# Introducción a la Ciencia de Datos

Guillermo Valdés Lozano

25 de septiembre de 2015

# Documento protegido por GFDL

# ¿Qué es la Ciencia de Datos?

- La Ciencia de Datos pretende abarcar a un conjunto de herramientas (basadas en la ciencia) y habilidades (humanas e informáticas) con un nombre muy atractivo.
- Se define como es la extracción de conocimiento a partir de grandes volúmenes de información estructurada o no estructurada.

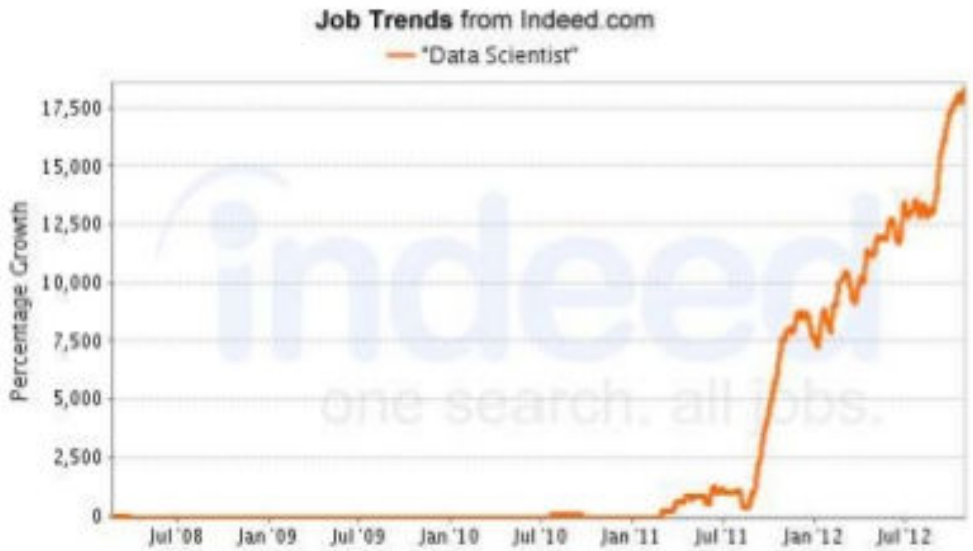
# Conceptos que involucra

# ¿Qué es un Científico de Datos?

Un Científico de Datos (Data Scientists) es una persona con habilidades estadísticas, computacionales (que sabe programar) y de visualización de datos que lo llevan a encontrar los patrones que le servirán a la empresa o institución para capitalizar la información recogida.

# Demanda de C. en D. en aumento

## Demand for Data Scientists surging




"Data Scientist"

Fastest growing term on
www.kdnuggets.com/jobs

1% of jobs in 2010

4% of jobs in 2011

19% of jobs in 2012

Data Scientist – sexiest job of the 21st Century (???)
say Thomas H. Davenport and D.J. Patil, (HBR, Oct 2012)

# ¿Qué se necesita saber para ser un científico de datos?

- Domine las matemáticas, la estadística y la informática.

# ¿Qué se necesita saber para ser un científico de datos?

- Domine las matemáticas, la estadística y la informática.
- Aprenda a programar.

# ¿Qué se necesita saber para ser un científico de datos?

- Domine las matemáticas, la estadística y la informática.
- Aprenda a programar.
- Conozca las Bases de Datos.

# ¿Qué se necesita saber para ser un científico de datos?

- Domine las matemáticas, la estadística y la informática.
- Aprenda a programar.
- Conozca las Bases de Datos.
- Sea ágil en herramientas de procesamiento y visualización.

# ¿Qué se necesita saber para ser un científico de datos?

- Domine las matemáticas, la estadística y la informática.
- Aprenda a programar.
- Conozca las Bases de Datos.
- Sea ágil en herramientas de procesamiento y visualización.
- De el salto al *Big Data*.

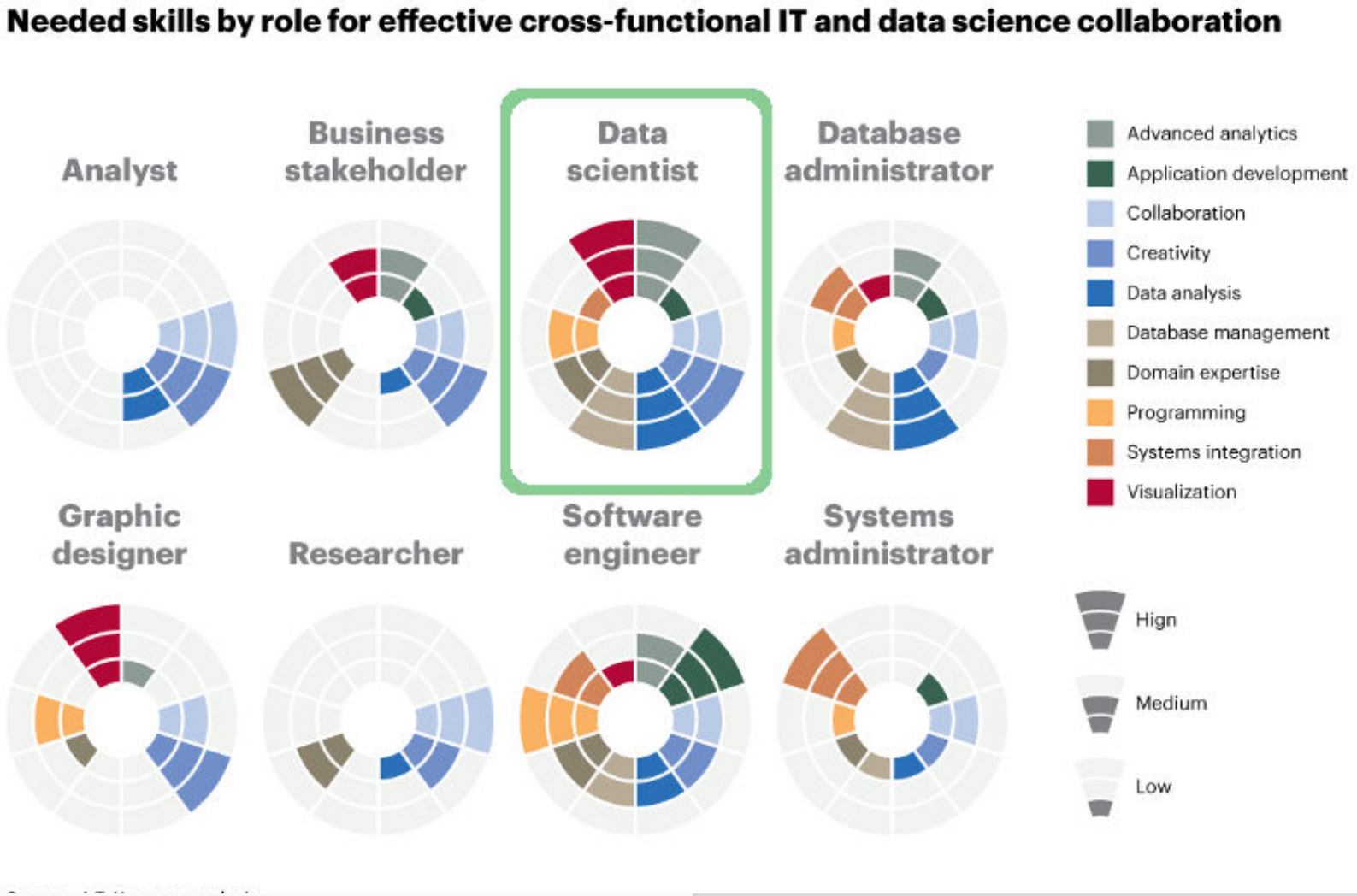
# ¿Qué se necesita saber para ser un científico de datos?

