
BIG DATA is GOING TO IMPACT US SOON!
CALL BRUCE WILLIS!!
DATA
www.nuxeo.com
Soulcié

Open Data + Mongo DB + R

Víctor Hugo Males
John Alexander Carvajal Agudelo

R Users Cali

Open Data + Mongo DB + R

- Introducción Open Data
- Ejemplo práctico OD + R
- Introducción MongoDB
- Ejemplo práctico MongoDB + R

Open Data

Open Data

Los datos abiertos u Open Data son información de libre acceso y gratuidad, que toman forma de archivos que respetan formatos interoperables.

Pueden ser de origen público: Servicios públicos, de colectividades o de comunas.
Pueden ser de origen privado: Provenientes de empresas o instituciones cuyos datos son concursables a proyectos de utilidad pública.

Características Open Data

- **Disponibilidad y acceso**: los datos deben estar disponibles mediante la descarga a través de Internet y en una forma legible.
- **Reutilización y redistribución**: los datos deben ser proporcionados bajo términos que permitan la reutilización y redistribución .
- **Participación universal**: Cualquier persona debe ser capaz de utilizar, reutilizar y redistribuirlos – no debe haber discriminación por campos de la actividad o en contra de personas o grupos.
- Descriptivos y explicativos: Deben contener una descripción de lo que contiene el archivo, condiciones de generación y descripción de cada campo del mismo

