Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación

Tema 2: Fuente de Datos

Prof. Jesús Lares

jesus.lares@ciens.ucv.ve

@cienciadedatos

+58 414 331.43.99

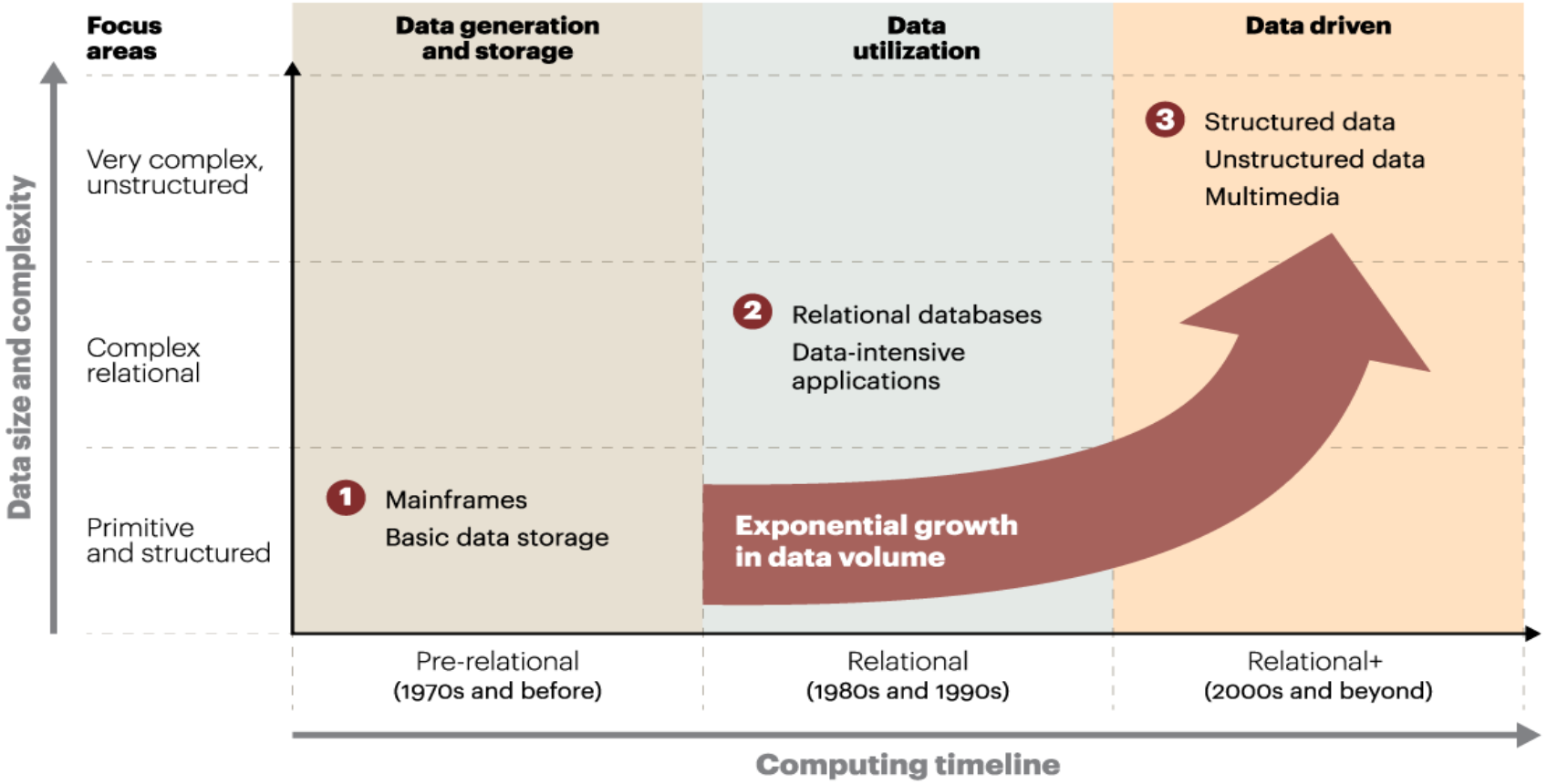
Agenda

- Evolución de los Datos
- Datawarehouse
- Business Intelligence
- OLAP
- Procesos ETL-C
- Fuentes de datos públicas
- API's para Redes Sociales

Agenda

- Evolución de los datos
- Datawarehouse
- Business Intelligence
- OLAP
- Procesos ETL-C
- Fuentes de datos públicas
- API's para Redes Sociales

Evolución de los datos



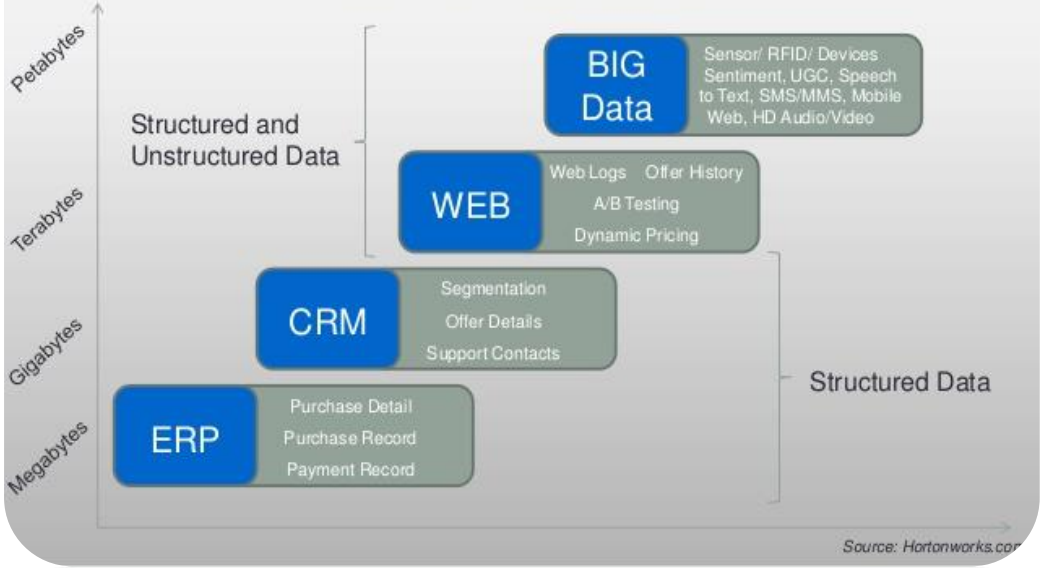
Source: A.T. Kearney analysis

Evolución de los datos

Hortonworks

Data Evolution …..

