



# Conceptos y Aplicaciones de Big Data

Presentación de la materia  
Conceptos de Big Data

Prof. Waldo Hasperué  
[whasperue@lidi.info.unlp.edu.ar](mailto:whasperue@lidi.info.unlp.edu.ar)

# Presentación de la materia

- Profesor: Waldo Hasperué  
whasperue@lidi.info.unlp.edu.ar
- Horario: martes de 15 a 18 en el aula 10B.
- Contacto con la cátedra a través del entorno IDEAS.
- La materia es de carácter introductorio y brinda un enfoque práctico del tema.
- Se presentaran diferentes frameworks y tecnologías usadas en aplicaciones de Big Data, donde se estudiará su funcionamiento y aplicación.
- Cada vez más empresas están utilizando tecnologías en Big Data.

# Presentación de la materia

- La materia se dicta bajo la **modalidad de taller**.
- Con cada framework estudiado se realizarán pequeños desarrollos en Java, Python o SQL.
- Durante la cursada los alumnos deberán realizar trabajos integradores. Para el desarrollo de las actividades la cátedra provee:
  - una máquina virtual de VirtualBox.
  - una imagen Docker.

# Evaluación

- **Modalidad presencial**
  - 70% de asistencia
  - Aprobar tres trabajos integradores durante la cursada (en grupo)
  - Examen parcial: evaluación reducida individual
- **Modalidad no presencial**
  - Aprobar los trabajos integradores durante la cursada y defenderlos mediante coloquio
  - Examen final: evaluación convencional

# Contenidos de la materia

- Fundamentos y conceptos de Big Data
- Frameworks para soluciones en Big Data
  - MapReduce
  - Spark

# Temario de la clase

- ¿Qué es Big Data?
  - Definición y dimensiones en Big Data.
  - Aplicaciones de Big Data.
  - Modelos de datos y modelos de procesamiento en Big Data
- Herramientas de Big Data
- Casos de uso

# ¿Qué es Big Data?

- Big Data no es fácil de definir, es un término que fue “inventado por el marketing” y que involucra múltiples tecnologías.
- Muy utilizado en las redes sociales por los departamentos de marketing.

# ¿Qué es Big Data?

- Existe un continuo crecimiento de las redes sociales, los sitios de "archivos multimediales" y los sitios de e-comercio
- Existe un crecimiento exponencial de datos científicos y sensores de tiempo real.



# Marea de información digital

- Hoy el universo digital está compuesto por 2.7 ZB de datos.
- IDC estimates que en 2020 habrá alrededor de 450 mil millones de transacciones por día.

# Marea de información digital

- Más de 5 mil millones de personas están llamando, escribiendo, tuiteando y navegando por internet en sus dispositivos móviles.
- Se estima que en 2020 cada ser humano generará 1.7MB por segundo.

