Escuela de Datos Vivos

Curso Data Engineer: Creando el pipeline de datos

MÓDULO B

Ingest

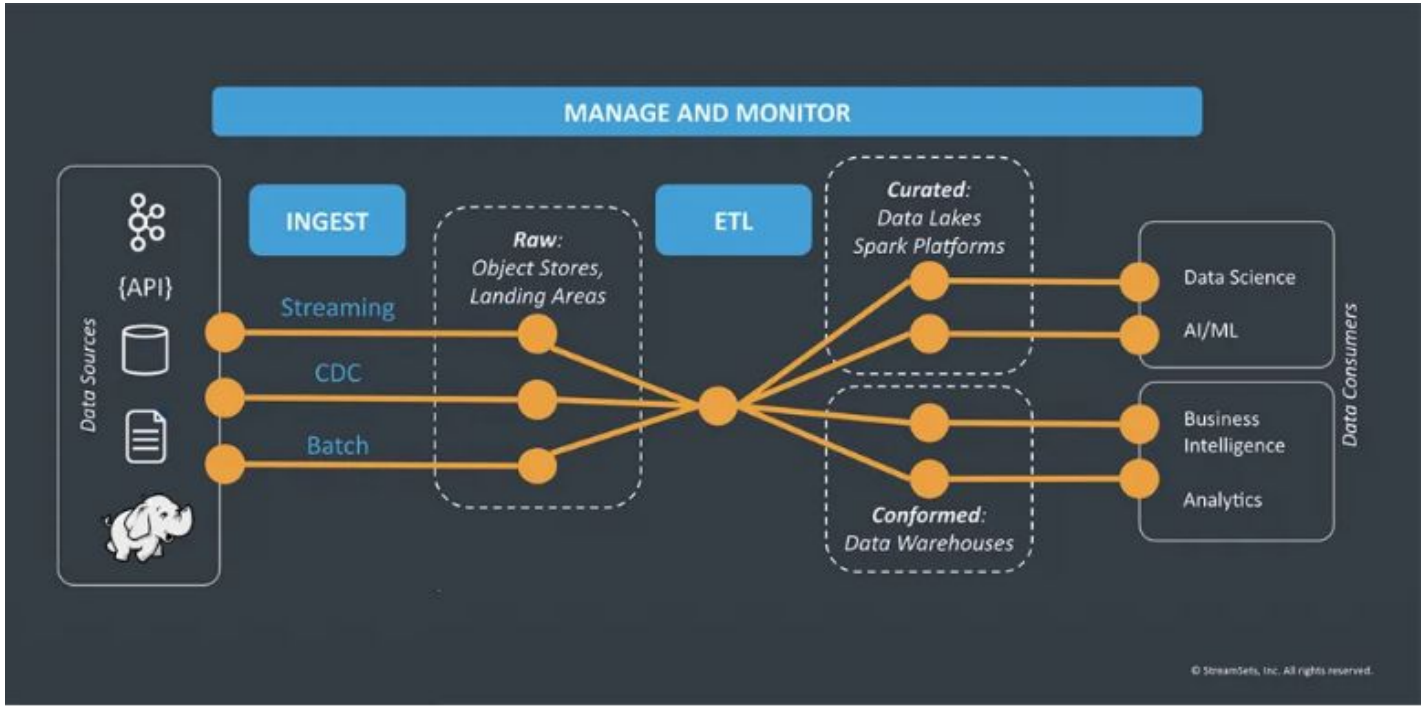
Ingest

Ingestion es el proceso de ingresar información a un proyecto de Big Data.

Esto se puede hacer mediante:

- **Batch data**
- **Streaming data**
- **CDC**

Ingest



fuente: streamsets.com

Batch Data

Batch data

- ¿Cómo se hace el ingest?
- Periodicidad
- Herramienta para hacer el ingest
- Ejemplos

Batch data

¿Cómo se hace el ingest?