10% are structured and 90% are unstructured like emails, videos, facebook posts, website clicks etc.



Agenda

- Evolución de Big Data
- Datawarehouse
- Business Intelligence
- OLAP
- Procesos ETL-C
- Fuentes de datos públicas
- API's para Redes Sociales

Datawarehouse (Almacén de Datos)

Datawarehouse

Un almacén de datos (Data warehouse) es "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis“. - Ralph Kimball

Datamart

Son subconjuntos de datos de un Datawarehouse para áreas específicas (Datawarehouse especializado).

Características de un Datamart:

- Usuarios limitados
- Área específica
- Tiene un propósito específico
- Tiene una función de apoyo

Conceptos básicos

Variables (Indicadores de Gestión)

Representan algún aspecto cuantificable o medible de los objetos o eventos a analizar.

Ejemplos:

- Ventas Bs.
- Ventas unidades
- % Ingresos
- % Egresos
- Inventario en unidades

Conceptos básicos

Dimensiones

Son atributos relativos a las variables, son las perspectivas de análisis de las variables.

Ejemplos:

- Ubicación Geográfica (Estado, Ciudad, Municipio)
- Tiempo (Año, Semestre, Trimestre, Mes, Fecha)
- Tiendas
- Productos

Agenda

- Evolución de Big Data
- Datawarehouse
- **Business Intelligence**
- OLAP
- Procesos ETL-C
- Fuentes de datos públicas
- API's para Redes Sociales

Business Intelligence (BI)

Business Intelligence (Inteligencia de Negocios)

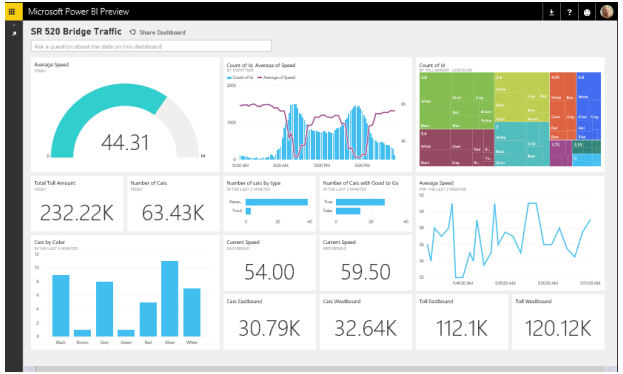
Conjunto de productos y servicios que permiten a los usuarios finales acceder y analizar de manera rápida y sencilla, la información para la toma de decisiones de negocio a nivel operativo, táctico y estratégico.

Business Intelligence (BI)

Cuadros de Mandos (Dashboard)

Son resúmenes visuales de información del negocio, que muestran de una mirada la comprensión del global de las condiciones del negocio mediante métricas e Indicadores Clave de Desempeño (KPIs).

Cuadros de Mandos (Dashboard)



* Elementos de Visualización

Agenda

- Evolución de Big Data
- Datawarehouse
- Business Intelligence
- **OLAP**
- Procesos ETL-C
- Fuentes de datos públicas
- API's para Redes Sociales

On-Line Analytical Processing (OLAP)

Es una solución utilizada en el campo la Inteligencia de Negocios cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

Se utilizan estructuras multidimensionales (Cubos OLAP) que contienen datos resumidos de grandes bases de datos o Sistemas Transaccionales.

On-Line Analytical Processing (OLAP)

Tipos de sistemas OLAP (Tradicionales)

- ROLAP (Relacional)
- MOLAP (Multidimensional o Cubos)
- HOLAP (Hibrido)

On-Line Analytical Processing (OLAP)

Otros tipos de sistemas OLAP

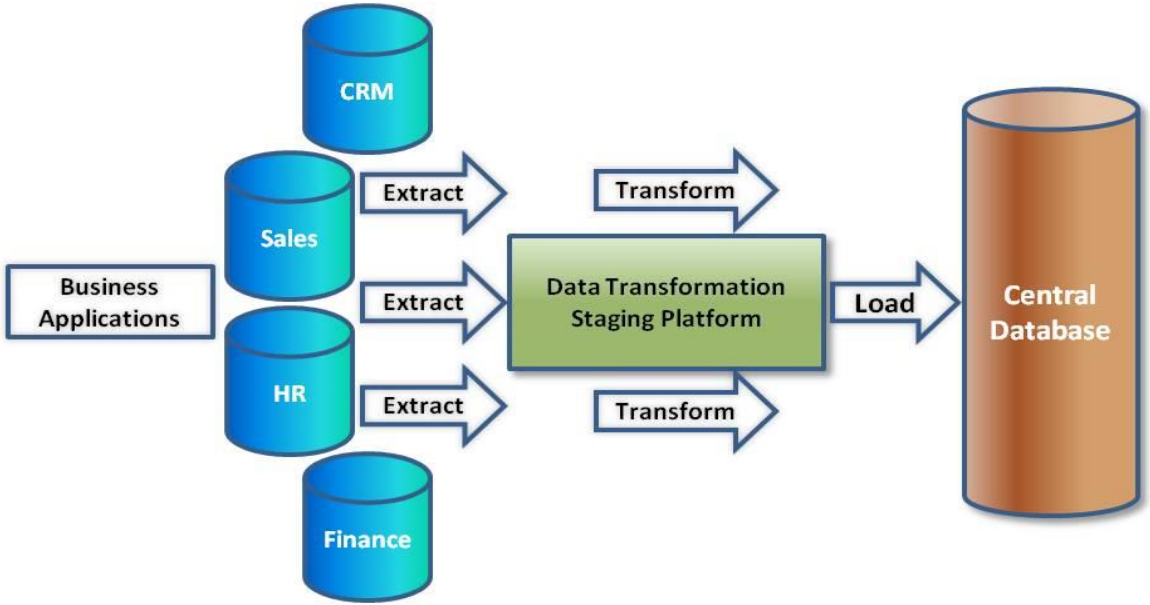
- WOLAP o Web OLAP: OLAP basado u orientado para la web.
- DOLAP o Desktop OLAP: OLAP de escritorio
- RTOLAP o Real Time OLAP: OLAP en tiempo real
- SOLAP o Spatial OLAP: OLAP espacial (GIS)