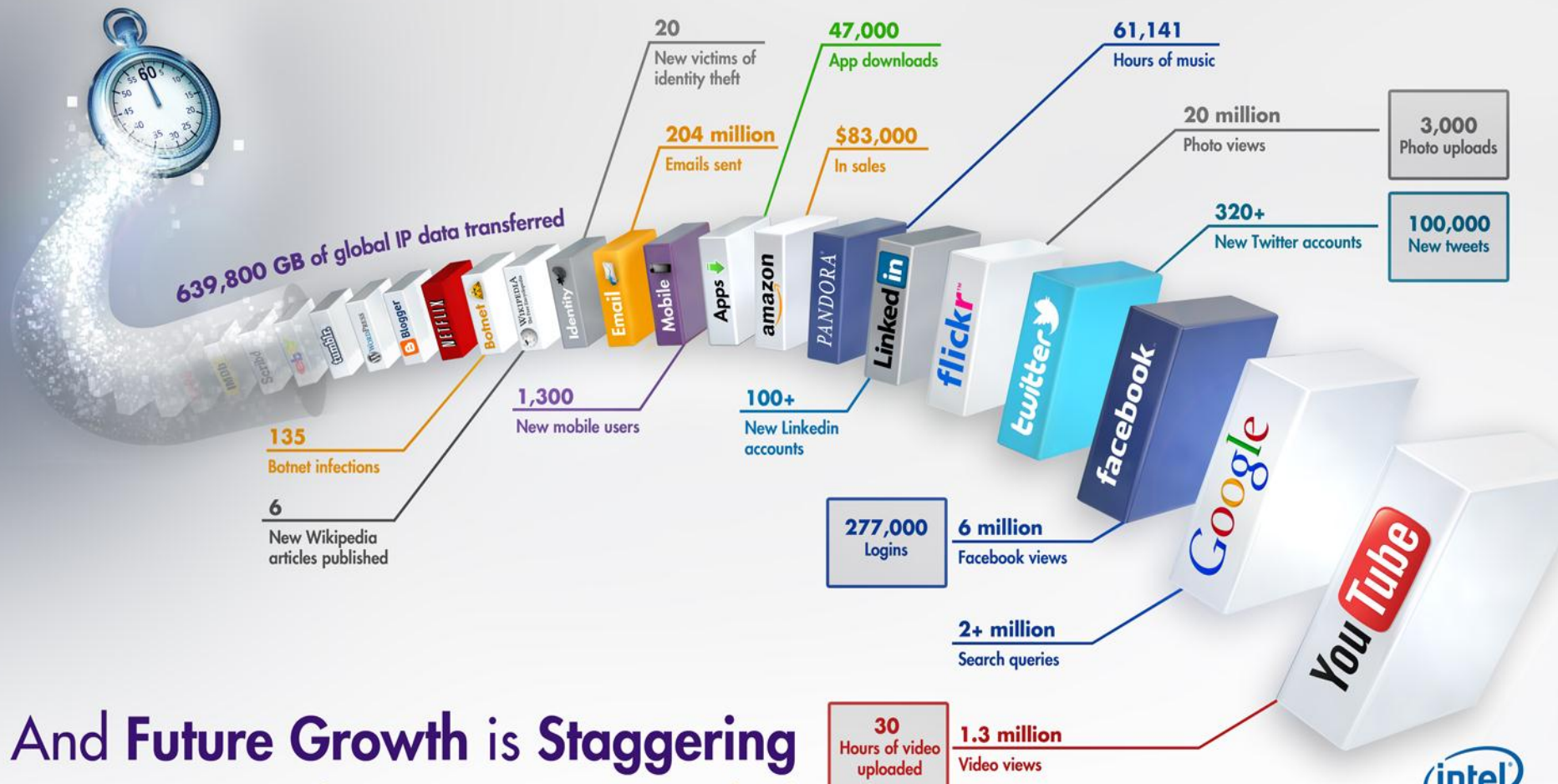
# Marea de información digital

- Facebook almacena y analiza más de 30PB de datos generados por usuarios.
- Akamai analiza 75 millones de eventos por día para mostrar avisos publicitarios.
- Walmart maneja más de un millón de transacciones por hora.
- En 2008 Google procesaba 20 PB por día.

# Marea de información digital

- Twitter gestiona más de 90 millones de tweets al día (8 terabytes de datos)
- El colisionador de partículas del CERN, puede llegar a generar 40 terabytes de datos por segundo durante los experimentos
- La base de datos más grande pertenece a AT&T. 312TB de datos. 1.9 billones de filas en una tabla.

# What Happens in an Internet Minute?



## And Future Growth is Staggering





**60GB**  
OF INFORMATION  
PER SECOND

expected to flow across  
British Telecom's networks  
(the equivalent of all of Wikipedia  
every 5 seconds)



**30%**

MORE RESULTS DATA  
will be processed during  
the 2012 London Games  
than during the 2008  
Beijing Games

**2000**  
HOURS  
of live sports media  
coverage (covering  
every single sport each  
day of the Games)

will be digitally broadcasted  
to more than...

**14,000+**  
TV and broadcast stations

with...

