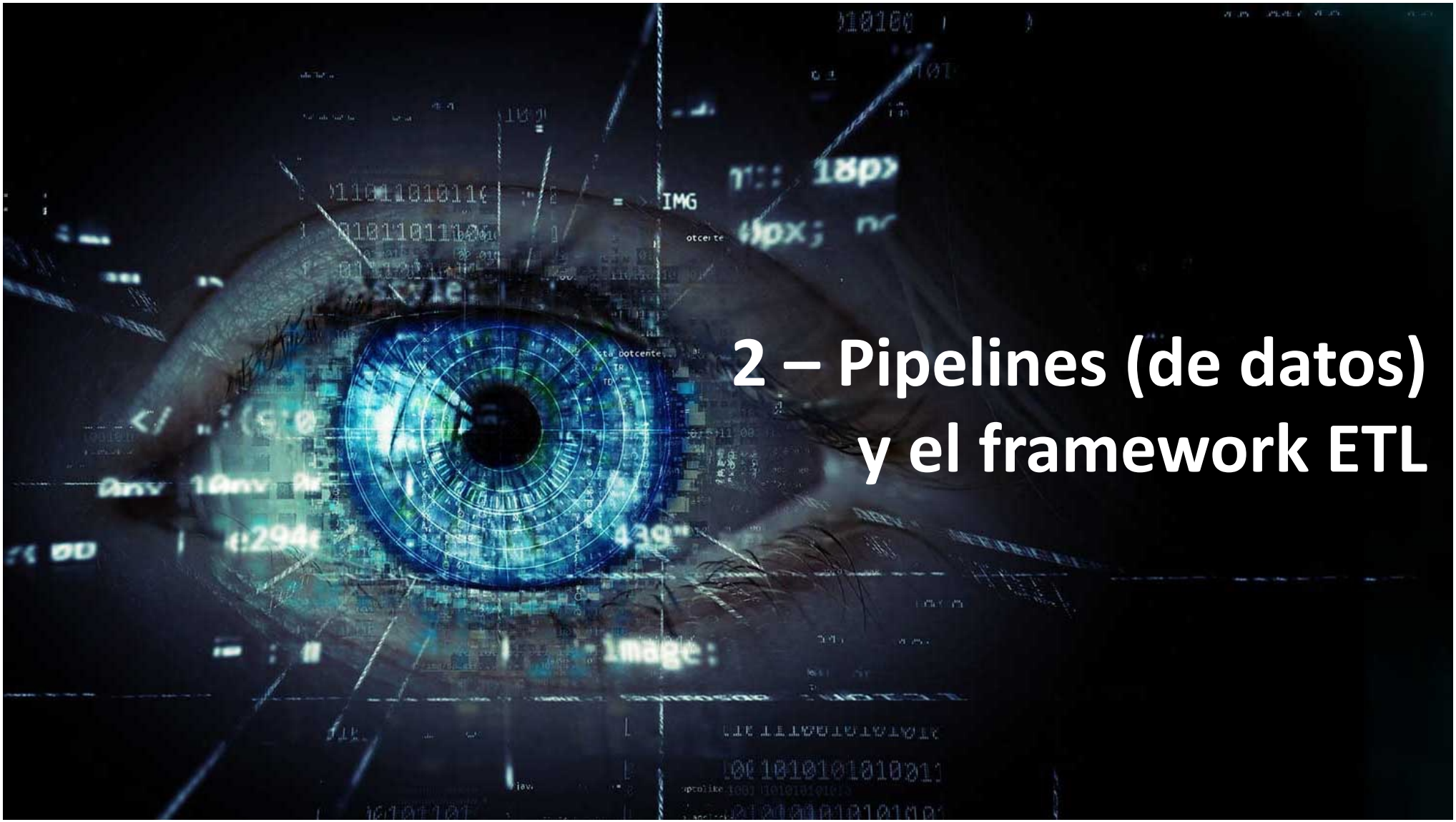


2 – Pipelines (de datos) y el framework ETL

Tubería de datos

- El objetivo del pipeline es poner en valor los resultados del análisis de datos.
- En proceso escalable y repetible
- Con un alto nivel de automatización

- Por ejemplo: un motor de recomendación para incentivar a los clientes a comprar más productos.

