Big Data

Qué es el Big Data?

Big data es un concepto que se refiere a grandes volúmenes de datos que son muy variados y veloces, al punto de que resulta muy complicado capturarlos y procesarlos con métodos tradicionales.

: IOPS

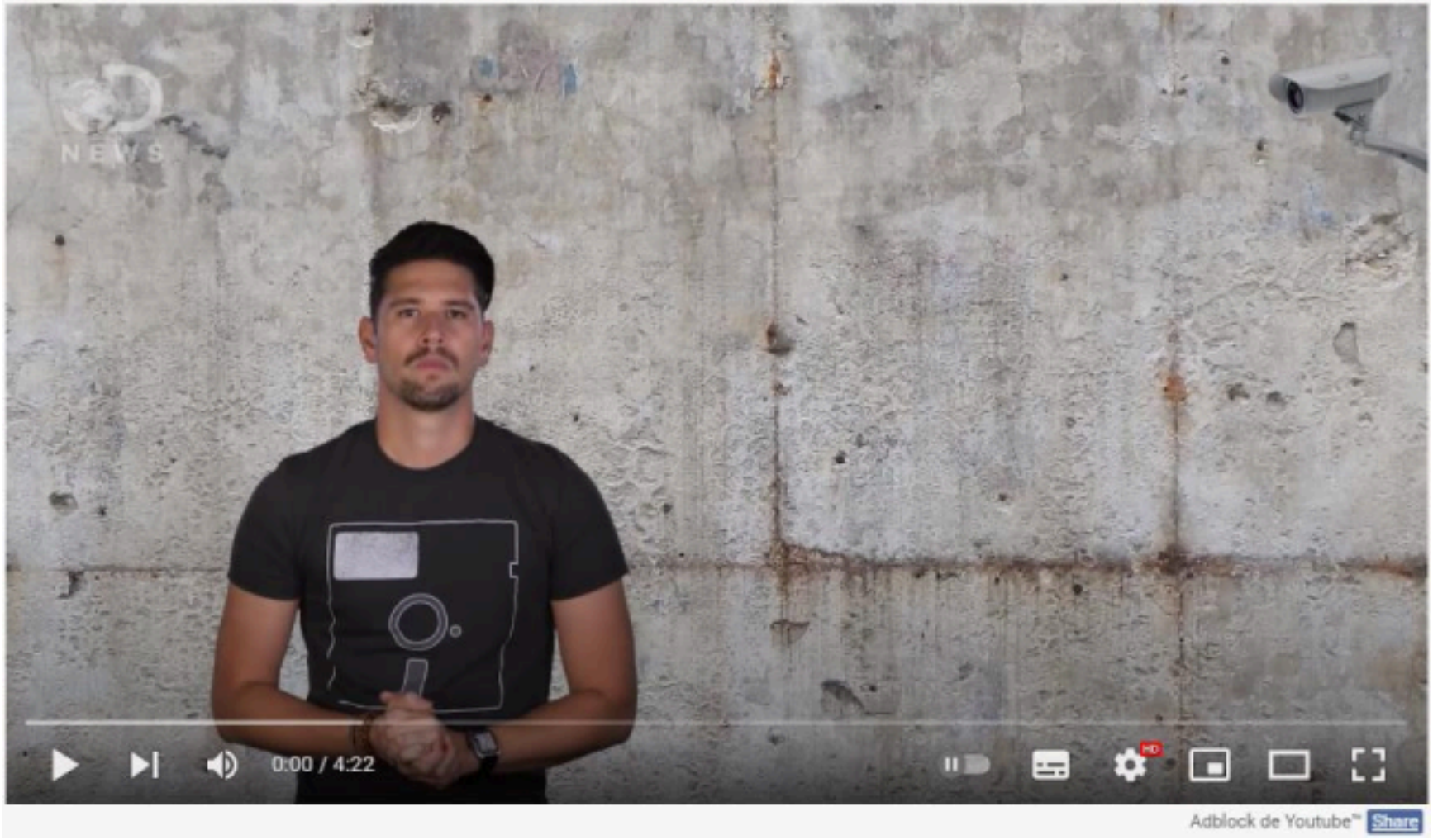
Activity

File / Object Size, Content Volume

Qué es el Big Data?

"Big Data" son datos cuyo *volumen*, *diversidad* y *complejidad* requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos ...

¿Sabes qué es BIG DATA? | Discovery en Español



Tomado de:

https://www.youtube.com/watch?v=Ju2oDsHAL-o

Algunas definiciones …

- “Big Data is the next generation of data warehousing and business analytics

and is poised to deliver top line revenues cost efficiently for enterprises”

- “Big Data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapse time

for its user population”

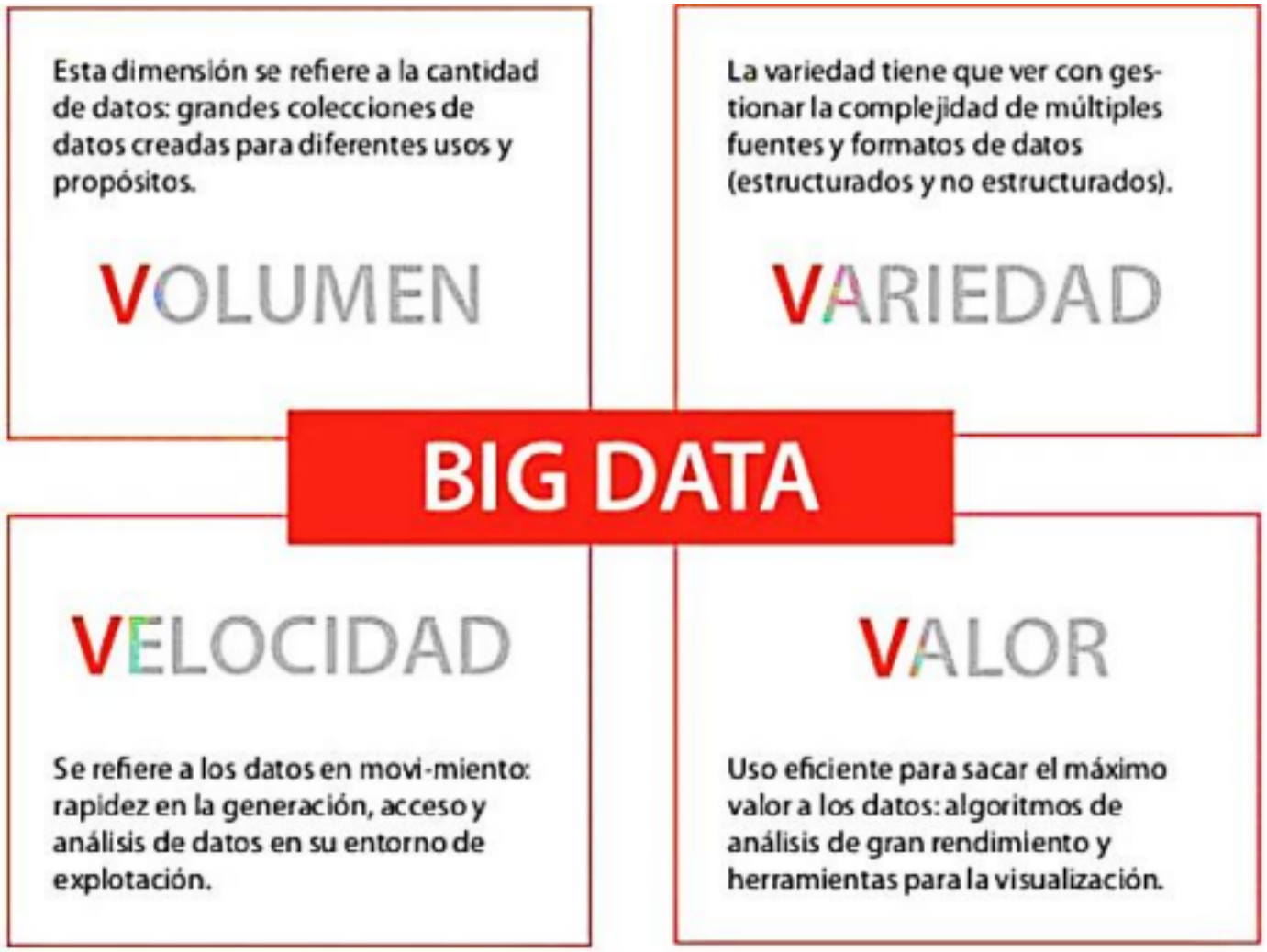
- “Disciplina que se ocupa de todas las actividades relacionadas con los sistemas que manejan grandes conjuntos de datos, principalmente, almacenamiento, búsqueda, compartición, análisis y visualización

Se considera Big Data si?

5V del Big Data 5V del Big Data	Definición Definición
Volumen	El almacenamiento de la masiva cantidad de datos que pueden ser recolectados de múltiples fuentes como páginas web, social media, IoT, etc.

Velocidad	Los datos se generan en tiempo real gracias a las interacciones con las fuentes mencionadas, por lo que deben ser procesados con la misma velocidad.
Variedad	Todo tipo de datos, ya sea estructurados o no estructurados. Podrían ser tablas, texto, imágenes, videos, audio, bases de datos, etc.
Veracidad	Es la calidad y confiabilidad de los datos. Al llegar de diversas fuentes, se vuelve complejo realizar su limpieza para evitar usar valores incorrectos.
Valor	Los datos deben poder proporcionar un valor o beneficio a la empresa que los est

Se considera Big Data si?



Predicción de propagación de la gripe

“Google puede predecir la propagación de la gripe (…) analizando lo que la gente busca en internet”

Más de 3.000M de búsquedas a diario

- 50 M de términos de búsqueda más utilizados
- Google comparó esta lista con los datos de los CDC sobre propagación de gripe entre 2003 y 2008

J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant.
Detecting influenza epidemics using search engine query data.
Nature 475 (2009) 1012-1014

Predicción de propagación de la gripe

- Encontraron una combinación de 45 términos de búsqueda que al usarse con un modelo matemático presentaba una correlación fuerte entre su predicción y las cifras oficiales de la enfermedad.

- Podían decir, como a los CDC, a dónde se había propagado la gripe pero casi en tiempo real, no una o dos semanas después,

con

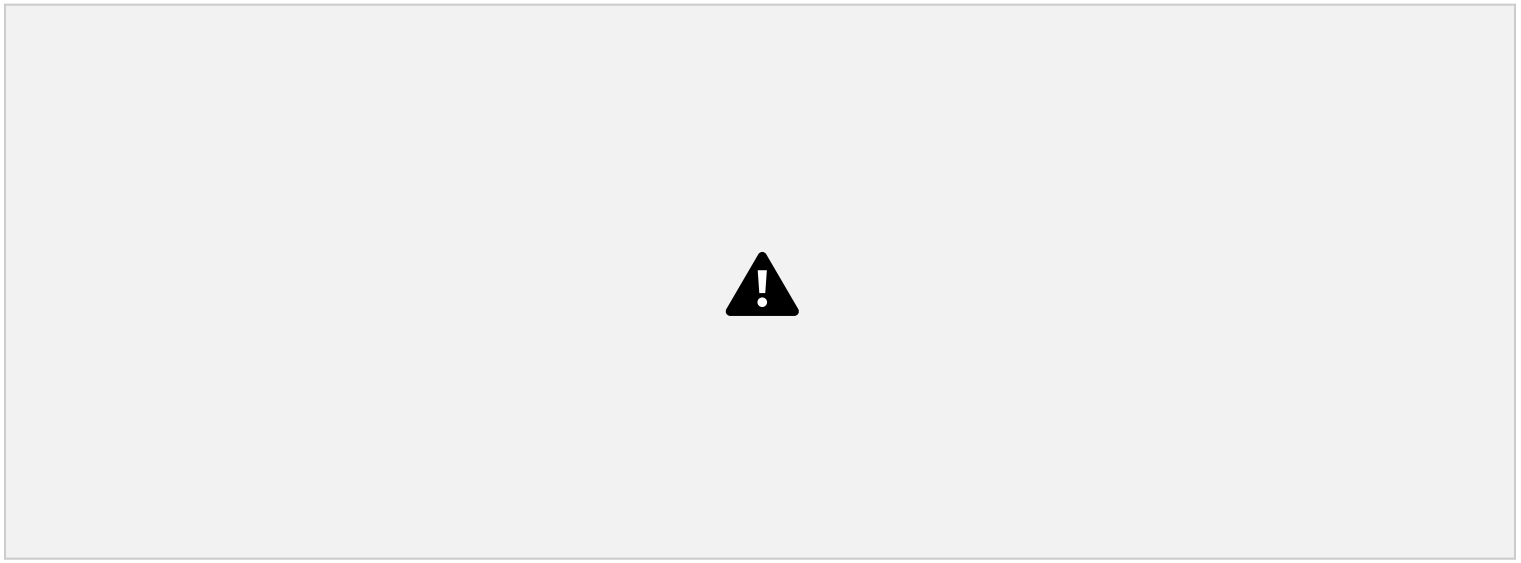
con un método basado en Big Data

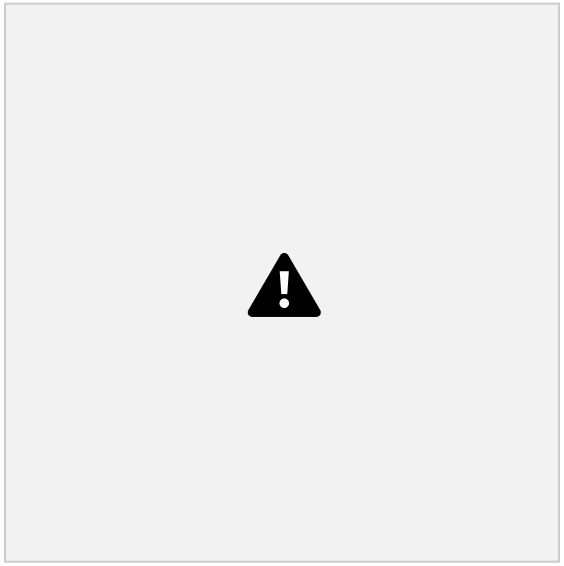
- Se ha extendido a 29 países

Predicción de propagación de la gripe

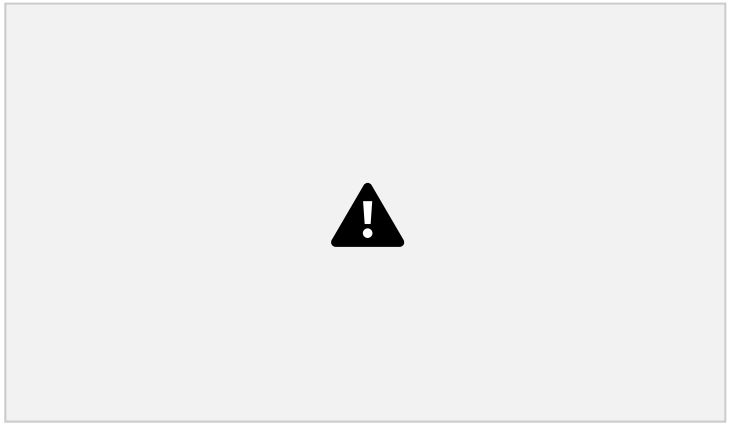
En 2013 sobreestimó los niveles de gripe (x2 la estimación CDC) o La sobreestimación puede deberse a la amplia cobertura mediática de la gripe que puede modificar comportamientos de búsqueda

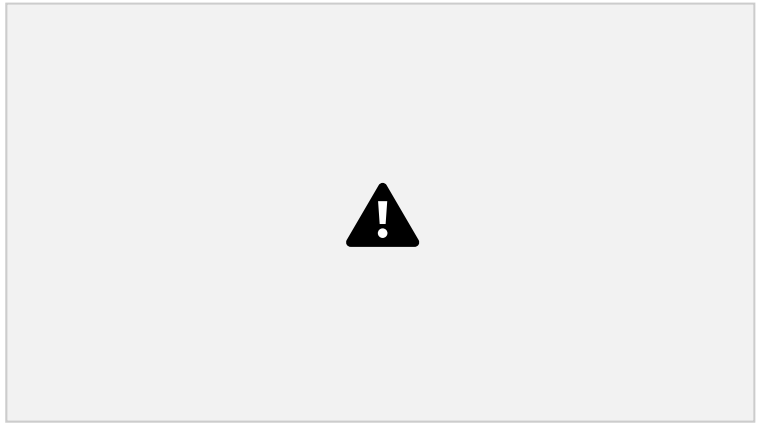
Los modelos se van actualizando anualmente