- Se transfiere la data desde una o más fuentes al proyecto de Big Data
- Archivos (json, avro, parquet, csv, etc.)
- Pueden provenir de la exportación de Apps, DBs, etc.
- Se pueden almacenar en HDFS, GCS, S3, Blob Storage para luego ser procesados

Batch data

Periodicidad

- Pocas veces por día
- Habitualmente se realiza por las noches
- Se envía un lote completo de información

Batch data

Herramienta para hacer el ingest

- Web
- CLI (HDFS DFS, gsutil, aws s3 cp, azcopy)
- API
- Servicios de Transferencia / Apliances
- Herramientas para hacer procesos en Batch (Apache Nifi)

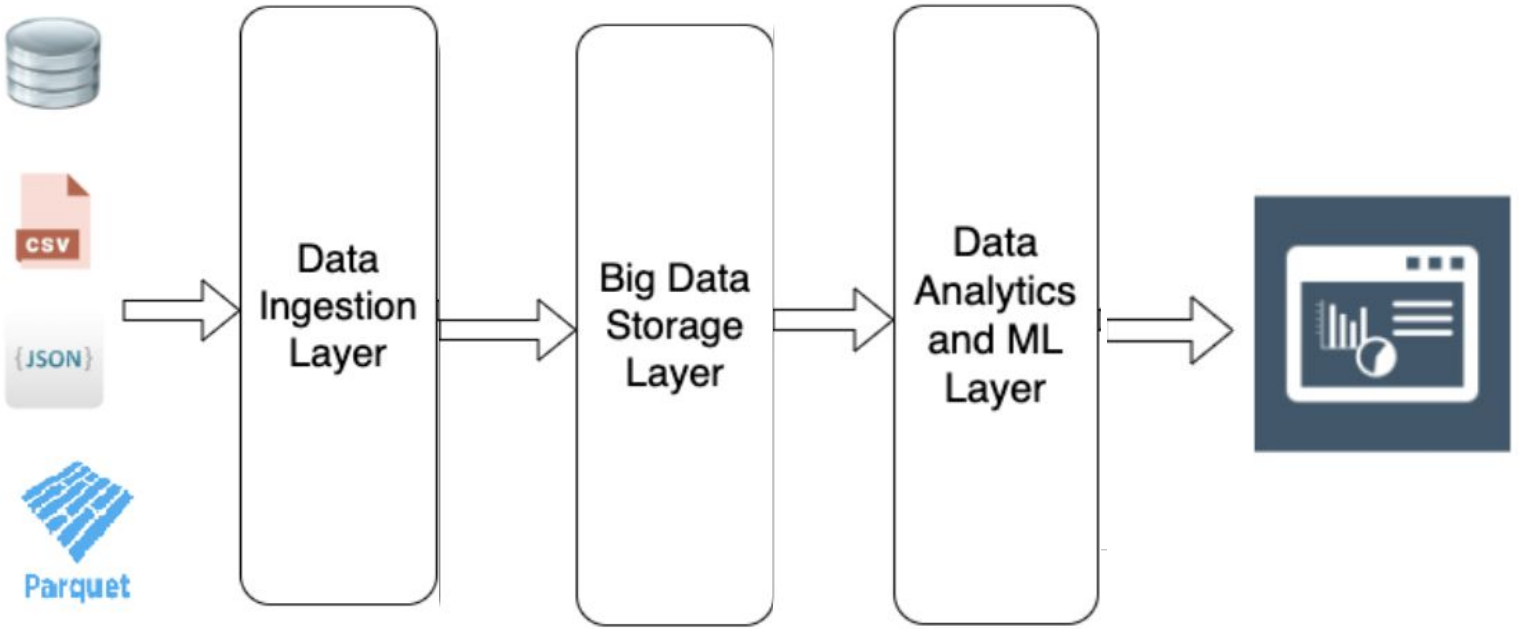
Batch data

Ejemplos

- Información generada en un retail es enviada al DataWarehouse
- Información transaccional es enviada al DW para realizar análisis
- Información generada en un mainframe es enviada al DW en la nube

Ingest

Batch data



fuente: medium.com

Streaming Data

Streaming Data

- ¿Cómo se hace el ingest?
- Periodicidad
- Herramienta para hacer el ingest
- Ejemplos