- Domine las matemáticas, la estadística y la informática.
- Aprenda a programar.
- Conozca las Bases de Datos.
- Sea ágil en herramientas de procesamiento y visualización.
- De el salto al *Big Data*.
- No deje de aprender y practicar.

# ¿Qué se necesita saber para ser un científico de datos?

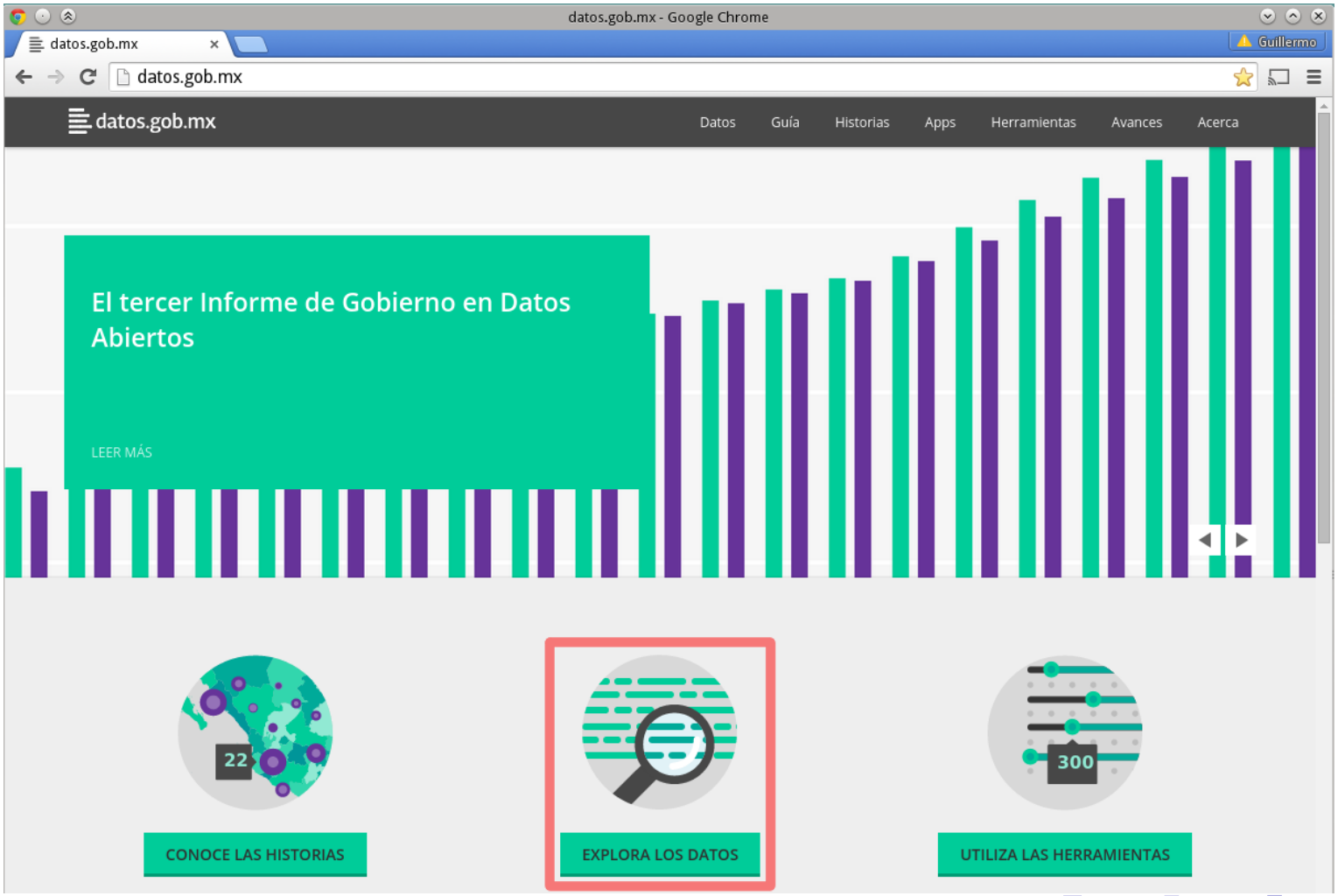
- Domine las matemáticas, la estadística y la informática.
- Aprenda a programar.
- Conozca las Bases de Datos.
- Sea ágil en herramientas de procesamiento y visualización.
- De el salto al *Big Data*.
- No deje de aprender y practicar.
- Colabore con la asociaciones, gobierno o con la iniciativa privada.

