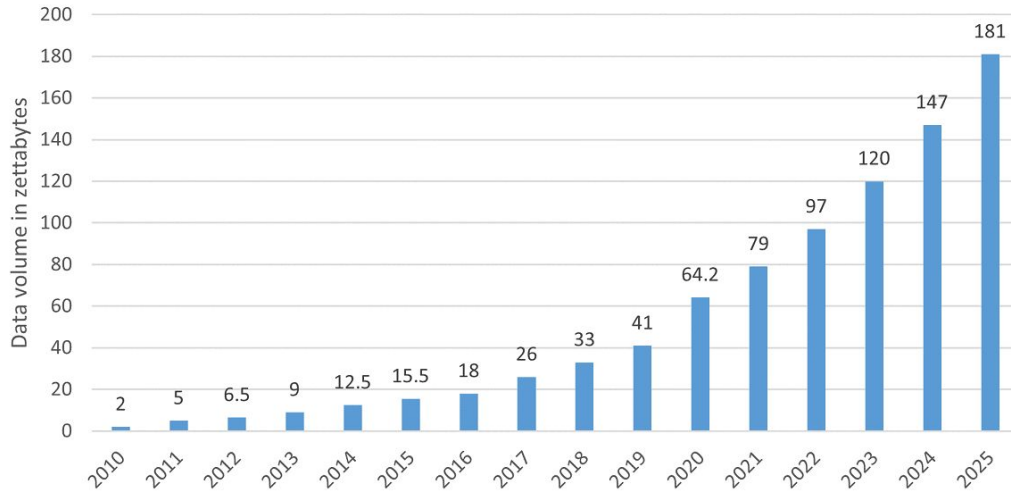


BD NOSQL

FIL AI

TAILLE DES DONNÉES

Volume of data created and replicated worldwide (source: IDC)

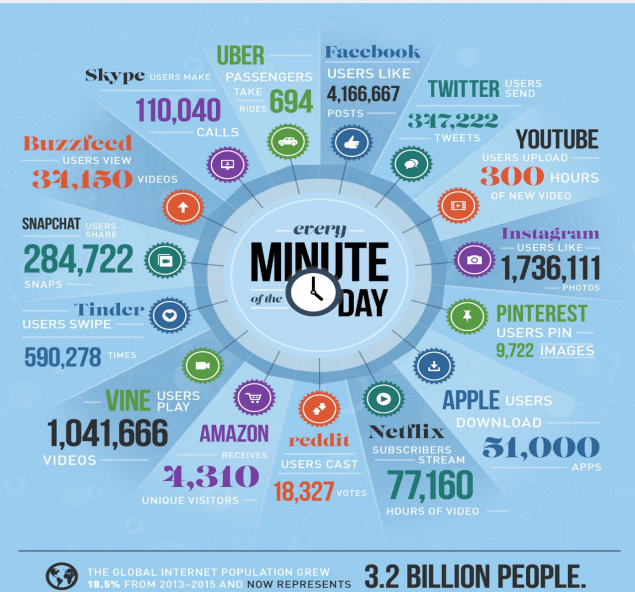


DOMO

DATA NEVER SLEEPS 3.0

How much data is generated every minute?

Data is being created all the time without us even noticing it. Much of what we do every day now happens in the digital realm, leaving an ever-increasing digital trail that can be measured and analyzed. Just how much data do our tweets, likes and photo uploads really generate? For the third time, Domo has the answer—and the numbers are staggering.



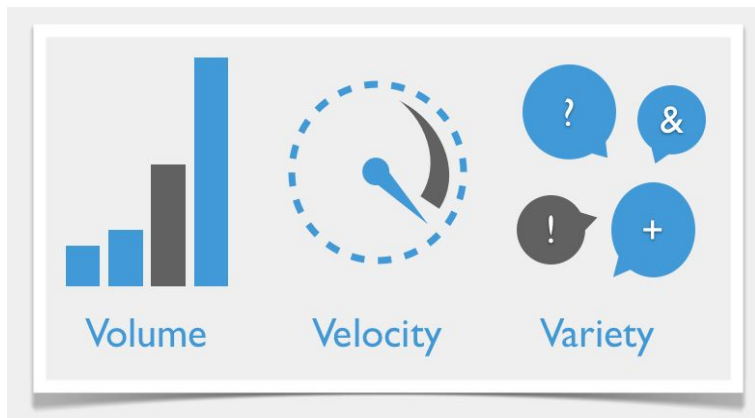
With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. [Learn more at www.domo.com](http://www.domo.com).