Framework ETL

Tuberías de datos

1
Mueven datos de un sistema a otro

2
Pueden seguir el framework ETL

4
Los datos pueden cargarse directamente en otras aplicaciones

3
Los datos pueden no necesitar transformación

Ejemplo pipeline

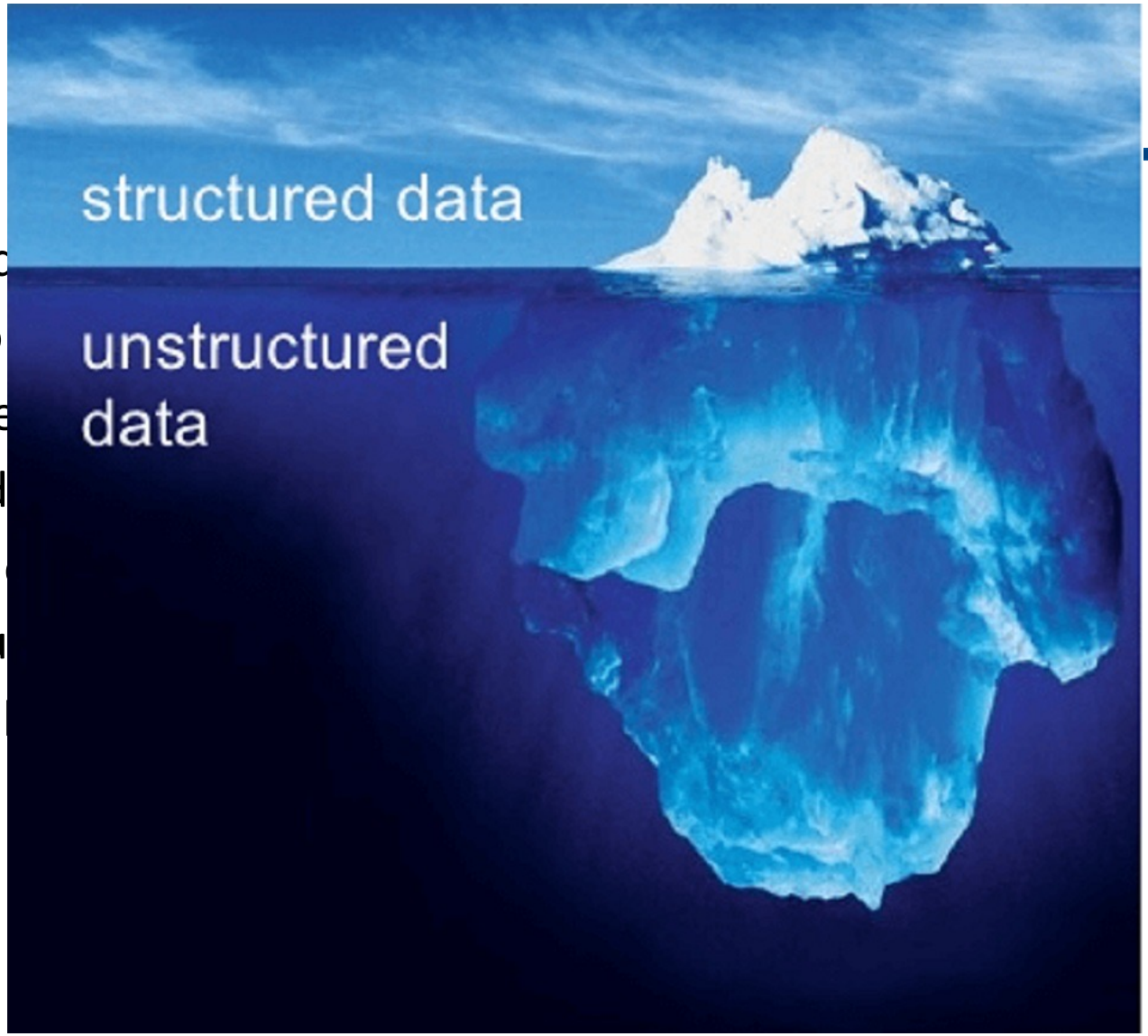
- Extract the songs Julian listened to the most over the past month
- Find other users who listened to these same songs a lot as well
- Load only the 10 top songs these users listened to the most over the past week into a table called "Similar profiles"
- Extract only songs these other users listen to that are of the same genre as the ones in Julian's listening sessions. These are our recommendations.
- Load the recommended songs into a new table. That's Julian's **Weekly Playlist**!