# Requerimientos para ser un C. en D.



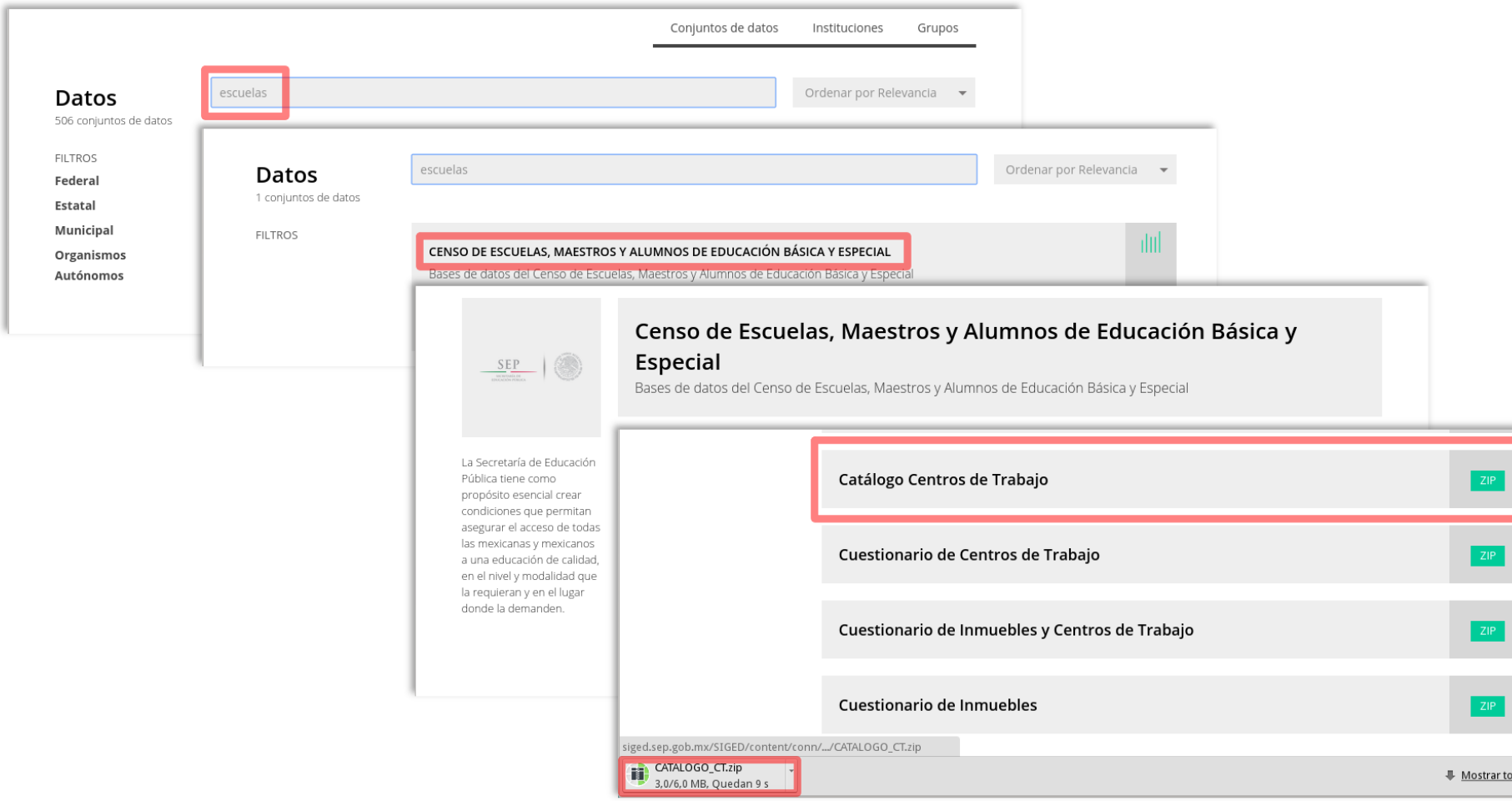
# Red México Abierto

En datos.gob.mx encontrará datos abiertos de nuestro país.