**4B**  
PEOPLE  
worldwide tuning in to  
watch the opening  
ceremonies



**200,000**  
HOURS

of big data will be  
generated while  
testing the IT systems  
before the Games  
even start  
(the equivalent of  
8,333 days of work)



**845**  
MILLION

monthly active Facebook  
users resulting in  
an average of

**15TB+**  
of data predicted  
to be collected  
each day during  
the Games



expected to visit the  
official Website  
of the 2012 Summer Games

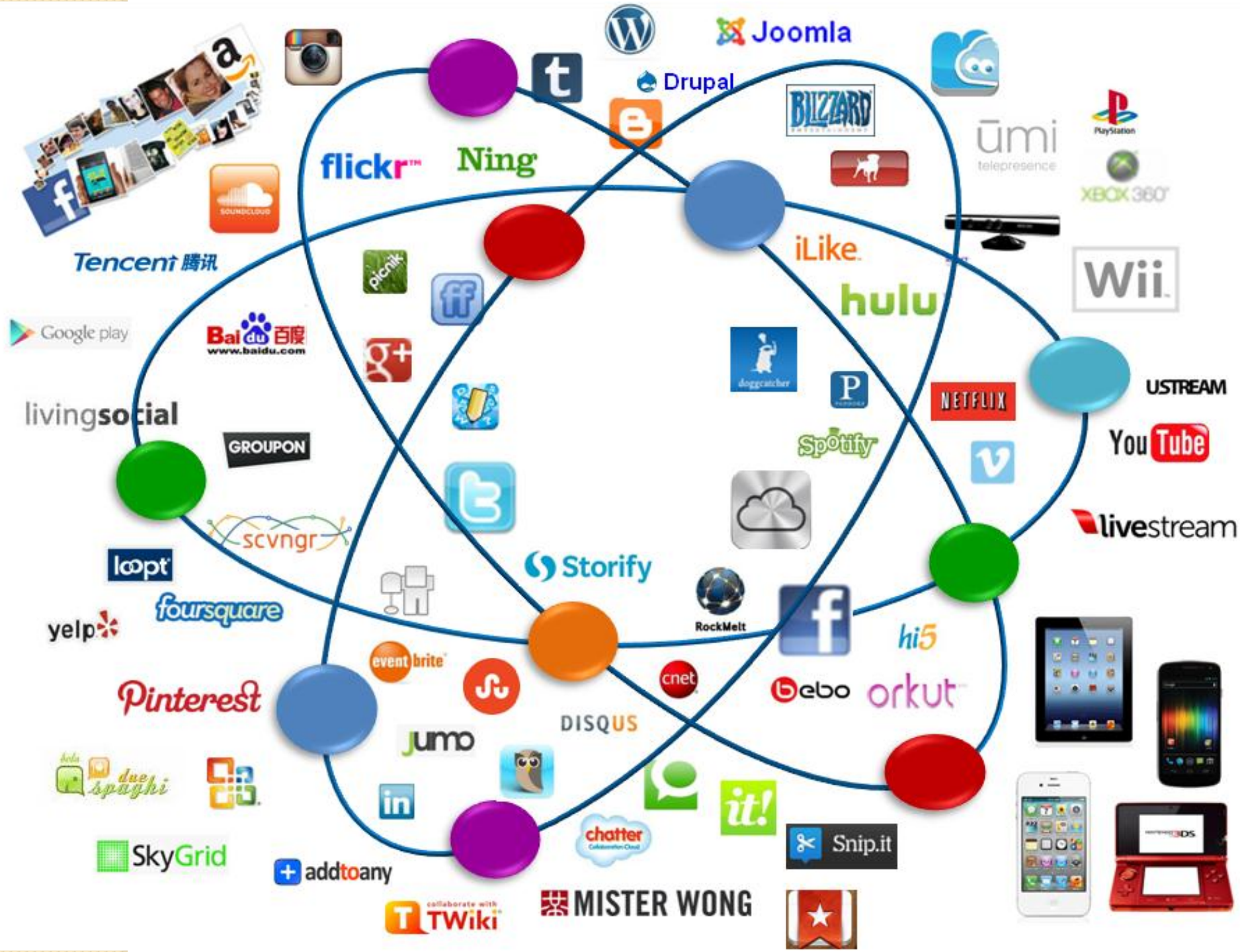
**13,000+**  
TWEETS  
PER SECOND

expected to  
be posted to  
Twitter during  
the Summer  
Games

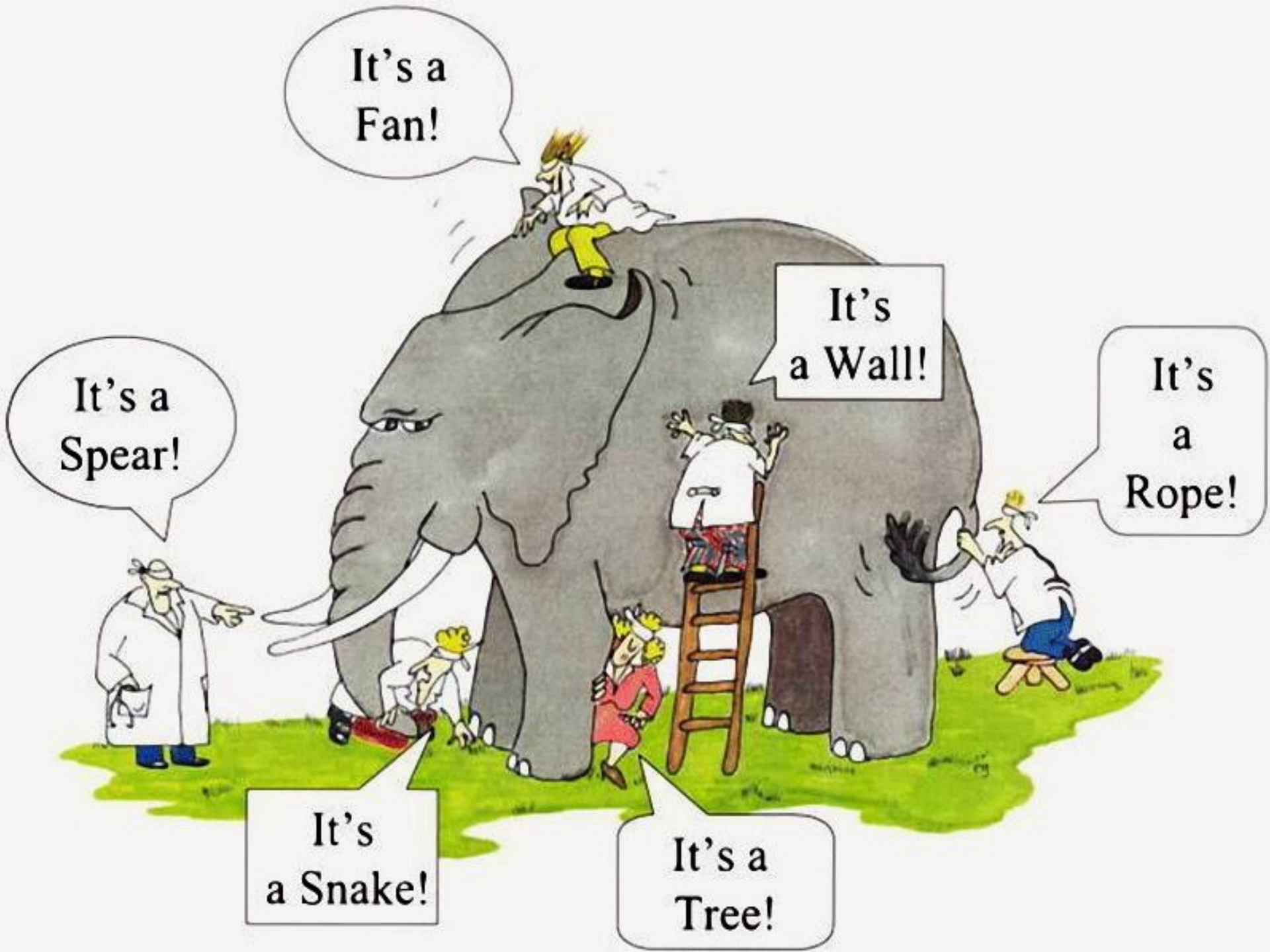
**8.5B**  
DEVICES



expected to be  
connected to the  
Internet in 2012







It's a  
Fan!

It's a  
Spear!

It's  
a Wall!