Tipos de archivos Open Data

Estadísticas

Gubernamentales

Cultural

Transporte

Públicos

Estado Financieros

Medio Ambiente

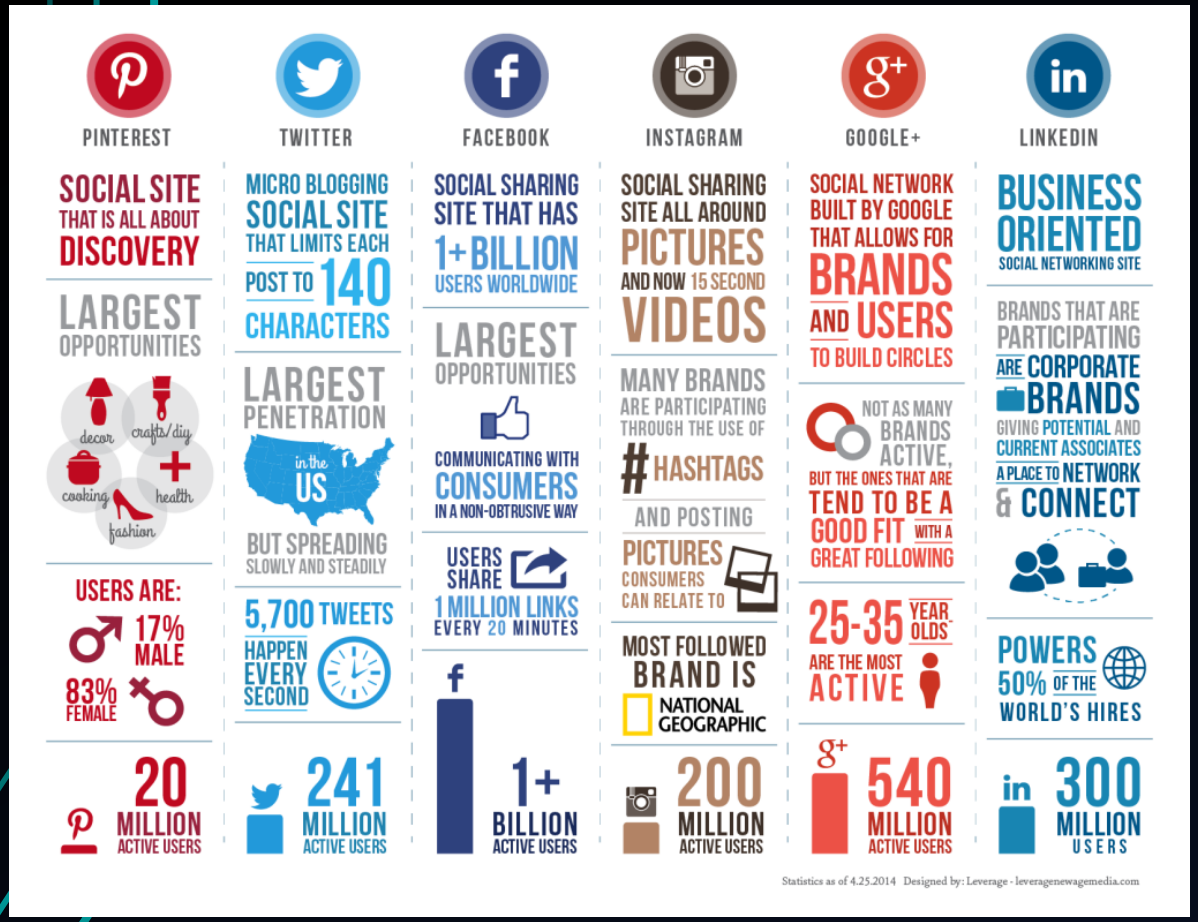
Políticos

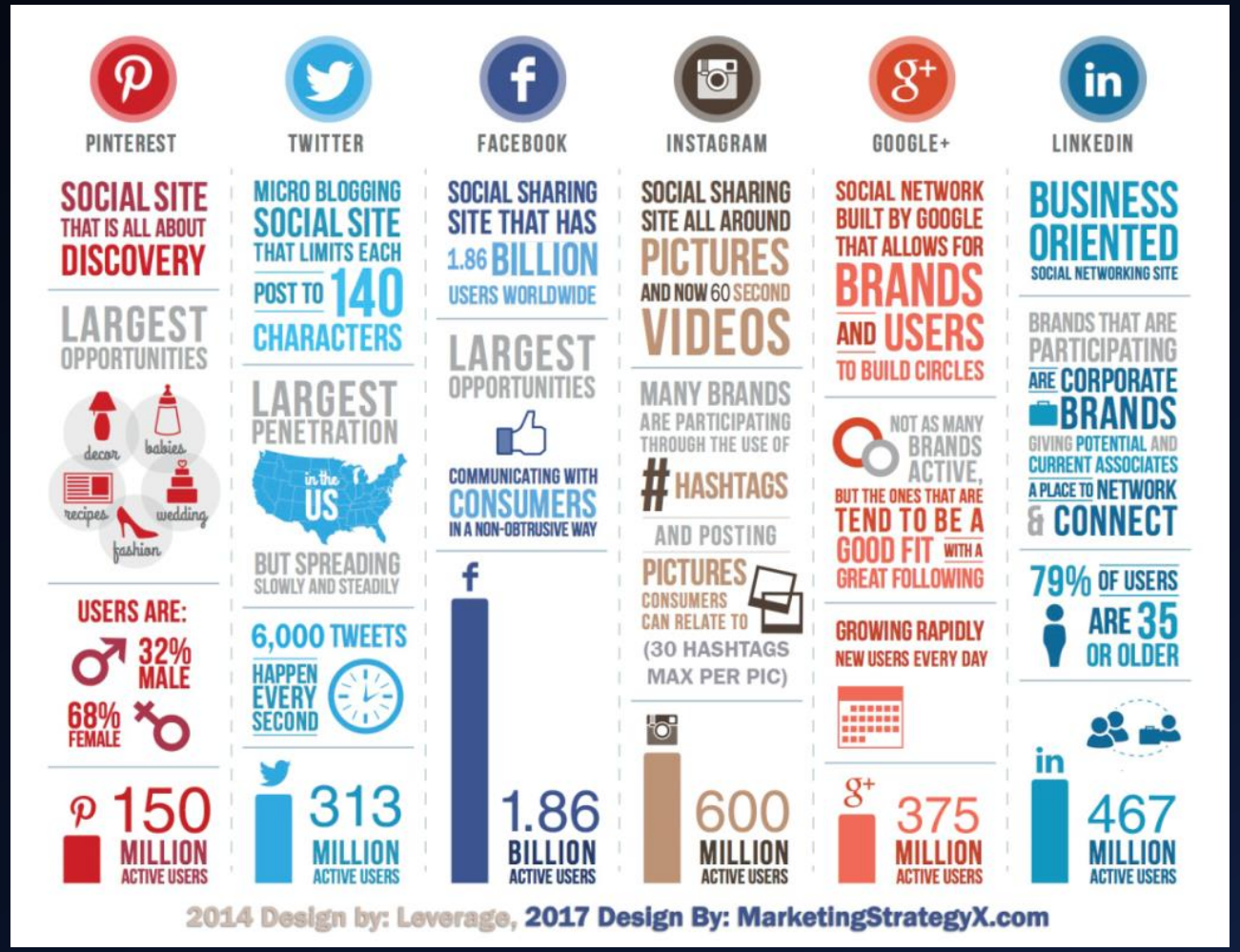
Open Data + R

MongoDB

Principales cambios en la tecnología

- Masificación del uso de internet.
- Surgimiento de las redes sociales.
- Crecimiento exponencial de dispositivos móviles.
- Posibilidad de distintos dispositivos conectados a internet (Internet de las cosas).
- Interfaces de usuario más simples e intuitivas.
- Cambios en las formas de procesamiento.
- Fuerte baja en los costos de almacenamiento.


