

BigData



Objectius



- Introducció al Bigdata
- V's del Bigdata

Evolució de les BD. El BigData

EL BIGDATA

S'anomena **Big Data**, dades massives o macrodades, als conjunts d'informació el volum de les quals sobrepassa les capacitats de les eines convencionals de processament de dades.

També es refereix a la recopilació de totes aquestes dades i la nostra capacitat d'utilitzar-les al nostre favor en una àmplia gamma d'àrees, des dels negocis fins a la política.



Bigdata

"Big data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."

Dan Ariely, Duke University

Big data fa referència a dades d'un mida molt gran, de forma que la seva manipulació i gestió presenten desafiaments significatius.

Oxford English Dictionary

Big data o dades a gran escala es un concepte que fa referència a un conjunts de dades tan grans que aplicacions informàtiques tradicionals de processat de dades no són suficients per tractar-les

Wikipedia



Big data fa referència al conjunt de dades que sobrepassa la capacitat del software tradicional per poder ser capturats, emmagatzemats, gestionats i analitzats.

McKinsey & Company

L'habilitat d'aprofitar la informació per aconseguir crear noves idees i amb això valor

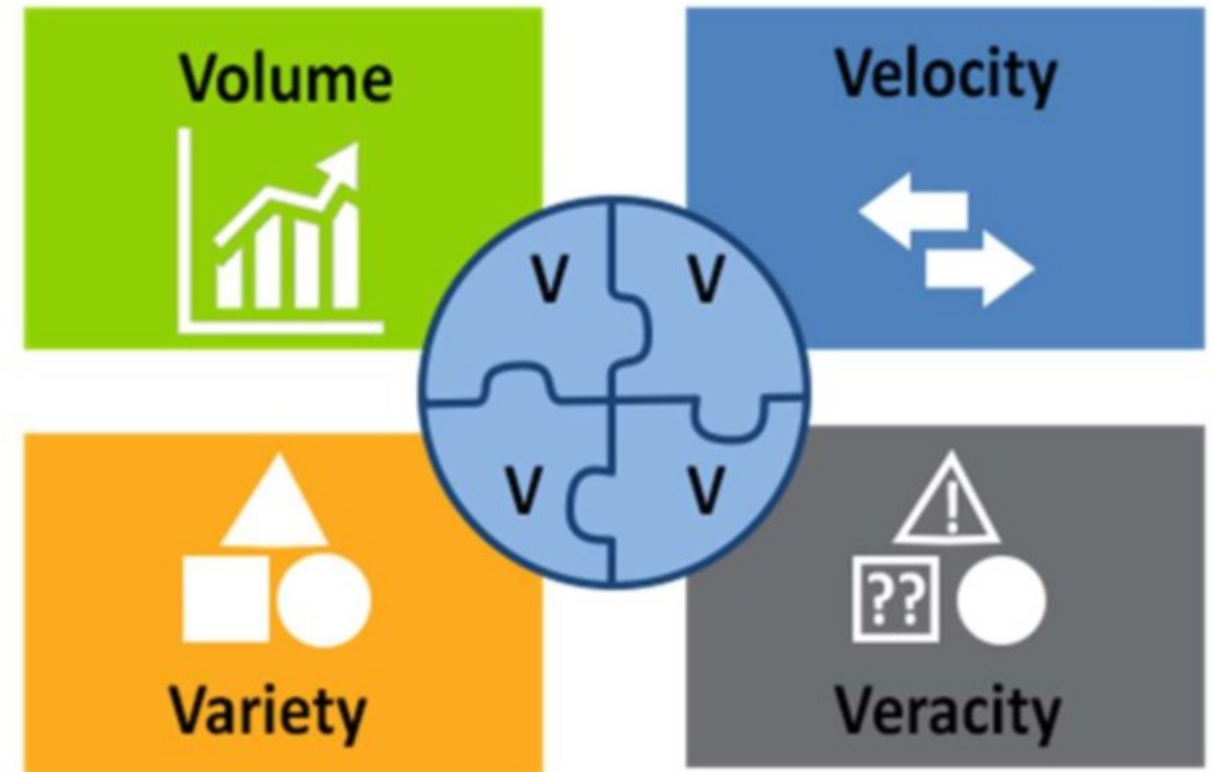
Viktor Mayer

V's del Bigdata

- **Volum:** Grans quantitats de dades a tractar. Ordre de magnitud de Petabyte – Zettabyte,...
- **Varietat:** Dades estructurades, no estructurades, textos, imatges, etc...
- **Velocitat:** Quantitat de temps que invertim en el processat de les dades. Terme important i a tenir en compte
- **Veracitat:** Fiabilitat de la informació que obtenim.

