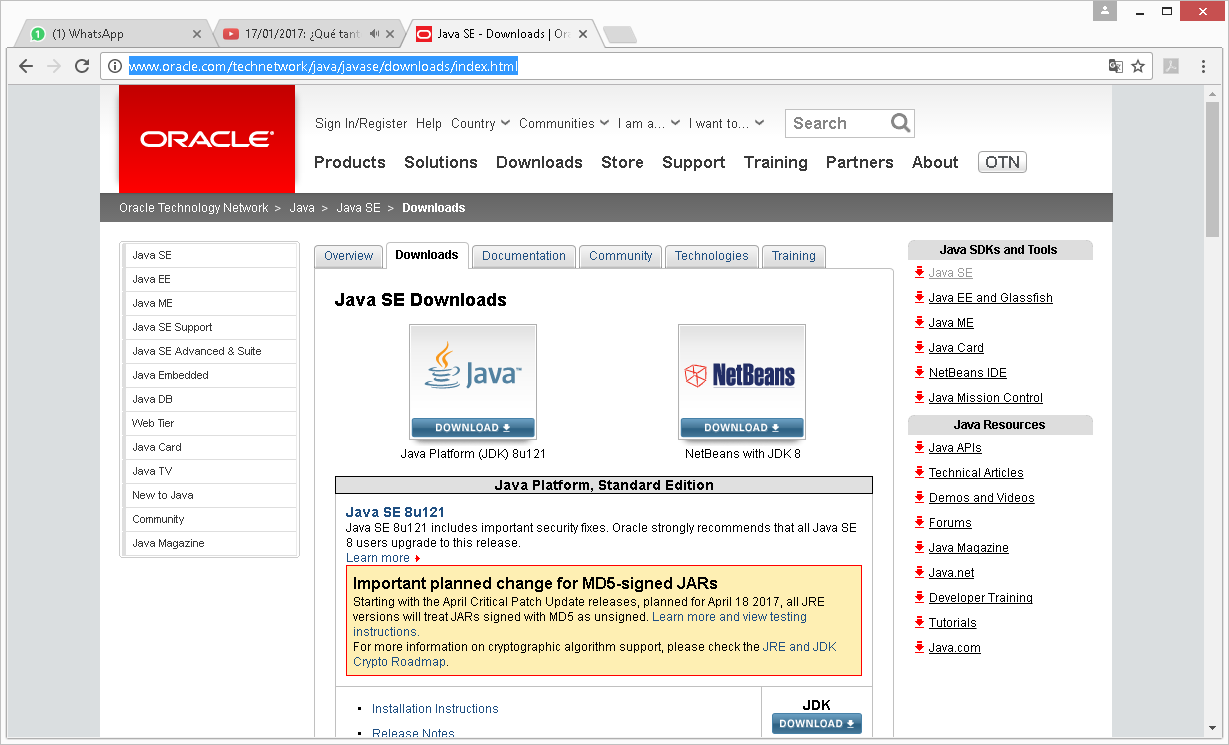
Una vez el archivo esté correspondientemente formado, debe procederse a la ejecución de la aplicación cliente JSE desarrollada para la construcción de forma automatizada de los archivos tanto para COUCHDB como para ElasticSearch. Desde la consola de comandos debe ejecutarse

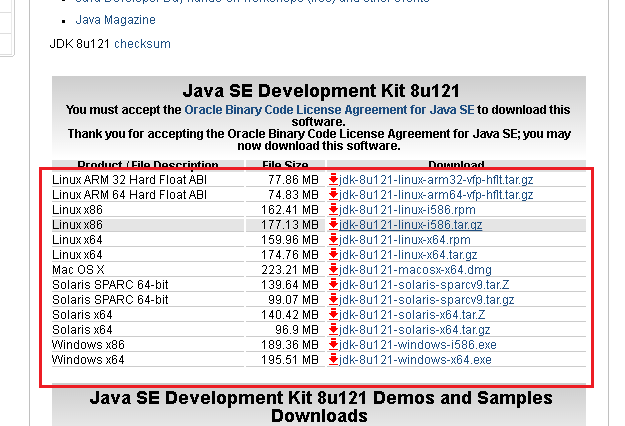
**java -jar GeneradorCouch.jar**

Previa validación de la instalación del ambiente JDK de Oracle y del HotSpot correspondiente para la ejecución de aplicaciones

