

# Introducción a la Ciencia de Datos

Guillermo Valdés Lozano

25 de septiembre de 2015

# Documento protegido por GFDL

Copyright (c) 2015 Guillermo Valdés Lozano.  
e-mail: [guillermo\(en\)movimientolibre.com](mailto:guillermo(en)movimientolibre.com)  
<http://www.movimientolibre.com/>

Se otorga permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.2 o cualquier otra versión posterior publicada por la Free Software Foundation; sin Secciones Invariantes ni Textos de Cubierta Delantera ni Textos de Cubierta Trasera.

Una copia de la licencia está en  
<http://www.movimientolibre.com/licencias/gfdl.html>

# ¿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](http://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 21<sup>st</sup> 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?

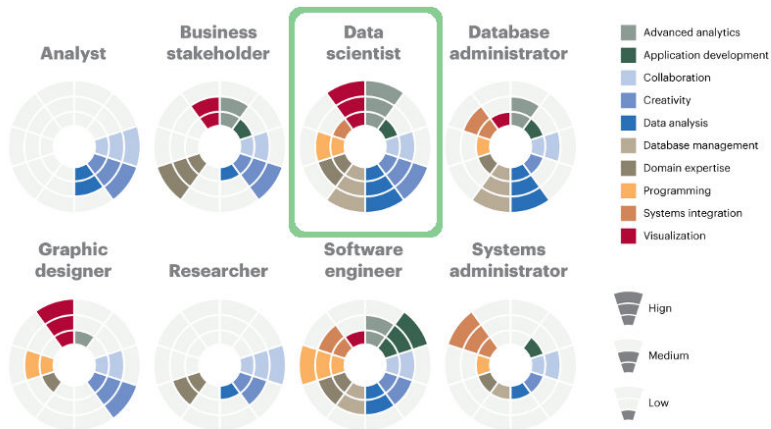
- Domine las matemáticas, la estadística y la informática.
- Aprenda a programar.
- Conozca las Bases de Datos.
- Sea ágil en herramientas de procesamiento y visualización.
- De el salto al *Big Data*.
- No deje de aprender y practicar.

# ¿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.

## Needed skills by role for effective cross-functional IT and data science collaboration



# Red México Abierto

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

The screenshot shows the homepage of datos.gov.mx. At the top, there is a navigation menu with links for Datos, Guía, Historias, Apps, Herramientas, Avances, and Acerca. The main content area features a large green banner with the text "El tercer Informe de Gobierno en Datos Abiertos" and a "LEER MÁS" link. Below the banner is a bar chart with alternating teal and purple bars. At the bottom, there are three circular icons with corresponding buttons: a globe icon with "22" and "CONOCE LAS HISTORIAS", a magnifying glass over a bar chart icon with "EXPLORA LOS DATOS" (highlighted with a red box), and a gear icon with "300" and "UTILIZA LAS HERRAMIENTAS".

# Busque, encuentre y descargue

Descargue el Catálogo de Centros de Trabajo de la SEP.

The screenshot shows a web interface for searching and downloading data from the SEP. The search results for 'escuelas' are displayed, including a table of data sets and a list of downloadable files. The 'Catálogo Centros de Trabajo' file is highlighted with a red box.

**Datos**  
506 conjuntos de datos

FILTROS  
Federal  
Estatal  
Municipal  
Organismos  
Autónomos

Conjuntos de datos   Instituciones   Grupos

escuelas   Ordenar por Relevancia

**Datos**  
1 conjuntos de datos

FILTROS

**CENSO DE ESCUELAS, MAESTROS Y ALUMNOS DE EDUCACIÓN BÁSICA Y ESPECIAL**  
Bases de datos del Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial

**Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial**  
Bases de datos del Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial

La Secretaría de Educación Pública tiene como propósito esencial crear condiciones que permitan asegurar el acceso de todas las mexicanas y mexicanos a una educación de calidad, en el nivel y modalidad que la requieran y en el lugar donde la demanden.

<b>Catálogo Centros de Trabajo</b>	ZIP
Cuestionario de Centros de Trabajo	ZIP
Cuestionario de Inmuebles y Centros de Trabajo	ZIP
Cuestionario de Inmuebles	ZIP

siged.sep.gob.mx/SIGED/conten/conv.../CATALOGO\_CT.zip

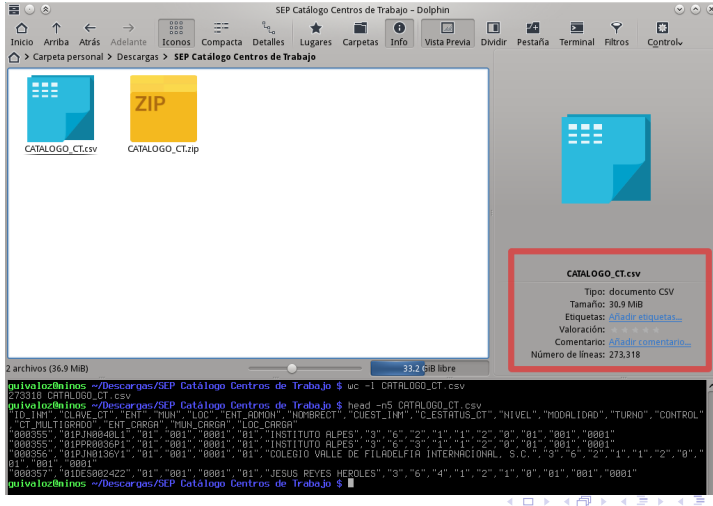
**CATALOGO\_CT.zip**  
3,676 MB. Quedan 9 s