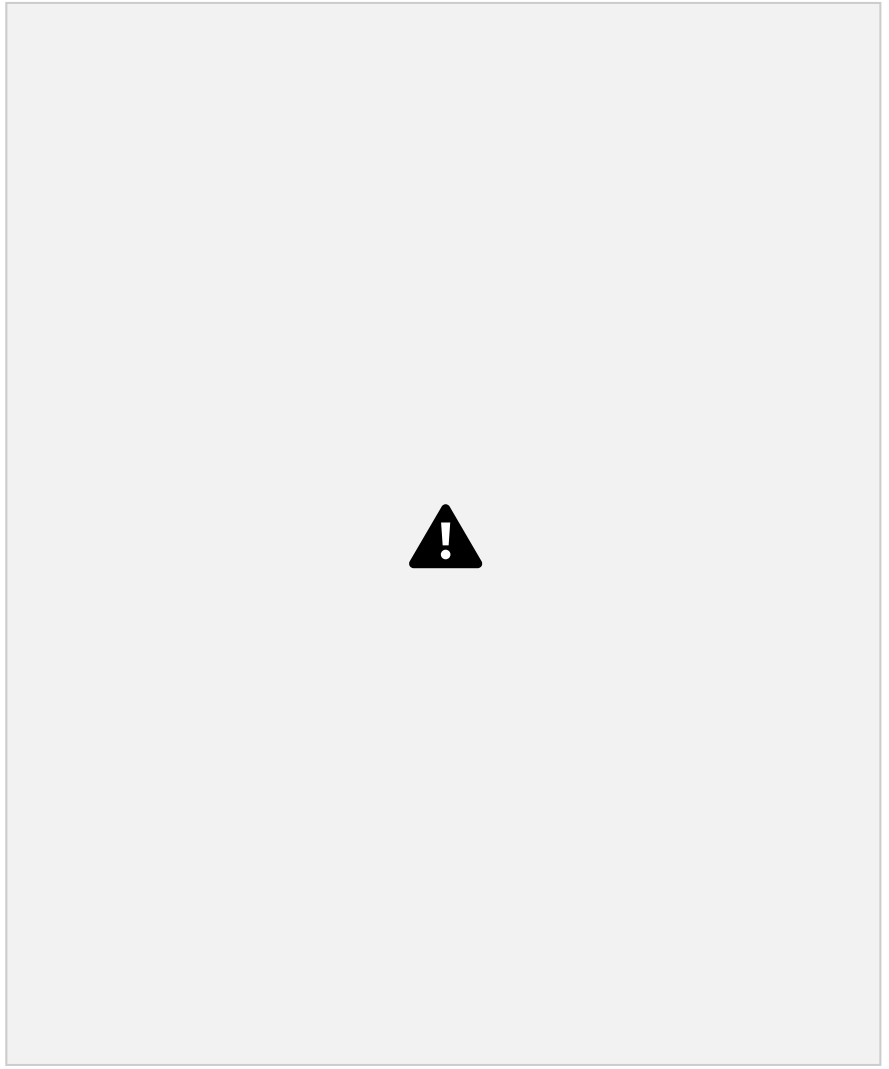




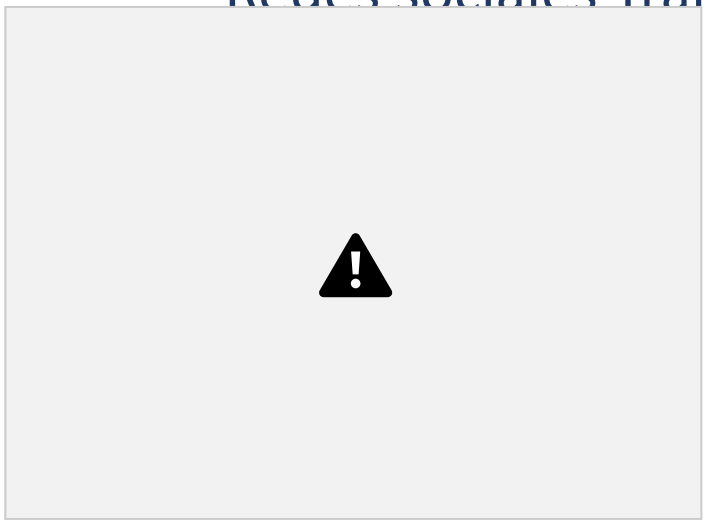
Generadores de datos masivos

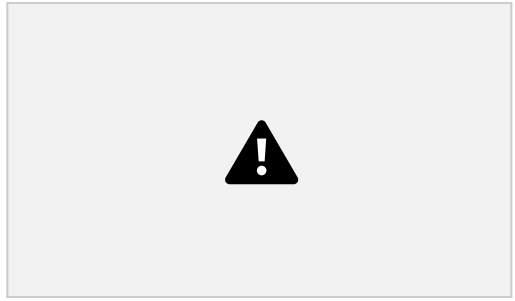


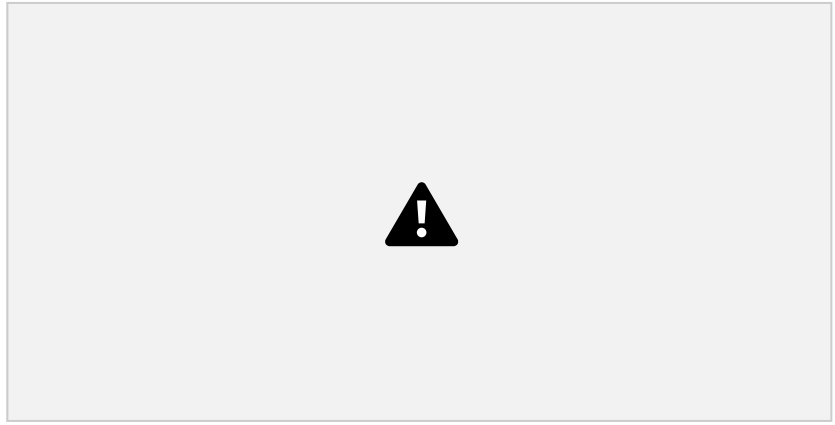




Redes sociales Transacciones bancarias



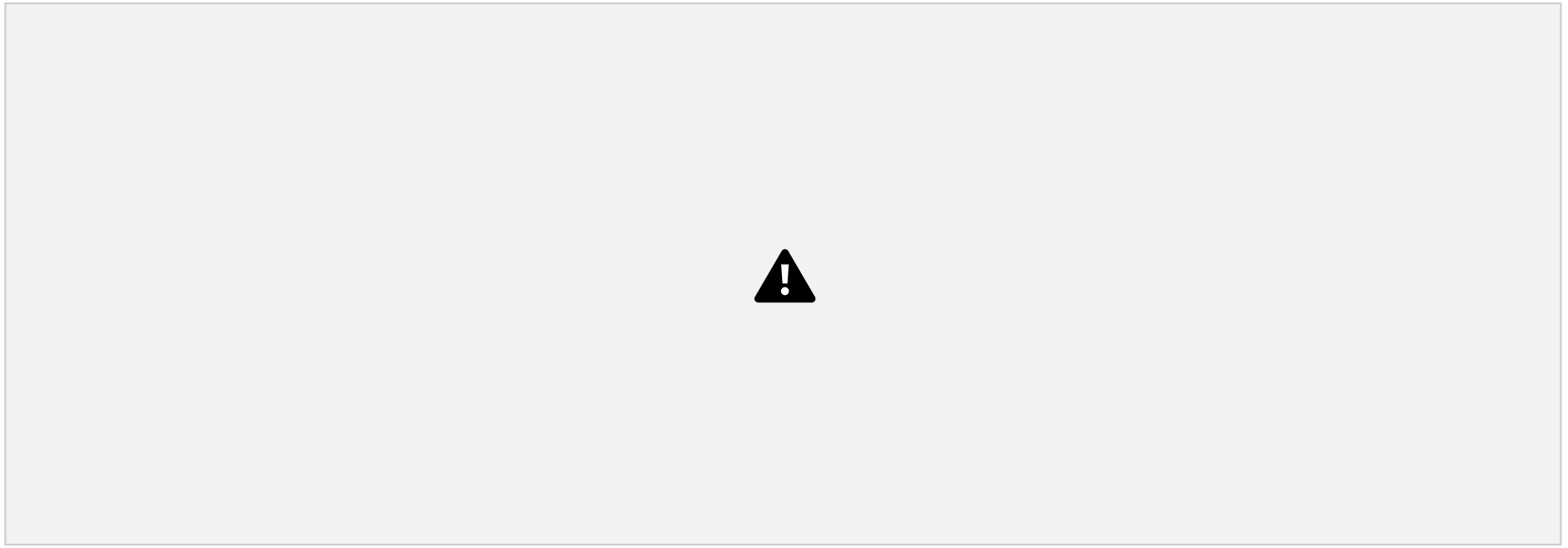


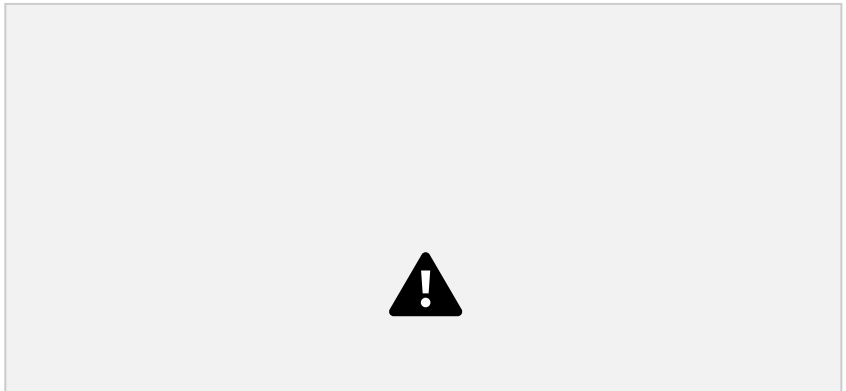


Dispositivos ubicuos de

Redes de sensores generación de contenido Instrumentos

Generadores de datos masivos





En el Colisionador de Hadrones (LHC) del CERN se pueden llegar a producir *600 millones de colisiones por segundo*, a punto de que sus *65.000 procesadores* para analizar *30 Petabytes* de datos, no son

suficientes

¿Cómo de “Big”?

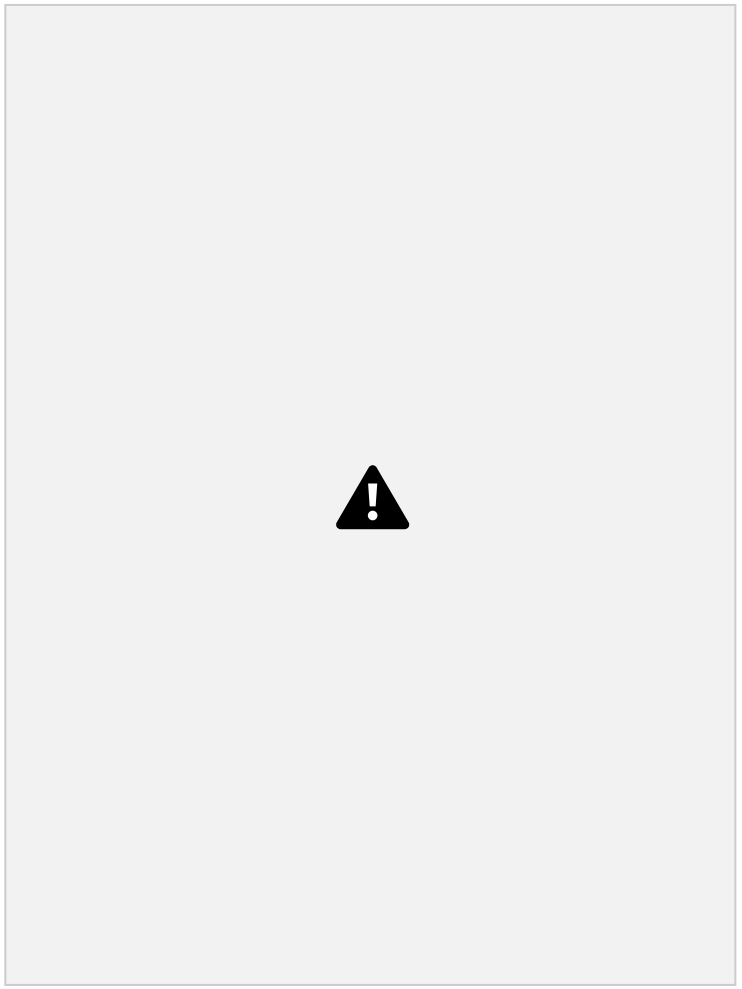
http://cruncher.ncl.ac.uk/bdcomp/index.pl?action=data

Evolutionary Computation for Big Data and Big Learning Workshop
Big Data Competition 2014: Self-deployment track

Objective: Contact map prediction
Details:

- 32 million instances
- 631 aNributes (539 real & 92 nominal values)
- 2 classes
- 98% of negative examples
- About 56.7GB of disk space

Evaluation:
True positive rate · True negative rate
TPR · TNR

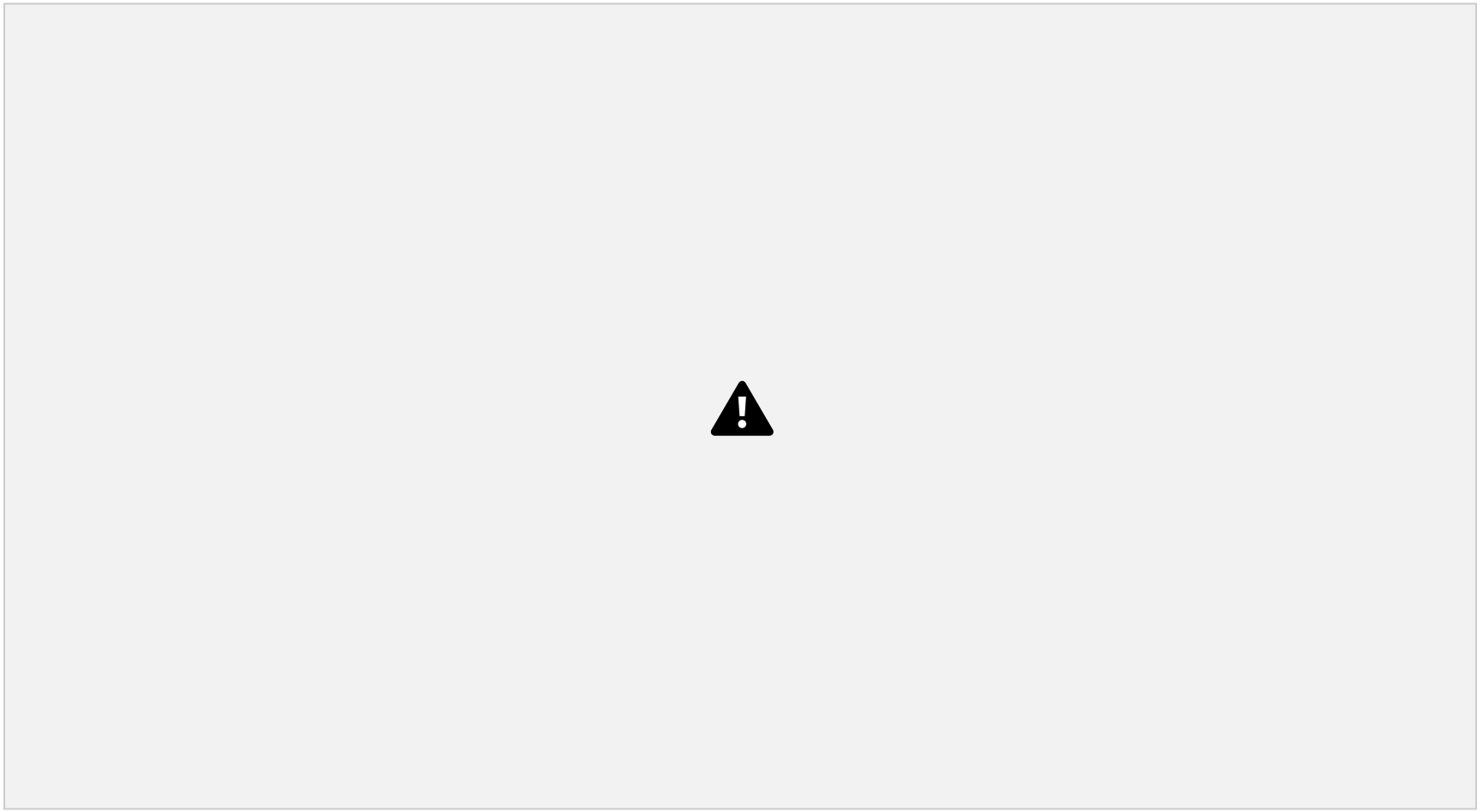


PDB

¿Cómo de “Big”?

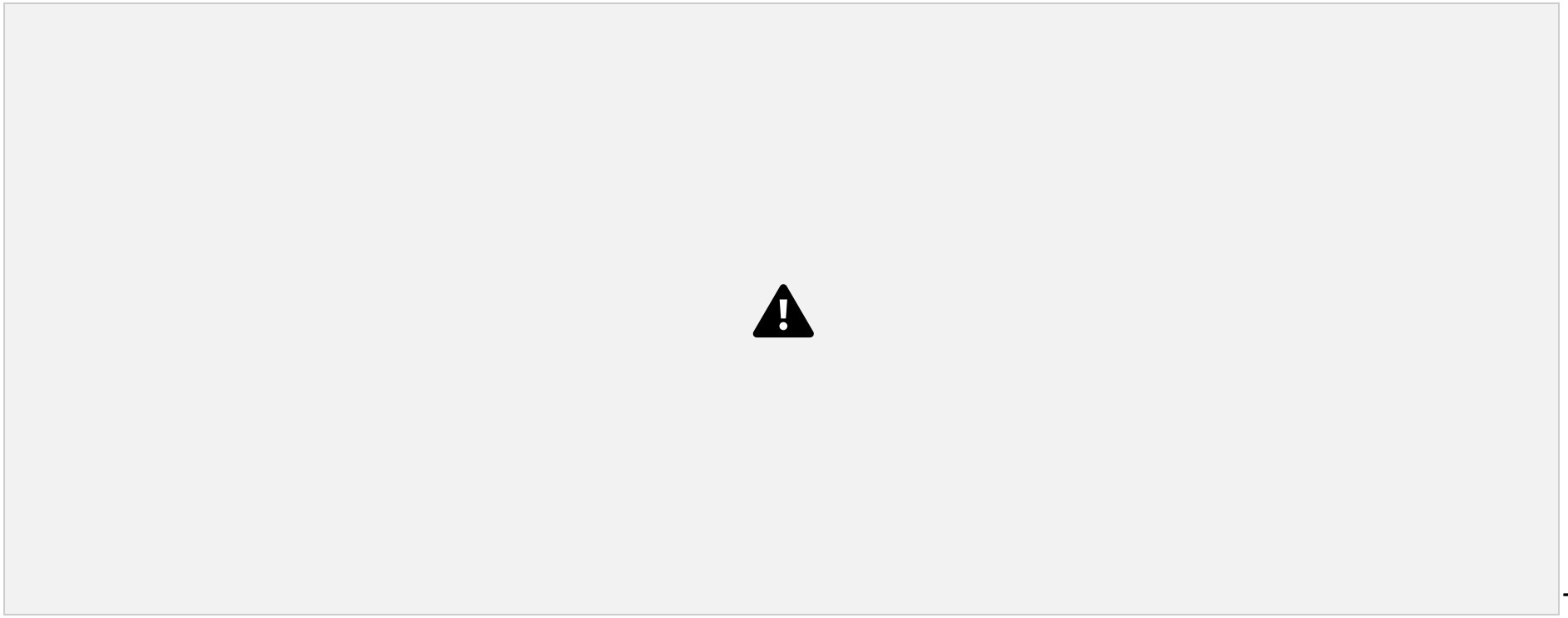
Tomado de: https://www.youtube.com/watch?v=3H20KsPlCBI