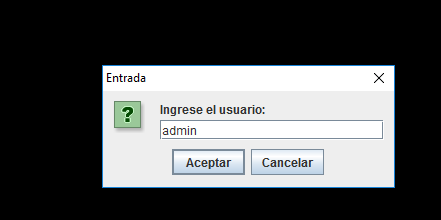


Si no se puede ejecutar el comando, debe descargarse desde la página oficial de Oracle en la sección DOWNLOADS/JSE TECHNOLOGY (<http://www.oracle.com/technetwork/java/javase/downloads/index.html>)

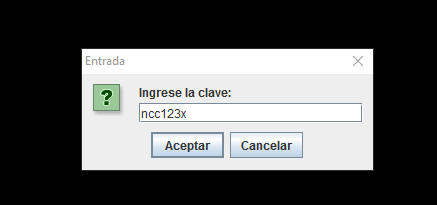




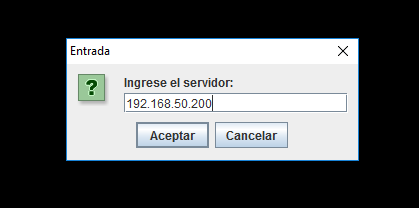
1. Lo primero que hará la aplicación será solicitar el usuario de COUCHDB que está configurado en la consola de FUTON:



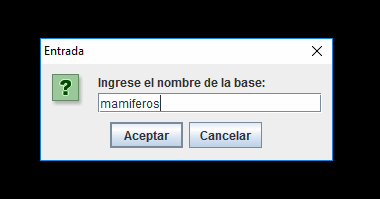
1. Posteriormente se ingresa la contraseña asignada a dicho administrador



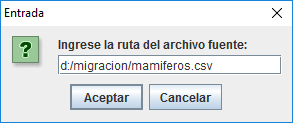
1. Acto seguido debe ingresarse la dirección IP o nombre DNS del servidor que aloja el servicio de CouchDB (No debe ingresarse el puerto, ya que este por defecto está configurado en el 5498).



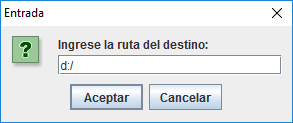
1. El aplicativo solicitará el nombre de la base (Esta se corresponde al nombre exacto (en letras minúsculas del índice o relatorio creado en COUCHDB) o a través del aplicativo de CHECKLIST.



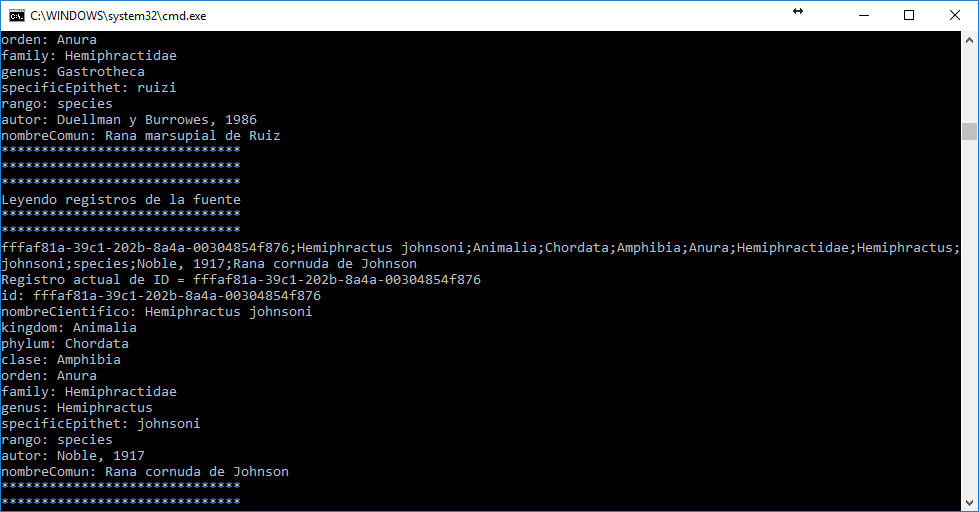
1. El aplicativo solicitará la ruta completa (Linux o Windows) de la ubicación del archivo fuente en formato CSV que contiene el listado completo de todas las especies. Se deberá digitar, como en el ejemplo que se ilustra a continuación:



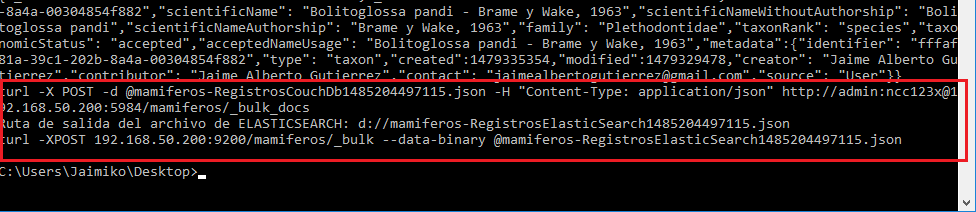
1. Posteriormente, digitar el path o ruta de destino en donde va a quedar el archivo destino tanto en COUCHDB como en Elasticsearch.



1. El aplicativo realiza todo el procesamiento y genera la salida en la carpeta destino de los archivos de la forma NOMBRE\_FUENTE\_RegistrosCouchDBXXXXXXXXXXXXXXXXXX.json y NOMBRE\_FUENTE\_RegistrosElasticXXXXXXXXXXXXXXXXXXXXX.json, donde NOMBRE\_FUENTE representa el nombre del archivo fuente de origen y XXXXXXXXXXXXXXXXXXXXXXXX un serial autonumérico extraído de la fecha hora del Sistema.



1. Al final de la ejecución de la aplicación, se generarán los correspondientes formatos de cargue de consola CURL, con su notación de acuerdo a los parámetros de entrada:



Ejemplo:

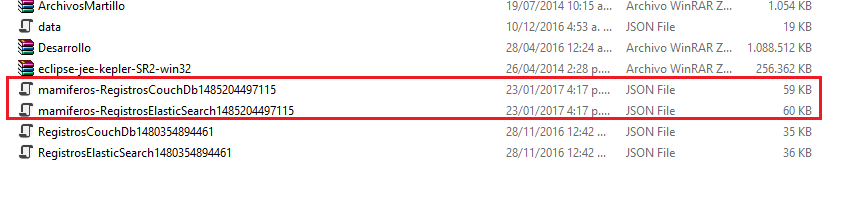
Cargue para COUCHDB:

**curl -X POST -d @mamiferos-RegistrosCouchDb1485204497115.json -H "Content-Type: application/json"** [**http://admin:ncc123x@192.168.50.200:5984/mamiferos/\_bulk\_docs**](http://admin:ncc123x@192.168.50.200:5984/mamiferos/_bulk_docs)

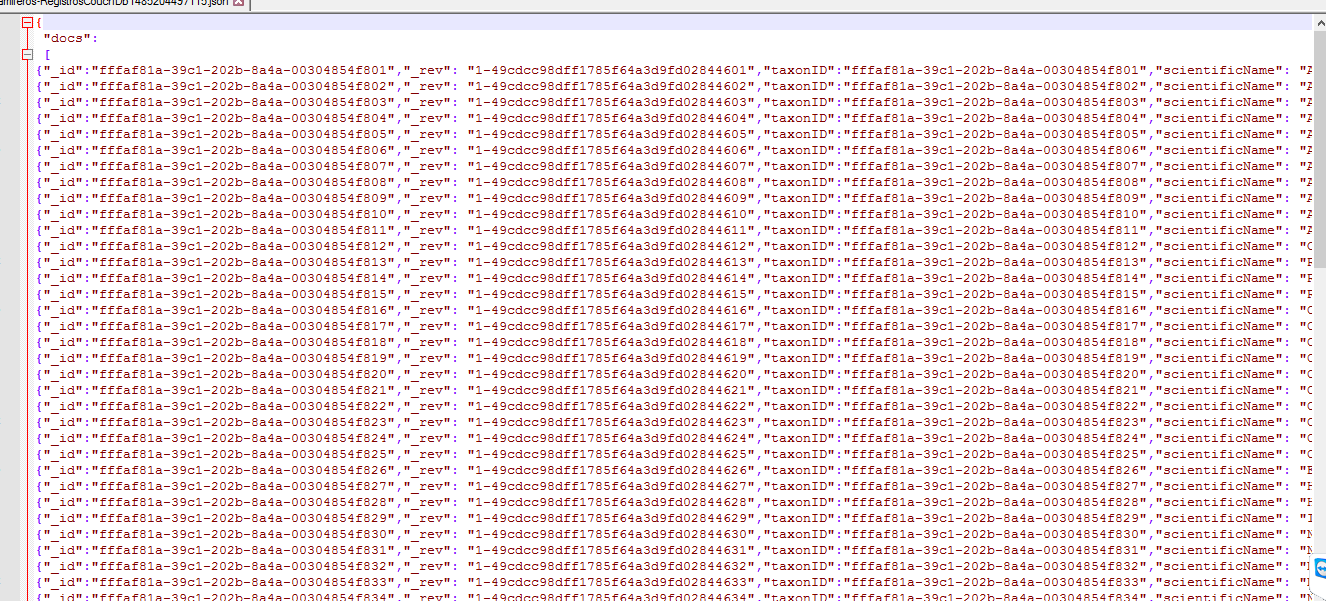
Cargue del Índice para ELASTICSEARCH:

**curl -XPOST 192.168.50.200:9200/mamiferos/\_bulk --data-binary @mamiferos-RegistrosElasticSearch1485204497115.json**

1. Los archivos resultantes en formato JSON (JAVASCRIPT NOTATION OBJECT), aparecerán en la carpeta de destino:

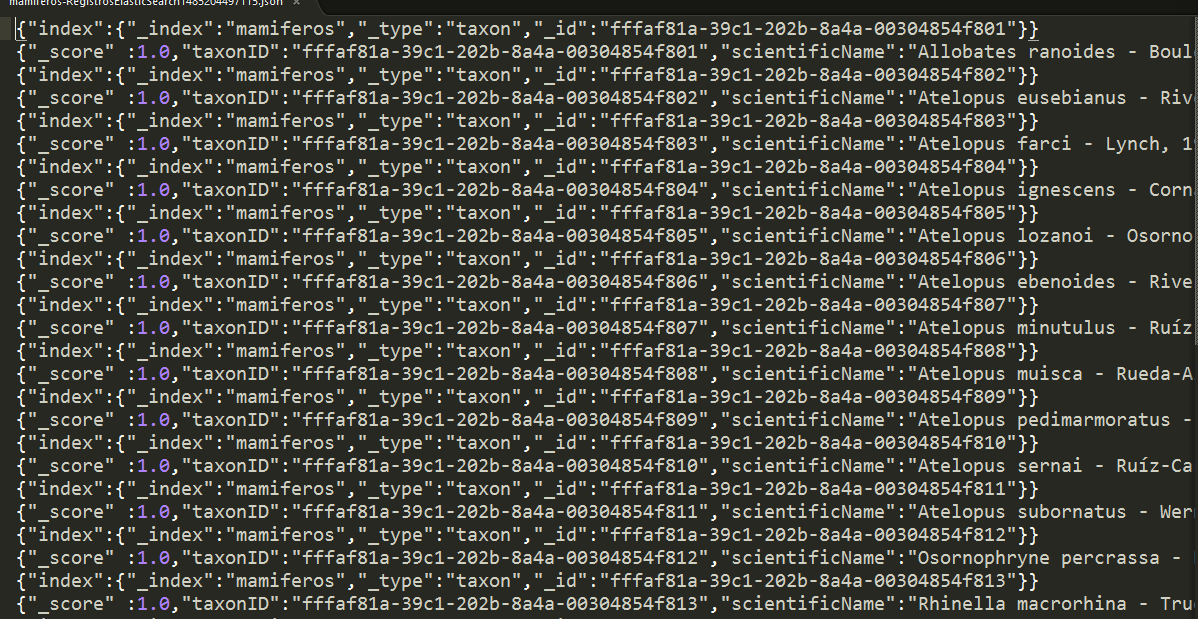


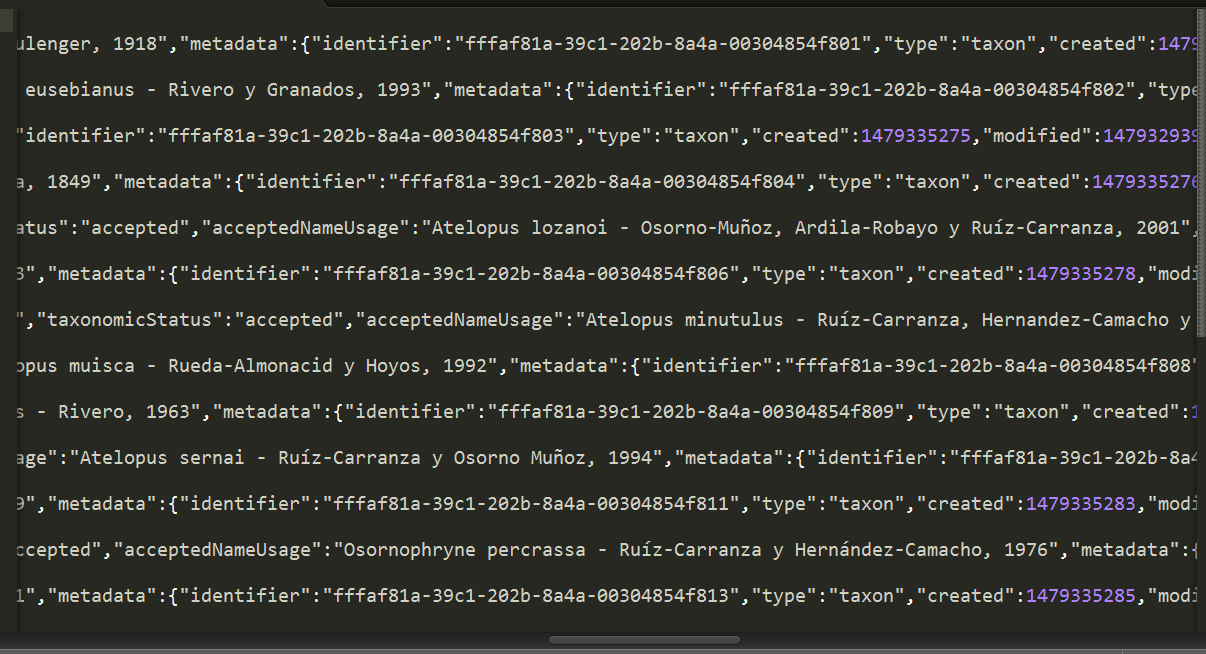
**Archivo de cargue COUCHDB:**





**Archivo de cargue ELASTICSEARCH:**





A continuación se presenta la estructura de cada entrada en el archivo de registro de alimentación de la base para COUCHDB:

|  |
| --- |
| {  "\_id":"fffaf81a-39c1-202b-8a4a-00304854f801",  "\_rev": "1-49cdcc98dff1785f64a3d9fd02844601",  "taxonID":"fffaf81a-39c1-202b-8a4a-00304854f801",  "scientificName": "Allobates ranoides - Boulenger, 1918",  "scientificNameWithoutAuthorship": "Allobates ranoides",  "scientificNameAuthorship": "Boulenger, 1918",  "family": "Aromobatidae",  "taxonRank": "species",  "taxonomicStatus": "accepted",  "acceptedNameUsage": "Allobates ranoides - Boulenger, 1918",  "metadata":{  "identifier": "fffaf81a-39c1-202b-8a4a-00304854f801",  "type": "taxon",  "created":1479335273,  "modified":1479329397,  "creator": "Jaime Alberto Gutierrez",  "contributor": "Jaime Alberto Gutierrez",  "contact": "jaimealbertogutierrez@gmail.com",  "source": "User"}  } |

|  |  |
| --- | --- |
| **Campo** | **Descripción** |
| id | Identificador único de la Especie |
| \_rev | Identificador del proceso de revisión |
| taxonID | Id del taxón asociado a la especie |
| scientificName | Nombre científico nominal de la Especie |
| scientificNameWithoutAuthorship | Nombre científico de la especie sin autor asociado |
| scientificNameAuthorship | Nombre científico dado por el investigador autor |
| family | Familia taxonómica a la que se asocia la especie |
| taxonRank | Rango del taxón |
| taxonomicStatus | Status taxonómico (aceptado o rechazado) |
| acceptedNameUsage | Nombre definido aceptado para la especie |
| identifier | Identificador general de metadatos de registro |
| type | Tipo (Especie, subespecie o derivación) |
| created | Fecha de creación del registro |
| modified | Fecha de modificación del registro |
| creator | Usuario responsable de la creación/modificación del registro |
| contributor | Persona responsable de la catalogación de la especie |
| contact | Correo electrónico o username de contacto del responsable del cargue |
| source | Fuente primaria de origen de la información de la especie |

