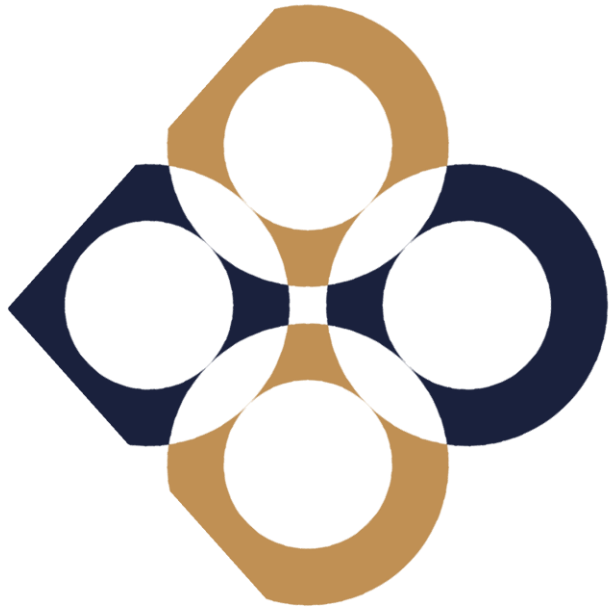
BUDAPESTI CORVINUS EGYETEM

Adatbázisok előadás 06

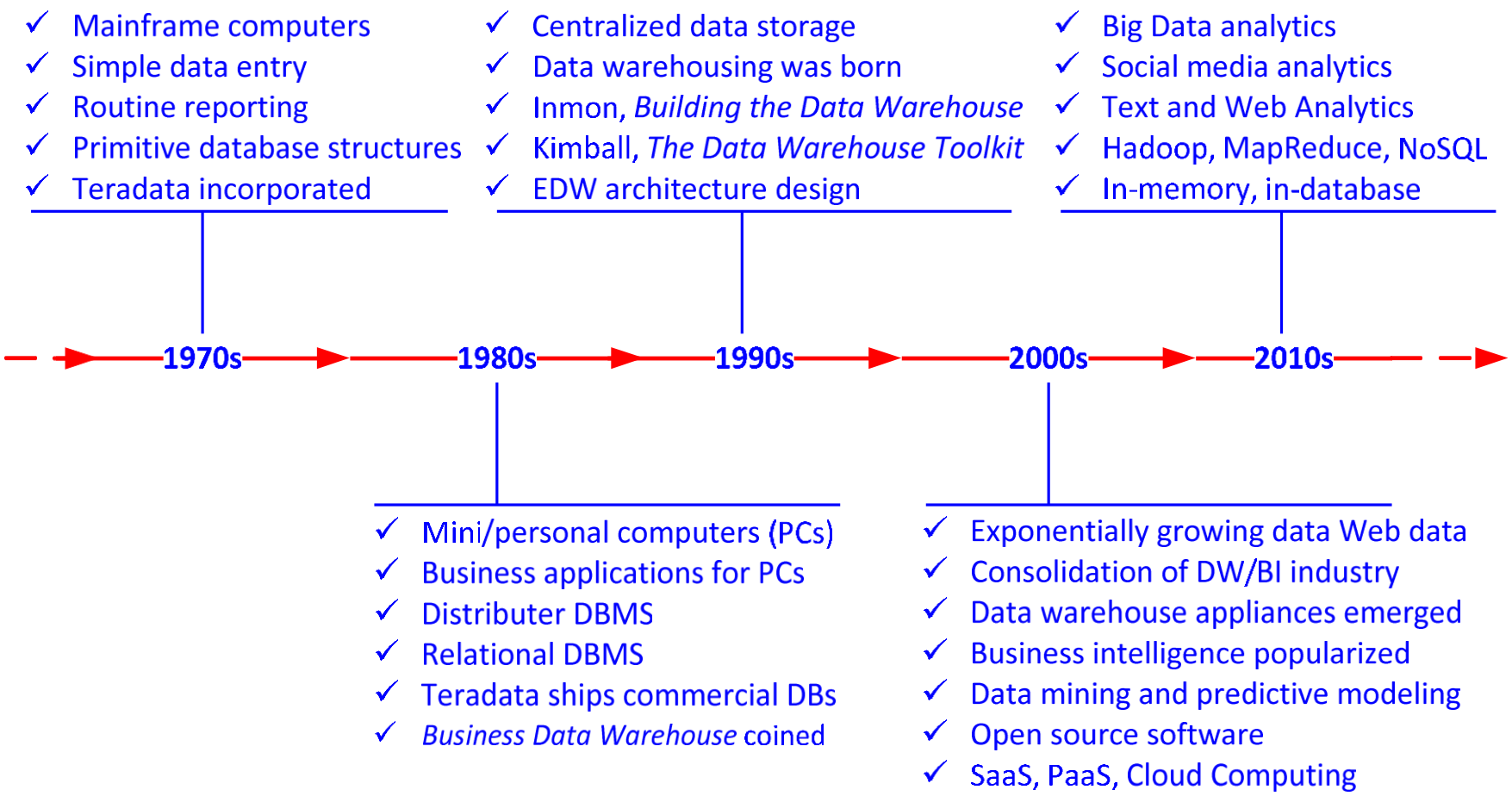
Adattárházak



Miről lesz szó?

- *Adattárház és a hozzá kapcsolódó fogalmak*
- *Kimball modellje*
- *Inmon modellje*
- *Kihívások, alternatív adattárház rendszerek*

Történeti áttekintés - címszavakban



Vezetői információs rendszerek

- 1980-as évek - Legelső vezetői információs rendszerek
  - Működési célú (tranzakciós, operatív) adatbázisok (OLTP)
  - Előre definiált riportok
  - Egyre több adat --> komoly terhelést jelentett az adatbázis rendszer számára
- Ötlet – készítsünk másolatot az adatbázisról, és azon futtassuk a riportokat
  - A működési és az elemzési célú adatbázisok egyre inkább különváltak
  - Az elemzési célú adatbázisokból fejlődtek ki az adattárházak (DWH, Data Warehouse)

Miért van szükség adattárházakra?

- ❑ Nehézkes és lassú riportolás
- ❑ Adatminőségi problémák
- ❑ Egységes metaadat kezelés hiánya
- ❑ Elemzési és adatbányászati igények megjelenése
- ❑ Körülményes a riportkészítés több adatforrás esetén
- ❑ Az adatok sokszor különböző formában állnak rendelkezésre

Milyen elven működnek az adattárházak?

- A különböző forrásokból származó adatokat egy helyre összegyűjtjük, majd
- aggregáljuk olyan mértékben, amilyen léptékben döntéseinket hozzuk, majd
- az üzleti gondolkodásnak megfelelő új struktúrát alakítunk ki, majd
- készítünk olyan eszközt vagy célalkalmazást, amely az elemzéshez szükséges funkciókat biztosítja, végül
- elérhetővé tesszük a vezetők és felhasználók számára

Adatgyűjtés
Aggregálás
Strukturálás
Eszköz készítése
Eszköz elérhetővé tétele

Adattárház fogalma

Az adattárház a tranzakciós adatok lekérdezési és elemzési célokból speciálisan strukturált másolata (Kimball)

Az adattárház témakör-orientált, integrált, időfüggő, de időben nem változó adatok gyűjteménye, amelyet a cég vezetői döntéshozatalának támogatására használnak (Inmon)