Unidades básicas de información y tratamiento de datos



Tomado de: https://www.iit.comillas.edu/documentacion/IIT-14153A/Big_data._Un_nuevo_paradigma_de_an%C3%A1lisis_de_datos.pdf

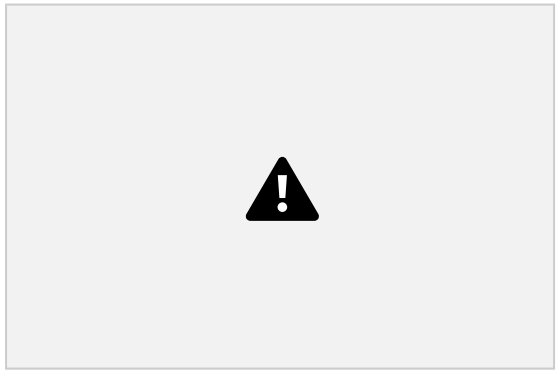
Tipos de datos en el paradigma big data

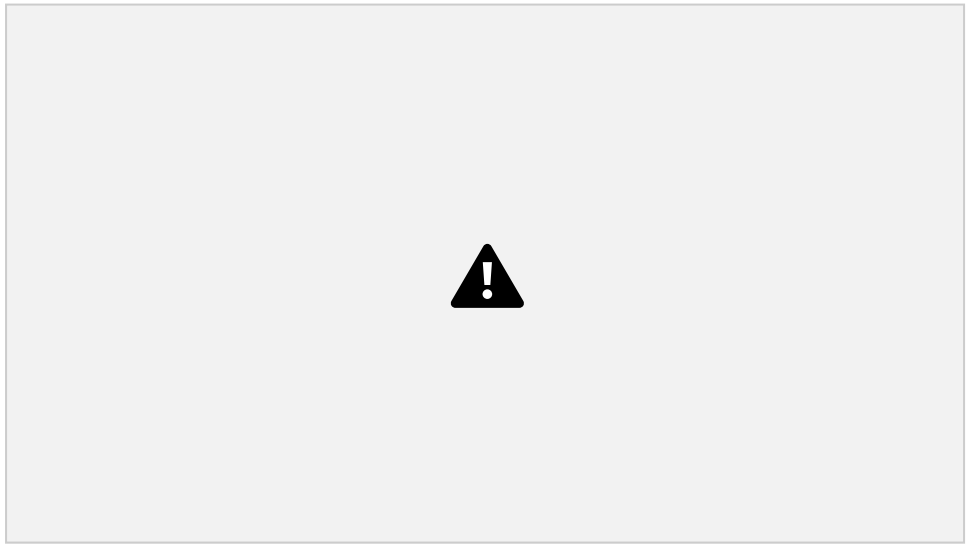


Tomado de:

https://www.iit.comillas.edu/documentacion/IIT-14 153A/Big_data._Un_nuevo_paradigma_de_an%C3%A1lisis_de_datos.pdf

Ejemplos de uso de Big Data





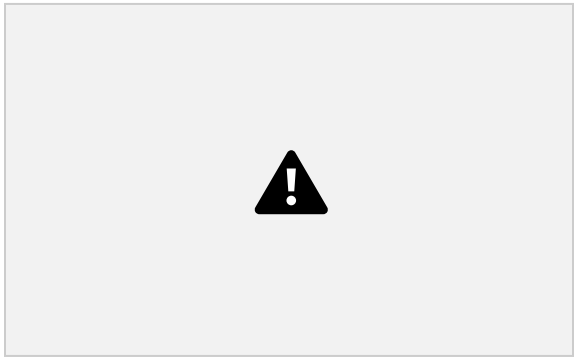
- Información personalizada (ofertas)
- Analítica de comportamiento
- Procesamiento de lenguaje natural (Sentimientos en marketing)

Beneficios del big data

- Reducción de coste: En el almacenamiento y gestión de grandes volúmenes de datos.

- Reducción de tiempo: Más eficiencia y eficacia en la toma de decisiones.

- Nuevos productos y servicios: Con la capacidad de medir y prever las necesidades y problemas de los usuarios (clientes y/o ciudadanos) se aumenta la satisfacción de los mismos.



¿Qué hacemos con estos datos?

de

Experto en computación
y desarrollo avanzados

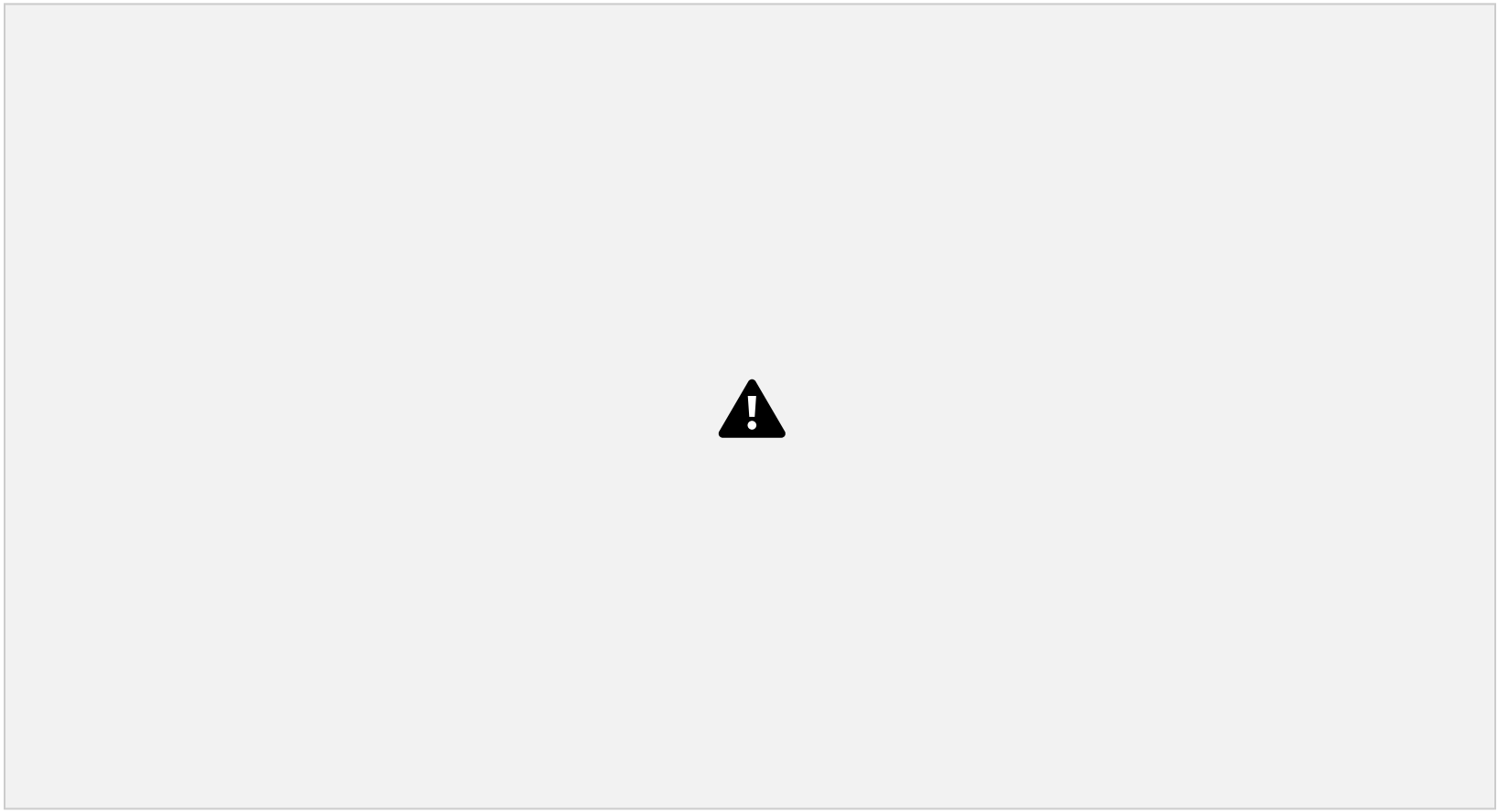
Científico

de datos

datos

Experto en el dominio

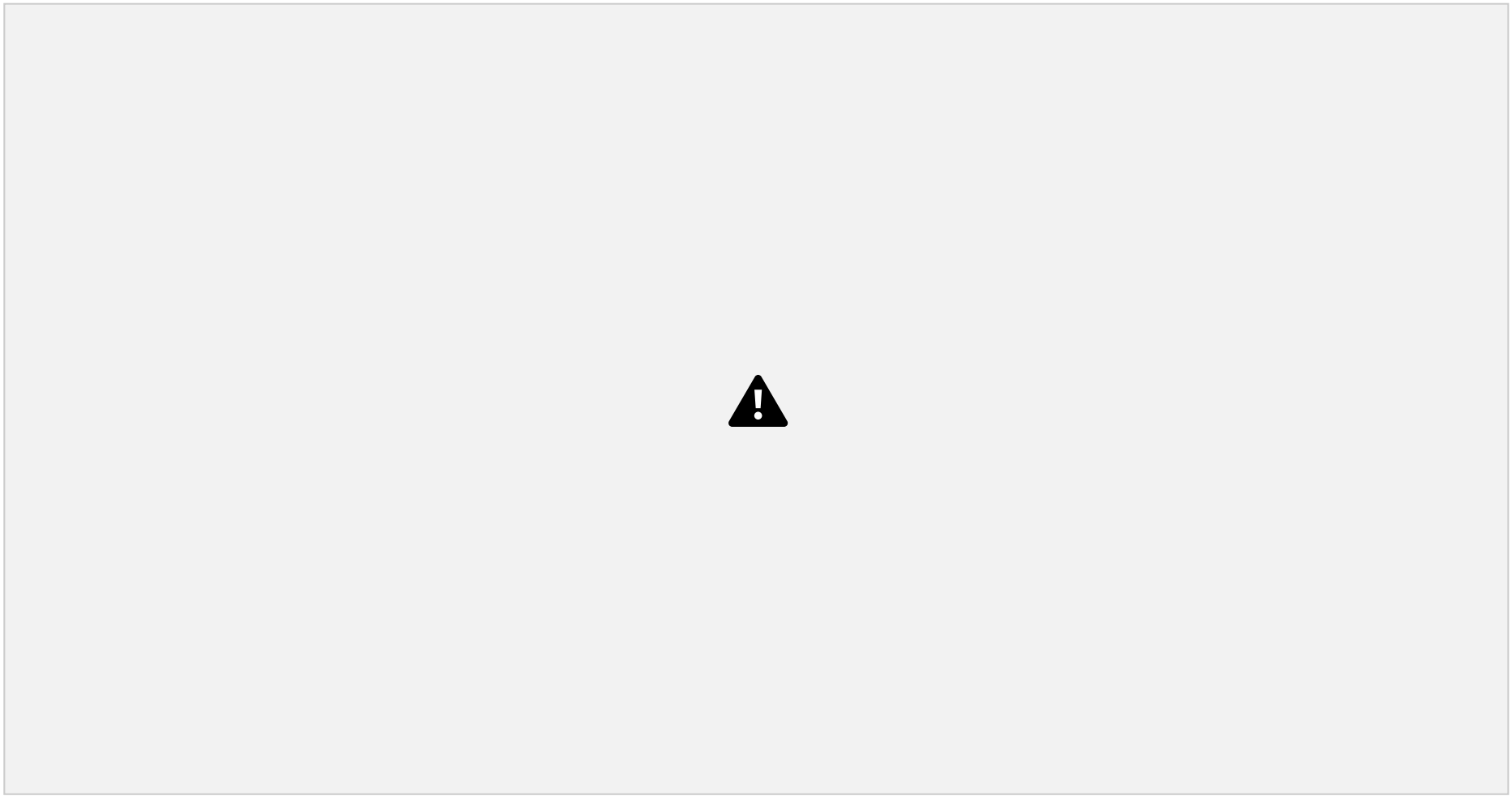
Aplicaciones del Big data



Aplicaciones del Big data

Tomado de: https://www.youtube.com/watch?v=l7-SKbcv6is

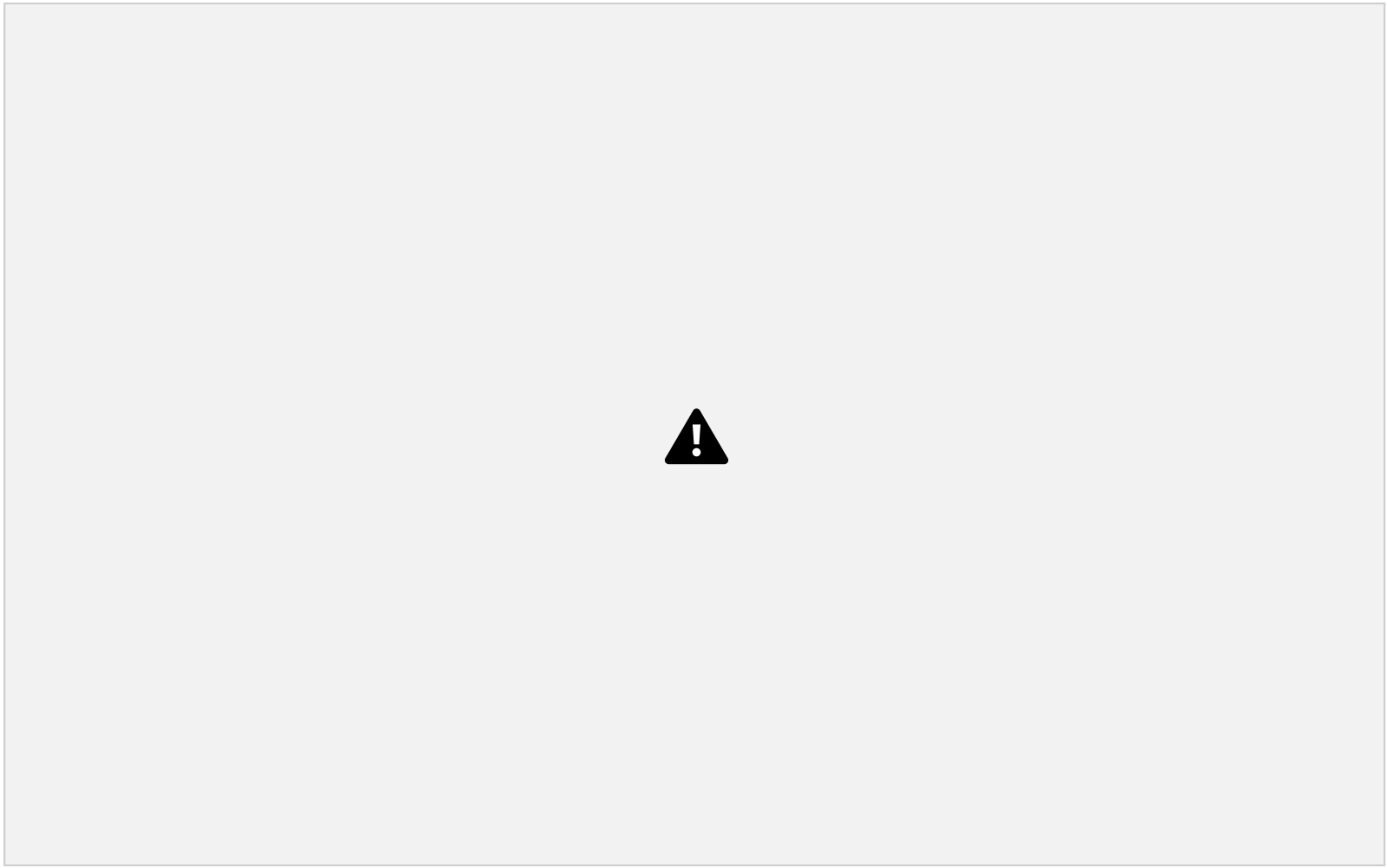
Sector financiero



Tomado de: https://es.statista.com/grafico/8484/aplicaciones-del-big-data-en-las-empresas-financieras/

Previsión de los ingresos de la industria de big data en el

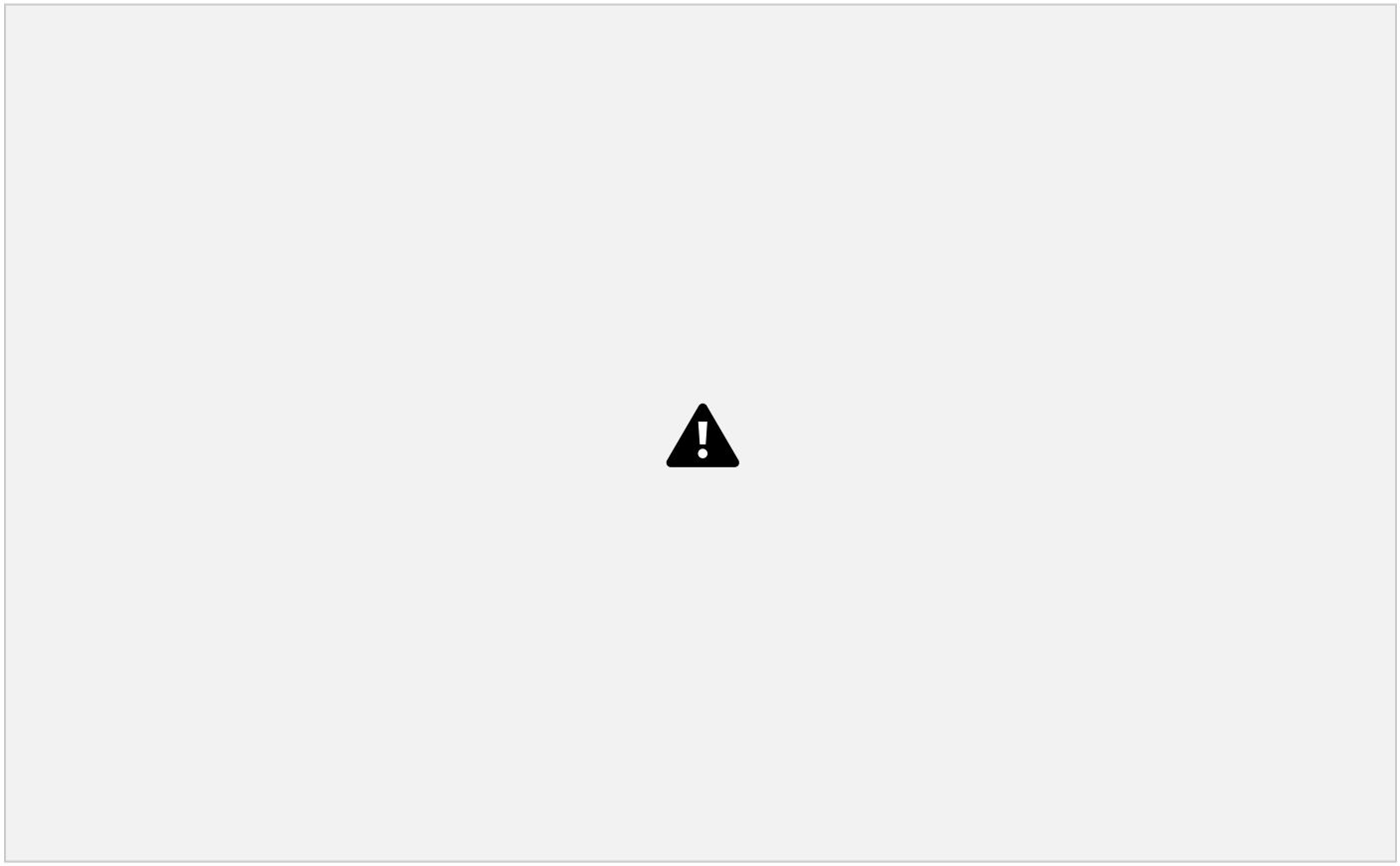
mundo entre 2017 y 2027(en miles de millones de dólares)