PINTEREST
SOCIAL SITE THAT IS ALL ABOUT DISCOVERY
LARGEST OPPORTUNITIES
decor
crafts/diy
cooking
health
fashion
USERS ARE:
17% MALE
83% FEMALE
20 MILLION ACTIVE USERS
TWITTER
MICRO BLOGGING SOCIAL SITE THAT LIMITS EACH POST TO 140 CHARACTERS
LARGEST PENETRATION
in the US
BUT SPREADING SLOWLY AND STEADILY
5,700 TWEETS HAPPEN EVERY SECOND
241 MILLION ACTIVE USERS
FACEBOOK
SOCIAL SHARING SITE THAT HAS 1+ BILLION USERS WORLDWIDE
LARGEST OPPORTUNITIES
COMMUNICATING WITH CONSUMERS IN A NON-OBTRUSIVE WAY
USERS SHARE 1 MILLION LINKS EVERY 20 MINUTES
1+ BILLION ACTIVE USERS
INSTAGRAM
SOCIAL SHARING SITE ALL AROUND PICTURES AND NOW 15 SECOND VIDEOS
MANY BRANDS ARE PARTICIPATING THROUGH THE USE OF #HASHTAGS
AND POSTING PICTURES CONSUMERS CAN RELATE TO
MOST FOLLOWED BRAND IS NATIONAL GEOGRAPHIC
200 MILLION ACTIVE USERS
GOOGLE+
SOCIAL NETWORK BUILT BY GOOGLE THAT ALLOWS FOR BRANDS AND USERS TO BUILD CIRCLES
NOT AS MANY BRANDS ACTIVE, BUT THE ONES THAT ARE TEND TO BE A GOOD FIT WITH A GREAT FOLLOWING
25-35 YEAR OLDS ARE THE MOST ACTIVE
540 MILLION ACTIVE USERS
LINKEDIN
BUSINESS ORIENTED SOCIAL NETWORKING SITE
BRANDS THAT ARE PARTICIPATING ARE CORPORATE BRANDS GIVING POTENTIAL AND CURRENT ASSOCIATES A PLACE TO NETWORK & CONNECT
POWERS 50% OF THE WORLD'S HIRES
300 MILLION USERS
Statistics as of 4.25.2014 Designed by: Leverage - leveragenewagemedia.com


PINTEREST
SOCIAL SITE THAT IS ALL ABOUT DISCOVERY
LARGEST OPPORTUNITIES
decor
babies
recipes
wedding
fashion
USERS ARE:
32% MALE
68% FEMALE
150 MILLION ACTIVE USERS
TWITTER
MICRO BLOGGING SOCIAL SITE THAT LIMITS EACH POST TO 140 CHARACTERS
LARGEST PENETRATION
in the US
BUT SPREADING SLOWLY AND STEADILY
6,000 TWEETS HAPPEN EVERY SECOND
313 MILLION ACTIVE USERS
FACEBOOK
SOCIAL SHARING SITE THAT HAS 1.86 BILLION USERS WORLDWIDE
LARGEST OPPORTUNITIES
COMMUNICATING WITH CONSUMERS IN A NON-OBTRUSIVE WAY
1.86 BILLION ACTIVE USERS
INSTAGRAM
SOCIAL SHARING SITE ALL AROUND PICTURES AND NOW 60 SECOND VIDEOS
MANY BRANDS ARE PARTICIPATING THROUGH THE USE OF #HASHTAGS
AND POSTING PICTURES CONSUMERS CAN RELATE TO (30 HASHTAGS MAX PER PIC)
600 MILLION ACTIVE USERS
GOOGLE+
SOCIAL NETWORK BUILT BY GOOGLE THAT ALLOWS FOR BRANDS AND USERS TO BUILD CIRCLES
NOT AS MANY BRANDS ACTIVE, BUT THE ONES THAT ARE TEND TO BE A GOOD FIT WITH A GREAT FOLLOWING
GROWING RAPIDLY NEW USERS EVERY DAY
375 MILLION ACTIVE USERS
LINKEDIN
BUSINESS ORIENTED SOCIAL NETWORKING SITE
BRANDS THAT ARE PARTICIPATING ARE CORPORATE BRANDS GIVING POTENTIAL AND CURRENT ASSOCIATES A PLACE TO NETWORK & CONNECT
79% OF USERS ARE 35 OR OLDER
467 MILLION ACTIVE USERS
2014 Design by: Leverage, 2017 Design By: MarketingStrategyX.com