Podem afegir altres com:

- **Variabilitat:** Capacitat de les dades de canviar constantment. El significat d'una paraula pot variar depenent del context en el qual ens trobem.
- **Valor:** Convertir les dades en brut en informació útil. De tot el que fem hem d'aconseguir-ne un valor.
- **Visualització:** Cal un mecanisme de facilitat de lectura i comprensió d'aquestes



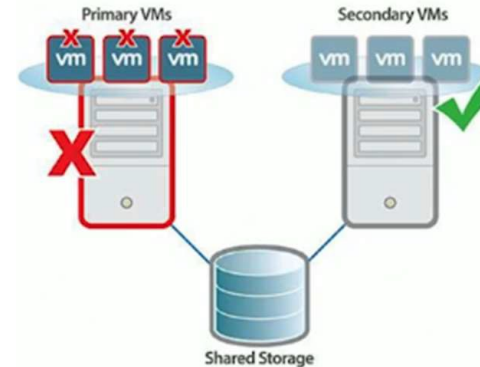
[What is big data? More than volume, velocity and variety - IBM Developer](#)

BigData: Conceptes bàsics



Processament distribuït

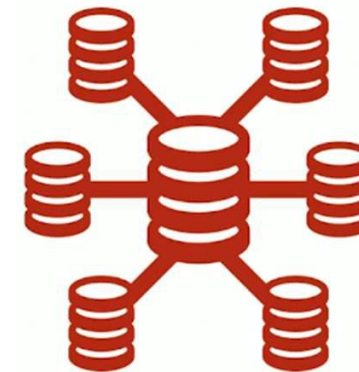
Les dades es processen en diferents màquines (nodes) aconseguint millorar la velocitat. No treballar amb un sol fitxer, sinó dividir les dades en diferents màquines i que aquestes processin la seva parcel·la de dades.



Tolerància a fallades

Cal que la infraestructura sigui tolerant a fallades. Si un node cau, el sistema ha de seguir funcionant i redistribuir la feina de les màquines que han caigut.

Si una màquina cau no cau tot el sistema i poder seguir processant les dades encara que en més temps.

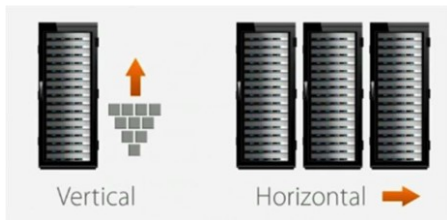


Localitat de les dades

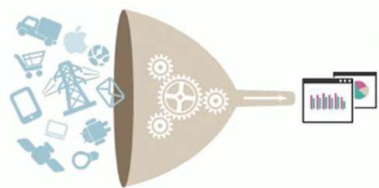
Cada node/màquina treballarà amb les dades que tingui més a prop. Així aconseguim evitar colls d'ampolla i minimitzar latències.

Escalabilitat horitzontal

Els sistemes creixen molt ràpid. Per tant necessitem sistemes que ens permetin afegir nous nodes i no afegir nou HW a les màquines existents. Amb això aconseguirem que mitjançant hardware econòmic puguem realitzar la feina sense necessitat de fer grans inversions.



BigData: 4 fases del Cicle de vida



1) Ingesta:

Les dades procedents d'una font externa (web, BBDD,..) entren en el nostre sistema per començar a treballar amb elles.

Per exemple un procés nocturn que reculli les dades de la BBDD i les tracti o un procés en streaming que vagi agafant tweets o dades IoT.



2) Persistència:

Cal guardar les dades i emmagatzemar-les en el nostre sistema. Normalment en una BD relacional o no relacional.

D'alguna manera les estructurarem com volem.



3) Processament:

Fase en la que processem i analitzem les dades realitzant operacions per tal d'aconseguir valor.



4) Visualització:

Mostrar gràficament o d'una forma amigable els resultats obtinguts per aquelles persones que n'extrauran conclusions.

BigData: Casos d'ús



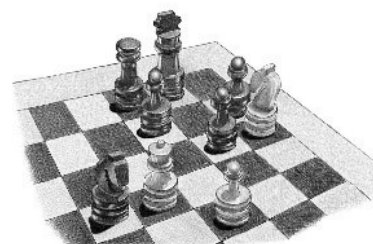
Processament de logs: Anàlisi de grans quantitats de dades desestructurades que es converteixen en valor. Per exemple, dades dels consums energètics d'una vivenda, processament de logs de servidors per trobar colls d'ampolla, a on hi ha més errors, etc...



Salut: Processament d'anàlisi clínics, ADN, hàbits alimentaris i ajuda a comprendre malalties. Si controlem que certs pacients amb els següents hàbits i han tingut certes malalties podem treure conclusions.



Sistemes de recomanació: Recomanació de productes gracies a l'anàlisi de les preferències i perfils d'altres usuaris



Esport: Estudiar i analitzar estadísticament i vídeos per treure patrons típics de joc, analitzar rivals, etc...