Agenda

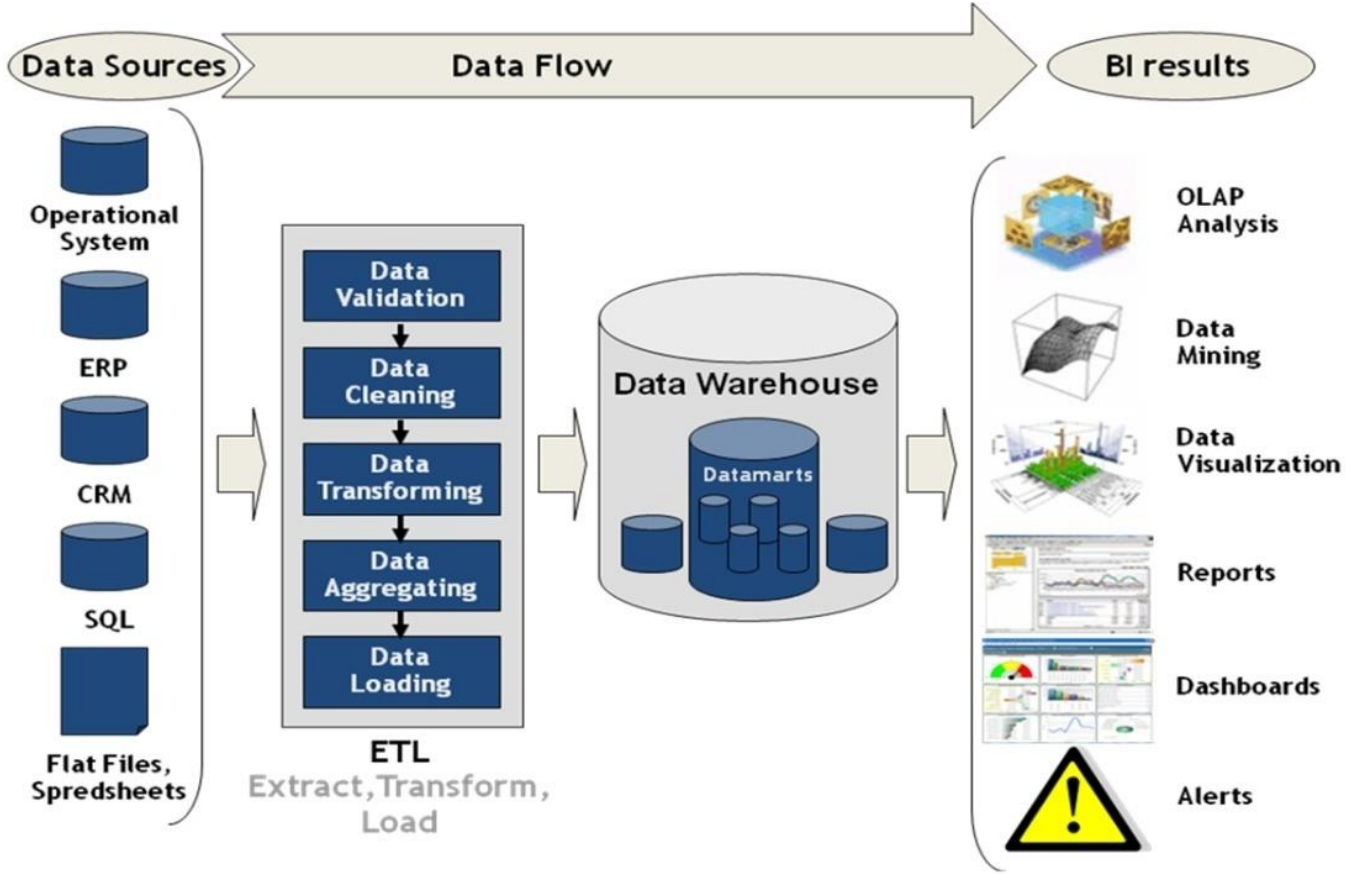
- Evolución de Big Data
- Datawarehouse
- Business Intelligence
- OLAP
- **Procesos ETL-C**
- Fuentes de datos públicas
- API's para Redes Sociales

Procesos ETL

Proceso que extrae datos desde múltiples fuentes origen, después los valida, normaliza, realiza determinadas transformaciones y los almacena en un entorno datawarehouse para su posterior análisis.



Etapas ETL


Data Sources
Data Flow
BI results
Operational System
ERP
CRM
SQL
Flat Files, Spredsheets
Data Validation
Data Cleaning
Data Transforming
Data Aggregating
Data Loading
ETL
Extract, Transform, Load
Data Warehouse
Datamarts
OLAP Analysis
Data Mining
Data Visualization
Reports
Dashboards
Alerts

Subprocesos ETL-C

Extracción: recuperación de los datos físicamente de las distintas fuentes de información.

- Base de datos de un ERP, CRM, etc.
- Hoja de cálculo, TXT
- XML, JSON
- NoSQL
- Data streaming

Subprocesos ETL-C

Limpieza: recuperación de los datos en bruto para comprobar su calidad, eliminar los duplicados y, cuando es posible, corrige los valores erróneos y completar los valores vacíos.

Subprocesos ETL-C

Problemas de Limpieza

- Ausencia de valores
- Campos que tienen distintas utilidades
- Valores crípticos
- Vulneración de las reglas de negocio
- Identificadores que no son únicos

Subprocesos ETL-C

Etapas de la Limpieza

La limpieza de datos se divide las siguientes etapas:

- Depurar los valores (parsing)
- Corregir (correcting)
- Estandarizar (standardizing)
- Relacionar (matching)
- Consolidar (consolidating)

Subprocesos ETL-C

Transformación: transformación de los datos de acuerdo a las reglas de negocio y estándares que han sido establecidos por el equipo de trabajo de una organización.

La transformación incluye:

- Cambios de formato
- Sustitución de códigos
- Valores derivados y agregados.

Subprocesos ETL-C

Integración: Este proceso valida que los datos que se van a cargar son consistentes con las definiciones y formatos del Datawarehouse.

Los datos se integran en los diferentes modelos de las áreas de negocio.

Subprocesos ETL-C

Carga: Este subproceso es el que permite cargar los nuevos datos al Datawarehouse o base de datos de destino.