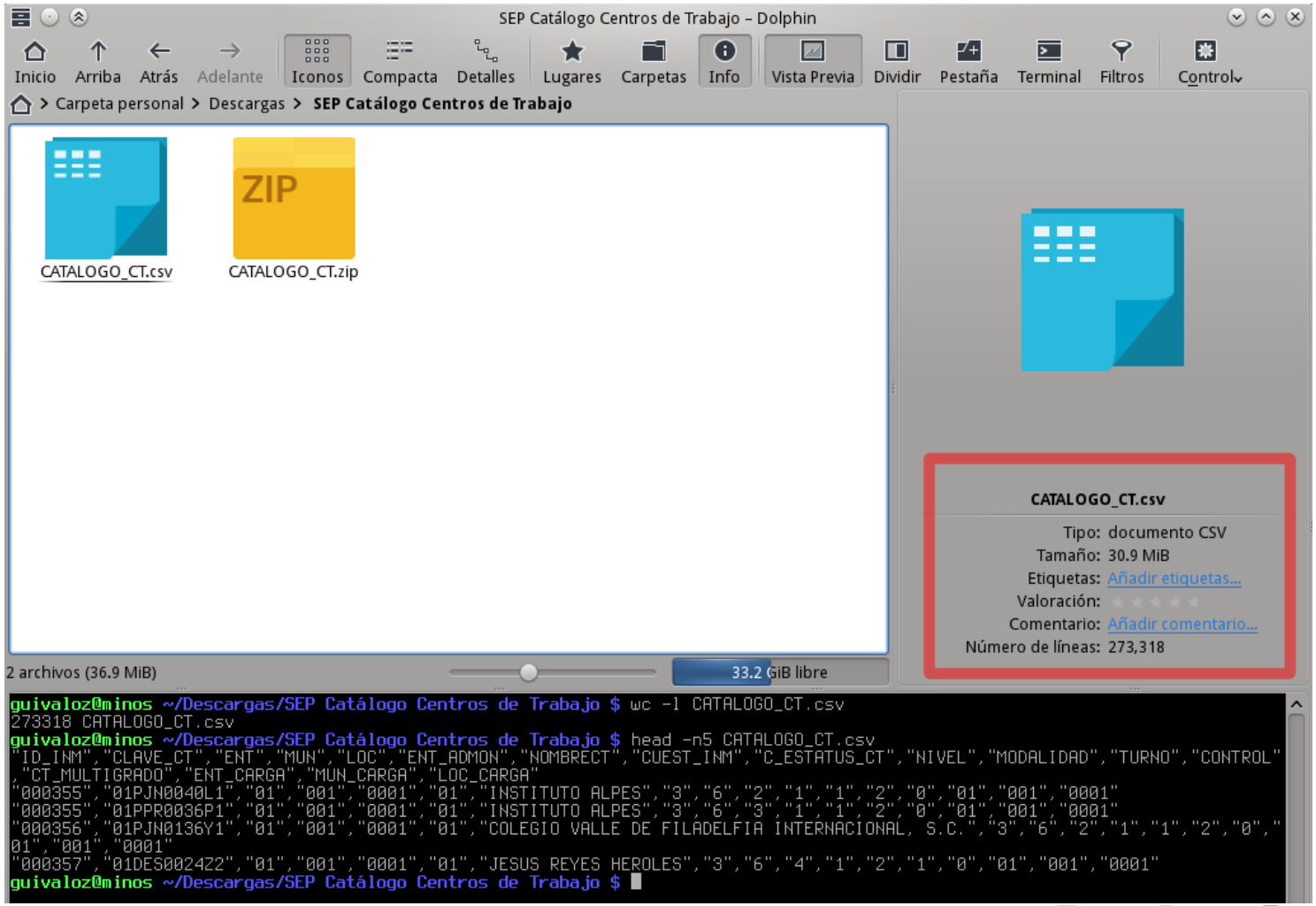
# Busque, encuentre y descargue

Descargue el Catálogo de Centros de Trabajo de la SEP.



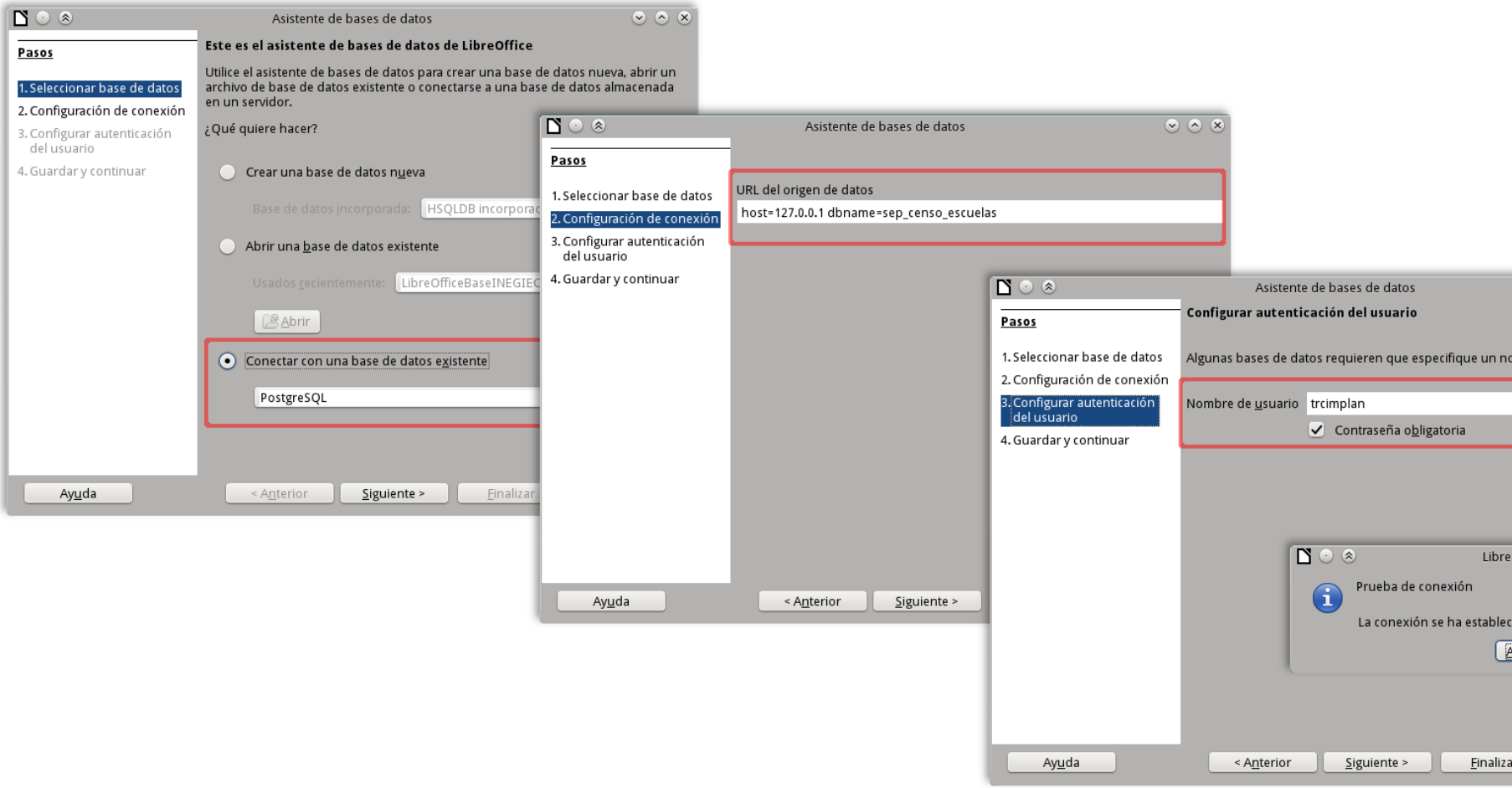
# Tipos de archivos recomendados

CSV para hojas de cálculo.