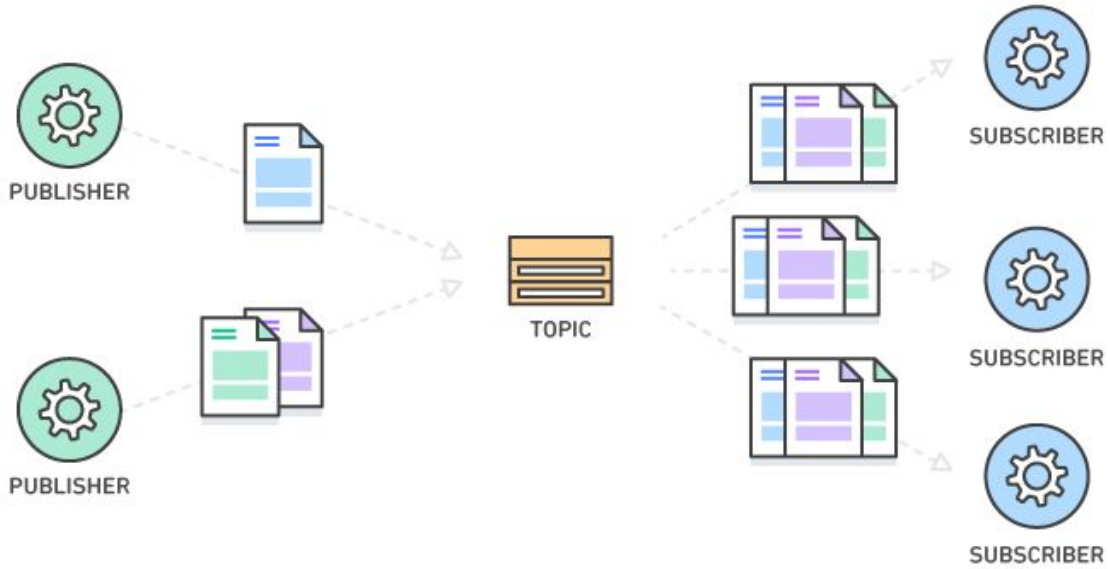
Streaming Data

¿Cómo se hace el ingest?

- Set de datos enviados en pequeños mensajes transmitidos constantemente
- Publishers envían información a un tópico
- Los tópicos guardan esa información por un tiempo determinado
- Subscribers toman información de ese tópico

Ingest

Streaming data



fuente: aws.com

Streaming Data

Periodicidad

- Constantemente

Streaming Data

Herramienta para hacer el ingest

- Apache Kafka
- Google pub/sub
- Amazon Simple Notification Service (SNS)
- Azure service bus

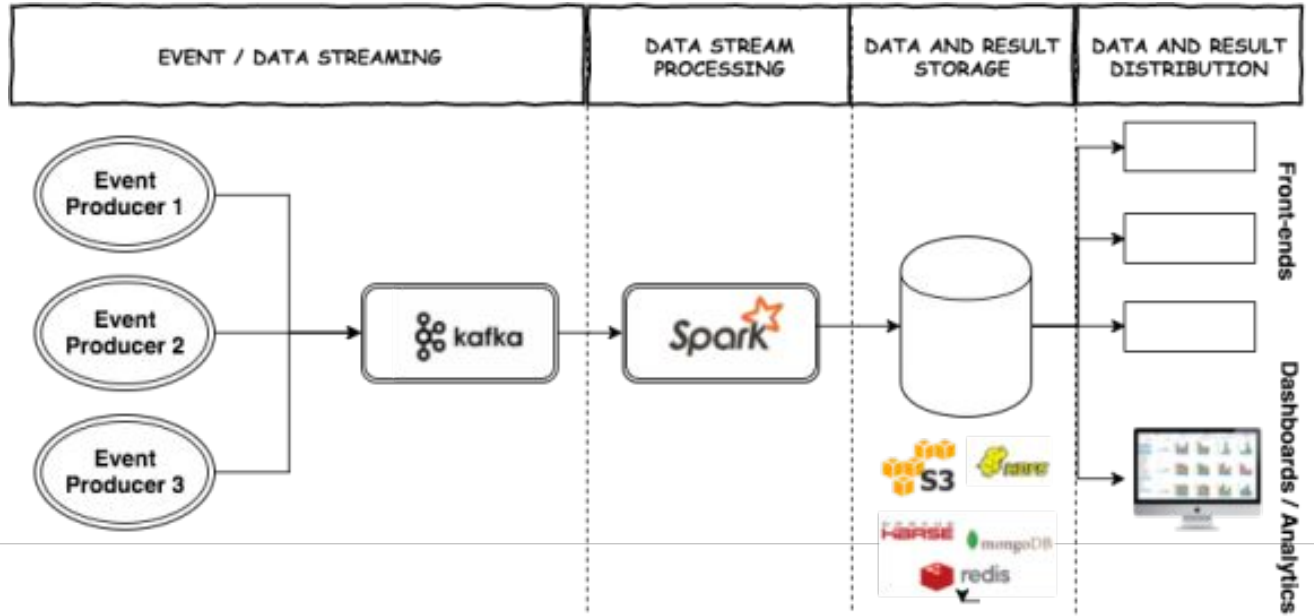
Streaming Data

Ejemplos

- Información generada en un depósito para contabilizar stock
- Información de dispositivos IOT es enviada al DW para realizar análisis
- Información de dispositivos móviles es enviada al DW en la nube