ETL sobre grandes volúmenes de datos (Big Data)

- Los procesos ETL consumen entre el 60% y el 80% del tiempo de un proyecto de BI y de Big Data.
- Cuando los volúmenes de datos son grandes los problemas de desempeño incrementan el tiempo de proyectos.
- Se necesitan herramientas de ETL que se ejecuten sobre plataformas de Big Data.

ETL sobre grandes volúmenes de datos (Big Data)

Apache Flume:

Es un servicio distribuido, fiable, y altamente disponible para recopilar, agregar, y mover eficientemente grandes cantidades de datos. Tiene una arquitectura sencilla y flexible basada en flujos de datos en streaming.

ETL sobre grandes volúmenes de datos (Big Data)

Apache Scoop:

Es una aplicación con interfaz de línea de comando para transferir datos entre bases de datos relacionales y Hadoop.

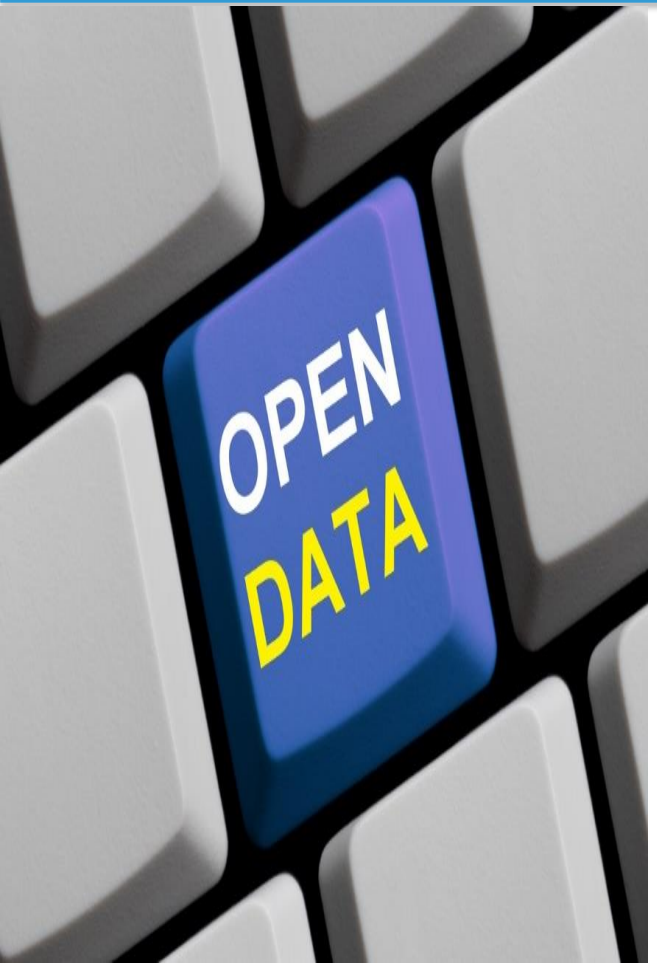
Las importaciones también pueden soler poblar tablas en Hive o HBase.

Las exportaciones pueden utilizarse para transferir datos desde Hadoop hacia a una base de datos relacional.

Agenda

- Evolución de Big Data
- Datawarehouse
- Business Intelligence
- OLAP
- Procesos ETL-C
- Fuentes de datos públicas
- API's para Redes Sociales

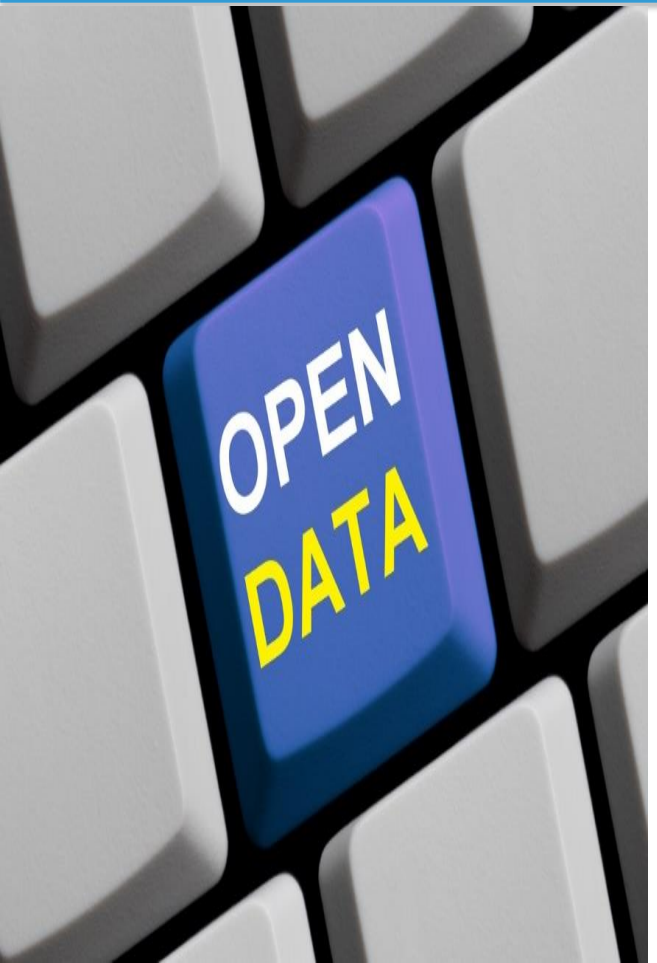
Fuentes de datos públicas (Open Data)



Open Data

Es una filosofía y práctica que persigue que determinados datos estén disponibles de forma libre a todo el mundo, sin restricciones de copyright (derecho de autor), patentes u otros mecanismos de control.