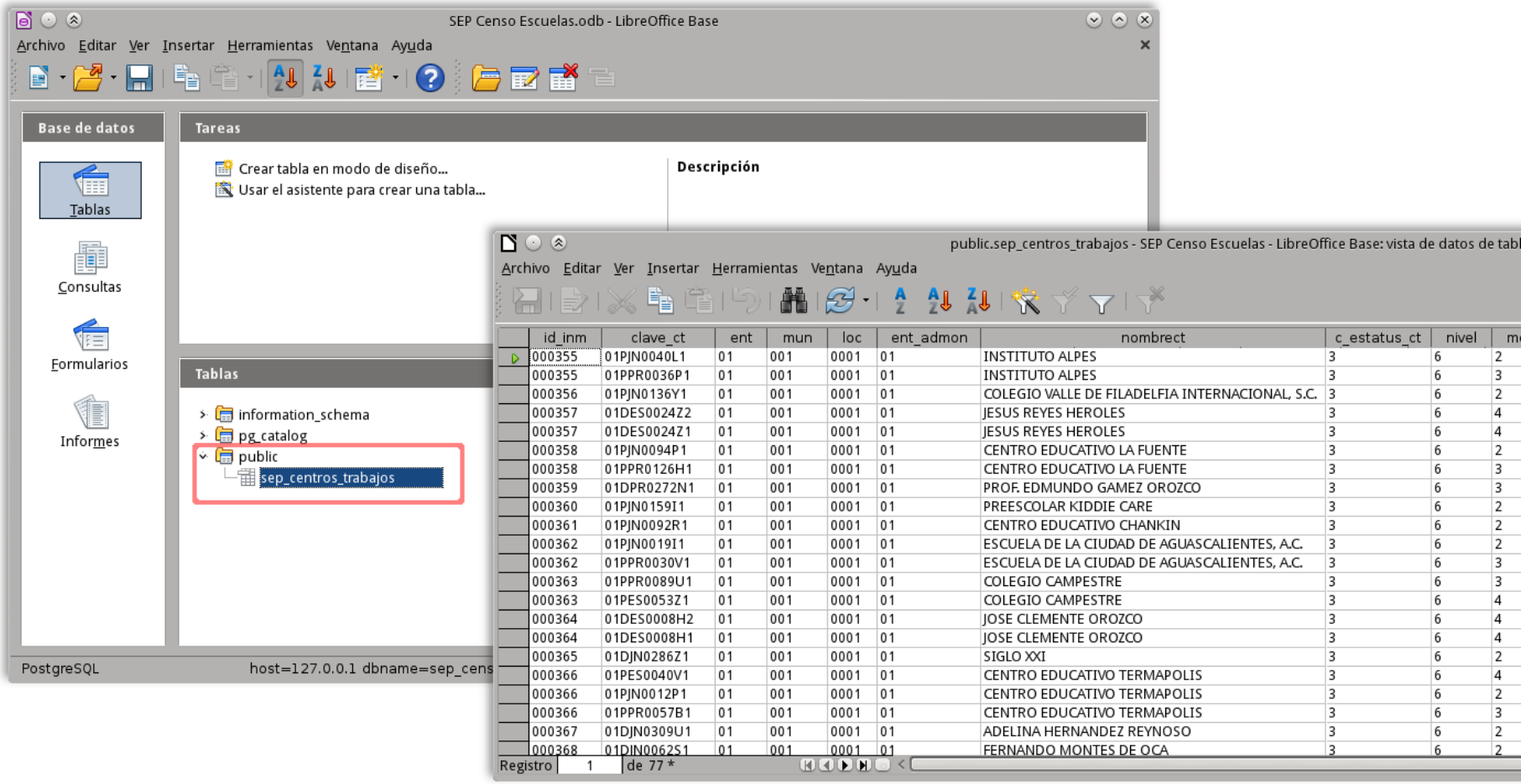
# LibreOffice Base

## Configure la comunicación con PostgreSQL.



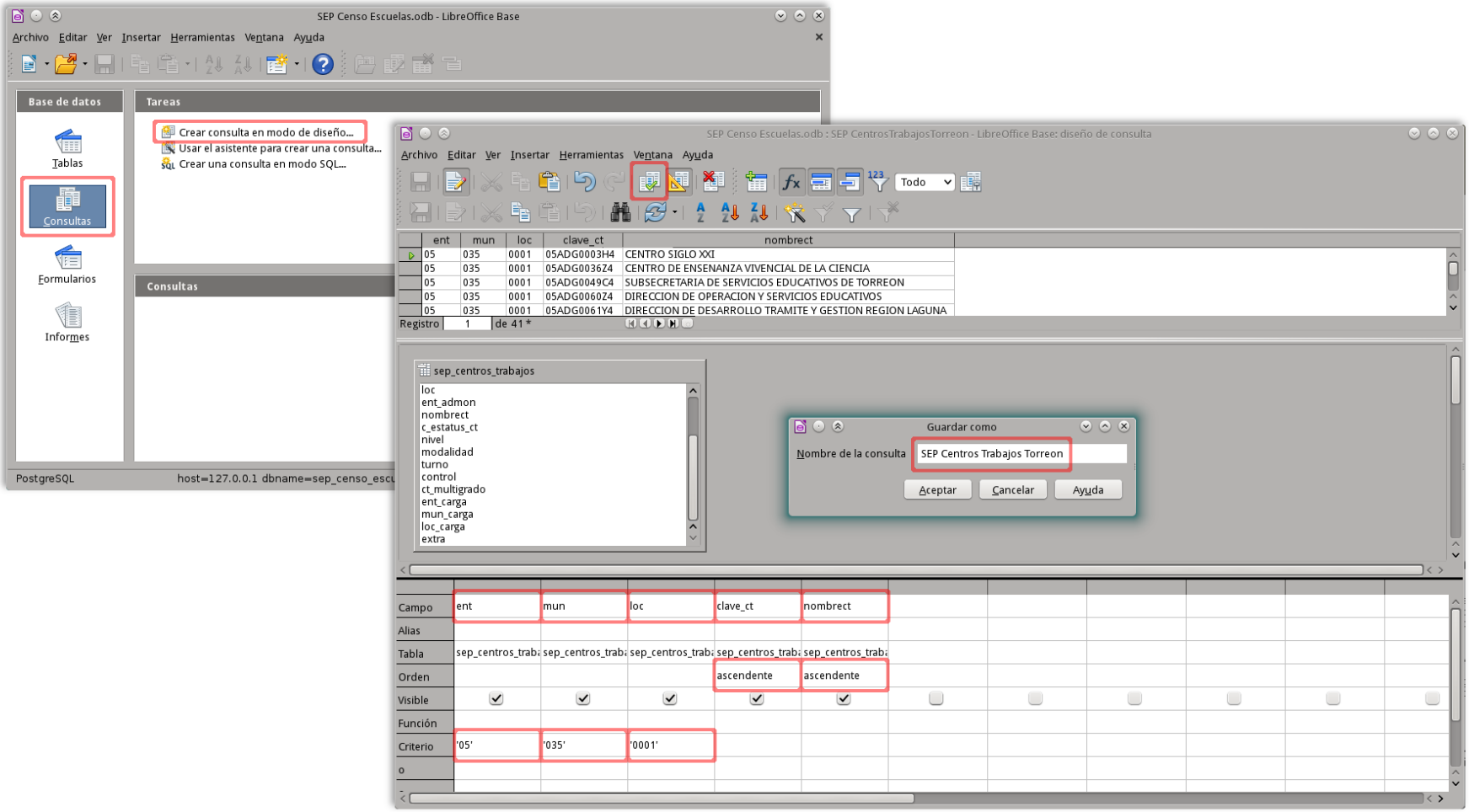
# Tablas en LibreOffice Base

Verifique que puede