3 – Datos estucturados,
no estructurados y
mediopensionistas

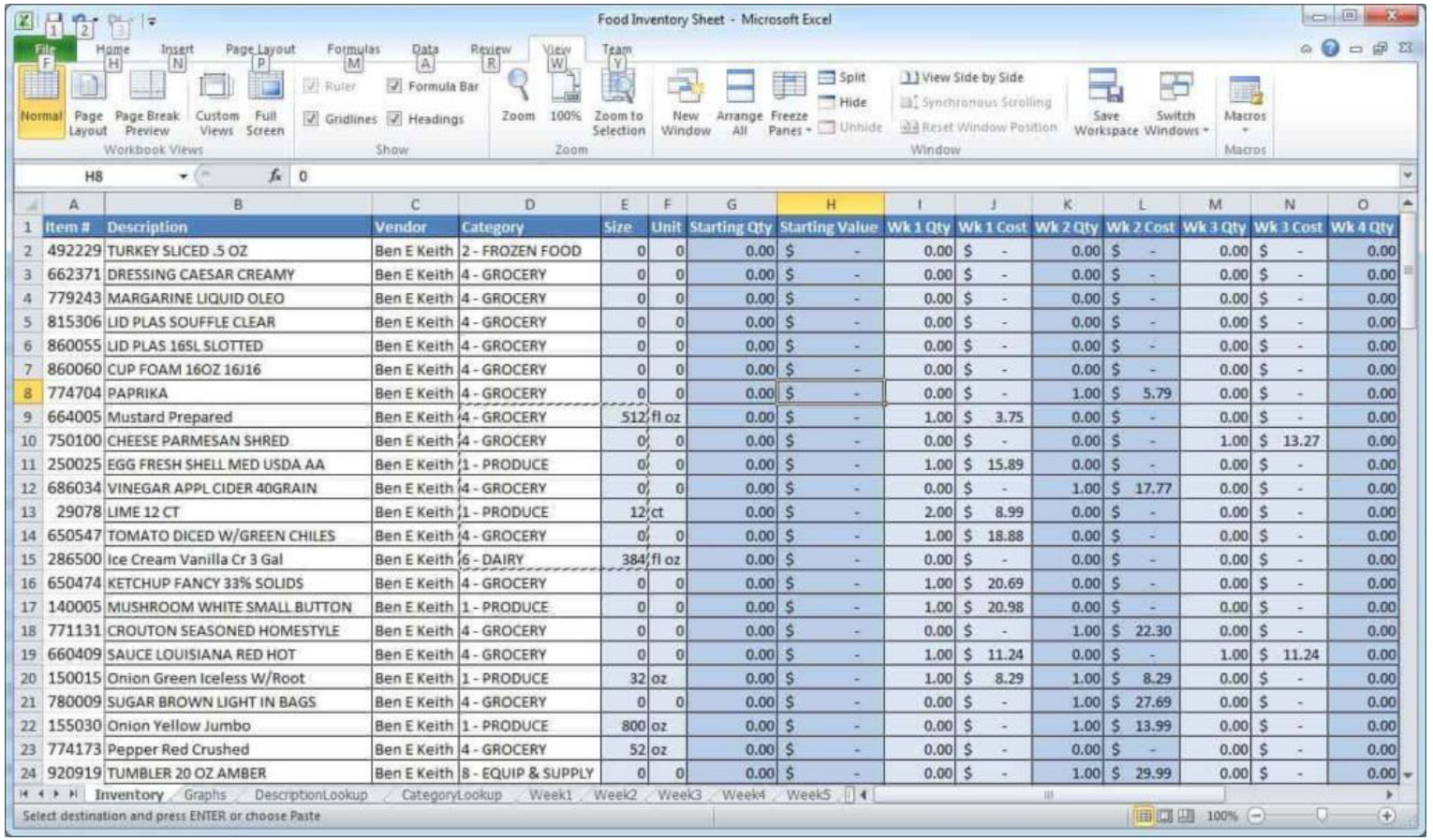
Datos estructurados

- Fáciles de buscar y organizar
- Modelo consistente: filas y columnas
- Tipos definidos
- Se pueden agrupar para formar relaciones
- Se almacenan en BBDD relacionales
- Se calcula que un 20% de los datos son estructurados
- Manipulados y consultados a través del lenguaje SQL