Tomado de: https://es.statista.com/estadisticas/517644/prevision-del-valor-de-mercado-del-big-data-en-el-mundo/

Volumen de ingresos del sector de big data a

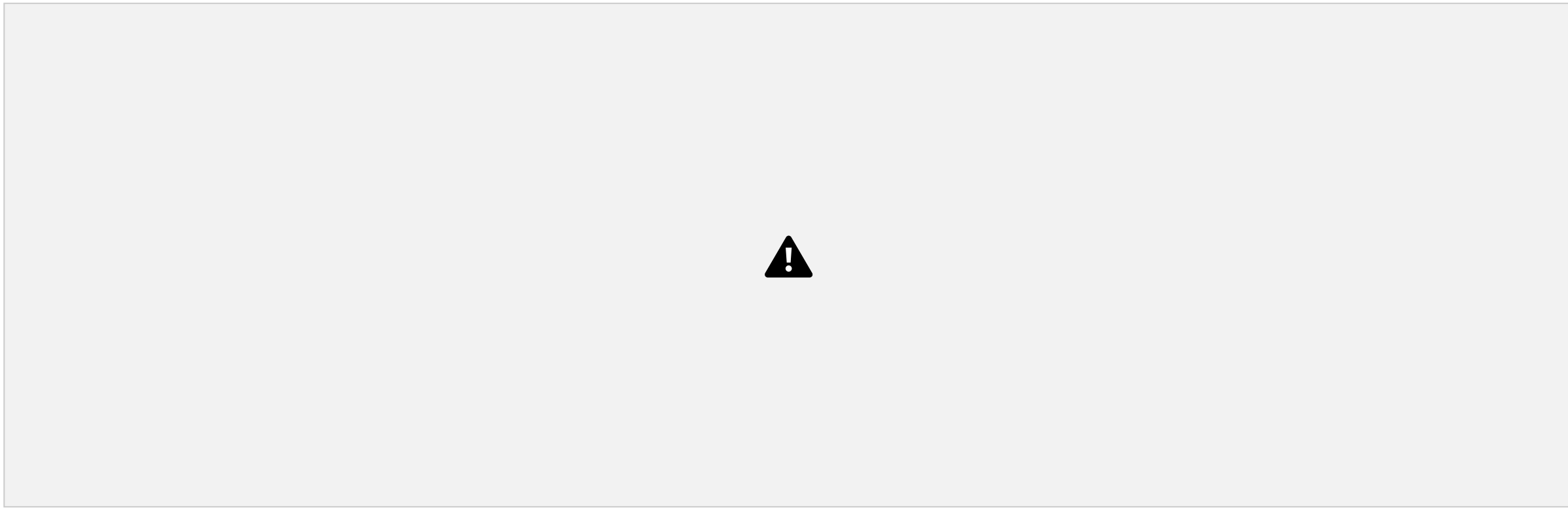
nivel mundial de 2016 y 2027, por área de negocio(en miles de millones de dólares)



En esta parte necesito de su ayuda

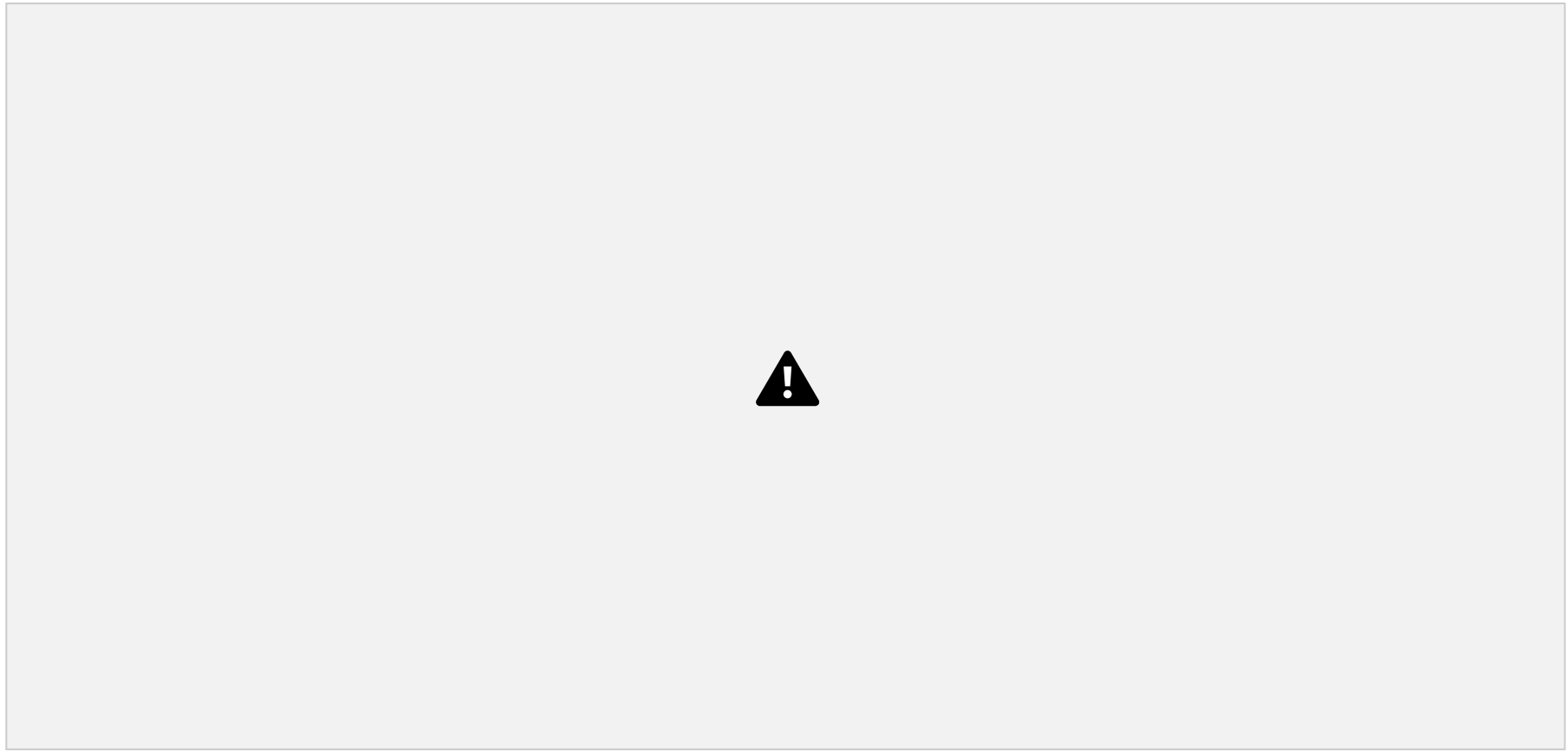
Por favor ingresen a los links de
cada diapositiva

Analítica de actualizaciones de contenido en Wikipedia en tiempo real con Spark Streaming



Ingresar a: https://bigdata.stratebi.com/spark-streaming/index.htm;jsessionid=077373E86C99D621AB8CBC691294DC11

Resultados de Tweets que contienen los términos "colombia“ Redes sociales

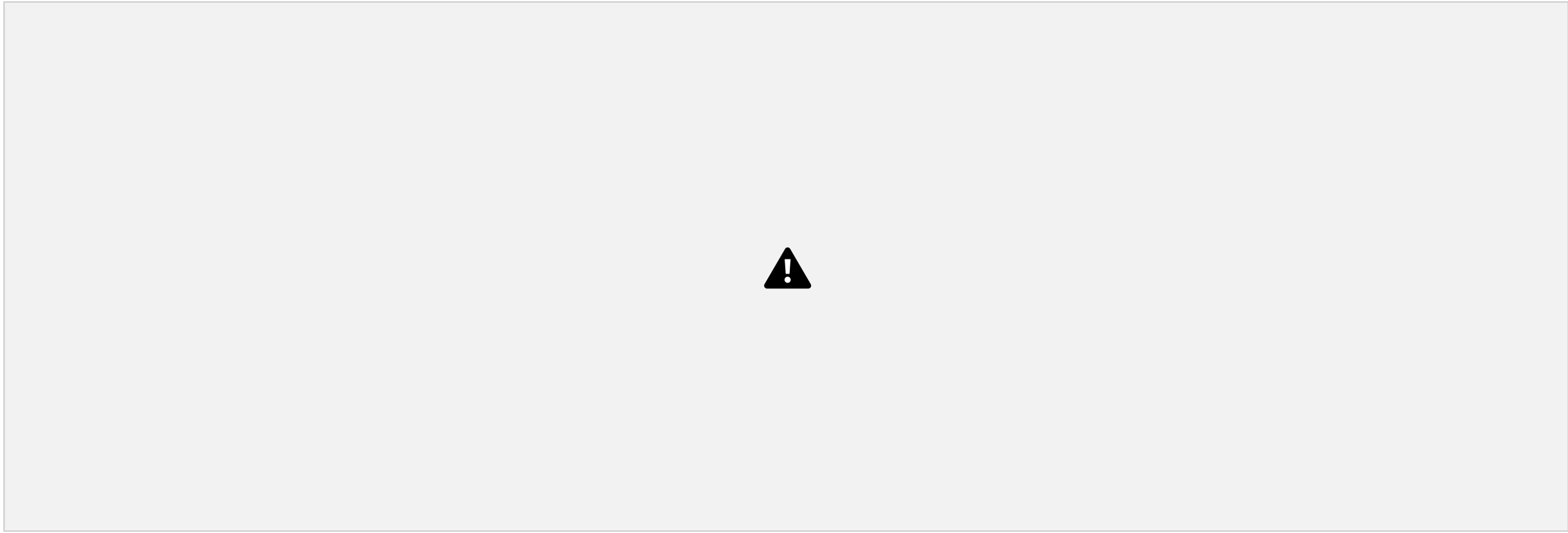


Ingresar a:

https://bigdata.stratebi.com/real-time/index.htm;jsessionid=077373E86C99D621AB8CBC691294DC11

Cuadro de Mando con Kylin y Power BI

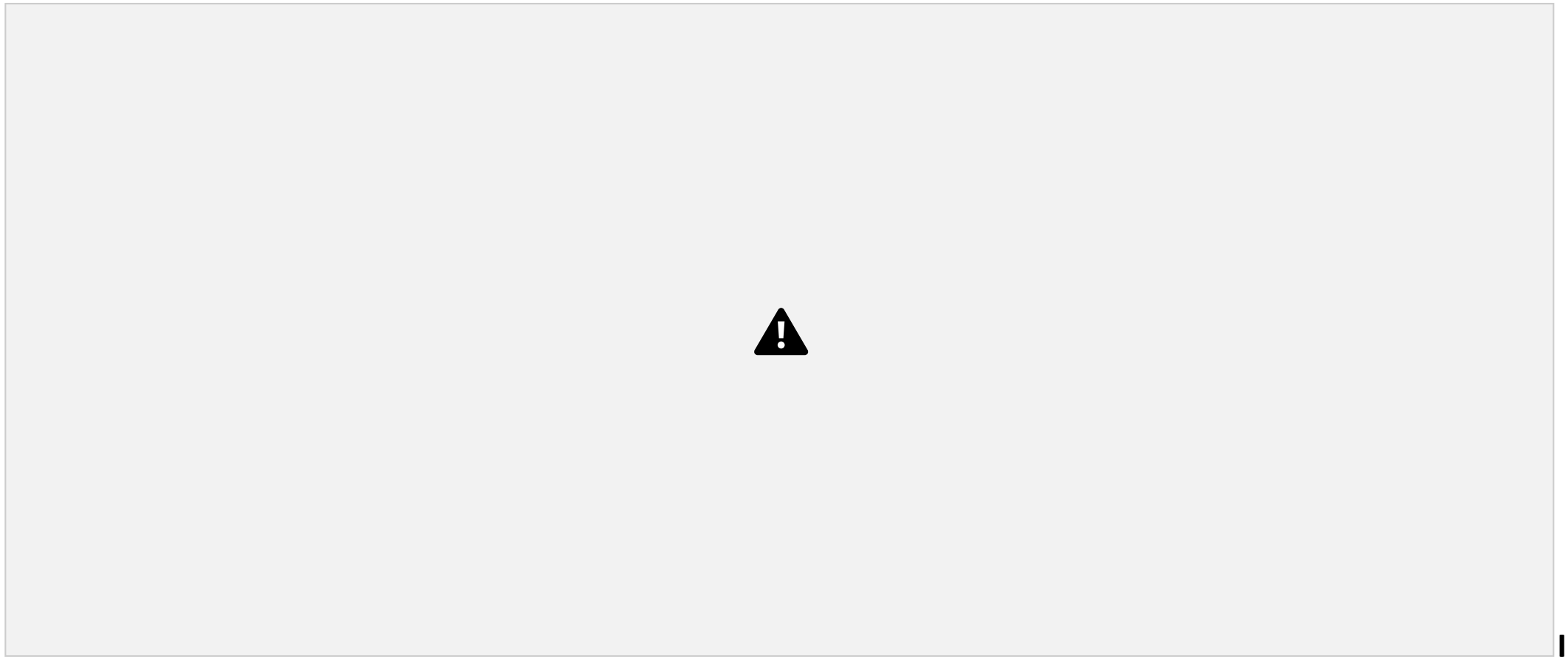
Analítica de datos



Ingresar a: https://bigdata.stratebi.com/kylin-power-bi/index.htm

IoT - Cuadro de Mando Power BI, con Talend & Vertica Internet de las cosas (Internet de las

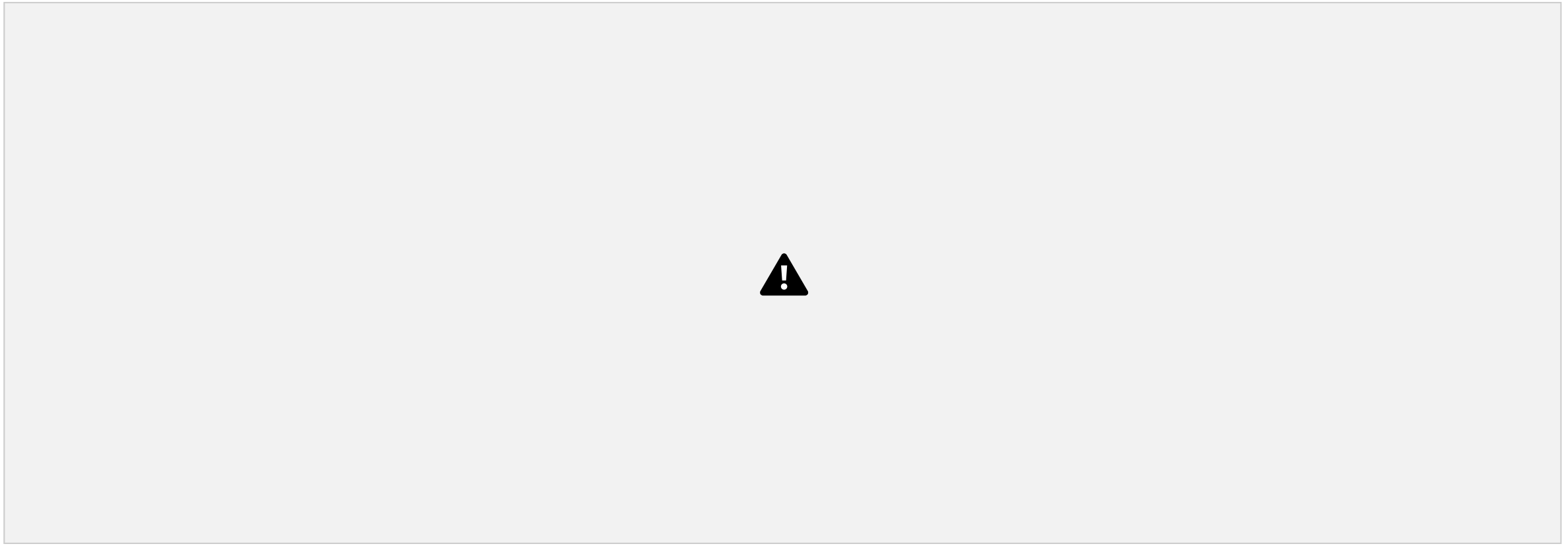
cosas)



Ingresar a:

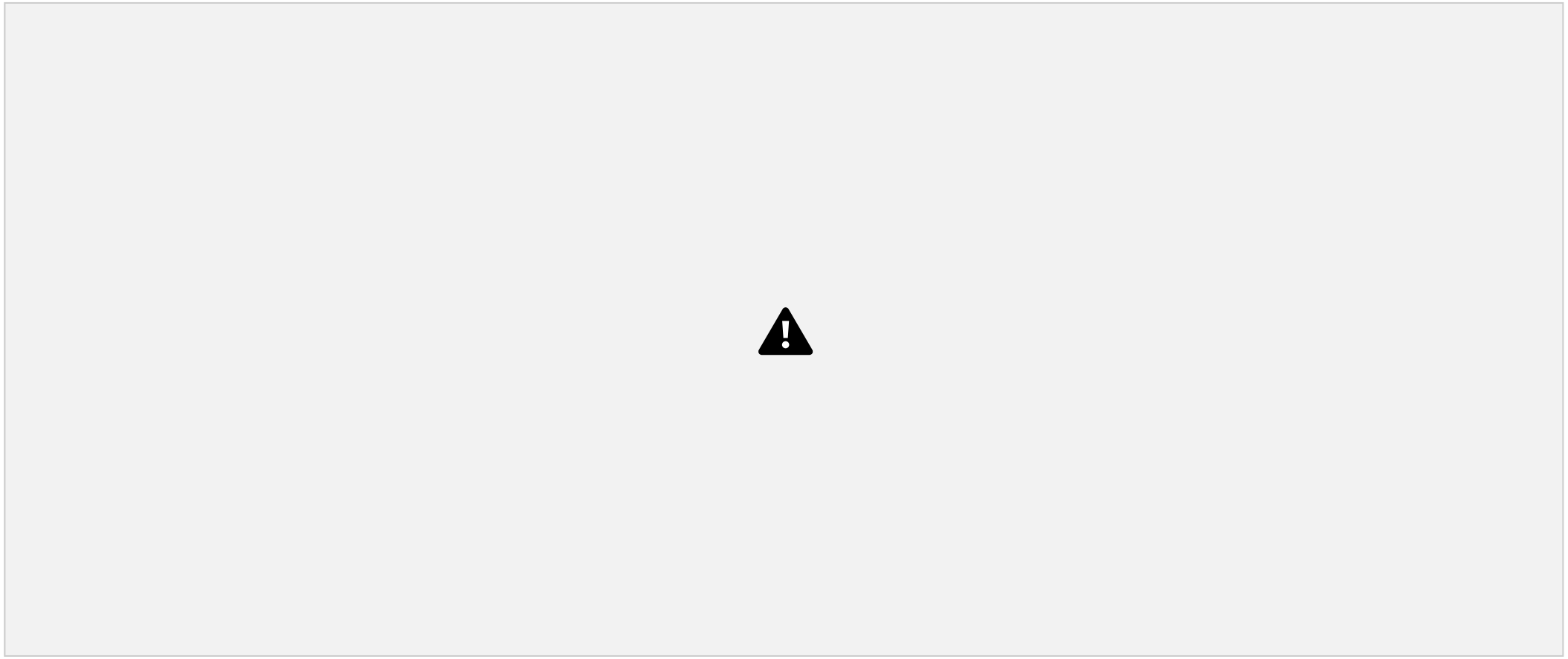
https://bigdata.stratebi.com/power-bi-agricultura/index.htm

Machine Learning - Mínimos Cuadrados



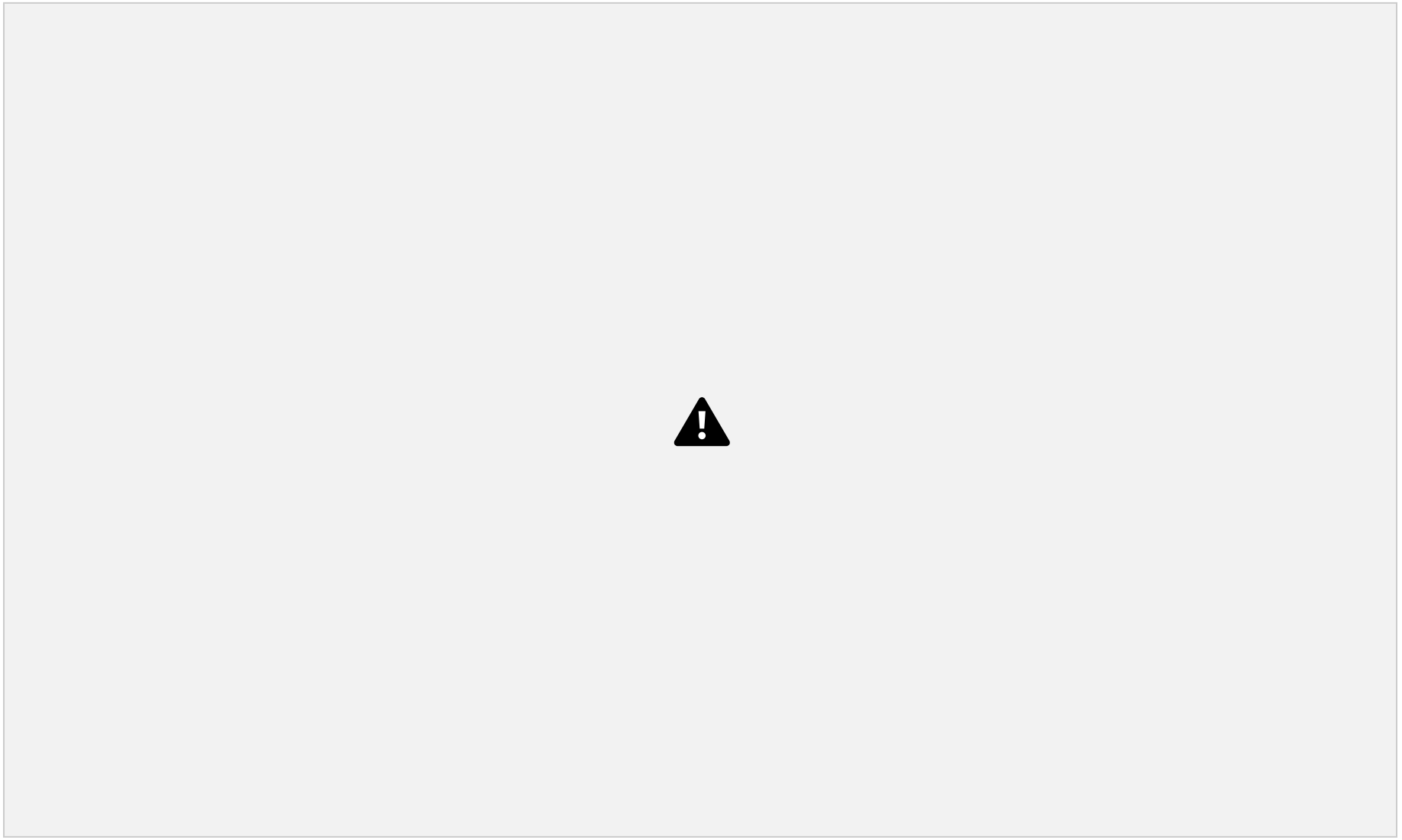
Ingresar a: https://bigdata.stratebi.com/alternating-least-squares/index.htm

Recuperación avanzada de información semántica en la web (Web search)



Ingresar a: https://bigdata.stratebi.com/web-search/index.htm

Sistemas de recomendación



Banca: Identificación de personas

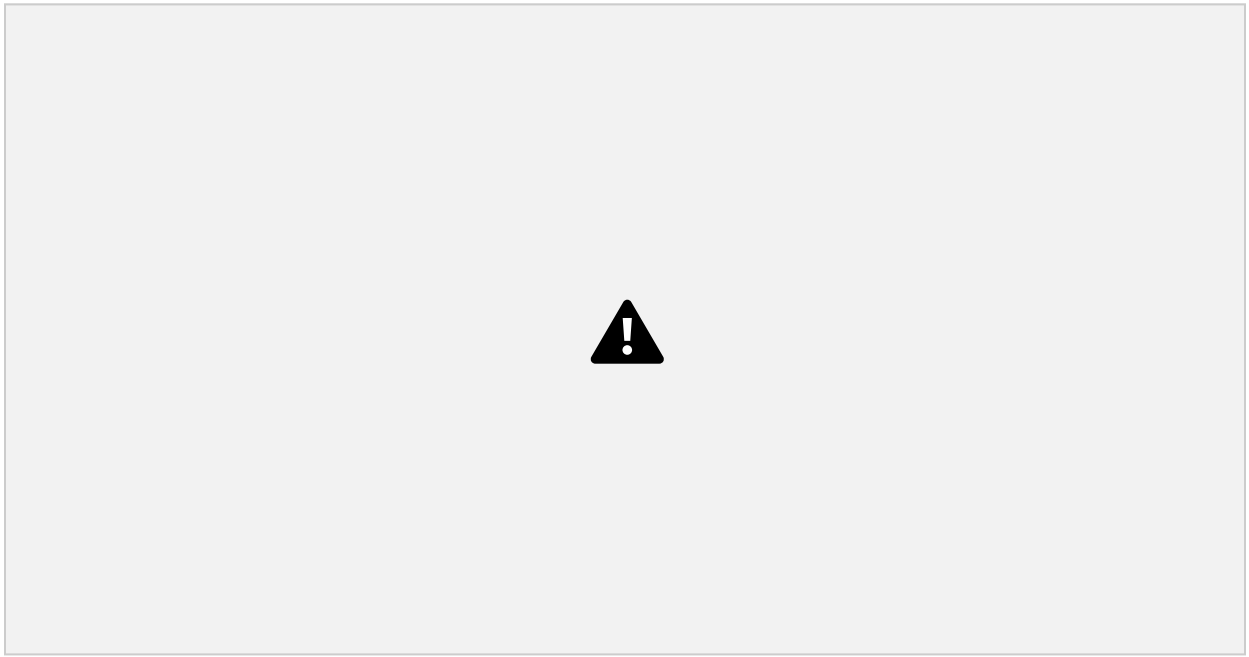
con las compras de tarjetas de crédito

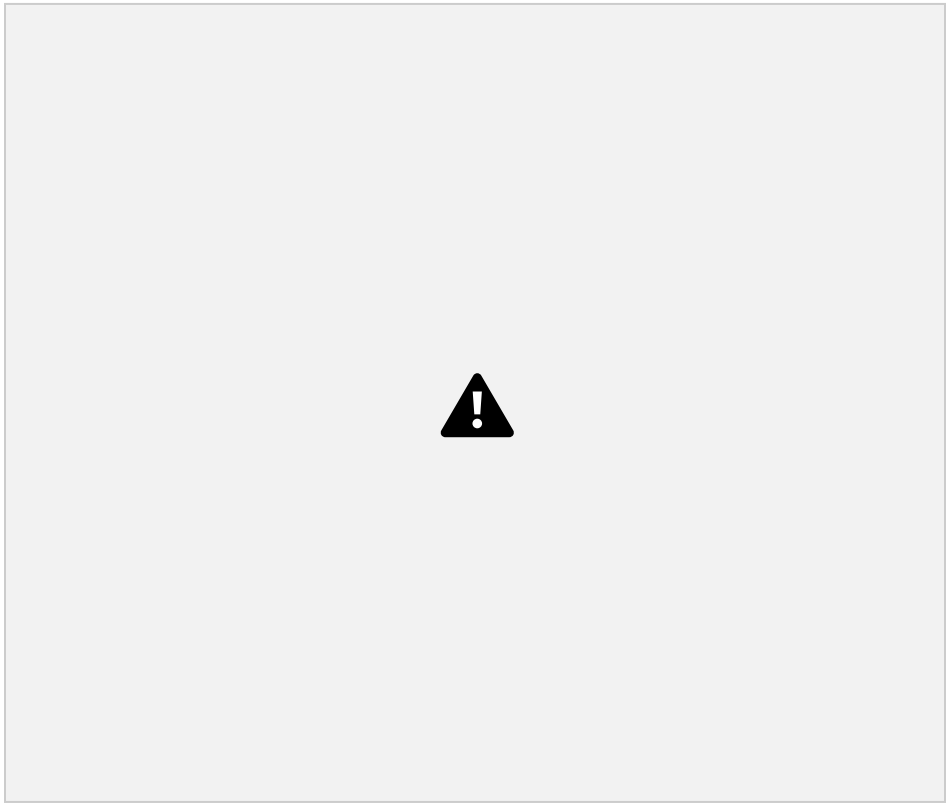
Identificación por el género

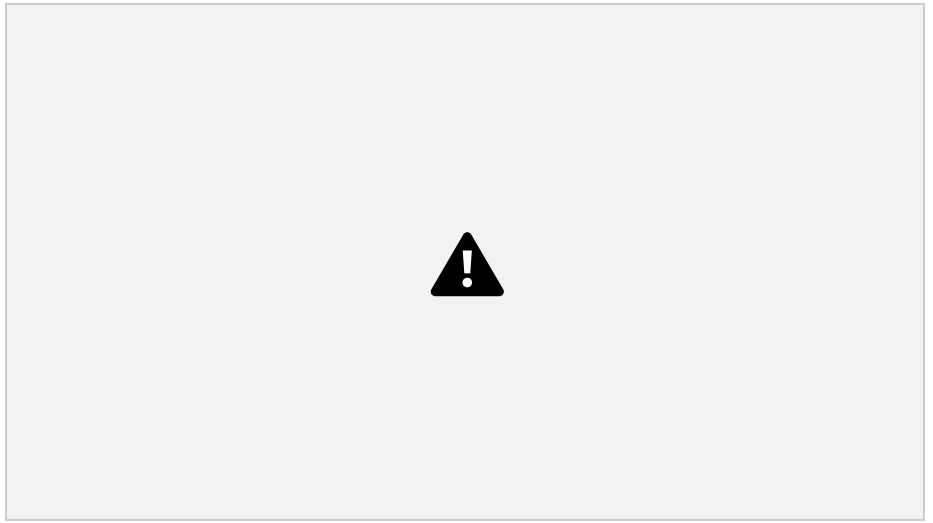
Identificación por el número de compras

Identificación por el poder adquisitivo

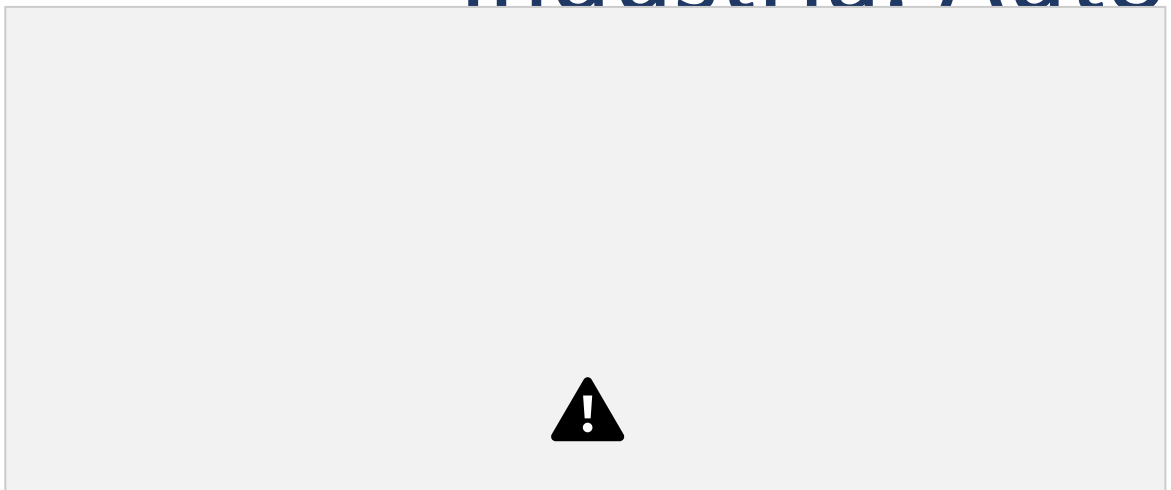
Industria. Automoción







Industria. Automoción: Colombia



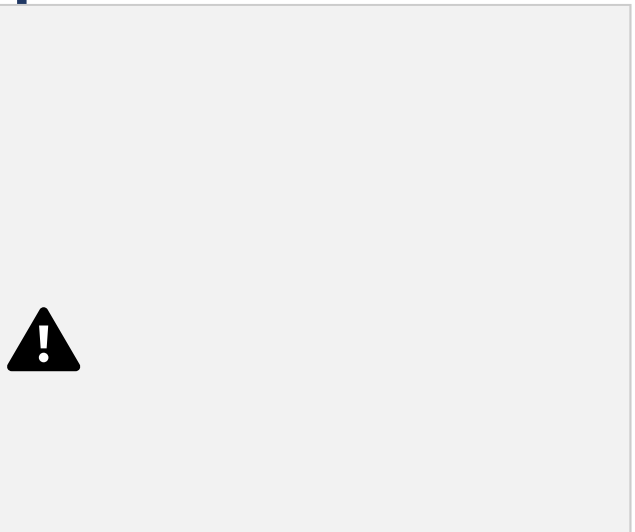
Seguimos esperando, los autónomos,
pero ya se vender eléctricos y híbridos

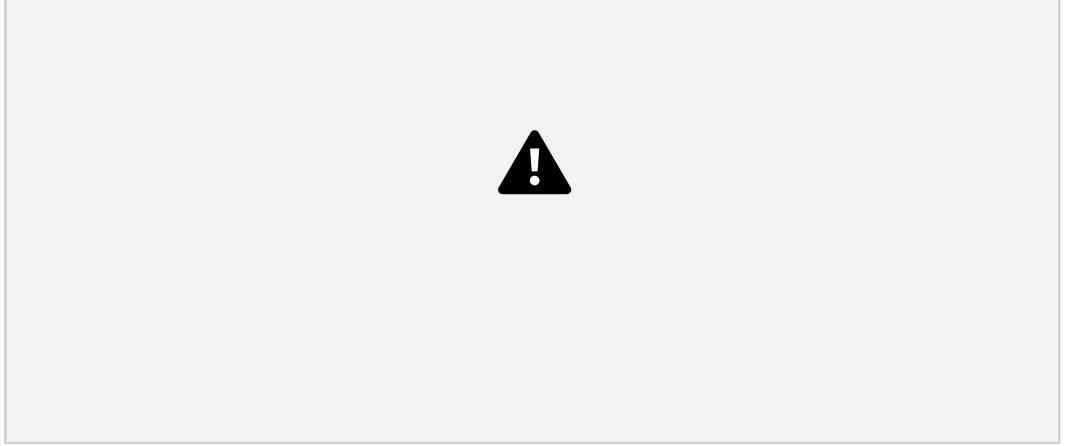
Probar maquinaria pesada para minería con tecnología 5g