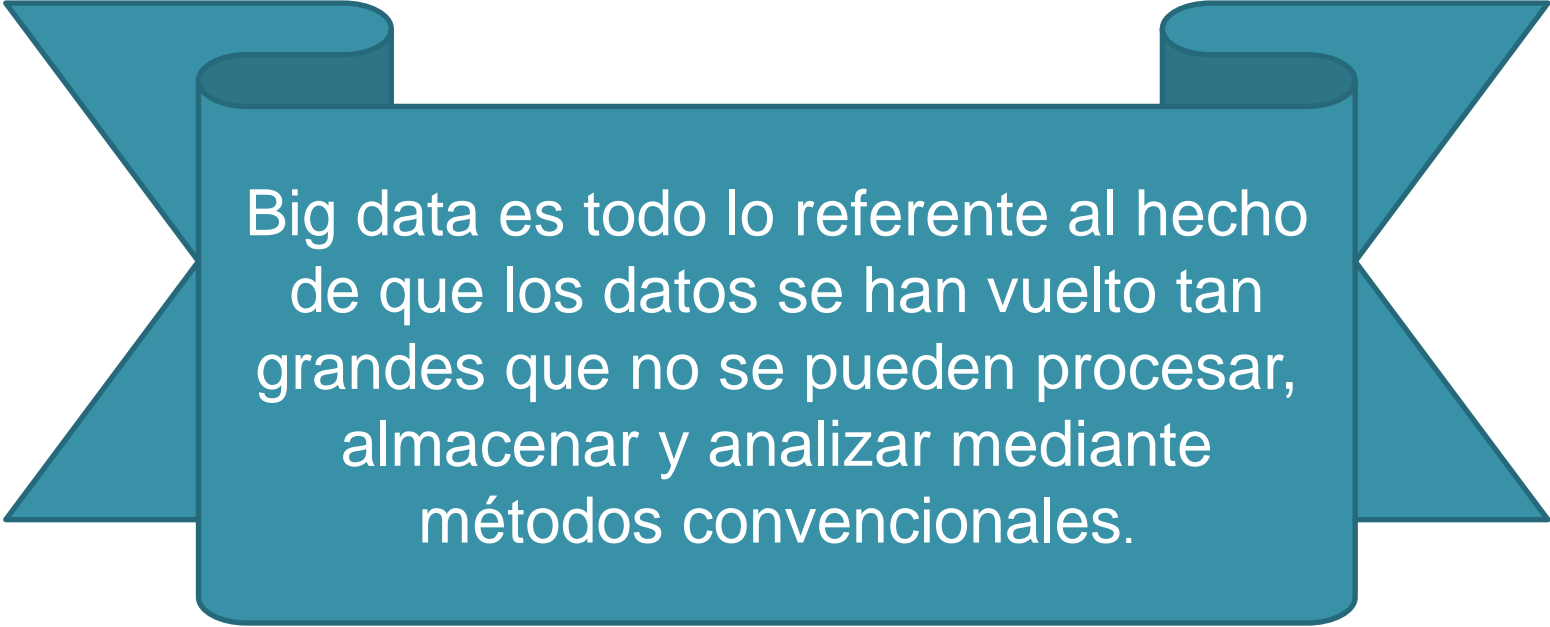
It's  
a  
Rope!

It's  
a Snake!

It's a  
Tree!



# Big Data



Big data es todo lo referente al hecho de que los datos se han vuelto tan grandes que no se pueden procesar, almacenar y analizar mediante métodos convencionales.

# Las cuatro 'V' de Big Data

- **Volumen:** el universo digital sigue expandiendo sus fronteras.
- **Velocidad:** la velocidad a la que generamos datos es muy elevada, y la proliferación de sensores es un buen ejemplo de ello. Además, los datos en tráfico –datos de vida efímera, pero con un alto valor para el negocio crecen más deprisa que el resto del universo digital.
- **Variedad:** los datos no solo crecen sino que también cambian su patrón de crecimiento, a la vez que aumenta el contenido desestructurado

# Las cuatro 'V' de Big Data

- **Valor:** Extraer valor de toda esta información marcará el futuro del manejo de información.
- El valor lo podremos encontrar en diferentes formas:
  - mejoras en el rendimiento del negocio
  - segmentación de clientes
  - tomas de decisiones
  - automatización de decisiones tácticas
  - etc.