Adattárház – néhány más elnevezés

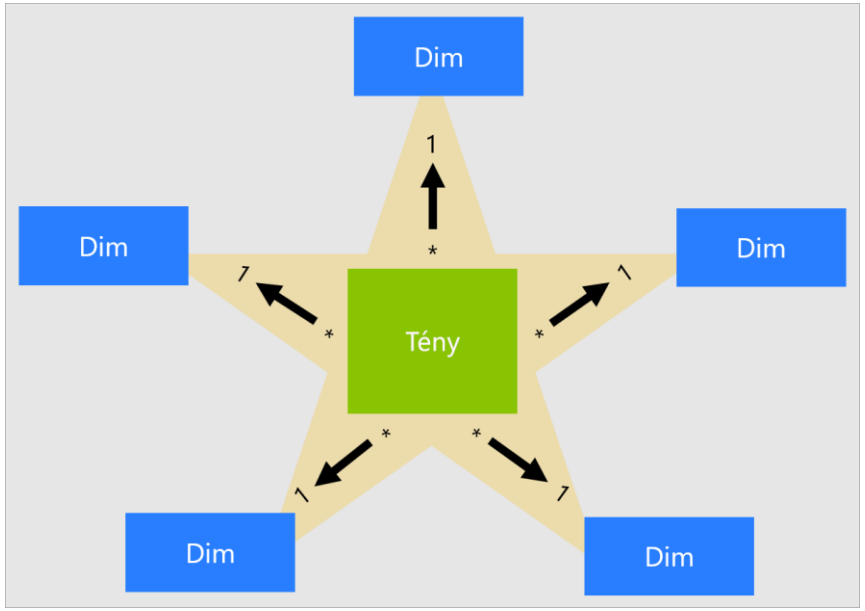
(OLTP) Adatbázis vs. adattárház

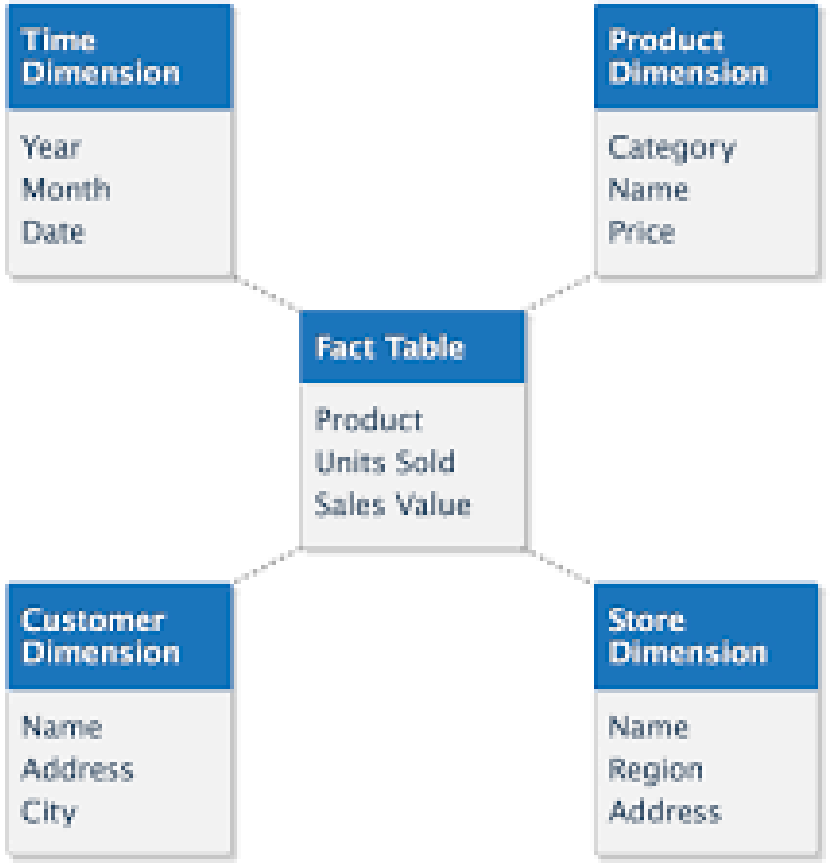
- Sok felhasználó
- Sok kicsi, konkurrens tranzakció
- Rengeteg SQL utasítás (pl. másodpercenként több ezer, esetenként még ennél is több)
- Az utasítások önmagukban leginkább egyszerűek
- A lekérdezések döntő többsége kevés sort érint
- A tranzakciókezelés a fő kihívás, az utasítások végrehajtása általában könnyű
- Elvárt a magasfokú normalizáltság: minél inkább elkerüljük az anomáliákat

- Viszonylag kis számú felhasználó (nevezik őket adatelemzőknek, adatbányászoknak is)
- Viszonylag kis számú, de gyakran igen nehéz lekérdezés
- Nem jellemzők az egyidejű tranzakciók – nem akkora gond a konzisztencia
- Gyakran nem jellemző az adatok online változtatása
- A lekérdezések jellemzően összesített adatokat kérnek
- Jellemzően nem cél az adatok normalizáltsága, hanem az ún. csillag séma (star schema) szerinti az adatmodell

OLTP (Online Transaction Processing): hagyományos (működési célú) adatbázis rendszert jelent