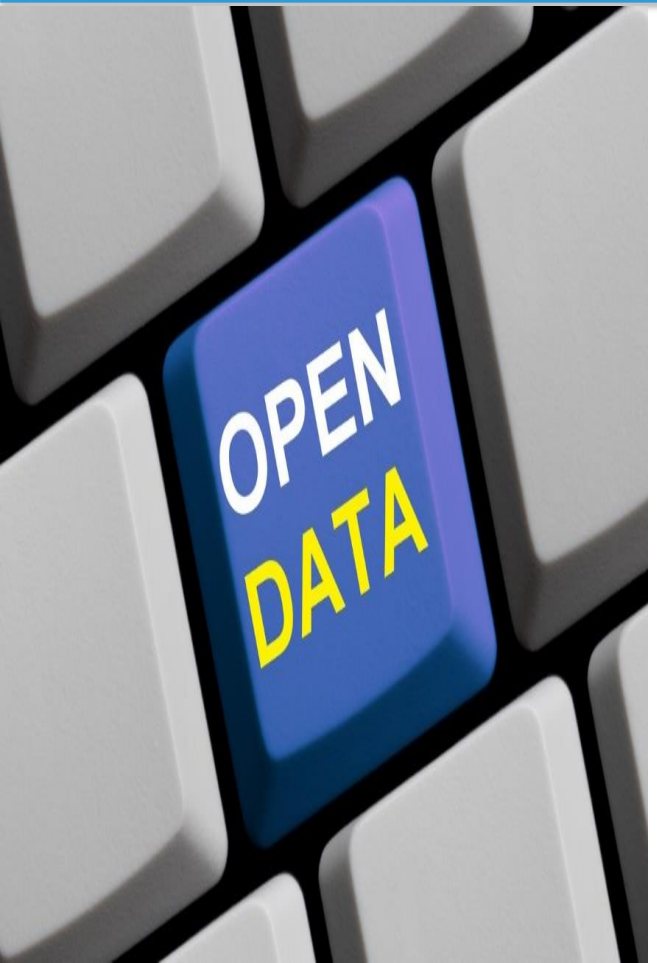
Fuentes de datos públicas (Open Data)



Tiene una ética similar a otros movimientos y comunidades abiertos como:

- Software libre
- Código abierto (open source)
- Acceso libre (open access).

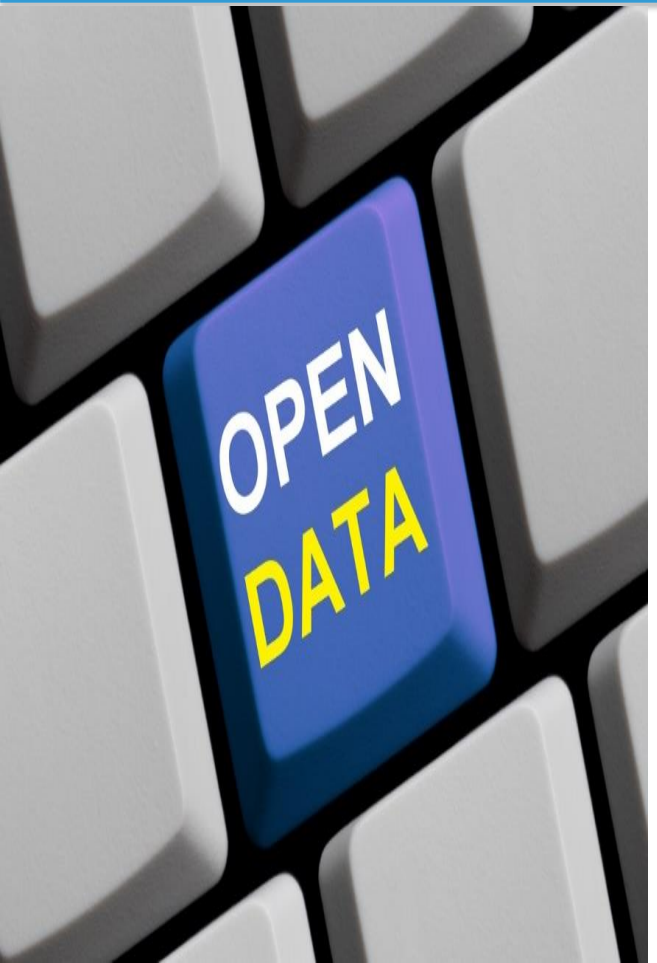
Fuentes de datos públicas (Open Data)



Son considerados datos abiertos todos aquellos datos accesibles y reutilizables, sin exigencia de permisos específicos.

No obstante, los tipos de reutilización pueden estar controlados mediante algún tipo de licencia.

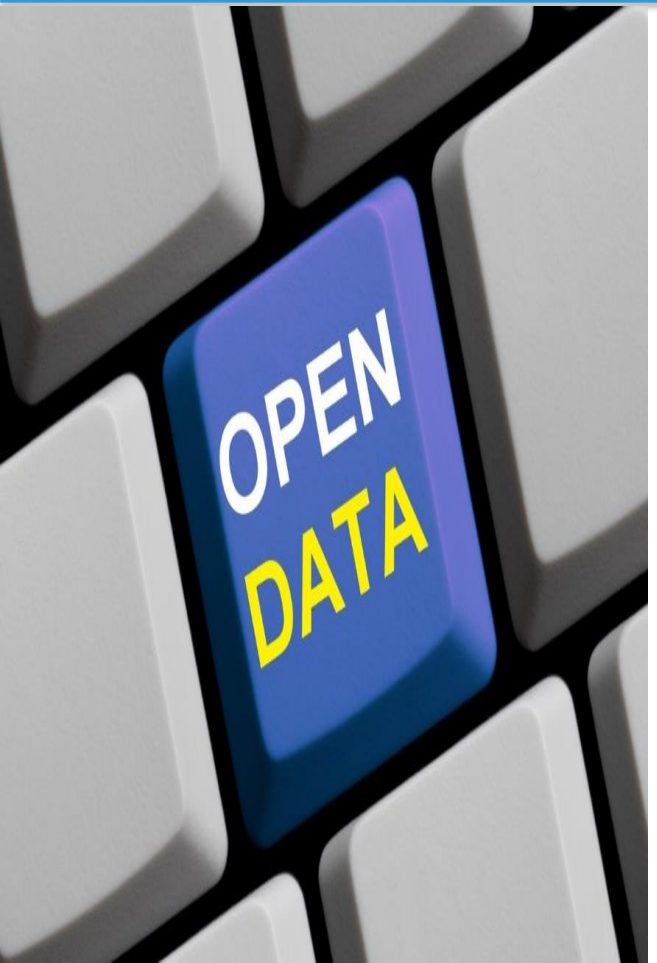
Fuentes de datos públicas (Open Data)



Los datos abiertos están centrados en material no documental como:

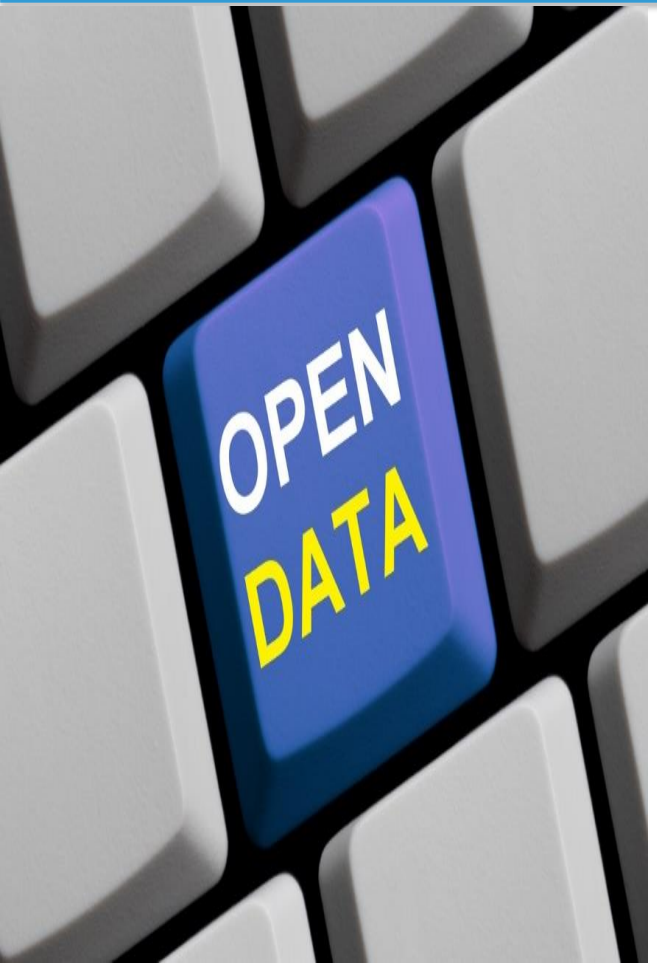
- Información geográfica
- El genoma
- Compuestos químicos
- Fórmulas matemáticas y científicas
- Datos médicos
- Biodiversidad, etc.

Fuentes de datos públicas (Open Data)



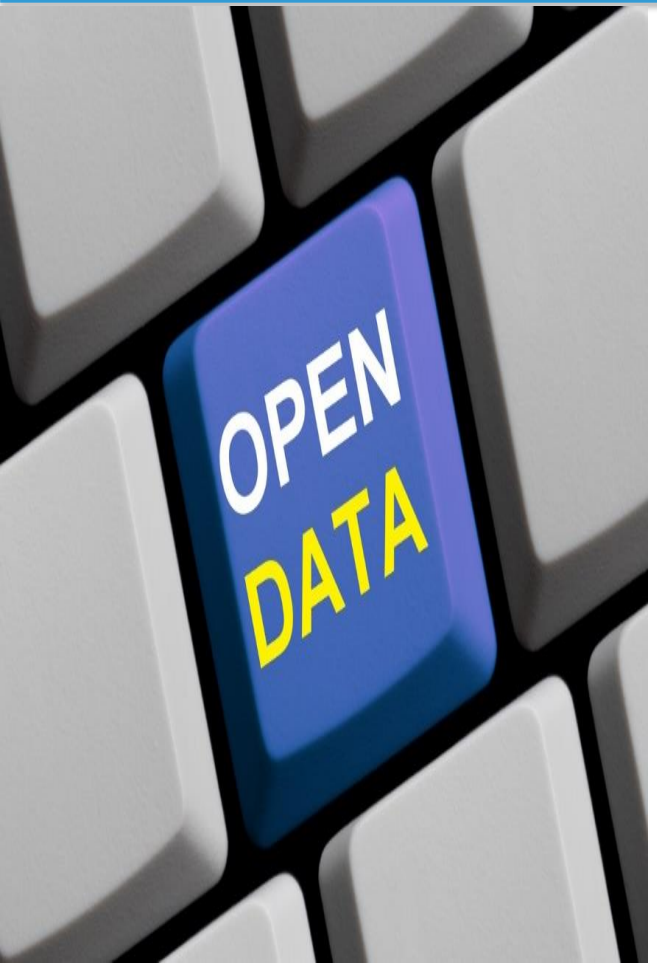
Se trata de fuentes de datos que históricamente han estado en control de organizaciones -públicas o privadas- y cuyo acceso ha estado restringido mediante limitaciones, licencias, copyright y patentes.

Fuentes de datos públicas (Open Data)



Los partidarios de los datos abiertos argumentan que estas limitaciones:

- Van en contra del bien común.
- Tienen que ser puestos en disposición del público sin limitaciones de acceso.
- Es información que pertenece a la sociedad
- Son datos que han sido creados u obtenidos por administraciones públicas financiadas por toda la ciudadanía.

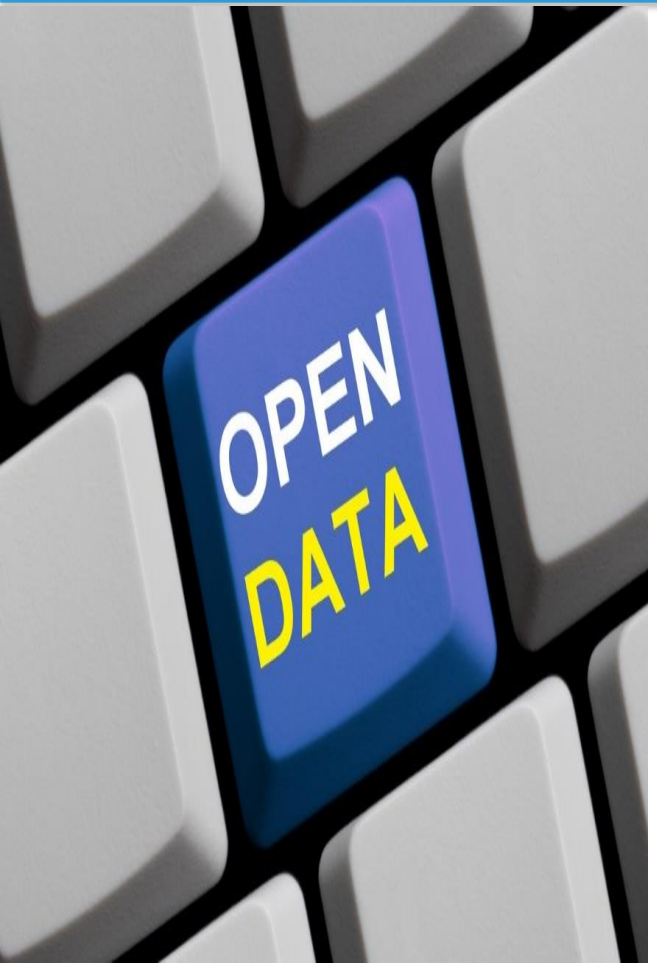


El 30 de septiembre de 2010

Fecha importante para la historia de los datos abiertos:

- El Archivo Nacional del Reino Unido libera una licencia gubernamental de re-utilización de los datos generados por esa nación.