Ingest

Streaming data



fuente: medium.com

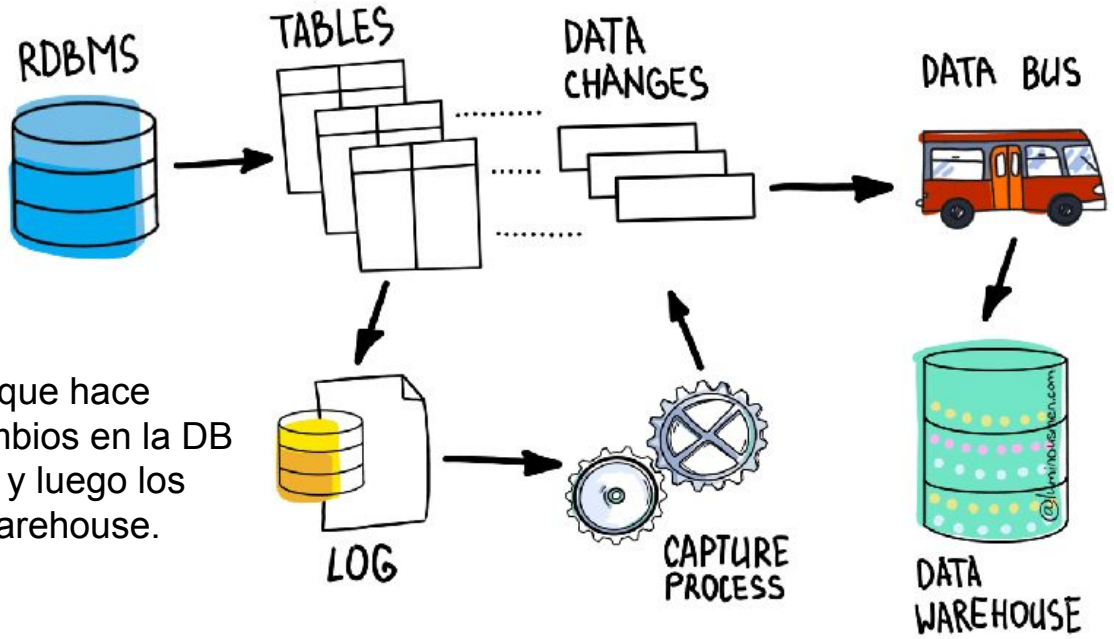
Change Data Capture

CDC

- ¿Cómo funciona?
- ¿Cuándo utilizarlo?
- Herramienta para hacer el ingest
- Ejemplos

CDC

¿Cómo funciona?



Es un mecanismo que hace tracking de los cambios en la DB origen, los captura y luego los aplica en el Datawarehouse.

fuente: medium.com

CDC

¿Cuándo utilizarlo?

- En casos que es importante conocer todos los cambios en el tiempo
- En casos que no queremos estresar la DB relacional
- En casos que necesitemos mantener actualizada la información en casi TR.
- Cuando no necesitemos hacer grandes transformaciones en el proceso