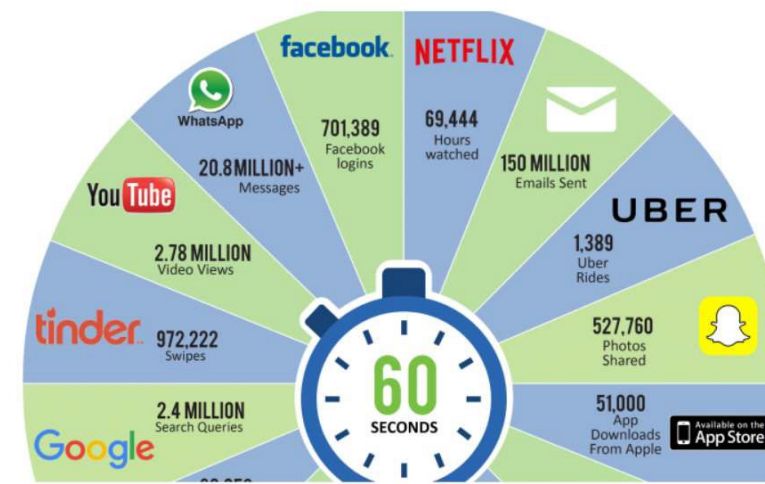
Cercadors web: Cerca i ordenació de milions de pàgines d'acord amb unes paraules clau.



Món Financer: Quan comprar, vendre, anàlisi de mercat, Etc ...

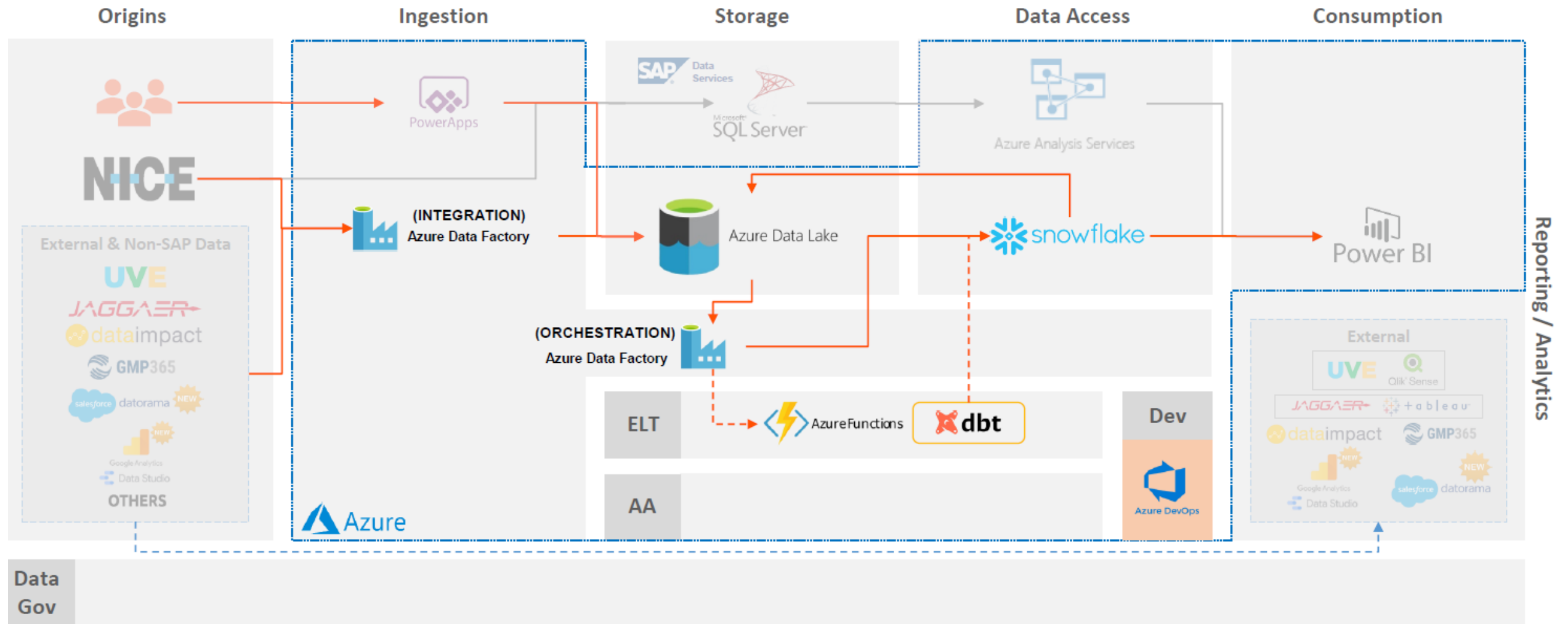
Evolució de les dades

- En el 2013 hi havia 4,4 zettabytes en el món i es preveu que en el 2023 hi hagi més de 50 zettabytes.
- El creixement de les dades són exponencials. En el 2023 cada ser humà crearà **2 MB d'informació cada segon**.
- **Facebook:**
 - Agost 2015 1 bilió d'usuaris actius
 - Juny 2017 1'32 bilions d'usuaris actius cada dia.
- Cada minut s'envien 31,25 milions de missatges i es visualitzen 2.77 milions de vídeos.
- A **Youtube** es pugen 300 hores de vídeos diaris.