- Fáciles d
- Modelo
- Tipos de
- Se pued
- Se alma
- Se calcu
- Manipul

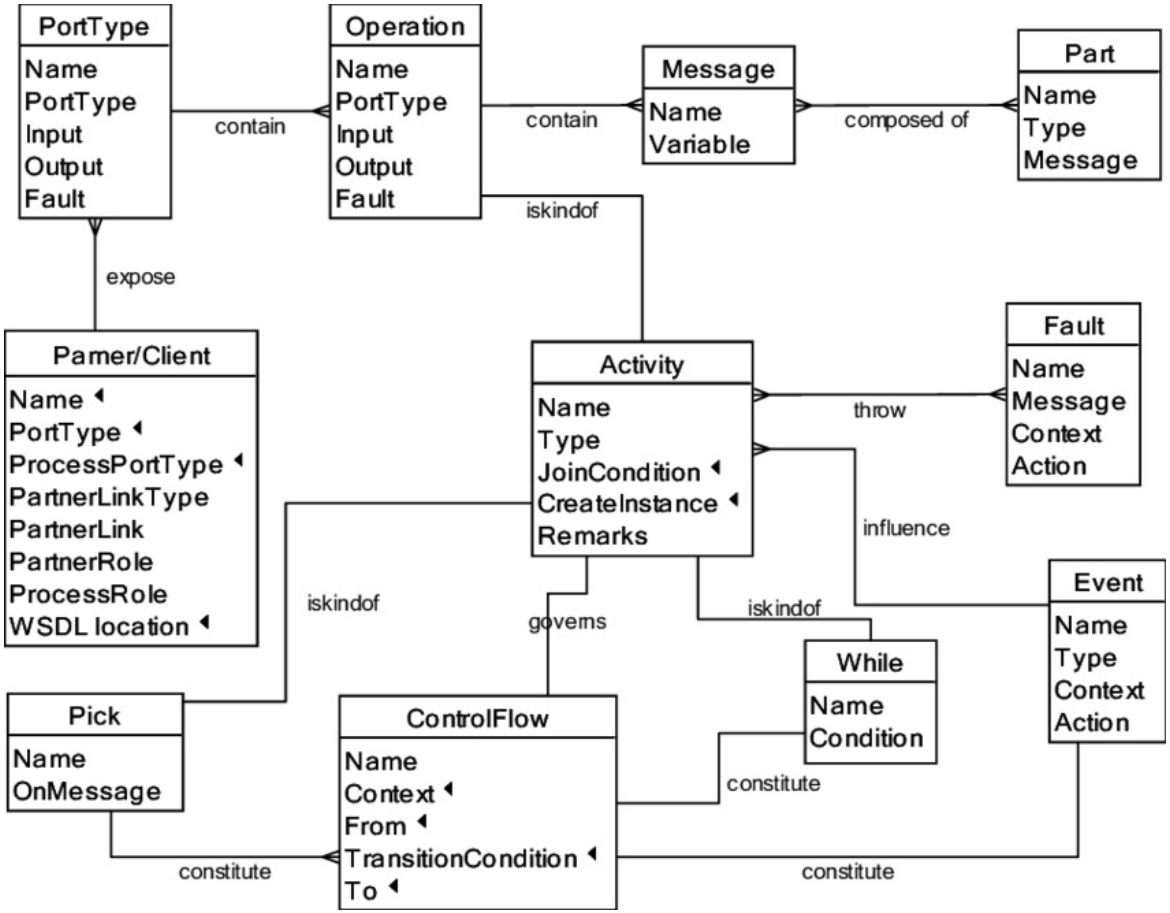


Ejemplo de datos relacionales (tabla)



Item #	Description	Vendor	Category	Size	Unit	Starting Qty	Starting Value	Wk 1 Qty	Wk 1 Cost	Wk 2 Qty	Wk 2 Cost	Wk 3 Qty	Wk 3 Cost	Wk 4 Qty
492229	TURKEY SLICED .5 OZ	Ben E Keith	2 - FROZEN FOOD	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
662371	DRESSING CAESAR CREAMY	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
779243	MARGARINE LIQUID OLEO	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
815306	LID PLAS SOUFFLE CLEAR	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
860055	LID PLAS 16SL SLOTTED	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
860060	CUP FOAM 16OZ 16J16	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
774704	PAPRIKA	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 5.79	0.00	$ -	0.00
664005	Mustard Prepared	Ben E Keith	4 - GROCERY	512	fl oz	0.00	$ -	1.00	$ 3.75	0.00	$ -	0.00	$ -	0.00
750100	CHEESE PARMESAN SHRED	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	0.00	$ -	1.00	$ 13.27	0.00
250025	EGG FRESH SHELL MED USDA AA	Ben E Keith	1 - PRODUCE	0	0	0.00	$ -	1.00	$ 15.89	0.00	$ -	0.00	$ -	0.00
686034	VINEGAR APPL CIDER 40GRAIN	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 17.77	0.00	$ -	0.00