Csillag (Star) séma



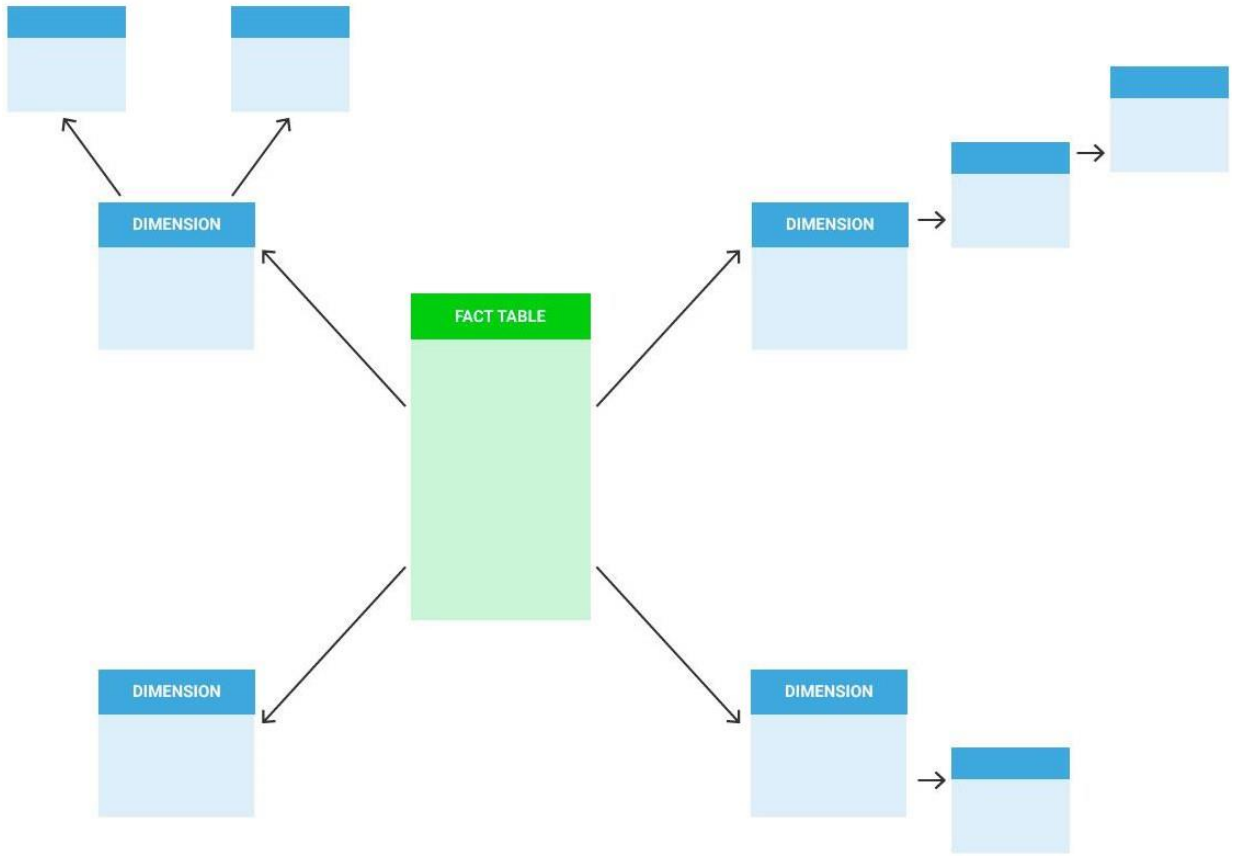


Tényadatok (ténytáblák) vs. Dimenzionális adatok (dimenziók)

Szempont	Tényadat	Dimenzió adat
Cél	Mérőszámok	Leíró adatok
Jellemző adattípus	Numerikus	Nem numerikus
Előfordulás	Sok rekordban	Kevés rekordban
Darabszám	Kevesebb	Több
Példa	Eladás összege	Termék kategóriája

- A ténytáblák tényadatokat és idegen kulcsokat tartalmaznak
- A dimenziótáblák dimenzió adatokat és elsődleges kulcsokat

Hópehely (Snowflake) séma



A dimenzió táblákhoz újabb dimenzió táblák kapcsolódhatnak

- Kisebb mértékű denormalizáltság
- Kisebb redundancia
- Kisebb helyfoglalás
- Bonyolultabb lekérdezések
- Lassúbb végrehajtás
- Nehezebben áttekinthető

A kép forrása: https://blog.sqldbm.com/bi-data-modeling-with-sqldbm/

OLAP – Online Analitical Processing

Codd (1992)

- Multidimenzionális nézet — Az adatokat több szempontból lehet vizsgálni, például időszak, termék, földrajzi hely szerint. Képzeld el, mintha egy kocka lenne, amelynek minden oldala egy másik szempont.
- Transzparencia — Az adatok könnyen elérhetők és érthetők, nem igényel mély technikai tudást.
- Jogosultságok beállíthatósága — Lehet szabályozni, hogy ki milyen adatokat láthat és módosíthat.
- Állandó lekérdezési teljesítmény — Függetlenül attól, hogy mekkora az adatmennyiség, a lekérdezések mindig gyorsan futnak.
- Kliens-szerver architektúra — Az OLAP egy szerveren fut, és a felhasználók különböző eszközökről csatlakozhatnak hozzá.
- Általános dimenzió fogalom — Az adatok rendszerezése rugalmasan kezelhető dimenziók segítségével, például idő, termék, régió.
- Egyidejűleg több felhasználó
- Korlátozás nélküli dimenzió műveletek — Az adatok bármilyen szempont szerint csoportosíthatók és elemezhetők.
- Intuitív adatkezelés — Az OLAP rendszerek általában könnyen kezelhetők, nem kell mély programozási tudás a használatukhoz.
- Rugalmas riportozás — Különböző formátumú jelentéseket és kimutatásokat lehet készíteni, attól függően, hogy mire van szükség.
- Korlátlan dimenziószám és aggregációs szint — Az adatok tetszőleges számú dimenzió mentén elemezhetők, és különböző szinteken lehet összesíteni őket (például napi, havi vagy éves szinten).