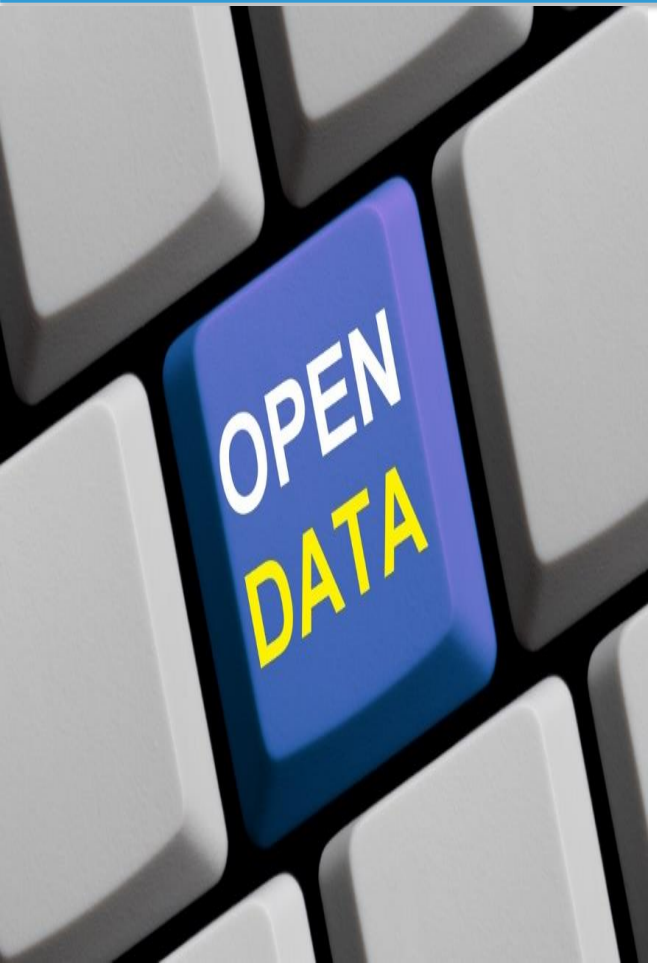
Fuentes de datos públicas (Open Data): Formatos



Formatos para Datos Abiertos (Open Data):

- Una estrella - Disponible en la Internet (en cualquier formato. Por ejemplo: PDF), siempre que sea con licencia abierta (de la data), para que sea considerado un Dato Abierto.
- Dos estrellas - Disponible en la Internet de manera estructurada (Ejemplo: archivo Excel).
- Tres estrellas - Disponible en la Internet de manera estructurada y en formato no propietario (Ejemplo: CSV en vez del Excel).

Fuentes de datos públicas (Open Data): Formatos



Formatos para Datos Abiertos (Open Data):

- Cuatro estrellas - Siguiendo todas las reglas anteriores, pero dentro de los estándares establecidos por el W3C (RDF e SPARQL): usar URI para identificar cosas y propiedades, de manera que las personas las puedan direccionar para sus publicaciones.
- Cinco estrellas - Todas las reglas ya mencionadas, y además: vincular sus datos a los de otras personas, de manera a proveer un contexto.

Agenda

- Evolución de Big Data
- Datawarehouse
- Business Intelligence
- OLAP
- Procesos ETL-C
- Fuentes de datos públicas
- API's para Redes Sociales

API's para Redes Sociales

Interfaz de Programación de Aplicaciones (API):

Es un conjunto de funciones previamente implementadas que brindan al programador una interfaz a través de la cual puede comunicarse con un sistema determinado, añadiéndole nuevas funcionalidades.

API's para Redes Sociales

API para Redes Sociales

- Twitter
- Facebook
- Youtube
- Instagram

API's para Redes Sociales

API para Twitter

- Existen límites de peticiones en un periodo de tiempo, es decir, existen intervalos de 15 minutos en donde se puede hacer un número máximo de peticiones.
- Los límites son por usuario, no por aplicación, de esta forma cada usuario es controlado de forma independiente.
- Dependiendo del tipo de recurso que sea solicitado, existen dos tipos de restricciones principales: 15 peticiones cada 15 minutos y 180 peticiones cada 15 minutos.

API's para Redes Sociales

API para Twitter

- Adicionalmente, no se permite recuperar información histórica, es decir, si se ejecuta una búsqueda sólo es posible obtener información generada siete días atrás como máximo.
- En caso de exceder el número máximo de peticiones en un periodo de tiempo, se obtiene un código de respuesta en donde se provee información acerca del recurso restringido temporalmente y el tiempo de espera para que el recurso se encuentre nuevamente disponible.

API’s para Redes Sociales

API para Twitter

- Otra particularidad de Twitter, es que cuenta con streaming, lo que permite obtener información en tiempo real sin restringir el número de peticiones por un periodo de tiempo.
- La versión actual del API de Twitter es la 1.1.

API's para Redes Sociales

API para Facebook

- Por todas las restricciones de privacidad que protegen la información del usuario, no es posible acceder a datos protegidos a menos que ambos usuarios tengan una relación de 'amistad' en la red social, en caso contrario sólo puede obtenerse acceso sin restricción al nombre y al género de los usuarios.

API's para Redes Sociales

API para Facebook

- Sin embargo, si un usuario da consentimiento explícito a que un tercero acceda a su información de perfil, correo electrónico, gustos, intereses, amigos, eventos, datos de amigos, entre otros más; entonces la aplicación prácticamente no tendrá restricciones para ese usuario en particular.

API's para Redes Sociales

API para Facebook

- El acceso a los mensajes públicos en Facebook es mediante su Open Graph que soporta también peticiones utilizando el Facebook Query Language.
- Facebook no tiene documentado el número máximo de peticiones en un periodo de tiempo o si las restricciones son por usuario, aplicación, dirección IP o la combinación de estos, sin embargo, si continuamente se envían peticiones hacia un recurso específico, con el tiempo se aprecia que se empieza a obtener menos información acerca del mismo.