MongoDB

Es una bases de datos Open Source orienta a documentos diseñada para facilitar el desarrollo y la escalabilidad de las aplicaciones:

- No usan SQL como lenguaje de consulta
- Los datos almacenados no requieren estructuras fijas como tablas
- No garantizan consistencia
- Escalan horizontalmente
- Distribución, Auto-Replicación y Particionamiento
- Los datos se almacenan de forma atómica
- Posee distintos tipos de modelamiento

Características

- Soporta JSON, XML, BSON Modelos documentales con esquema dinámico
- Soporte de Índices
- Consultas Complejas
- Manejo de Seguridad Avanzada
- Almacenamiento de archivos de gran tamaño en su file system interno GridFS
- Búsquedas de texto(Full Text Search)
- Aggregation Framework y Map Reduce Nativo o con Hadooop

Tamaño máximo de un documento 16 MB
Tamaño máximo de una base de datos en 32 Bit es 2GB

```
        "_id" : NumberLong("722951307997401088"),
        "Text" : "RT @TaxisLibres4: Accidente de transito en la Cl 16 con Cr 100
. #traficocali @elpaiscali @TwiterosCali",
        "RetweetCount" : 1,
        "GeoLocation" : null,
        "Cuenta" : 1,
        "DateCreate" : "2016-04-20T21:53:18.000Z",
        "DateInsert" : "2016-04-27T00:15:47.066Z",
        "User" : {
                "isFavorited" : 0,
                "isRetweet" : 1,
                "isRetweetedByMe" : 0,
                "UserID" : NumberLong(185627001),
                "ScreenName" : "@HEFZIBA_CALI",
                "CreatedAt" : "2010-09-01T10:05:36.000Z",
                "Name" : "Luz Marina Ramirez",
                "Description" : "",
                "Followers" : 73,
                "Friends" : 206,
                "Location" : "Cali",
                "ProfileImageURL" : "http://pbs.twimg.com/profile_images/1127225
509/Karen__tia__y_mami__hot__normal.jpg",
                "TimeZone" : null,
                "URL" : null,
                "isGeoEnabled" : 0,
                "isVerified" : 0
        }
```

```
{
  "_id" : ObjectId("57310f4072f120a28c1d8e31"),
  "cliente" : "GOMESA S.A.",
  "cuit" : "30404884831",
  "region" : [
          "CENTRO"
  ],
  "direccion" :[{
          "calle" : "9 de julio",
          "numero" : 1010,
          "localidad" : "TANDIL",
          "provincia" : "BUENOS AIRES"},

          {"calle" : "25 de mayo",
           "numero" : 941,          e32"),
           "piso" : "2do",
           "dto" : "B",
           "localidad" : "CABA"},

          {"calle" : "Reconquista",
           "numero" : 654,
           "piso" : "1er",
           "dto" : "A",
           "localidad" : "CABA"}]

}
```

```
> db.reservas.find().limit(3).pretty()
{
        "_id" : ObjectId("59924e03c049295f1dd5b3d7"),
        "Cliente" : {
                "edad" : 28,
                "profesion" : "Blogger",
                "paisOrigen" : "Chile",
                "nombre" : "Doug Cutting",
                "aerolinea" : "Qatar Airways"
        },
        "fechaReserva" : ISODate("2014-05-08T12:51:00Z"),
        "mes" : "May",
        "diasHospedaje" : 4,
        "origen" : "Agencia de Viajes",
        "Equipaje" : 8,
        "calificacion" : 2,
        "monedaPago" : "Euro"
}
{
        "_id" : ObjectId("59924e03c049295f1dd5b3d8"),
        "Cliente" : {
                "edad" : 48,
                "profesion" : "Medico",
                "paisOrigen" : "Australia"
        },
        "fechaReserva" : ISODate("2010-01-01T20:42:00Z"),
        "mes" : "Ene",
        "diasHospedaje" : 7,
        "origen" : "Portal",
        "Equipaje" : 9,
        "calificacion" : 4,
        "monedaPago" : "Euro"
}
{
        "_id" : ObjectId("59924e03c049295f1dd5b3d9"),
        "Cliente" : {
                "edad" : 38,
                "profesion" : "Estudiante",
                "paisOrigen" : "España",
                "idiomas" : [
                        "Ingles",
                        "Español",
                        "Portugues"
                ]
        },
        "fechaReserva" : ISODate("2015-08-21T14:56:00Z"),
        "mes" : "Ago",
        "diasHospedaje" : 1,
        "origen" : "Llamada",
        "Equipaje" : 1,
        "calificacion" : 1,
        "monedaPago" : "Dolar"
}
>
```