DOMO

SOURCES:

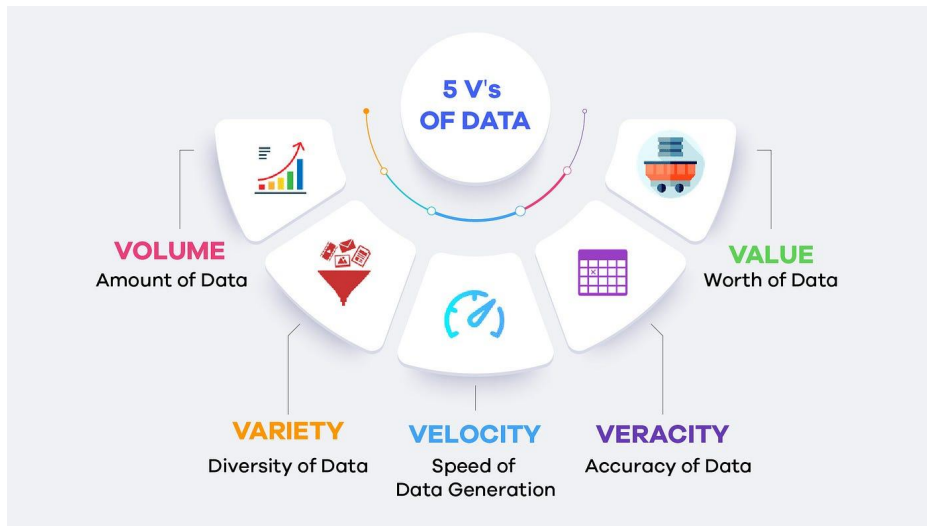
FACEBOOK, TWITTER, YOUTUBE, INSTAGRAM, PINTEREST, APPLE, NETFLIX, REDDIT, AMAZON, TINDER, BUZZFEED, STATISTA, INTERNET LIKE STATS, STATISTICBRAIN.COM

LES 3 V DU BIGDATA



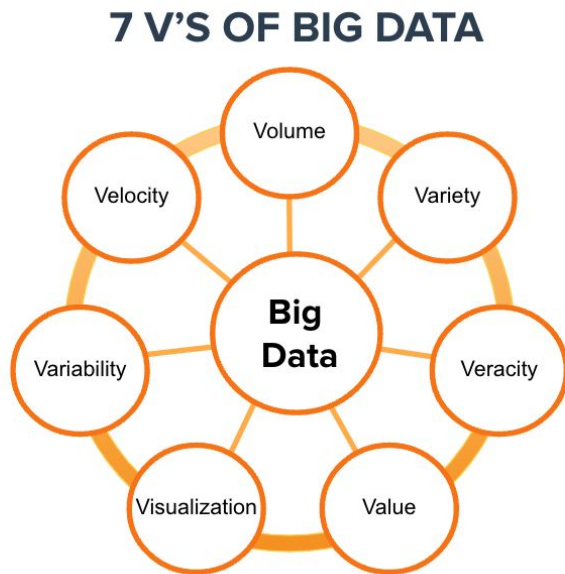
- **Volume**
 - La masse de données générées est de plus en plus grande
- **Vélocité**
 - La fréquence à laquelle sont produites les données est de plus en plus grande
- **Variété**
 - une grande variété de données collectées

LES 5 V DU BIGDATA



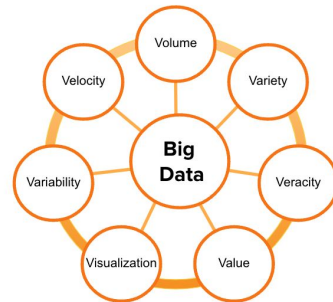
- Volume
- Vélocité
- Variété
- **Véracité**
 - Lié à la qualité de l'information, à son intégrité, à la fiabilité de la source
- **Valeur**
 - Les données brutes ont souvent peu d'intérêt, il faut réussir à créer de la valeur à partir des données