ver la tabla con los Centros de Trabajo.



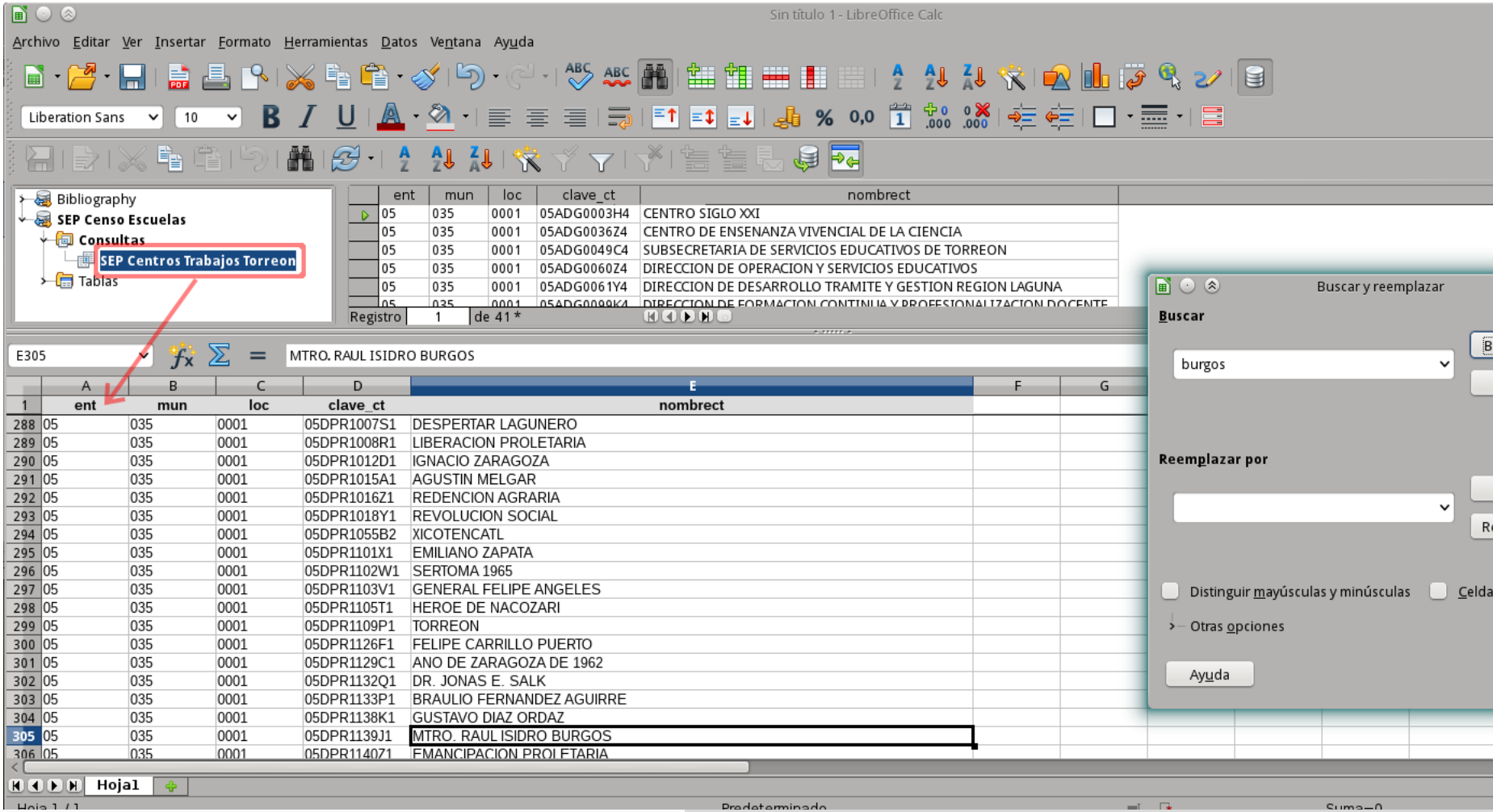
# Consultas con filtros en LibreOffice Base

Criterio: ent 05 (Coah.), mun 035 (Torreón) y loc 0001 (Torreón).