# Datos

- Datos estructurados
  - Bases de datos relacionales
- Datos semiestructurados
  - Archivos de texto plano, planillas de cálculo
- Datos no estructurados
  - Texto escrito en lenguaje natural
  - Contenido multimedia, imágenes, fotos, audio y video

# Datos estructurados

- Generados por humanos
  - Ingreso de datos
  - Actividad web (sites, pages, clicks)
  - Datos generados por juegos
- Generados por computadoras
  - Sensores
  - Logs de aplicaciones o servidores
  - Productos con códigos de barra
  - Operaciones bancarias

# Datos no estructurados

- Generados por humanos
  - Informes, reportes
  - Redes sociales
- Generados por computadoras
  - Imágenes satelitales
  - Monitoreo (sísmicos, atmosféricos)
  - Fotografía
  - Video
  - Radares

# Datos



# DBMS

- Relacionales
  - MySQL
  - PostgreSQL
  - Derby
- No relacionales noSQL (Not only SQL)
  - MongoDB



# DBMS no relacional

- Clave/valor
  - No requieren un esquema
  - No son tipadas (por lo general todo se almacena como string)
  - Ofrecen el manejo de colecciones de clave/valor
  - Ej: Riak

# DBMS no relacional

- Documentos
  - La estructura de los documentos se almacena en formato JSON
  - Útiles cuando se generan muchos reportes
  - Ej: MongoDB, CouchDB

# DBMS no relacional

- Orientadas a columnas
  - Permite el agregado simple de columnas, estas se pueden ir llenando fila a fila
  - Es modelado usando BigTable de Google
    - Cada elemento se indexa con una fila, una columna y un timestamp
  - Ej: Hbase
- Orientadas a grafos
  - Su elemento básico es el nodo-relación
  - Se navega de nodo a nodo siguiendo las relaciones
  - Orientado a problemas con naturaleza de grafos
  - Ej: Neo4J

# ¿Tiempo real o no tiempo real?

- Problemas de tiempo real
  - Detección de fraudes
  - Detección de fallas
  - Determinar eventos en redes sociales para detectar alertas tempranas
  - Publicidad web
- Problemas de no tiempo real (batch)
  - Segmentación de clientes
  - Tomas de decisiones (semanales, mensuales, anuales)

# Big Data - Desafíos

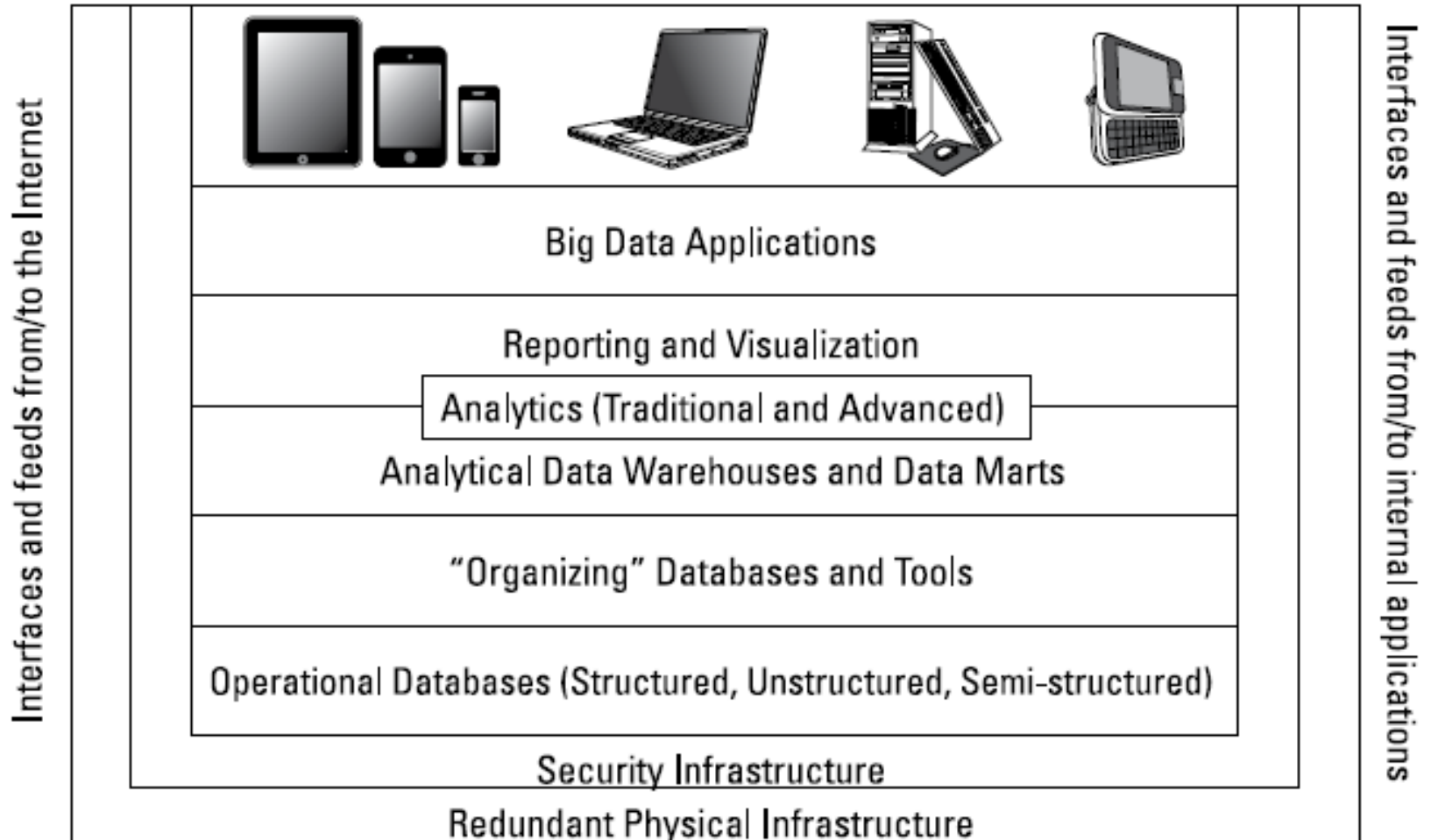
- Almacenamiento
- Procesamiento (debe ser rápido y efectivo)
- Diversidad de los datos (estructurados, no estructurados, semiestructurados)