CDC

Ejemplos

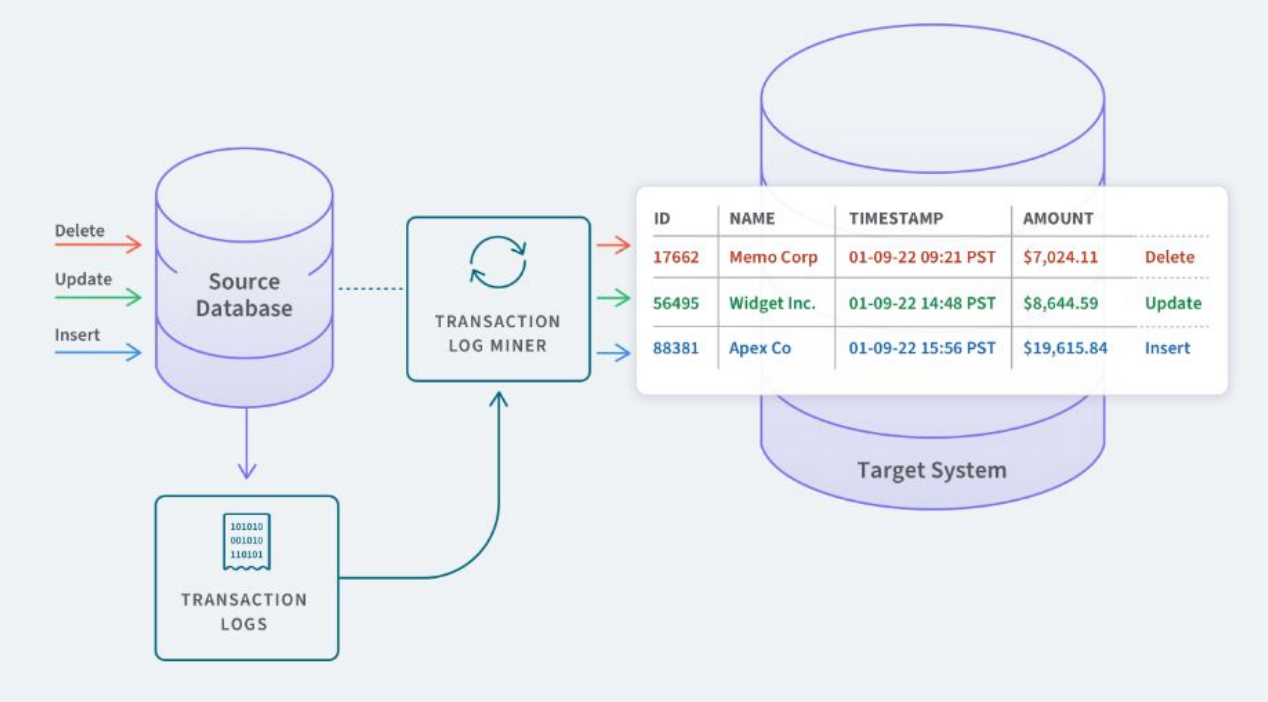
- Mantener actualizado un stock de un depósito
- Cuando la RDBMS está comprometida de performance
- Mantener sincronizada una DB en otro destino (Disaster Recovery)
- Múltiples fuentes de consulta

CDC

TimeStamp	Usuario	Tipo de transacción	Cantidad Monedas
8/21/2021 15:30:15	User1	Compra	1000
8/21/2021 16:01:22	User1	Gasta	300
8/21/2021 16:05:59	User2	Compra	500
8/21/2021 16:06:22	User2	Gasta	200
8/21/2021 20:37:46	User1	Gasta	200

fuente: qlik.com

CDC



fuente: qlik.com

Scraping Data

Scraping data

- ¿Qué es y cómo se hace scraping data?
- ¿Cuándo utilizarlo?
- Herramienta para hacer el ingest
- Ejemplos

Streaming Data

¿Qué es y cómo se hace scraping data?

- Navegar automáticamente en una web y extraer la información de esta
- Mediante herramientas de scraping tomar información para ingestar
- También conocido como scraper, bot o spider
- Es muy útil para tomar información externa y enriquecer la información existente

Scraping data

¿Cuándo utilizarlo?

- Cuando necesitemos información que no cuento internamente
- Para tomar tendencias de precios, cambio de moneda, etc.
- En muchos casos se utiliza para comparar precio contra la competencia
- Información de productos por parte del fabricante