LES 7 V DU BIGDATA



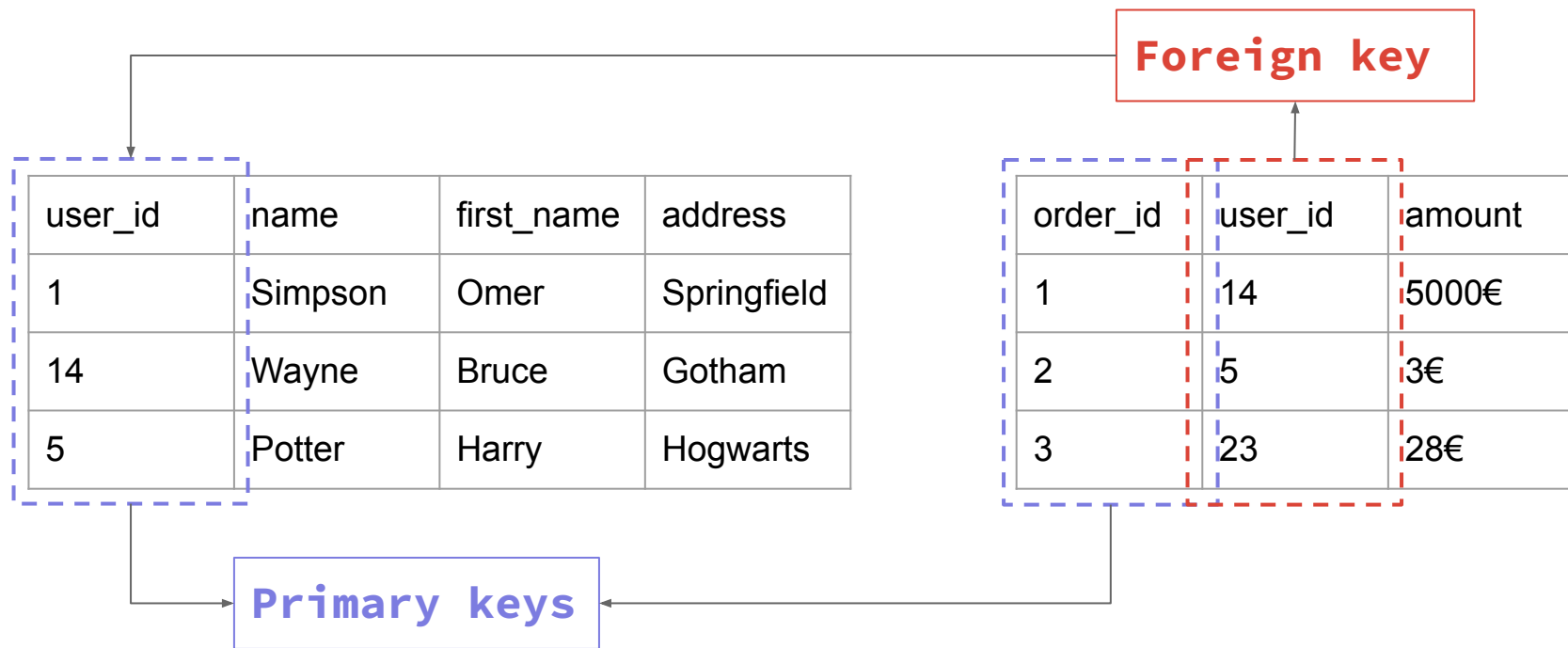
- Volume
- Vélocité
- Variété
- Véracité
- Valeur
- **Visualisation**
 - Rendre l'information exploitable par le plus grand nombre
- **Variabilité**
 - Nature changeante de la donnée dont le format ou la valeur peut varier avec le temps