Prueba de 10 días, septiembre 2021, Jericó, Antioquia

Netflix y Amazon

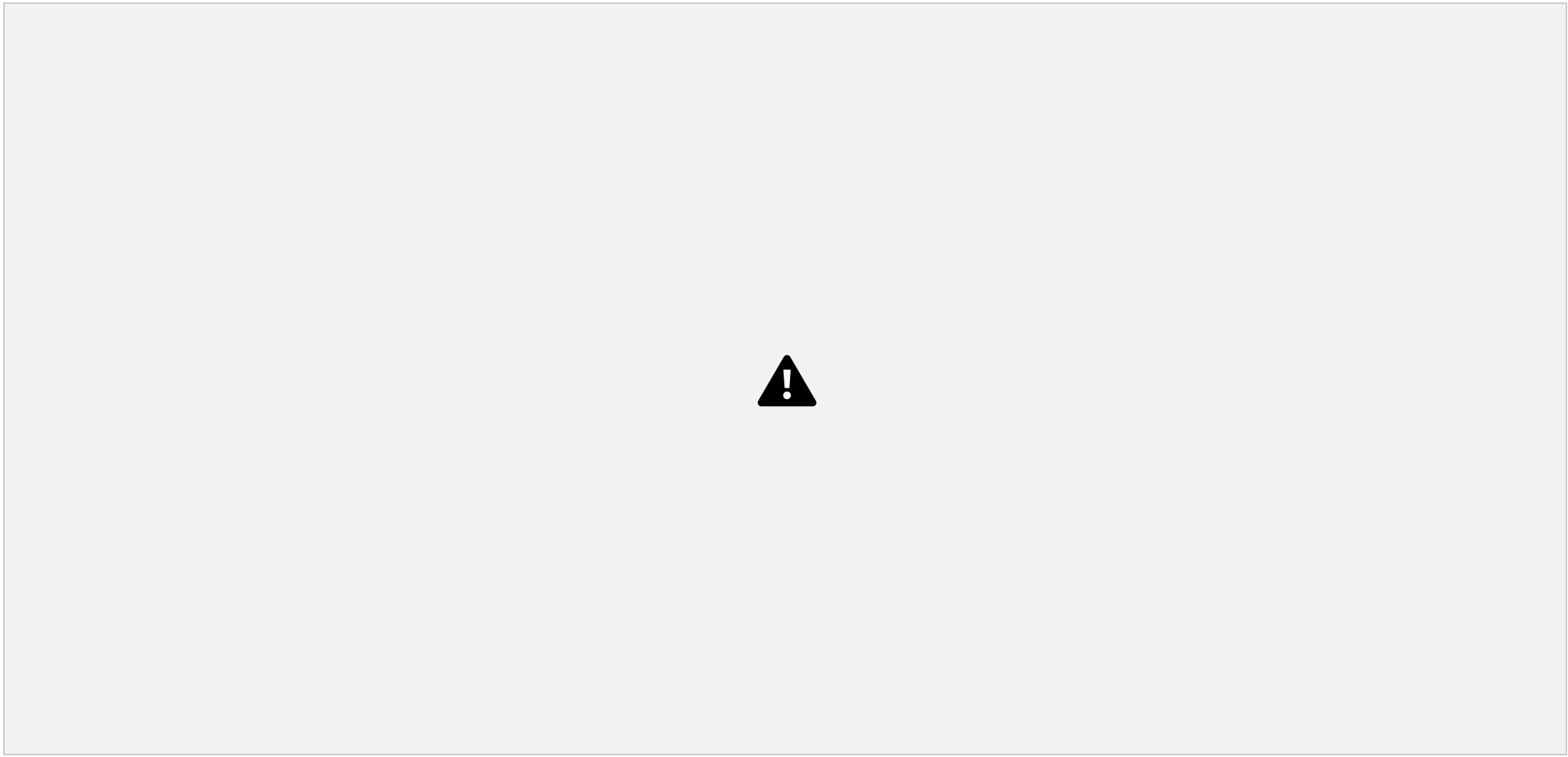
Para Netflix, compañía de alquiler de películas online, las tres cuartas
partes de los pedidos nuevos



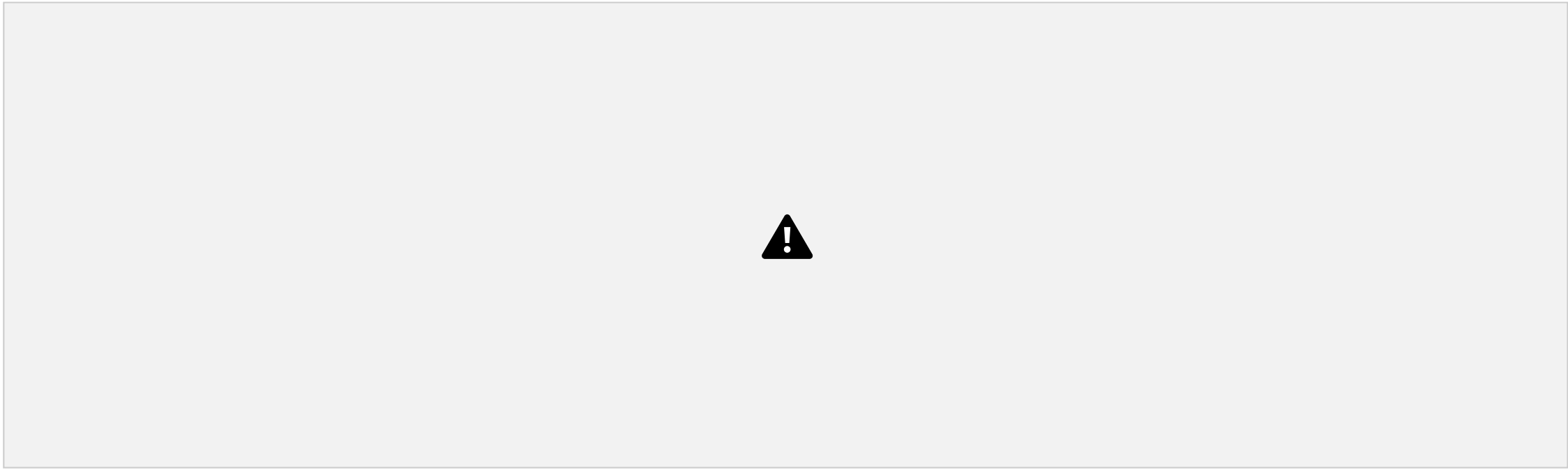


Netflix y Amazon son dos empresas cuyo plan de negocio está basada en big data y sistemas de recomendación

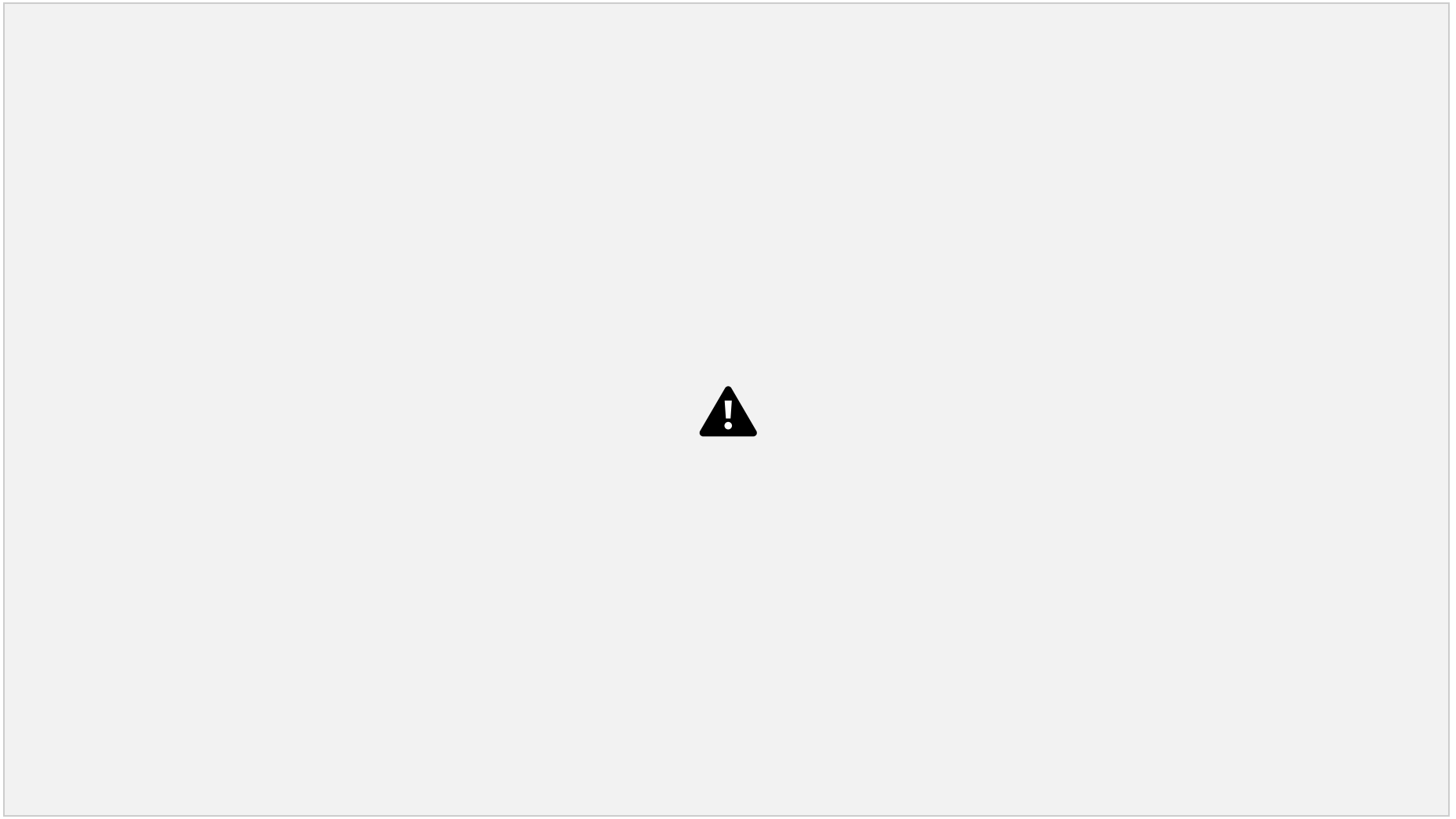
Herramientas para procesar Big Data



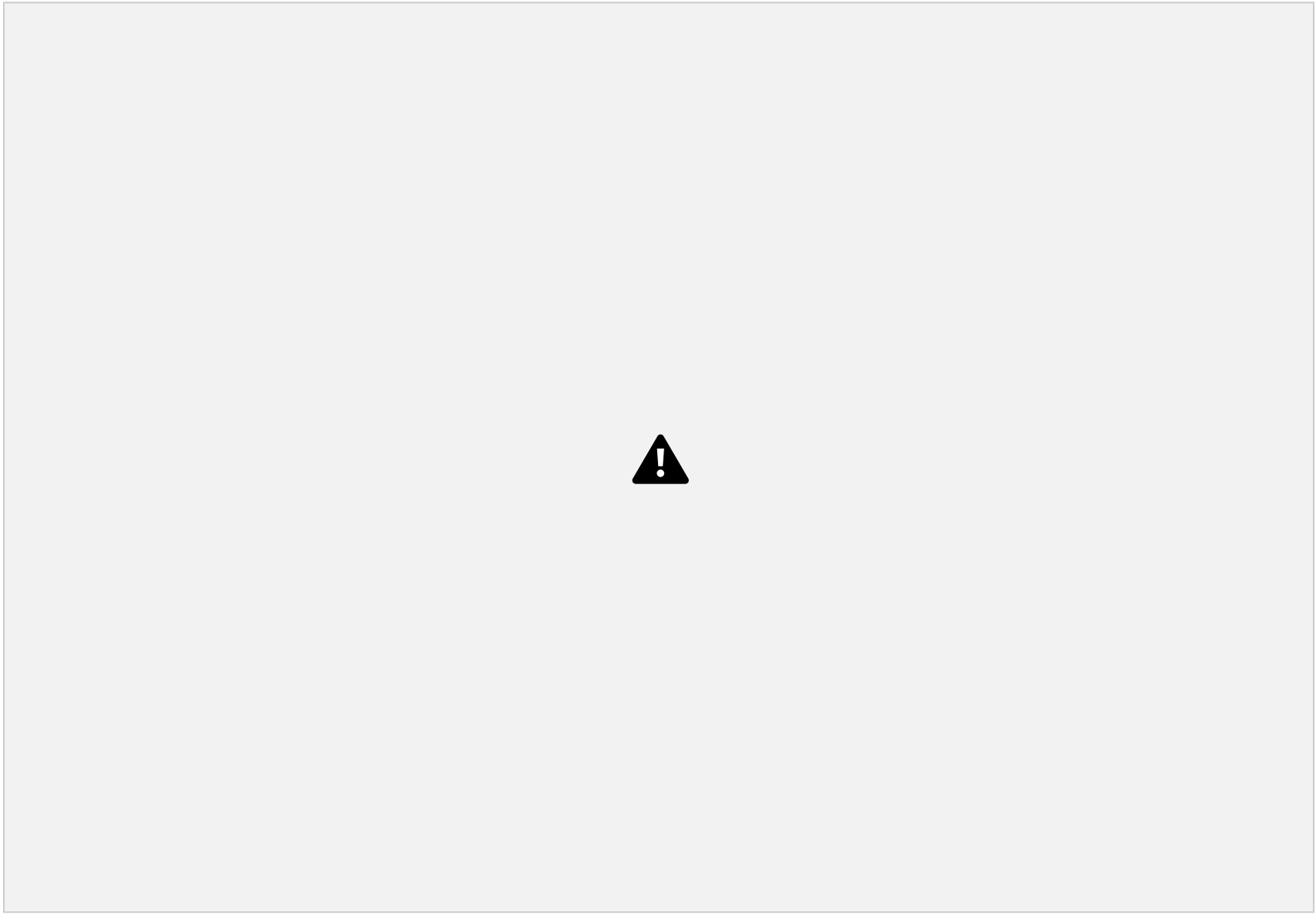
Aplicaciones de Software Libre/ Código abierto para Big Data