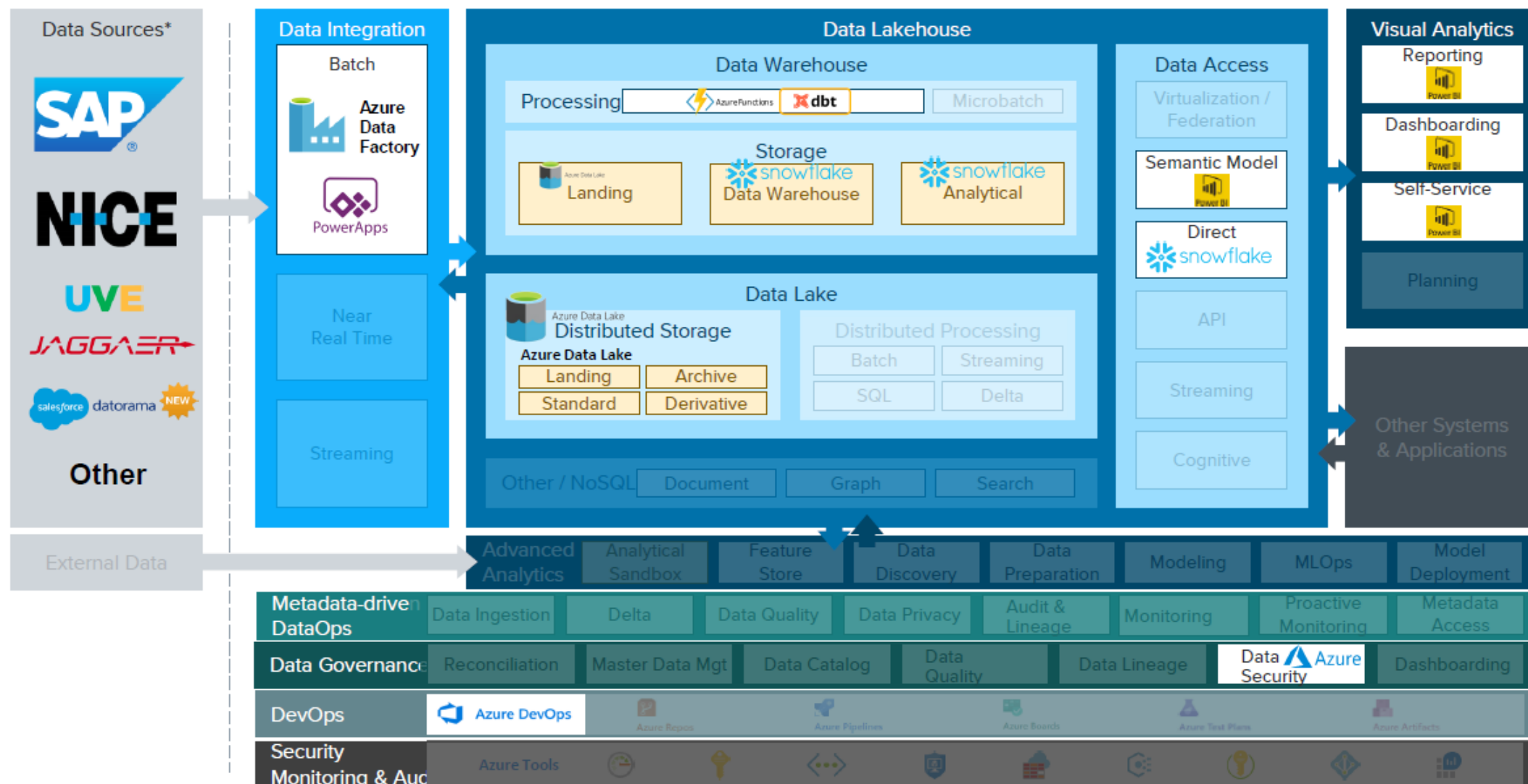


Actualment menys del 0,5% de les dades s'estan analitzant.

BigData: Cas real



BigData: Cas real



WEBGRAFIA

- Batini, C.; Ceri, S.; Navathe, S.B. (1992). Conceptual Database Design: An Entity-Relationship Approach. Reading, Massachusetts: Addison Wesley.
- Teorey, T.J. (1999). Database Modeling & Design. The Fundamental Principles (3a ed.). San Francisco: Morgan Kaufmann Publishers, Inc.