# Tecnologías

- Big Data no es una tecnología, es la combinación de varias tecnologías para hacer más fácil el tratamiento de los datos con los que contamos hoy en día.
- Para la ejecución de aplicaciones de Big Data es necesario contar con hardware y software específico
- Clusters, sistemas distribuidos, etc.
- Cloud computing

# Tecnologías

## Big Data Tech Stack



# Casos reales

- Segmentación de clientes
  - Marketing
  - Ventas
  - *Churn* de clientes
- ¿Quién lo hace?
  - Empresas de comunicación
  - Hipermercados
  - Aseguradoras
  - Campañas electorales



# Casos reales

- Optimizando procesos de negocio
  - Manejo de stock
  - Manejo de recursos humanos
  - Optimización de rutas de reparto
- ¿Quién lo hace?
  - Cadena de puntos de venta
  - Correo

# Casos reales

- Optimización de rendimiento personal
  - Consumo de calorías
  - Nivel de condición física
  - Patrones de sueño
- ¿Quién lo hace?
  - Google Fit
  - Apple Swatch
  - Jawbone (recolecta 60 años de sueño en una sola noche)

# Casos reales

- Salud
  - Codificación de material genético
  - Dietas y alimentos adecuados
  - Descubrir la activación de genes
- ¿Quién lo hace?
  - Laboratorios
  - Farmacias
  - Hospitales

# Casos reales

- Rendimiento deportivo
  - Patrones de juego
  - Análisis del juego.
  - Imágenes y sensores
- ¿Quién lo hace?
  - SlamTracker (Tenis)
  - BNA
  - Beisbol

# Casos reales

- Seguridad
  - Fraudes
  - Cyber-ataques
  - Perfil criminal.
- Optimización de ciudades
  - Tráfico
  - Optimización de suministro (electricidad)

# Casos reales

## Ciencia



## Trading financiero



## Auto autónomo



# Herramientas

- Hadoop MapReduce
- Spark
- Gridgane
- HPCC
- Storm
- Hana
- Hive
- Kafka
- Flume