Mostrar



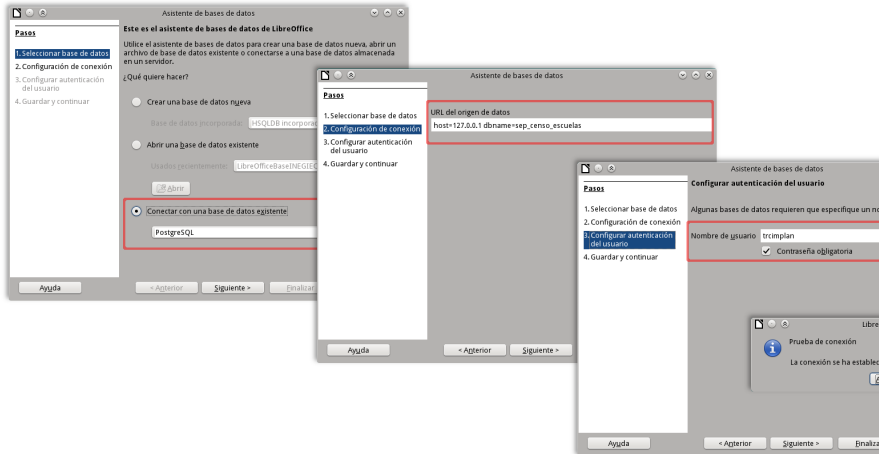
# Tipos de archivos recomendados

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.

The image shows two overlapping windows from LibreOffice Base. The background window is titled 'SEP Censo Escuelas.oddb - LibreOffice Base' and shows the 'Base de datos' sidebar with 'pg\_catalog' expanded to show the 'public' schema, where the table 'sep centros trabajos' is highlighted with a red box. The foreground window is titled 'public.sep\_centros\_trabajos - SEP Censo Escuelas - LibreOffice Base: vista de datos de tabla' and displays a data grid with the following columns: id\_inm, clave\_ct, ent, mun, loc, ent\_admon, nombrect, c\_estatus\_ct, nivel, and m. The data grid contains 28 rows of school records.

id_inm	clave_ct	ent	mun	loc	ent_admon	nombrect	c_estatus_ct	nivel	m
000355	01PJN0040L1	01	001	0001	01	INSTITUTO ALPES	3	6	2
000355	01PPR0036P1	01	001	0001	01	INSTITUTO ALPES	3	6	3
000356	01PJN0136Y1	01	001	0001	01	COLEGIO VALLE DE FILADELFA INTERNACIONAL, S.C.	3	6	2
000357	01DES0024Z2	01	001	0001	01	JESUS REYES HEROLES	3	6	4
000357	01DES0024Z1	01	001	0001	01	JESUS REYES HEROLES	3	6	4
000358	01PJN0094P1	01	001	0001	01	CENTRO EDUCATIVO LA FUENTE	3	6	2
000358	01PPR0126H1	01	001	0001	01	CENTRO EDUCATIVO LA FUENTE	3	6	3
000359	01DPR0272N1	01	001	0001	01	PROF. EDMUNDO GAMEZ OROZCO	3	6	3
000360	01PJN0159I1	01	001	0001	01	PREESCOLAR KIDDIE CARE	3	6	2
000361	01PJN0092R1	01	001	0001	01	CENTRO EDUCATIVO CHANKIN	3	6	2
000362	01PJN0019I1	01	001	0001	01	ESCUELA DE LA CIUDAD DE AGUASCALIENTES, A.C.	3	6	2
000362	01PPR0030V1	01	001	0001	01	ESCUELA DE LA CIUDAD DE AGUASCALIENTES, A.C.	3	6	3
000363	01PPR0089U1	01	001	0001	01	COLEGIO CAMPESTRE	3	6	3
000363	01PES0053Z1	01	001	0001	01	COLEGIO CAMPESTRE	3	6	4
000364	01DES0008H2	01	001	0001	01	JOSE CLEMENTE OROZCO	3	6	4
000364	01DES0008H1	01	001	0001	01	JOSE CLEMENTE OROZCO	3	6	4
000365	01DJN0286Z1	01	001	0001	01	SIGLO XXI	3	6	2
000366	01PES0040V1	01	001	0001	01	CENTRO EDUCATIVO TERMAPOLIS	3	6	4
000366	01PJN0012P1	01	001	0001	01	CENTRO EDUCATIVO TERMAPOLIS	3	6	2
000366	01PPR0057B1	01	001	0001	01	CENTRO EDUCATIVO TERMAPOLIS	3	6	3
000367	01DJN0309U1	01	001	0001	01	ADELINA HERNANDEZ REYNOSO	3	6	2
000368	01DJN0062S1	01	001	0001	01	FERNANDO MONTES DE OCA	3	6	2

# Consultas con filtros en LibreOffice Base

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

The screenshot shows the LibreOffice Base interface with a query filter configuration window open. The main window displays a table with columns: ent, mun, loc, clave\_it, and nombrect. The filter configuration window is titled 'Guardar como' and shows the query name 'SEP Centros Trabajos Torreón'. Below the window, a table shows the filter criteria for the 'ent', 'mun', and 'loc' columns.