CARACTÉRISTIQUES DES DONNÉES DU BIGDATA



- **volume**
 - il faut pouvoir **partitionner** les données
 - il faut **éviter les mécanismes qui ralentissent les requêtes**
 - il faut pouvoir faire des calculs parallèles
- **variété**
 - les données peuvent être **structurées semie- ou non-structurées**
 - la **nature** des données est très **variée**
- **variabilité**
 - les **schémas sont un frein** si les données évoluent dans le temps
- **vélocité**
 - on ne peut pas tout stocker, il faut traiter à la volée

RÉSUMÉ BD RELATIONNELLES



Données structurées / Données qui doivent valider le schéma

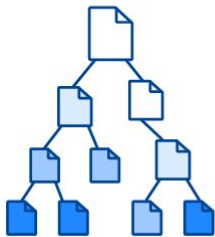
NOT ONLY SQL (NoSQL)

Utilisent des modèles de données dont la **structure est différente** de celle du modèle relationnel en table, lignes, et colonnes

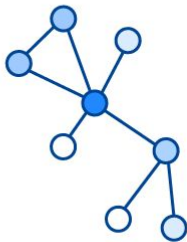
- Première apparition du terme en 1998
 - Carl Stroz - bases légères et open source
- Popularisé par les GAFAM
 - **2000** Neo4J
 - 2004 couchDB
 - **2008** Cassandra
- Adaptées au **BigData** pour différentes raisons...

4 TYPES PRINCIPAUX

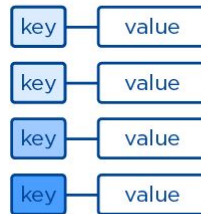
Document



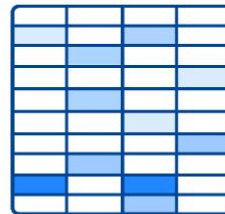
Graph



Key-Value



Wide-column



ORIENTÉES COLONNES

- Chaque colonne est traitée séparément, et les valeurs sont stockées de façon contigüe
- Hautes performances pour les requêtes d'agrégation comme SUM, COUNT, AVG et MIN



Google
BigTable



amazon
REDSHIFT



cassandra

order_id	user_id	amount
1	14	5000€
2	5	3€
3	23	28€
...		
2Billions		6859€

SUM

ORIENTÉES DOCUMENTS



- Ensemble de collections contenant des documents
- Un document est typiquement un objet JSON associé à une clé unique
- CRUD sur des documents JSON

```
{"id": "iuhd768", "type": "mobile", "name": "iPhone", "version": 15, ...}
```

```
{"id": "leo8", "type": "camera", "name": "Sony aR7", "optics": "mirrorless", ...}
```

```
{"id": "po65h", "type": "DVD", "name": "Harry Potter and the goblet of fire", "year": "2005", ...}
```

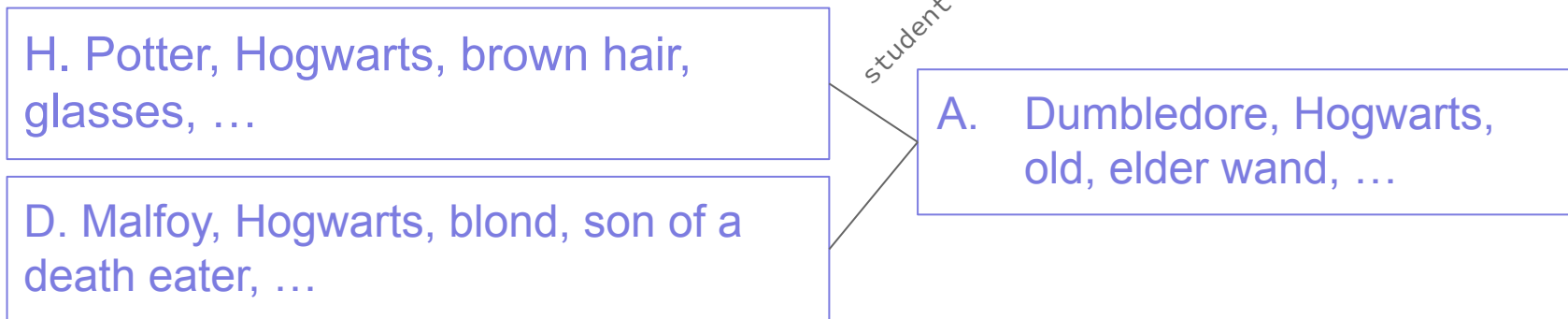
Object storage est similaire mais
les objets peuvent être non-structurés



ORIENTÉES GRAPHES



- L'entité est stockée sous forme de noeud, et les relations comme arêtes
- Facilite la visualisation des relations entre les noeuds
- On l'utilise principalement pour les réseaux sociaux, la logistique, etc.