29078	LIME 12 CT	Ben E Keith	1 - PRODUCE	12	ct	0.00	$ -	2.00	$ 8.99	0.00	$ -	0.00	$ -	0.00
650547	TOMATO DICED W/GREEN CHILES	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 18.88	0.00	$ -	0.00	$ -	0.00
286500	Ice Cream Vanilla Cr 3 Gal	Ben E Keith	6 - DAIRY	384	fl oz	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
650474	KETCHUP FANCY 33% SOLIDS	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 20.69	0.00	$ -	0.00	$ -	0.00
140005	MUSHROOM WHITE SMALL BUTTON	Ben E Keith	1 - PRODUCE	0	0	0.00	$ -	1.00	$ 20.98	0.00	$ -	0.00	$ -	0.00
771131	CROUTON SEASONED HOMESTYLE	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 22.30	0.00	$ -	0.00
660409	SAUCE LOUISIANA RED HOT	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	1.00	$ 11.24	0.00	$ -	1.00	$ 11.24	0.00
150015	Onion Green Iceless W/Root	Ben E Keith	1 - PRODUCE	32	oz	0.00	$ -	1.00	$ 8.29	1.00	$ 8.29	0.00	$ -	0.00
780009	SUGAR BROWN LIGHT IN BAGS	Ben E Keith	4 - GROCERY	0	0	0.00	$ -	0.00	$ -	1.00	$ 27.69	0.00	$ -	0.00
155030	Onion Yellow Jumbo	Ben E Keith	1 - PRODUCE	800	oz	0.00	$ -	0.00	$ -	1.00	$ 13.99	0.00	$ -	0.00
774173	Pepper Red Crushed	Ben E Keith	4 - GROCERY	52	oz	0.00	$ -	0.00	$ -	0.00	$ -	0.00	$ -	0.00
920919	TUMBLER 20 OZ AMBER	Ben E Keith	8 - EQUIP & SUPPLY	0	0	0.00	$ -	0.00	$ -	1.00	$ 29.99	0.00	$ -	0.00