Tecnologías del Big Data

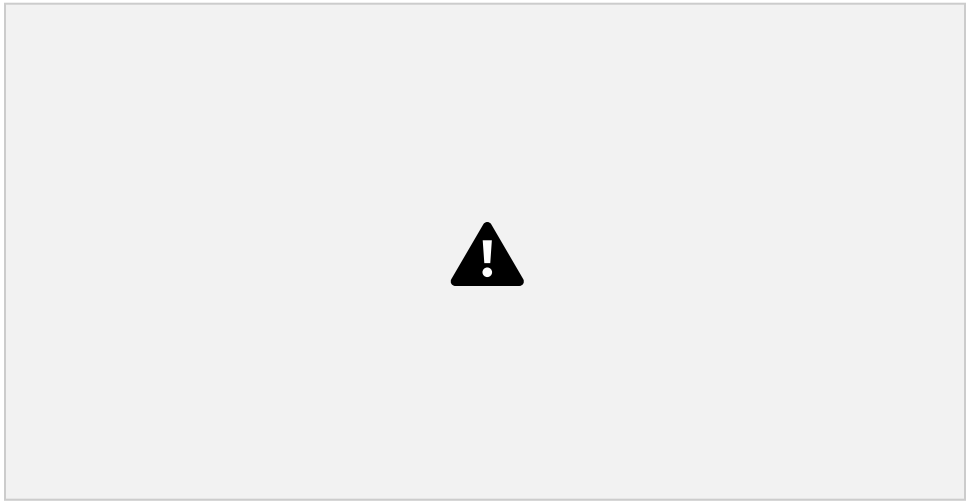


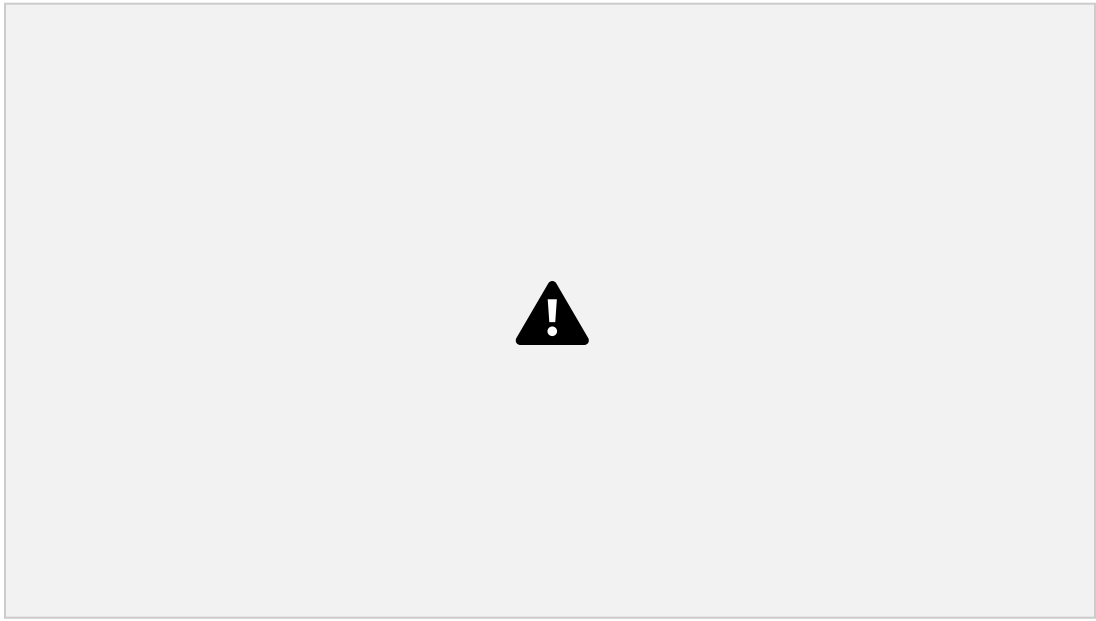
Ecosistema Big Data



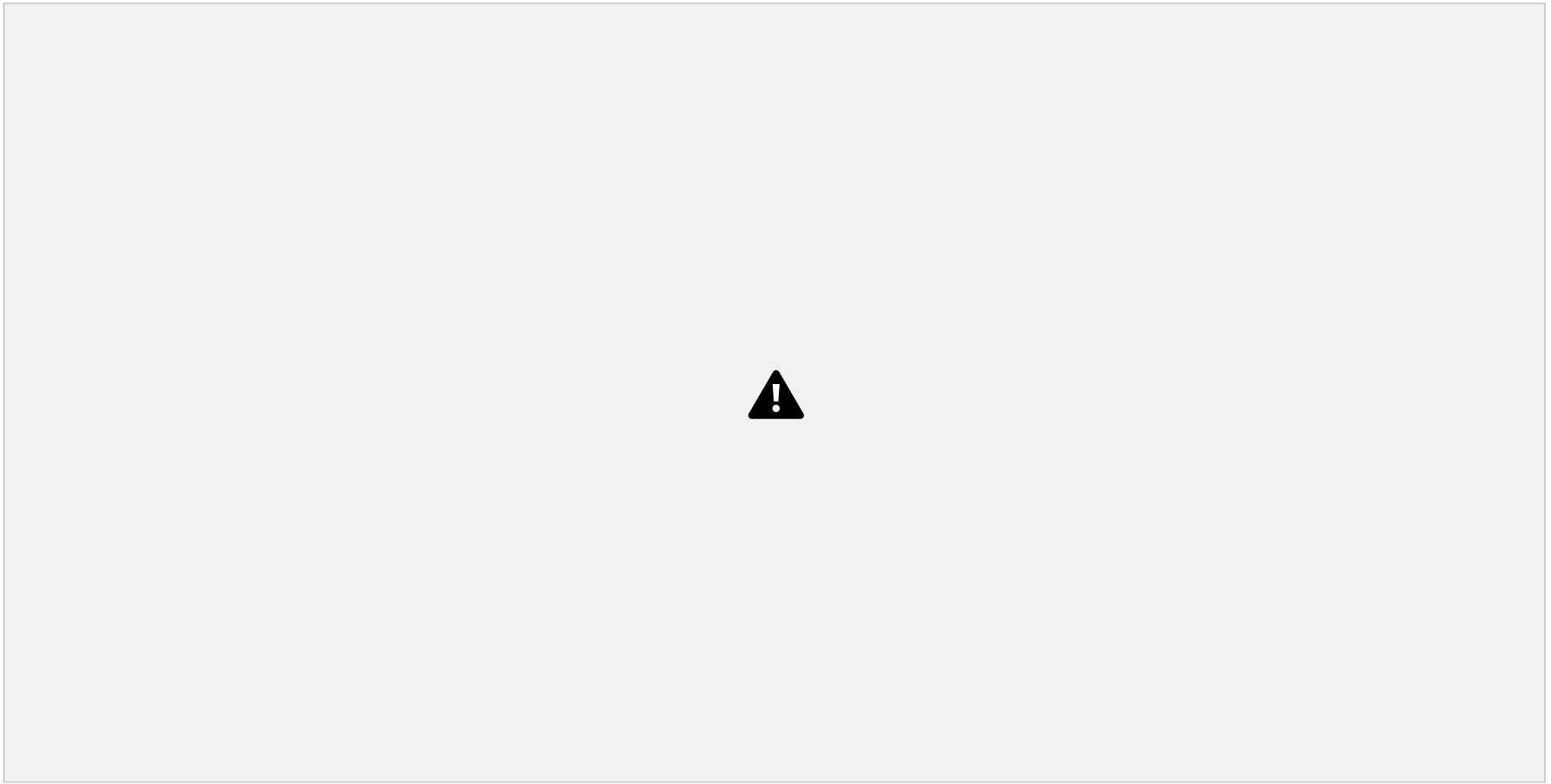
MapReduce

MapReduce es un framework, un modelo de programación que Google lanzó en el año 2004. Lo realmente innovador en este caso es que permite realizar computación en paralelo. Explicado de forma sencilla, en lugar de realizar el procesamiento desde una única máquina, distribuye las enormes cantidades de datos entre varios servidores que los procesan al unísono.

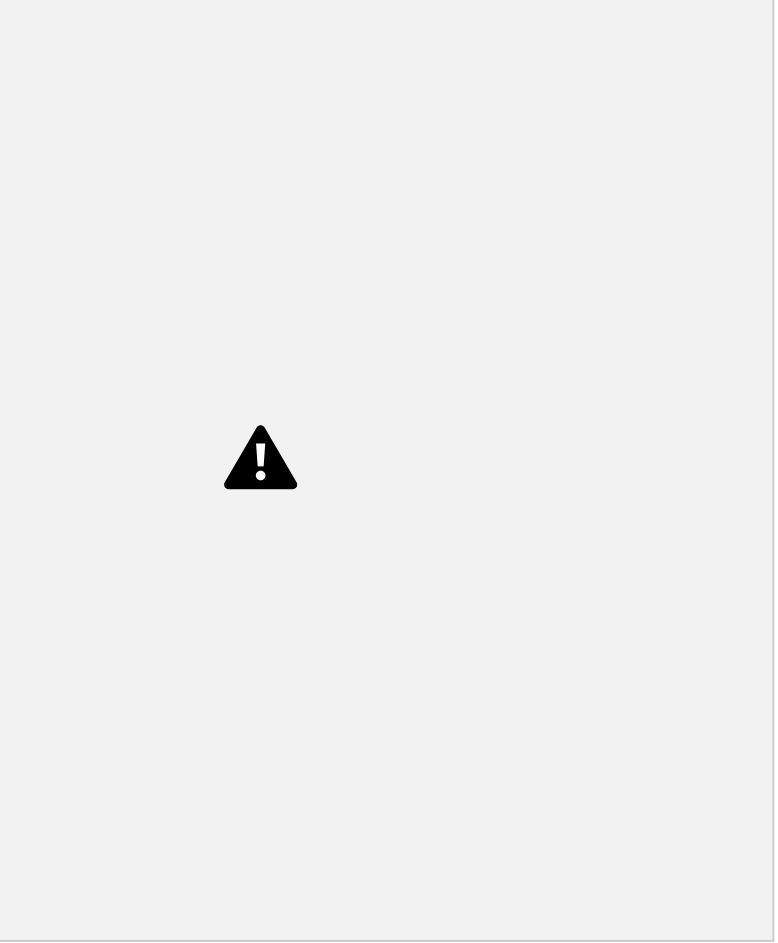


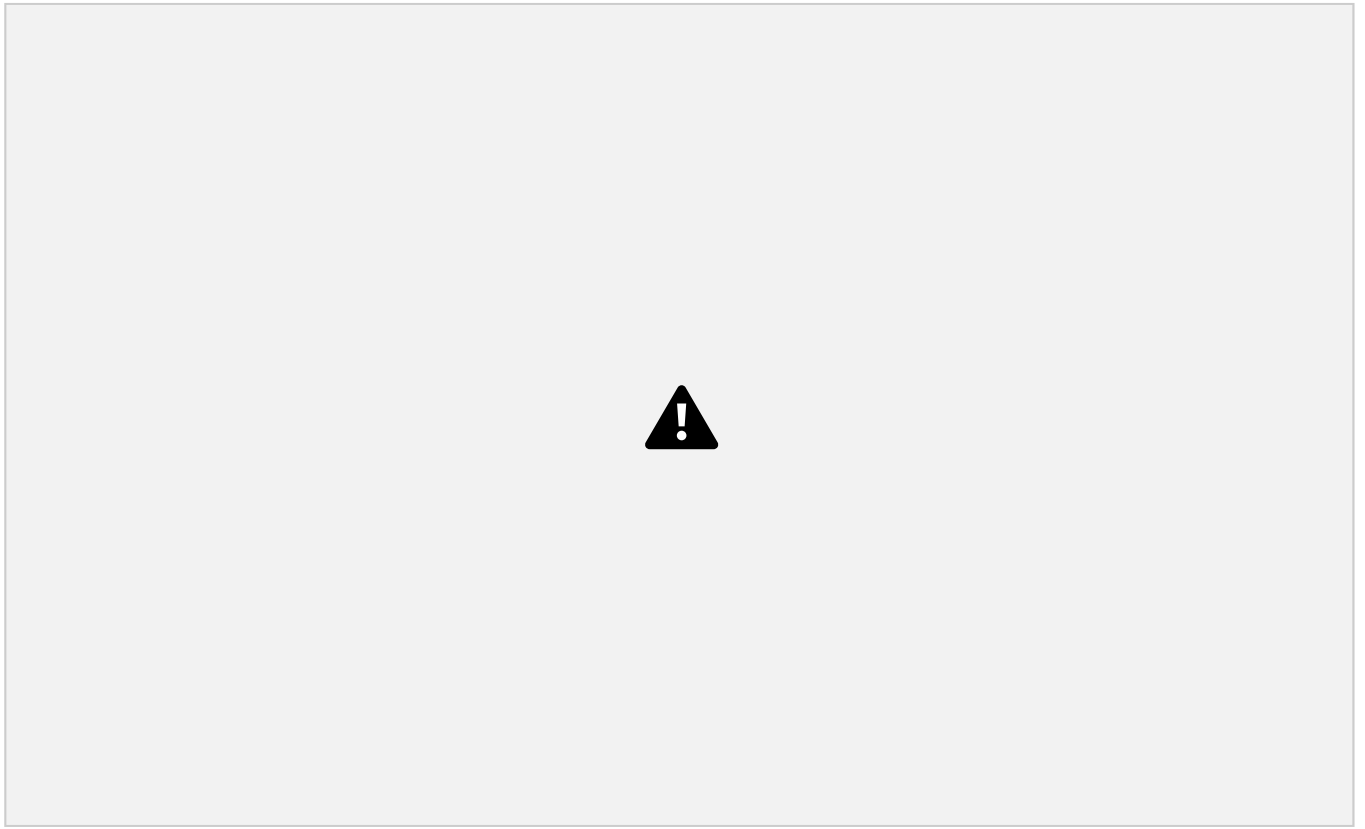


MapReduce



MapReduce: Flujo de datos





Ficheros de entrada

Ficheros Intermedio[S]

Esquema de fases de MapReduce

Ficheros de salida

Características de MapReduce

Paralelización automática:

• Dependiendo del tamaño de ENTRADA DE DATOS → se crean múltiples tareas MAP • Dependiendo del número de intermedio <clave, valor> particiones → se crean tareas REDUCE

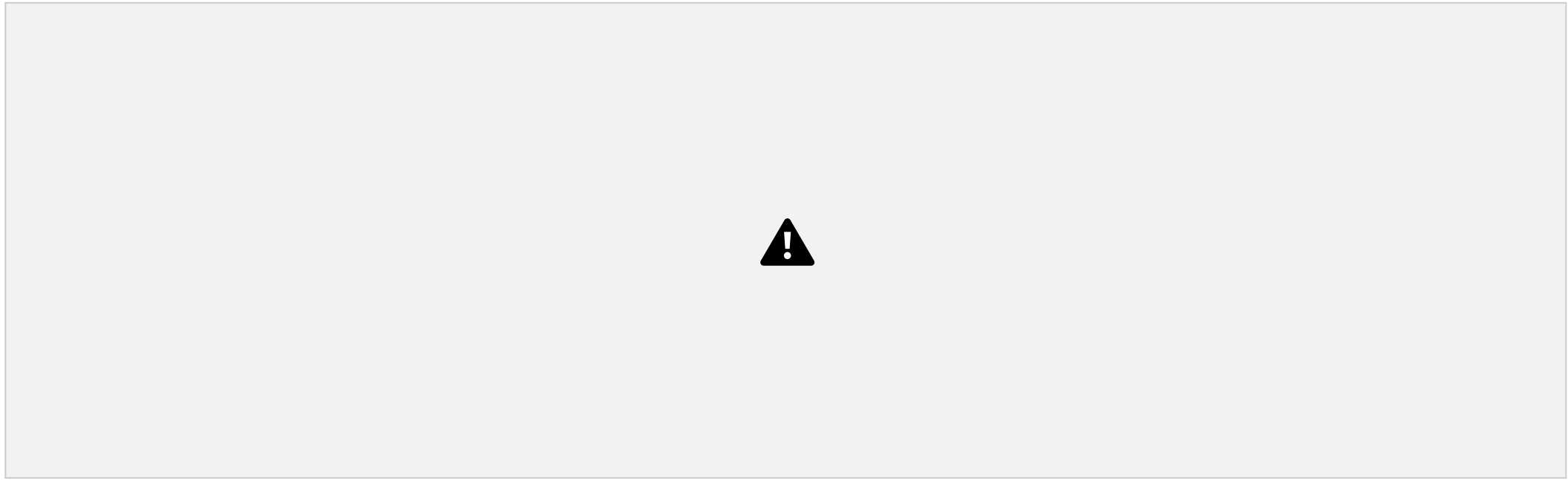
Escalabilidad:

• Funciona sobre cualquier *cluster* de nodos/procesadores
• Puede trabajar desde 2 a 10.000 máquinas

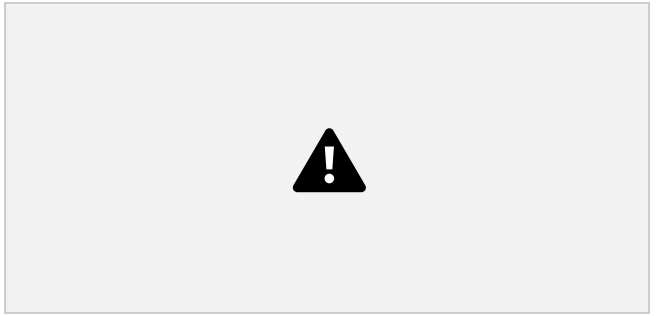
Transparencia programación:

• Manejo de los fallos de la máquina
• Gestión de comunicación entre máquinas

Ejemplo de MapReduce



Hadoop



Hadoop es una implementación de código abierto del paradigma de programación de Mapreduce, incluye:

- MapReduce (motor de cálculo offline).
- HDFS (sistema de ficheros distribuidos de Hadoop).
- HBase (acceso de datos online).

Esquema de Hadoop y MapReduce

Características de Hadoop

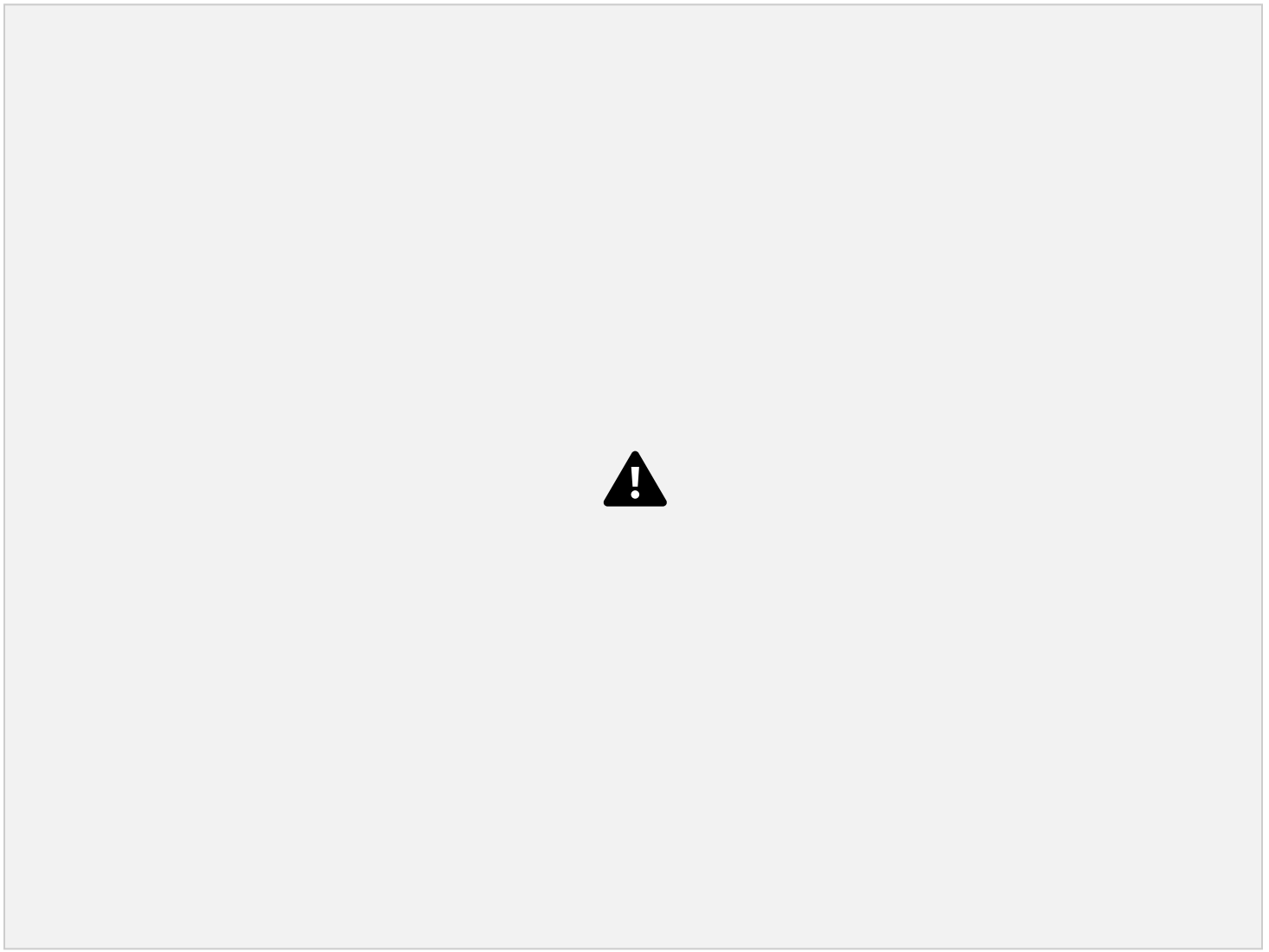
- Escalable: diseñado para escalar de servidores individuales a miles de máquinas, cada una ofreciendo cálculo local y almacenamiento; puede llegar a procesar y almacenar petabytes de manera fiable.

- Económico: distribuye los datos y los procesa a través de clústers de ordenadores comúnmente disponibles (en miles).

- Eficiente: al distribuir los datos puede procesarlos en paralelo sobre los nodos donde los datos están localizados.