Scraping data

Herramienta para hacer el ingest

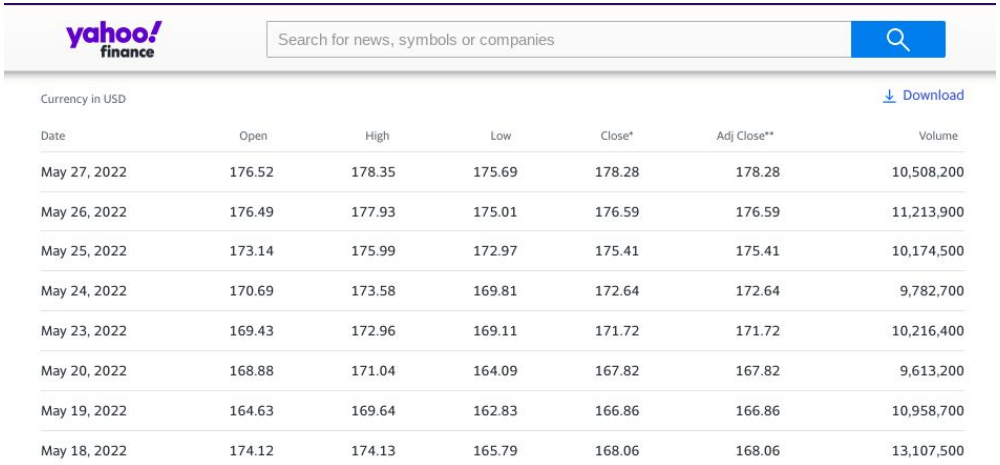
- Mayormente se realiza con librerías de Python3
- BeautifulSoup
- Scrapy
- Selenium

Scraping data

Ejemplos

- Comparar los precios de mi producto con los que hay en un e-comerce
- Mantener la cotización actualizada de el cambio de monedas
- Precio de stocks (acciones) de empresas
- Buscar en la página del fabricante características técnicas de un disp. móvil

Scraping data



Date	Open	High	Low	Close*	Adj Close**	Volume
May 27, 2022	176.52	178.35	175.69	178.28	178.28	10,508,200
May 26, 2022	176.49	177.93	175.01	176.59	176.59	11,213,900
May 25, 2022	173.14	175.99	172.97	175.41	175.41	10,174,500
May 24, 2022	170.69	173.58	169.81	172.64	172.64	9,782,700
May 23, 2022	169.43	172.96	169.11	171.72	171.72	10,216,400
May 20, 2022	168.88	171.04	164.09	167.82	167.82	9,613,200
May 19, 2022	164.63	169.64	162.83	166.86	166.86	10,958,700
May 18, 2022	174.12	174.13	165.79	168.06	168.06	13,107,500

Tipos de cambio frente al dólar

	Fecha	Cambio	Var.%	Var. Mes %	Var. Año %
Libras esterlinas [+]	27/05/2022	0,7942	0,19%	-0,35%	12,55%
Euros [+]	27/05/2022	0,9327	-0,23%	-1,61%	13,73%
Yenes japoneses [+]	27/05/2022	127,0100	-0,37%	-1,11%	15,66%
Yuanes chinos [+]	27/05/2022	6,7291	-0,05%	2,59%	5,43%
Leks [+]	27/05/2022	112,8000	0,01%	-0,87%	11,71%
Kwanzas [+]	26/05/2022	419,6469	0,01%	3,37%	-34,75%
Riales [+]	26/05/2022	3,7500	0	0	0
Dinares argelinos [+]	27/05/2022	145,0990	-0,24%	0,62%	8,71%
Pesos argentinos [+]	26/05/2022	119,4200	0,27%	3,99%	26,38%
Drams [+]	27/05/2022	448,1800	0,23%	-2,57%	-13,94%
Dólares australianos [+]	27/05/2022	1,4017	-0,95%	-0,14%	8,44%
Manat azeríes [+]	27/05/2022	1,7000	0	0	0
Takas [+]	16/03/2022	86,0000	0	0	1,41%
Dinares [+]	25/01/2019	0,3760	0	0	0
Dólares Beliceños [+]	26/05/2022	1,9988	0	0	0
Rublos bielorrusos [+]	27/02/2022	2,7977	5,90%	6,63%	7,52%
Bolivianos [+]	27/05/2022	6,8600	0	0	0
Marcos convertibles [+]	27/05/2022	1,8330	-0,38%	-0,22%	14,61%
Pulas [+]	27/05/2022	12,0337	0	-0,24%	13,24%
Reales brasileños [+]	26/05/2022	4,7961	-0,81%	-3,79%	-9,61%
Dólares de Brunéi [+]	27/05/2022	1,3705	-0,29%	-0,47%	3,50%

Desafíos de Ingestion

Ingest

Desafíos

- Fuentes cambian constantemente
- Páginas que contienen anti bot
- Cada vez exigen mayor cantidad de fuentes
- Monitoreo constante

Ingest