API's para Redes Sociales

API para Facebook

- Facebook cuenta con actualizaciones en tiempo real, con esto es posible subscribirse a ciertos recursos, como el muro de un usuario, y recibir notificaciones si hay cambios. Esta funcionalidad evita estar haciendo peticiones de forma periódica.
- Actualmente el Open Graph se encuentra en la versión 2.0 y de acuerdo a la documentación oficial de la misma, ya no será posible acceder a los mensajes públicos que actualmente están disponibles en su versión 1.0.

API's para Redes Sociales

API para YouTube

- En el caso de YouTube el acceso a los datos por medio de su API es muy similar a Twitter dado que los videos están disponibles públicamente.
- Las restricciones de acceso no son medidas por intervalos cortos de tiempo donde se permite un número máximo de peticiones, sino que es controlado por día.

API's para Redes Sociales

API para YouTube

- YouTube asigna diariamente un número de 'Unidades' para cada aplicación (no es por usuario) y cada operación que se realiza tiene un costo de esas unidades.
- Por ejemplo, subir un video representa el uso de mil 600 unidades que son restadas del total disponible del día.

API's para Redes Sociales

API para YouTube

- Ahora bien, cuando se requiere realizar una búsqueda de videos, YouTube tiene una restricción que permite recuperar un máximo de 500 videos por búsqueda y no tiene una restricción de temporalidad, es decir, pueden recuperarse videos que fueron subidos recientemente o de meses atrás. Si la búsqueda devuelve más de 500 videos, YouTube sugiere que se utilicen otros criterios de búsqueda para refinar los resultados y de esta forma obtener la información deseada.

API's para Redes Sociales

API para YouTube

- YouTube hace uso principalmente de XML para compartir sus datos y para enviar respuestas o mensajes de error.
- Sólo algunos recursos pueden ser configurados para que la respuesta sea devuelta en formato JSON.
- La versión actual del API de YouTube es la 3.0.

API's para Redes Sociales

API para Instagram

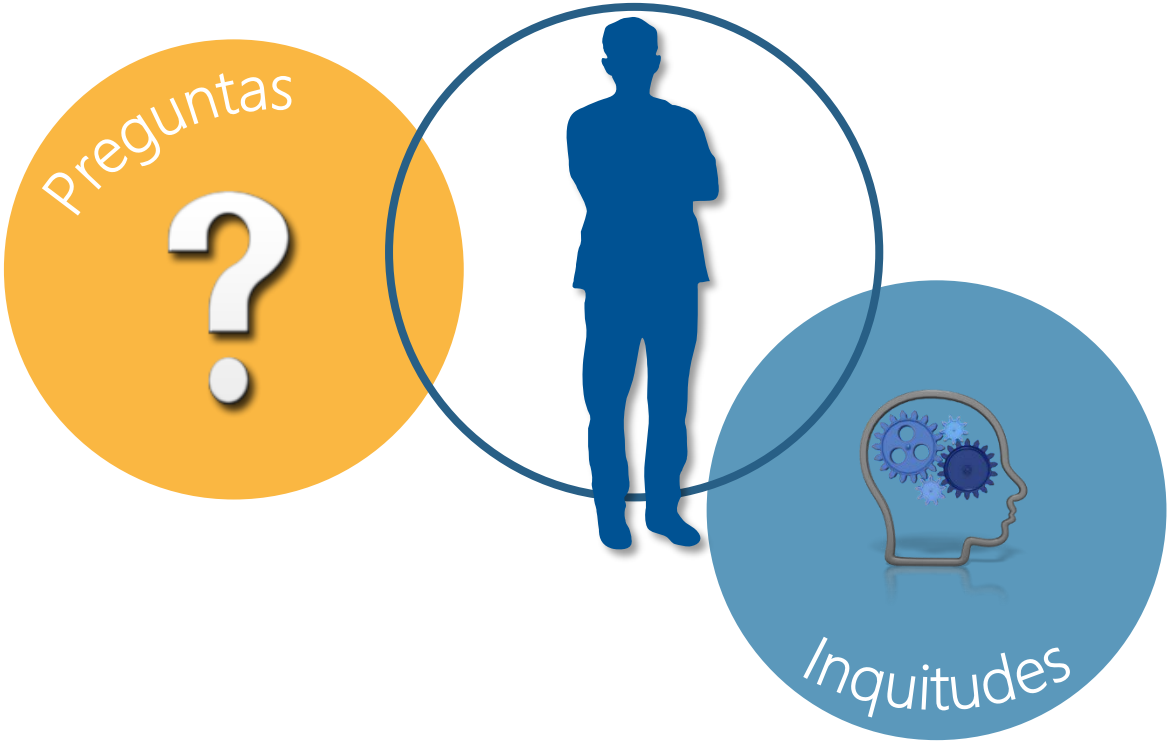
- Instagram cuenta con una API con mayores limitantes desde el punto de vista de interacción, dado que a diferencia de las redes sociales descritas anteriormente, con Instagram no es posible subir contenido, esta acción solamente es permitida mediante su aplicación oficial.
- De forma adicional, se tiene una restricción de 5 mil peticiones por hora por "token" de acceso.

API's para Redes Sociales

API para Instagram

- El API de Instagram solamente permite realizar búsquedas sobre mensajes que contengan etiquetas que existan en la red social.
- Este tipo de búsqueda se realiza sobre los mismos mensajes y sobre sus comentarios.

Preguntas/Inquietudes




thank you
danke
merci
gracias
obrigado
dziękuję
bedankt
go raibh maith agat
mochchakkeram
teşekkür ederim
ngiyabonga
شكراً جزيلاً
谢谢
спасибо
Баярлалаа
рахмат
dank je
tapadh leat
mauruuru
hvala
sukriya
kop khun krap
terima kasih
감사합니다
ευχαριστώ
arigatô
dakujem
takk
grazie
taiku
sagolun
tack
mersi
barka
welalin
spas
faafetai lava
vinaka
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
kiitos
dankie
dhanyavad
köszönöm
akun
dankon
aciu
enkosi
bayarlalaa
grâcie
nanii
nandri
djiere dieuf
tau
mamnun
дякую
chnorakaloutioun
gratias ago
grâcies
sulpáy
sobodi
dēkuji
mèsi
didi madloba
kam sah hamnida
তোমাকে ধন্যবাদ
najis tuke
rahmat
ありがとう
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
merci
мерси
trugarez
asante
manana
obrigada
murakoze
tenki
chokrane
хвала
xiexie