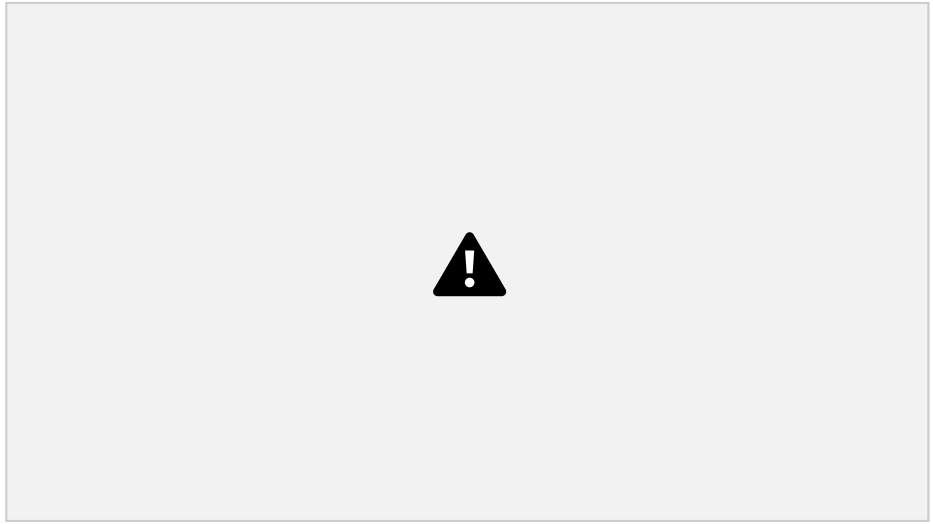
- Fiable: automáticamente mantiene copias de datos y también de manera automática realiza de nuevo tareas de computación basadas en fallos

Arquitectura Hadoop



Spark

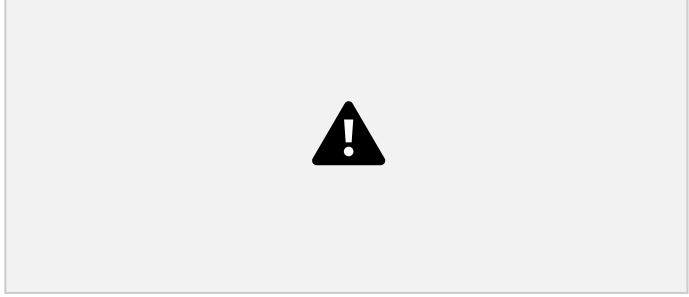
Entorno de trabajo (gestión de ejecución, API) genérico y rápido *(hasta 100 veces más que Hadoop)* para procesamiento de datos masivos.



Centrado en una estructura de datos distribuida denominada “Resilient Distributed Dataset” (RDD).

Desarrollado en Scala. Interfaces en Scala, Java, Python, …

Spark



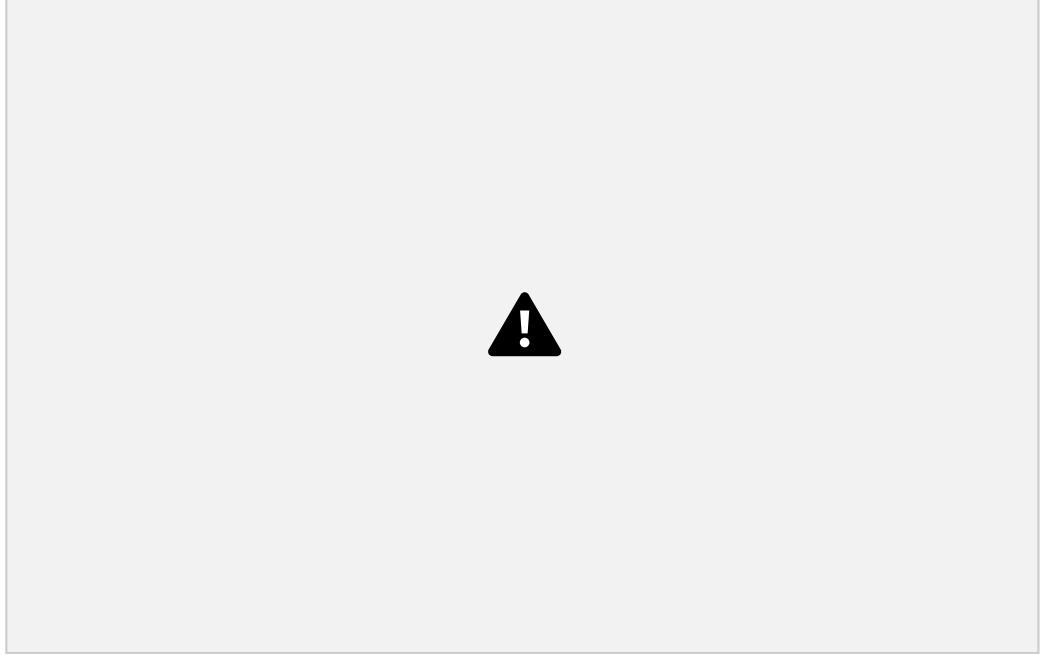
- Big Data “in-memory”. Spark *permite realizar trabajos paralelizados totalmente en memoria*: Reducción de tiempos Procesos iterativos

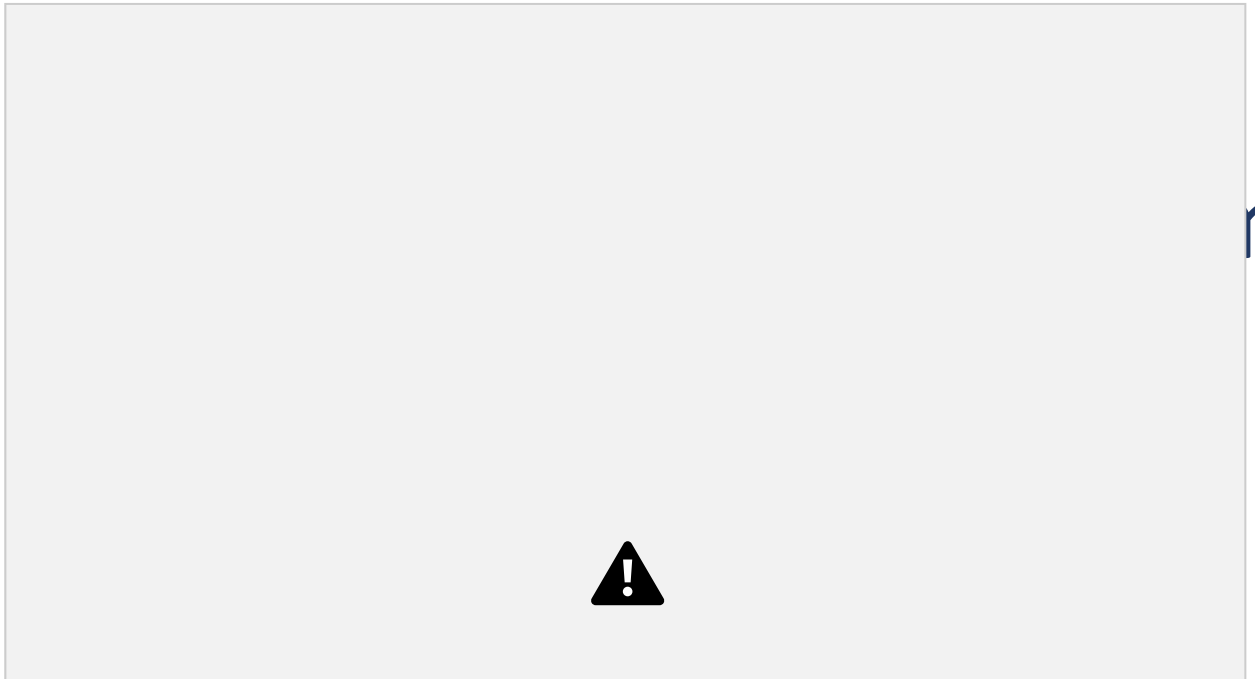
- Esquema de computación más flexible que MapReduce. Permite la flujos acíclicos de procesamiento de datos

Spark vs hadoop

Cloud Computing

Modelo de prestación de servicios de negocio y tecnología, que *permite al usuario acceder a un catálogo de servicios estandarizado y responder a las necesidades del negocio*, de forma flexible y adaptativa, […] pagando únicamente por el consumo efectuado.



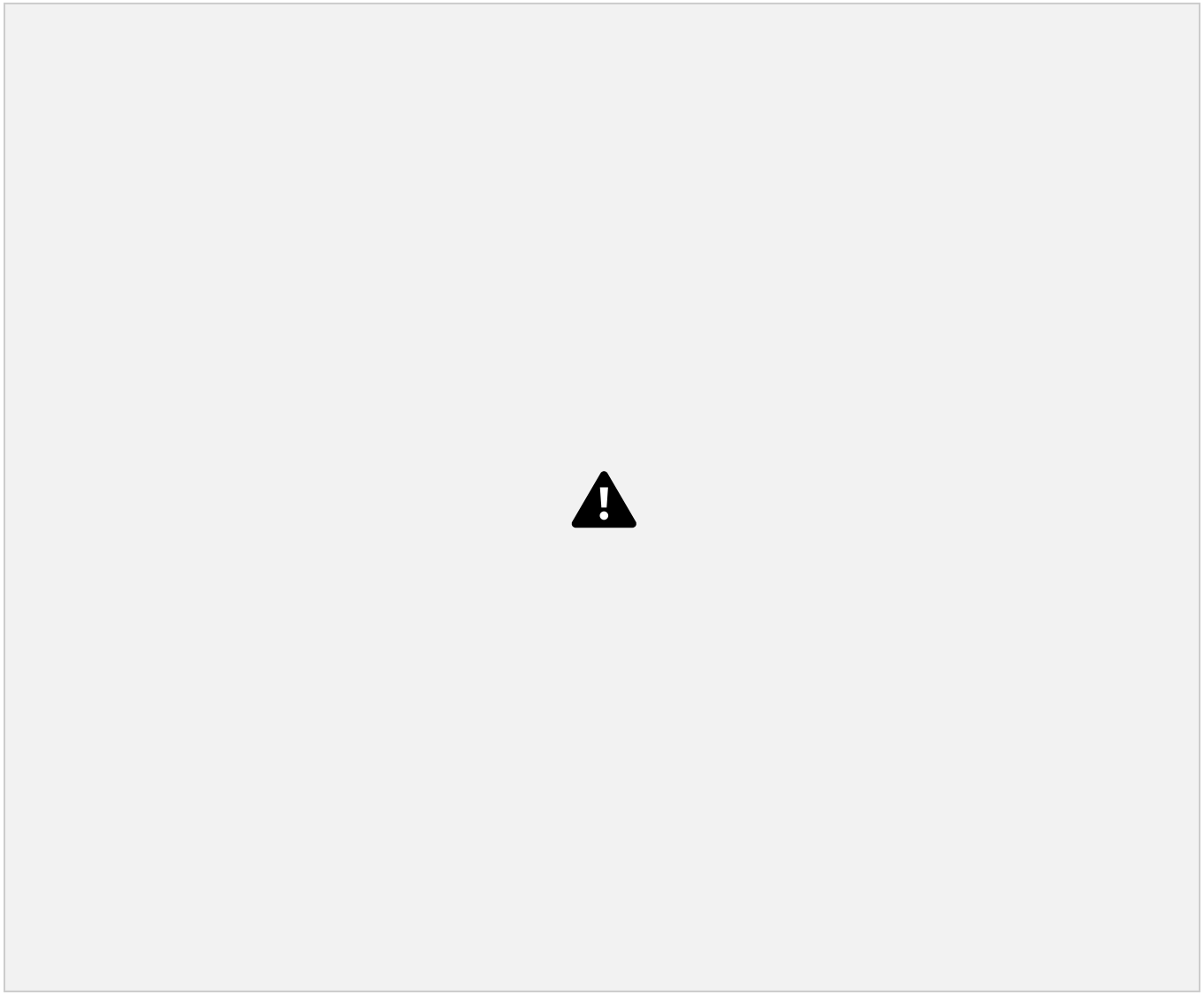


nputing

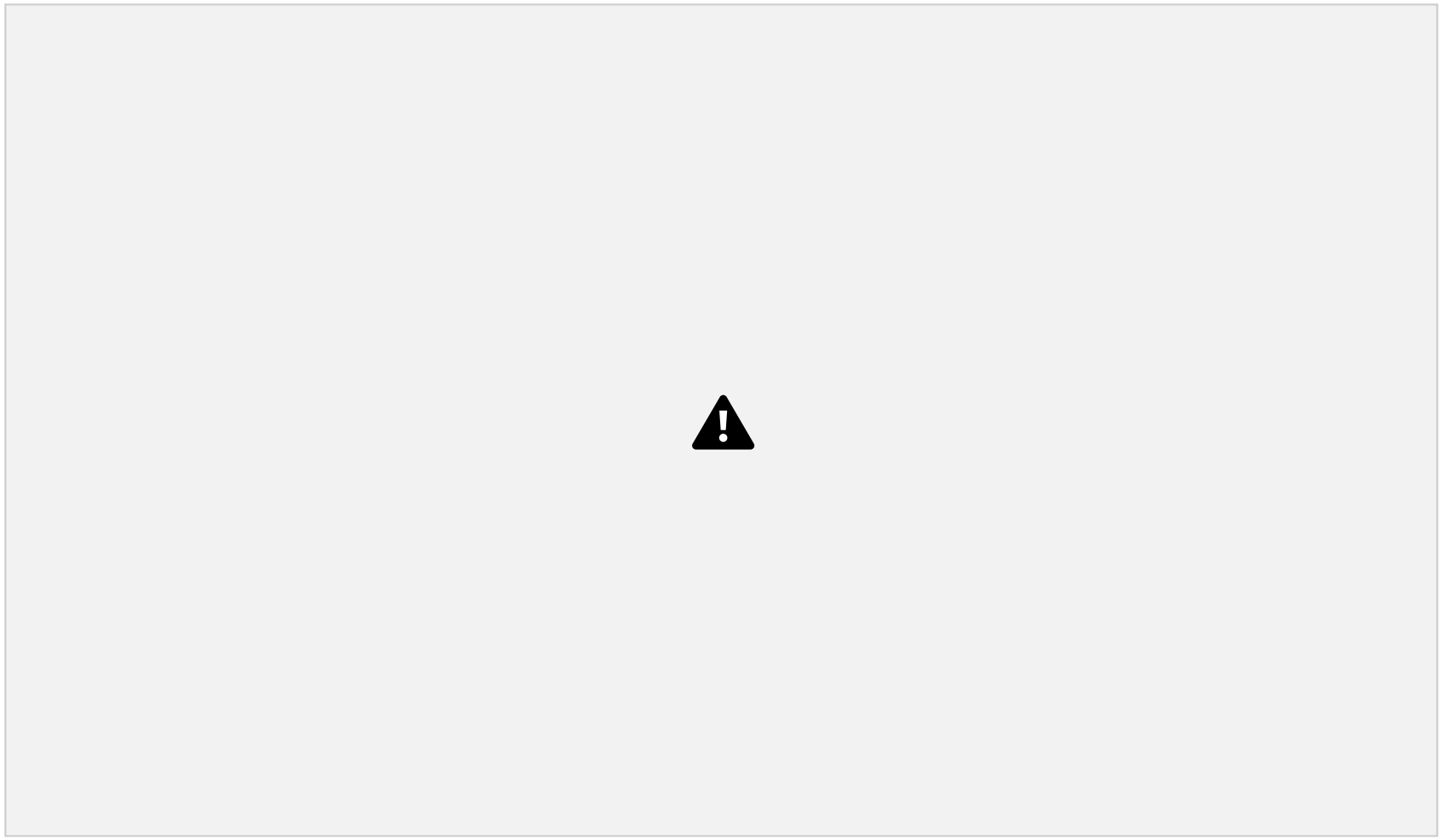
La idea básica es que toda la

información se almacena de forma distribuida en servidores, siendo accesible en cualquier momento por el usuario sin que éste se preocupe de nada

Estructura Cloud Computing



Tecnologías Big Data



Tomado de: https://www.youtube.com/watch?v=EfOMesB7sMQ

Aplicaciones de SL/CA para Big Data

Apache Hadoop: Plataforma de código abierto compuesta por Hadoop Distributed File System (HDFS), Hadoop MapReduce y Hadoop Common.

Avro: Proyecto de Apache que provee servicios de serialización.

Cassandra: Base de datos no relacional distribuida y basada en un modelo de almacenamiento de <clave-valor>, desarrollada en Java.

Chukwa: Software diseñado para la colección y análisis a gran escala de registros de eventos (logs).

Aplicaciones de SL/CA para Big Data

Flume: Software cuya tarea principal es dirigir los datos de una fuente hacia alguna otra localidad.

HBase: Base de Datos columnar (column-oriented database) que se ejecuta en HDFS.

Hive: Infraestructura de «Data Warehouse» que facilita la administración de grandes volúmenes de datos que se encuentran almacenados en un ambiente distribuido.

Jaql: Lenguaje funcional y declarativo que permite la explotación de datos en formato JSON diseñado para procesar grandes volúmenes de información.

Aplicaciones de SL/CA para Big Data

Lucene: Software que provee de librerías para la indexación y búsqueda sobre texto.

Oozie: Proyecto de código abierto que simplifica los flujos de trabajo y la coordinación entre cada uno de los procesos.

Pig: Software que permitir a los usuarios de Hadoop enfocarse más en analizar todos los conjuntos de datos y dedicar menos tiempo en construir los programas MapReduce.

ZooKeeper: Infraestructura centralizada y de servicios que puede ser utilizada por aplicaciones para asegurarse de que los procesos a través de un clúster sean serializados o sincronizados.

Independientes

Otros igual de conocidos, pero no relacionados con la plataforma de código abierto

Hadoop son:

Elasticsearch: Motor de búsqueda y análisis basado en texto completo.

MongoDB: Base de datos NoSQL basada en el modelo de datos de

documentos.

Cassandra: Proyecto de código abierto de Apache diseñado para la administración de bases de datos NoSQL.

CouchDB: Base de datos NoSQL de código abierto basada en estándares comunes para facilitar la accesibilidad y compatibilidad web con una diversidad.

Solr: Motor de búsqueda de código abierto basado en la biblioteca Java del proyecto Lucene.