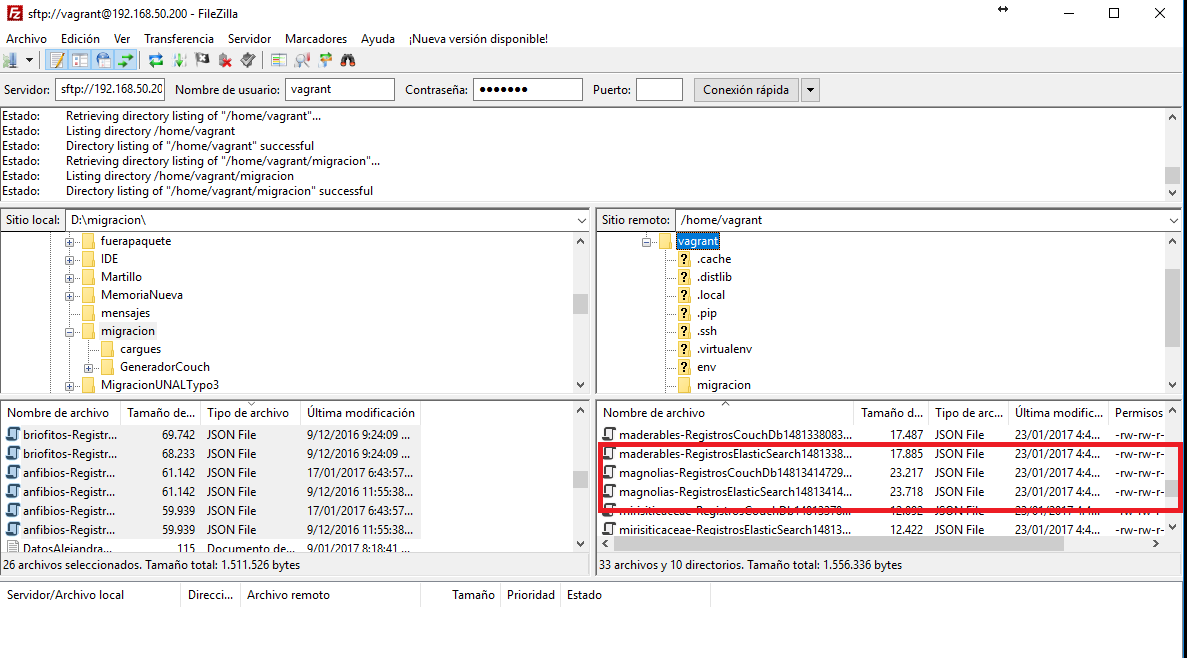
Dentro de la estructura de indización para cada registro, se presenta la forma del archivo de cargue masivo de Elastic, de la siguiente manera:

|  |
| --- |
| {  "index":  {"\_index":"mamiferos",  "\_type":"taxon",  "\_id":"fffaf81a-39c1-202b-8a4a-00304854f801"}  }  {  "\_score" :1.0,  "taxonID":"fffaf81a-39c1-202b-8a4a-00304854f801",  "scientificName":"Allobates ranoides - Boulenger, 1918",  "scientificNameWithoutAuthorship":"Allobates ranoides",  "scientificNameAuthorship":"Boulenger, 1918",  "family":"Aromobatidae",  "taxonRank":"species",  "taxonomicStatus":"accepted",  "acceptedNameUsage":"Allobates ranoides - Boulenger, 1918",  "metadata":  {"identifier":"fffaf81a-39c1-202b-8a4a-00304854f801",  "type":"taxon",  "created":1479335273,  "modified":1479329397,  "creator":"Jaime Alberto Gutierrez",  "contributor":"Jaime Alberto Gutierrez",  "contact":"jaimealbertogutierrez@gmail.com",  "source":"User"},"rev":null  } |

Que cómo se observa, comparte los campos comunes con el archivo de COUCHDB.