ORIENTÉES CLÉS-VALEURS

- Les données sont stockées sous forme de paires clé / valeur (~table de hachage)
- Permet la prise en charge de larges volumes de données
- Les données sont entreposées dans un tableau de “hash” au sein duquel chaque clé est unique
- Stocker facilement des données sans schéma
- On récupère la valeur entière, pas de requêtes complexes



Amazon
DynamoDB

"dummy"	{"test": "ok", "nothing": True}
786	100111001011
"oih78zz"	42

AUTRES TYPES

SÉRIES TEMPORELLES



- Bases faites pour les événements ou mesures enregistrés dans le temps et associés à un timestamp
- Exemples : monitoring, données de capteurs IoT, enregistrement de clics, transactions financières etc.

timestamp	city (tag key)	country (tag key)	field	field value
2022-01-01T12:00:00Z	London	UK	temperature	12.1
2022-02-01T12:00:00Z	London	UK	temperature	12.5
2022-04-01T12:00:00Z	London	UK	temperature	5.4

ORIENTÉES RECHERCHE

- Requêtes pour faire de la recherche d'information dans des objets/documents JSON (orienté Read, pas Update)
- Données semi-structurées
- Optimisé pour la recherche d'information



elastic

ts |



algolia

```
{“id”: “iuhd768”, “type”: “mobile”, “name”: “iPhone”, “version”: 15, ...}
```

```
{“id”: “dkhb789”, “type”: “mobile”, “name”: “Sony aR7”, “optics”: “mirrorless”, ...}
```

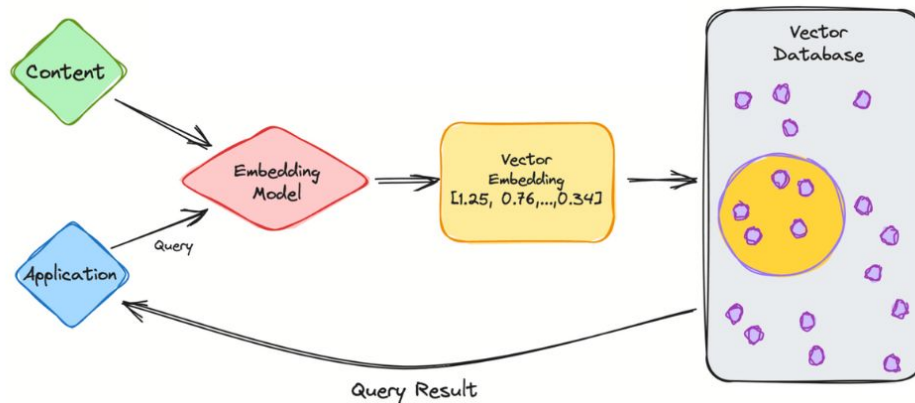
```
{“id”: “po65h”, “type”: “DVD”, “name”: ...}
```

type	ids
“mobile”	[“iuhd768”, “dkhb789”]

inverted index

VECTORIELLES

- Recherche de similarités dans des données non-structurées
- Calcul d'un vecteur représentant la donnée pour comparer des données de nature différentes par des **distances**
- **Indexation** des données pour accélérer les requêtes
- **Retrieval Augmented Generation (RAG)** : LLM utilise des bases vectorielles pour dynamiquement enrichir sa génération



VISUALISATION DE LA DONNÉE

- Prometheus, Grafana pour les time-series
- Elastic Kibana
- Neo4j Browser, Neo4J Bloom, Neovis.js
- ...

Mais il faudra découvrir par vous même, peut être dans votre tutoriel ou votre projet ?