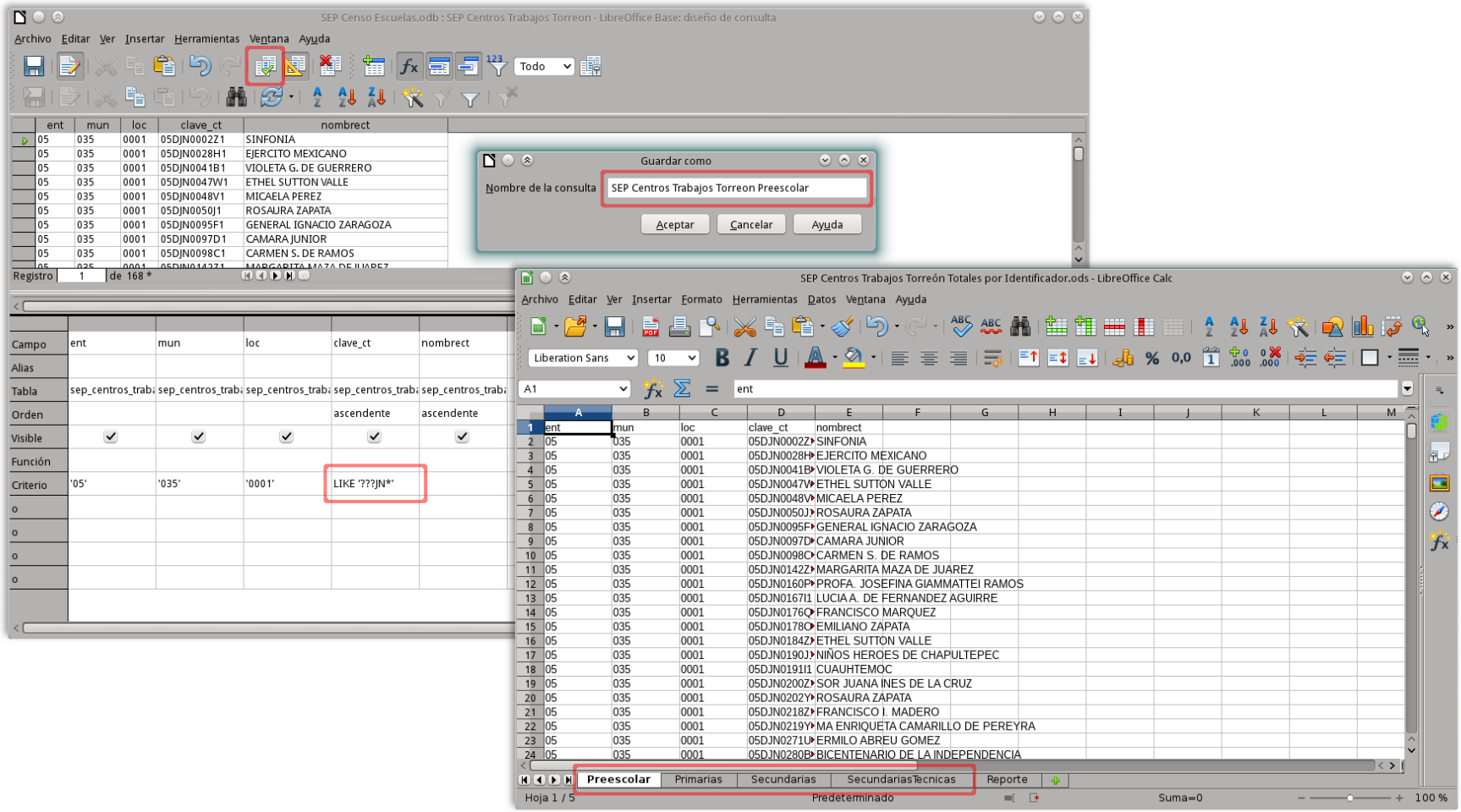
# Importe la consulta a LibreOffice Calc

Arrastre desde Orígenes de Datos (F4) la consulta.