Donde podemos usar MongoDB?

- Aplicaciones donde el crecimiento de registros sea grande y a gran velocidad
- Aplicaciones donde la estructura de los datos varíe sin previo aviso
- Aplicaciones donde distintas entidades se puedan relacionar entre si
- Aplicaciones CMS o blogging
- Debido a sus características escalables y soporte en velocidad es perfecto para Web-analytics y Real - Time analytics.
- Por el comportamiento variable de órdenes de compras y productos es una buena opción para E-Commerce

Ventajas

- En una sola consulta se puede obtener mucha información
- Los conceptos entre un RDBMS y MongoDB no difieren mucho
- Facilidad en la integración con aplicaciones
- Posee variedad de drivers que facilitan su utilización en distintos ambientes
- Permite trabajar JavaScript
- A pesar de ser Open Source, posee un gran soporte y escalabilidad
- Manejo sencillo e intuitivo

Rank			DBMS	Database Model	Score		
Aug 2017	Jul 2017	Aug 2016			Aug 2017	Jul 2017	Aug 2016
1.	1.	1.	Oracle	Relational DBMS	1367.88	-7.00	-59.85
2.	2.	2.	MySQL	Relational DBMS	1340.30	-8.81	-16.73
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1225.47	-0.52	+20.43
4.	4.	↑ 5.	PostgreSQL	Relational DBMS	369.76	+0.32	+54.51
5.	5.	↓ 4.	MongoDB	Document store	330.50	-2.27	+12.01
6.	6.	6.	DB2	Relational DBMS	197.47	+6.22	+11.58
7.	7.	↑ 8.	Microsoft Access	Relational DBMS	127.03	+0.90	+2.98
8.	8.	↓ 7.	Cassandra	Wide column store	126.72	+2.60	-3.52
9.	9.	↑ 10.	Redis	Key-value store	121.90	+0.38	+14.57
10.	10.	↑ 11.	Elasticsearch	Search engine	117.65	+1.67	+25.16
11.	11.	↓ 9.	SQLite	Relational DBMS	110.85	-3.02	+0.99
12.	12.	12.	Teradata	Relational DBMS	79.23	+0.86	+5.59
13.	↑ 14.	↑ 14.	Solr	Search engine	66.96	+0.93	+1.18
14.	↓ 13.	↓ 13.	SAP Adaptive Server	Relational DBMS	66.92	+0.00	-4.13
15.	15.	15.	HBase	Wide column store	63.52	-0.10	+8.01
16.	16.	↑ 17.	Splunk	Search engine	61.46	+1.17	+12.56
17.	17.	↓ 16.	FileMaker	Relational DBMS	59.65	+1.00	+4.64
18.	18.	↑ 20.	MariaDB	Relational DBMS	54.70	+0.33	+17.82
19.	19.	19.	SAP HANA	Relational DBMS	47.97	+0.03	+5.24
20.	20.	↓ 18.	Hive	Relational DBMS	47.30	+1.10	-0.51
21.	21.	21.	Neo4j	Graph DBMS	38.00	-0.52	+2.43
22.	22.	↑ 25.	Amazon DynamoDB	Document store	37.62	+1.16	+11.02
23.	23.	↑ 24.	Couchbase	Document store	32.97	-0.05	+5.57
24.	24.	↓ 23.	Memcached	Key-value store	29.96	+1.43	+2.27
25.	25.	↓ 22.	Informix	Relational DBMS	27.42	-0.25	-1.63

MongoDB + R

Muchas Gracias!!!

También puedes visitar nuestros blogs

blog.jacagudelo.com

datavictor.wordpress.com