Datos estructurados: esquema antes que los datos



Datos semi estructurados

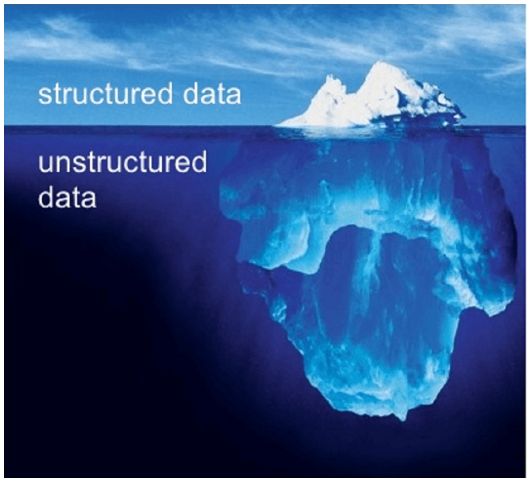
- Relativamente fáciles de buscar y organizar
- Modelo consistente con implementación menos rígida: diferentes observaciones pueden tener distinto tamaño (más o menos datos).
- Diferentes tipos
- Pueden ser agrupados, pero requieren más trabajo.
- Típicamente en BBDD NoSQL: JSON, XML, YAML

Ejemplo semi estructurado: JSON de Twitter

{
"id": "1212092628029698048",
"text": "We believe the best future version of our API will come from building it with YOU. Here’s to another great year … ",
"possibly_sensitive": false,
"referenced_tweets": [{ "type": "replied_to", "id": "1212092627178287104" }],
"entities": {
"urls": [{ "start": 222, "end": 245, "url": "https://t.co/yvxdK6aOo2", "display_url": "pic.twitter.com/yvxdK6aOo2" }],
"annotations": [{ "start": 144, "end": 150, "probability": 0.626, "type": "Product", "normalized_text": "Twitter" }] },
"author_id": "2244994945",
"public_metrics": { "retweet_count": 8, "reply_count": 2, "like_count": 40, "quote_count": 1 },
"lang": "en",
"created_at": "2019-12-31T19:26:16.000Z",
"source": "Twitter Web App",
"in_reply_to_user_id": "2244994945",
…
}

Datos no estructurados

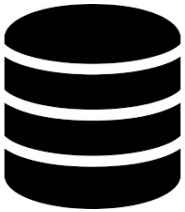
- No siguen un modelo, no pueden ser organizados en filas y columnas.
- Difíciles de buscar y organizar.
- Típicamente son textos, sonidos, imágenes o vídeos.
- Normalmente están almacenados en lagos de datos (*data lakes*), pero pueden aparecer en almacenes de datos (*data warehouses*) y bases de datos.
- La mayoría de los datos son no estructurados.
- Pueden ser extremadamente valiosos



Datos no estructurados (II)

- Normalmente se recurre al aprendizaje automático o la inteligencia artificial para buscar y organizar datos no estructurados.
- Se puede agregar información para convertirlos en semi estructurados.

Resumen



{“key” : “value”}

Estructurados (BBDD con esquema)

No estructurados (estilo ficheros)

BBDD relacionales

JSON

Videos, fotos

Gestión de Datos

Fin

Alvaro Ortigosa

alvaro.ortigosa@uam.es