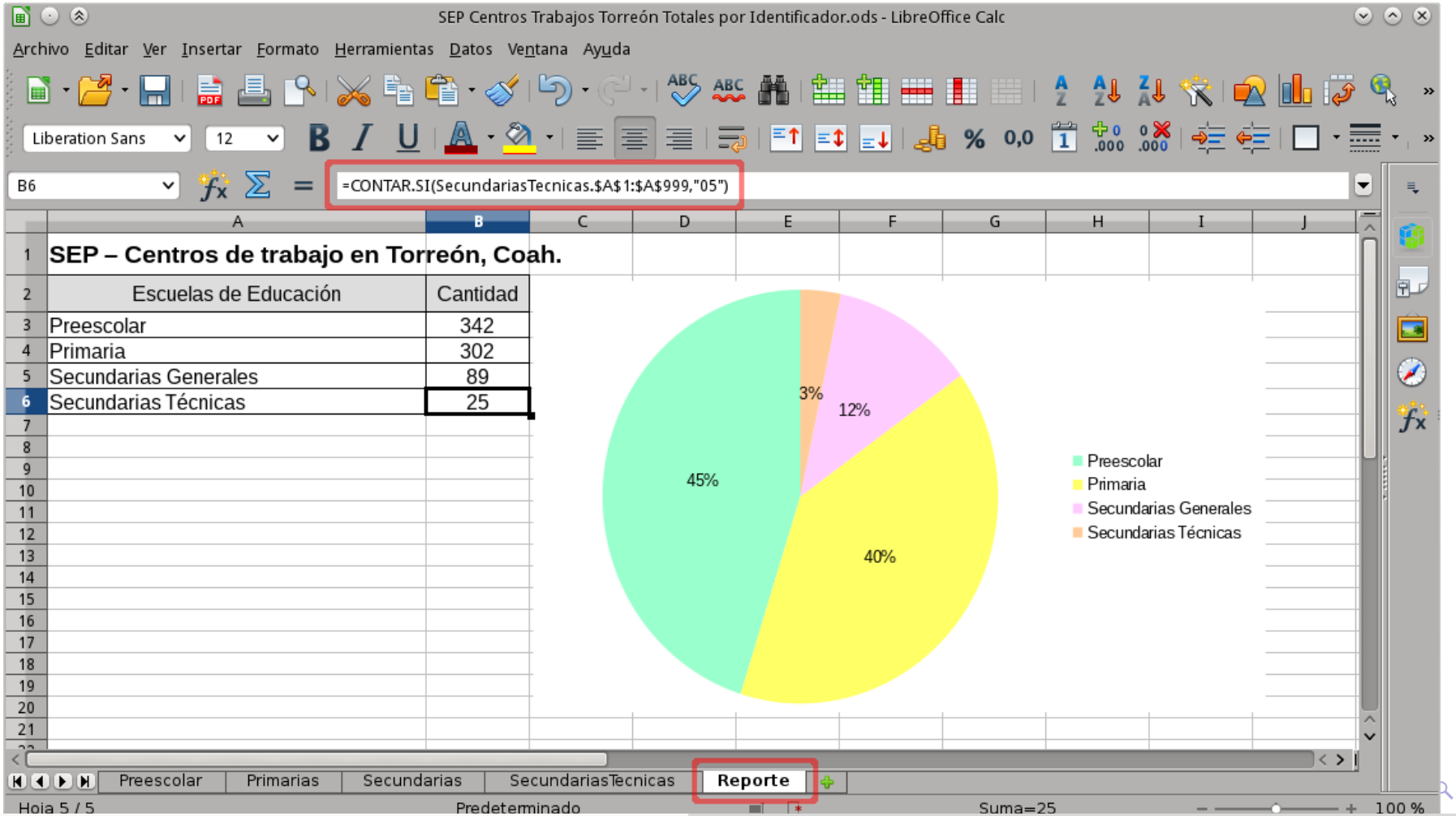
# Cree consultas por tipo de centro de trabajo

Filtre en Base y ponga cada consulta en su hoja de cálculo.



# Analize los datos

Calcule la cantidad de filas en cada hoja y grafique. Luego analize.



# Python

- Python es un lenguaje de programación creado por Guido van Rossum a principios de los años 90 cuyo nombre está inspirado en el grupo de cómicos ingleses *Monty Python*.

# Python

- Python es un lenguaje de programación creado por Guido van Rossum a principios de los años 90 cuyo nombre está inspirado en el grupo de cómicos ingleses *Monty Python*.
- Es un lenguaje interpretado o de script, con tipado dinámico, fuertemente tipado, multiplataforma y orientado a objetos.

# Python

- Python es un lenguaje de programación creado por Guido van Rossum a principios de los años 90 cuyo nombre está inspirado en el grupo de cómicos ingleses *Monty Python*.
- Es un lenguaje interpretado o de script, con tipado dinámico, fuertemente tipado, multiplataforma y orientado a objetos.
- Python es un lenguaje que todo el mundo debería conocer. Su sintaxis simple, clara y sencilla.

# Python

- Python es un lenguaje de programación creado por Guido van Rossum a principios de los años 90 cuyo nombre está inspirado en el grupo de cómicos ingleses *Monty Python*.
- Es un lenguaje interpretado o de script, con tipado dinámico, fuertemente tipado, multiplataforma y orientado a objetos.
- Python es un lenguaje que todo el mundo debería conocer. Su sintaxis simple, clara y sencilla.
- Disponibilidad Windows, Mac, Linux.

# R

- Inspirado por el lenguaje S. Desarrollado por John Chambers en los laboratorios Bell.

# R

- Inspirado por el lenguaje S. Desarrollado por John Chambers en los laboratorios Bell.
- R es un lenguaje de script para manipulación de datos, análisis estadístico y visualización.

# R

- Inspirado por el lenguaje S. Desarrollado por John Chambers en los laboratorios Bell.
- R es un lenguaje de script para manipulación de datos, análisis estadístico y visualización.
- Es comparable y a menudo superior en poder a productos comerciales. Lenguaje de propósito general.

# R

- Inspirado por el lenguaje S. Desarrollado por John Chambers en los laboratorios Bell.
- R es un lenguaje de script para manipulación de datos, análisis estadístico y visualización.
- Es comparable y a menudo superior en poder a productos comerciales. Lenguaje de propósito general.
- Disponibilidad Windows, Mac, Linux.

# Hadoop

- Creado por Apache Software Foundation. Fuertemente desarrollado por Yahoo.

# Hadoop

- Creado por Apache Software Foundation. Fuertemente desarrollado por Yahoo.
- Es un framework de software que soporta aplicaciones distribuidas.

# Hadoop

- Creado por Apache Software Foundation. Fuertemente desarrollado por Yahoo.
- Es un framework de software que soporta aplicaciones distribuidas.
- Puede usarse en granjas de computadoras y entornos de alto rendimiento.

# Hadoop

- Creado por Apache Software Foundation. Fuertemente desarrollado por Yahoo.
- Es un framework de software que soporta aplicaciones distribuidas.
- Puede usarse en granjas de computadoras y entornos de alto rendimiento.
- Hadoop implementa un paradigma computacional llamado map/reduce, donde la aplicación se divide en muchos pequeños fragmentos de trabajo, cada uno de los cuales se pueden ejecutar o volver a ejecutar en cualquier nodo del clúster.