# REALIZACIÓN DEL CARGUE AL AMBIENTE DE BASES DE DATOS NO RELACIONALES

Una vez se hayan realizado los procesos de generación de cargue de todas las bases para las familias de especies, el administrador debe realizar el cargue de los archivos mediante su herramienta favorita de SSH y administración remota de archivos. Se recomienda usar FILEZILLA para ello.



Una vez los archivos se encuentren cargados, se debe ingresar por la consola y realizar el proceso de cargue con los comandos CURL generados por el aplicativo de transformación y migración explicado en el punto 6.



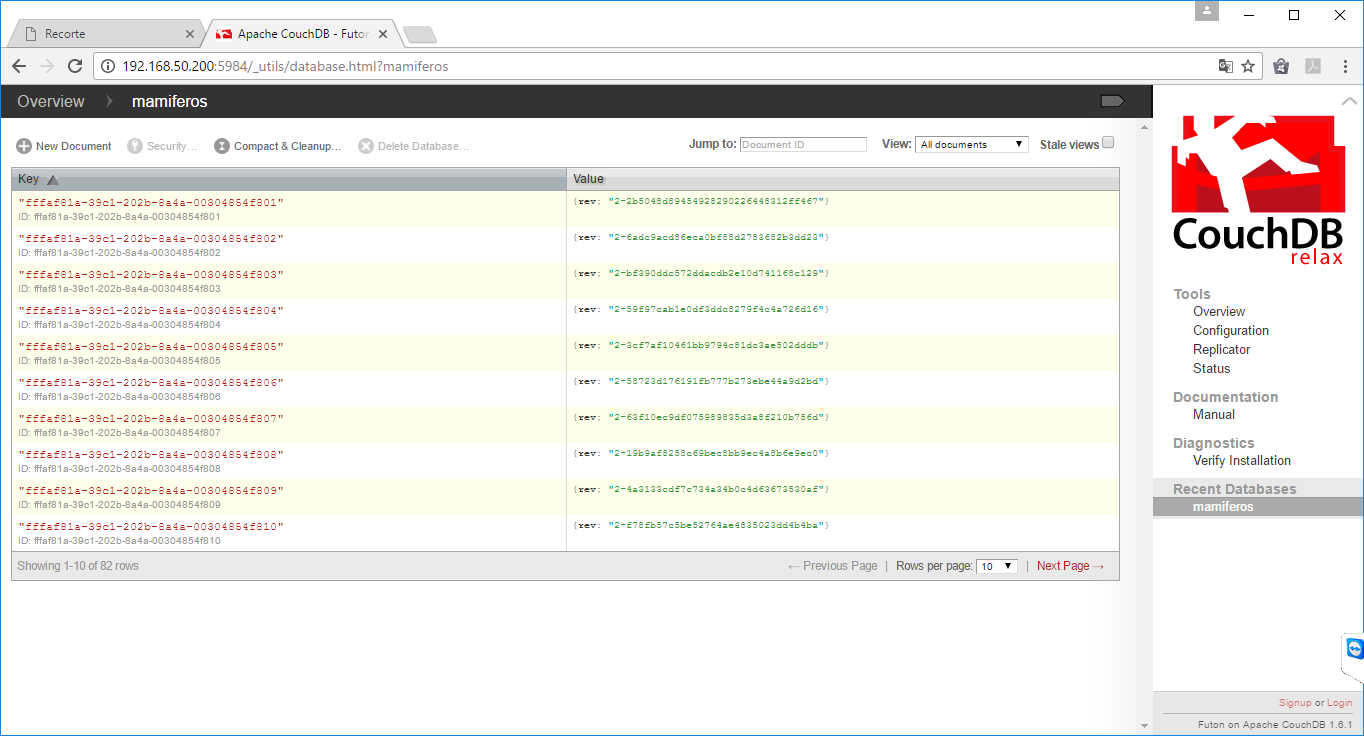
Se debe ejecutar el primer comando para insertar los registros en la base de COUCHDB, así:

**curl -X POST -d @mamiferos-RegistrosCouchDb1485204497115.json -H "Content-Type: application/json"** [**http://admin:ncc123x@192.168.50.200:5984/mamiferos/\_bulk\_docs**](http://admin:ncc123x@192.168.50.200:5984/mamiferos/_bulk_docs)

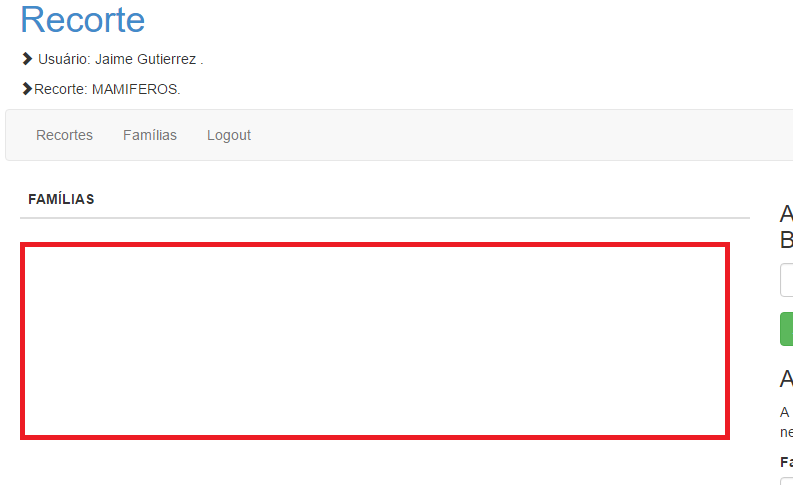
Se deberá generar en pantalla una salida de la siguiente manera:



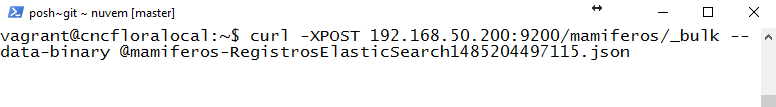
Para verificar, ingresar por FUTON y validar los documentos cargados en la base:



No obstante, hasta que no se realice la configuración del índice, los registros no serán visibles desde la aplicación de CHECKLIST, tal como se ilustra a continuación:



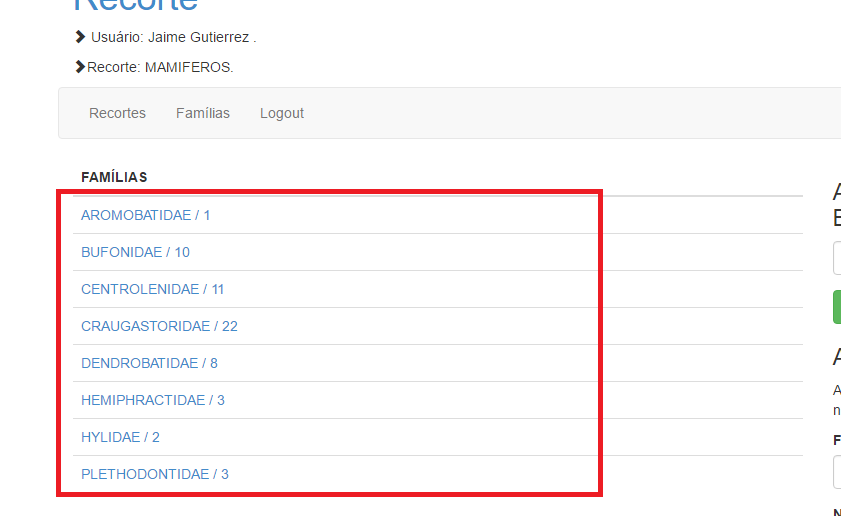
Para realizar este proceso, con los registros de Elastic generados, debe procederse a ejecutar el comando que sigue:



El cual debe generar el siguiente resultado luego del cargue:



Lo cual, al validar en la aplicación de CHECKLIST, ya debe permitir visualizar los datos cargados:



**Elaboró:**

**Ingeniero**

**JAIME ALBERTO GUTIÉRREZ MEJÍA**

**CONSTRATISTA INSTITUTO ALEXANDER VON HUMBOLDT**

**DICIEMBRE 2016**