Campo	ent	mun	loc	clave_it	nombrect
Alias					
Tabla	sep_centros_trab;	sep_centros_trab;	sep_centros_trab;	sep_centros_trab;	sep_centros_trab;
Orden				ascendente	ascendente
Visible	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Función					
Criterio	05	035	0001		

# Importe la consulta a LibreOffice Calc

Arrastre desde Orígenes de Datos (F4) la consulta.

The screenshot shows the LibreOffice Calc interface with a data table imported from a query. The table has columns: ent, mun, loc, clave\_ct, and nombrect. The data includes various educational institutions. A red box highlights the 'SEP Centros Trabajos Torreón' query in the 'Consultas' folder. A red arrow points to the 'ent' column header in the spreadsheet. A search dialog is open on the right.

ent	mun	loc	clave_ct	nombrect
05	035	0001	05ADG0003H4	CENTRO SIGLO XXI
05	035	0001	05ADG00036Z4	CENTRO DE ENSEÑANZA VIVENCIAL DE LA CIENCIA
05	035	0001	05ADG00049C4	SUBSECRETARÍA DE SERVICIOS EDUCATIVOS DE TORREÓN
05	035	0001	05ADG0006GZ4	DIRECCION DE OPERACION Y SERVICIOS EDUCATIVOS
05	035	0001	05ADG0006V4	DIRECCION DE DESARROLLO TRAMITE Y GESTION REGION LAGUNA
05	035	0001	05ADG0008A4	DIRECCION DE FORMACION CONTINUA Y PROFESIONALIZACION DOCENTE

# Cree consultas por tipo de centro de trabajo

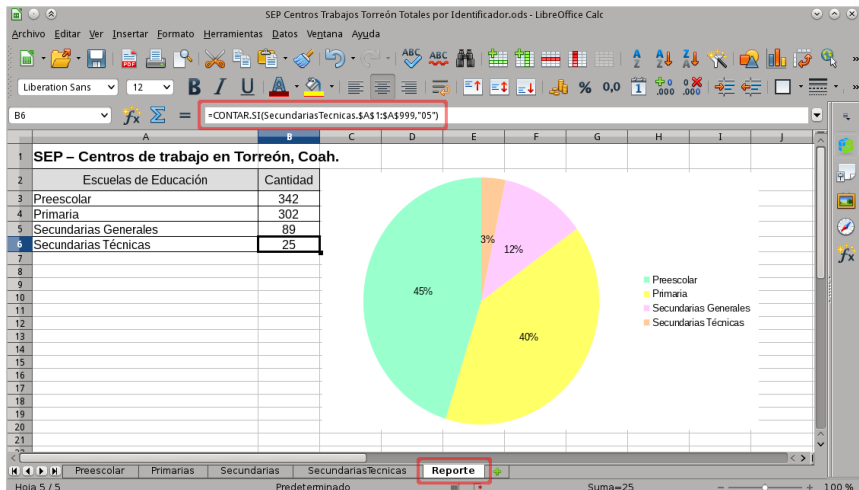
Filtre en Base y ponga cada consulta en su hoja de cálculo.

The image shows two overlapping windows from the LibreOffice suite. The top window is LibreOffice Base, displaying a table of school data. A dialog box titled 'Guardar como' (Save as) is open, with the name 'SEP Centros Trabajos Torreon Preescolar' entered. The bottom window is LibreOffice Calc, showing a spreadsheet with the same data. A red box highlights the formula bar containing the SQL query: `LIKE '177N*'`. The spreadsheet has columns for 'ent', 'mun', 'loc', 'clave\_ct', and 'nombrect'.

ent	mun	loc	clave_ct	nombrect
05	035	0001	ES0N0022	SINFONIA
05	035	0001	ES0N0028H	EJERCITO MEXICANO
05	035	0001	ES0N0041B	VIOLETA G. DE GUERRERO
05	035	0001	ES0N0047W	ETHEL SUTTON VALLE
05	035	0001	ES0N0048V	MICHAELA PEREZ
05	035	0001	ES0N0050J	ROSALBA ZAPATA
05	035	0001	ES0N0055F	GENERAL IGNACIO ZARAGOZA
05	035	0001	ES0N0057D	CAMARA JUNIOR
05	035	0001	ES0N0060C	CARMEN S. DE RAMOS
05	035	0002	ES0N0143Z	MARGARITA MAZA DE JUAREZ

# Analyze los datos

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



# 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.