# Cloud Computing

Servers



Virtual  
Desktop



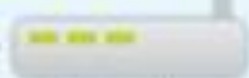
Software  
Platform



Applications



Storage/  
Data



Router



Switch



End  
User





# Cloud computing



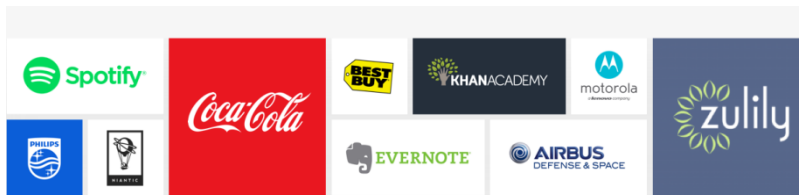
Google Cloud Platform



SOFTLAYER®



# ¿Quién usa Big Data?

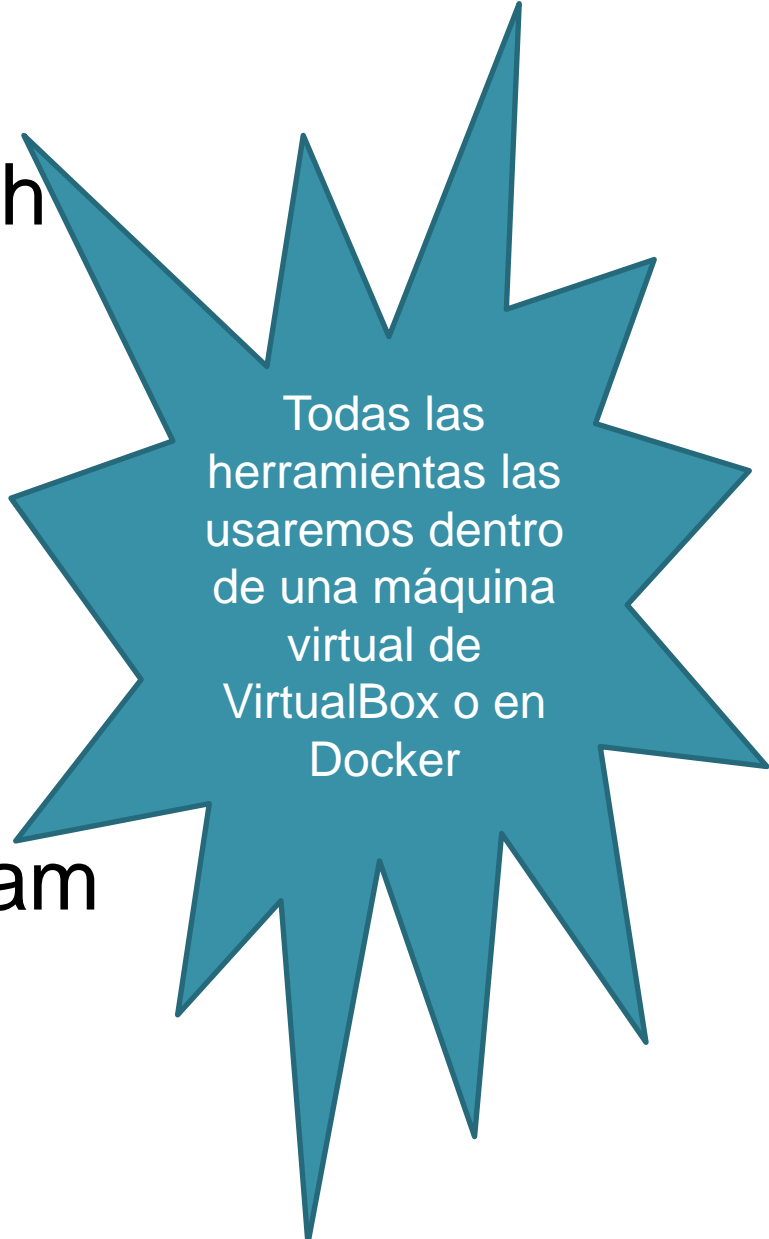


# ¿Qué veremos?

- Procesamiento batch
  - Hadoop MapReduce
  - Apache Spark
- Procesamiento stream
  - Spark streaming

# ¿Qué veremos?

- Procesamiento batch
  - Hadoop MapReduce
  - Apache Spark
- Procesamiento stream
  - Spark streaming



Todas las herramientas las usaremos dentro de una máquina virtual de VirtualBox o en Docker