OLAP vs. OLTP*

Szempont	OLTP	OLAP
Cél	Napi működés	Elemzések
Felépítés	Relációs adatbázis	Adattárház
Lekérdezés	INSERT, DELETE, UPDATE	SELECT
Tábla	Normalizált	Nem normalizált
Forrás	Egy forrásrendszer	Több OLTP forrásrendszer
Válaszidő	Rövid	Nagy

* Nem véletlen, hogy az adatbázis vs. adattárház összehasonlításhoz hasonló szempontok vannak itt is. Az adatbázisokat sokszor az OLTP, adattárházakat pedig az OLAP névvel is azonosítják

Adatpiac (Data Mart) – nincs egységes meghatározás

Az adatpiac nem normalizált, aggregált adatokat tartalmaz, és valamely szervezeti egység működési, felhasználói követelményei által meghatározott (Inmon, 2002)

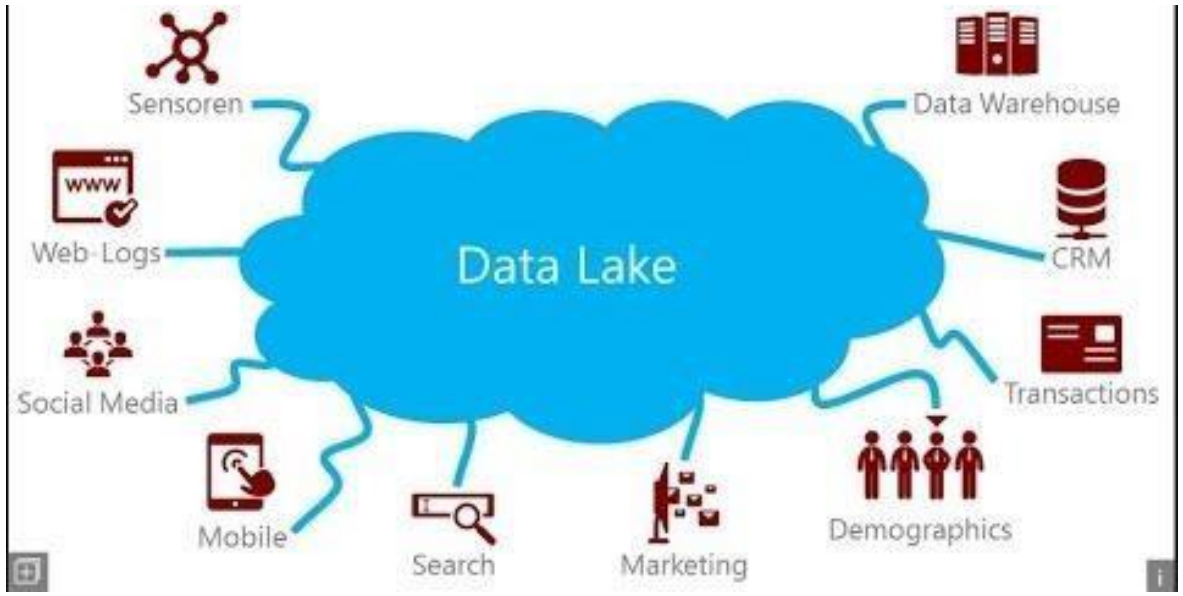
Az adatpiac az adattárház egy tárgyterülethez (pl. marketing) köthető része (Turban, 2014)

Az adatpiacok az adattárházakhoz hasonló adatkezelési képességekkel rendelkeznek, de egy-egy szervezeti egység speciális információs igényeinek megfelelően optimalizáltak (Wikipédia)

Az adatpiac lehet egy önállóan létező, az adattárháztól független megoldás, vagy annak része

Adattavak (Data Lake)

Nagyvállalati szintű adatmenedzsment-platform, amelyen a különböző forrásokból származó adatok natív formátumukban érhetők el elemzésre



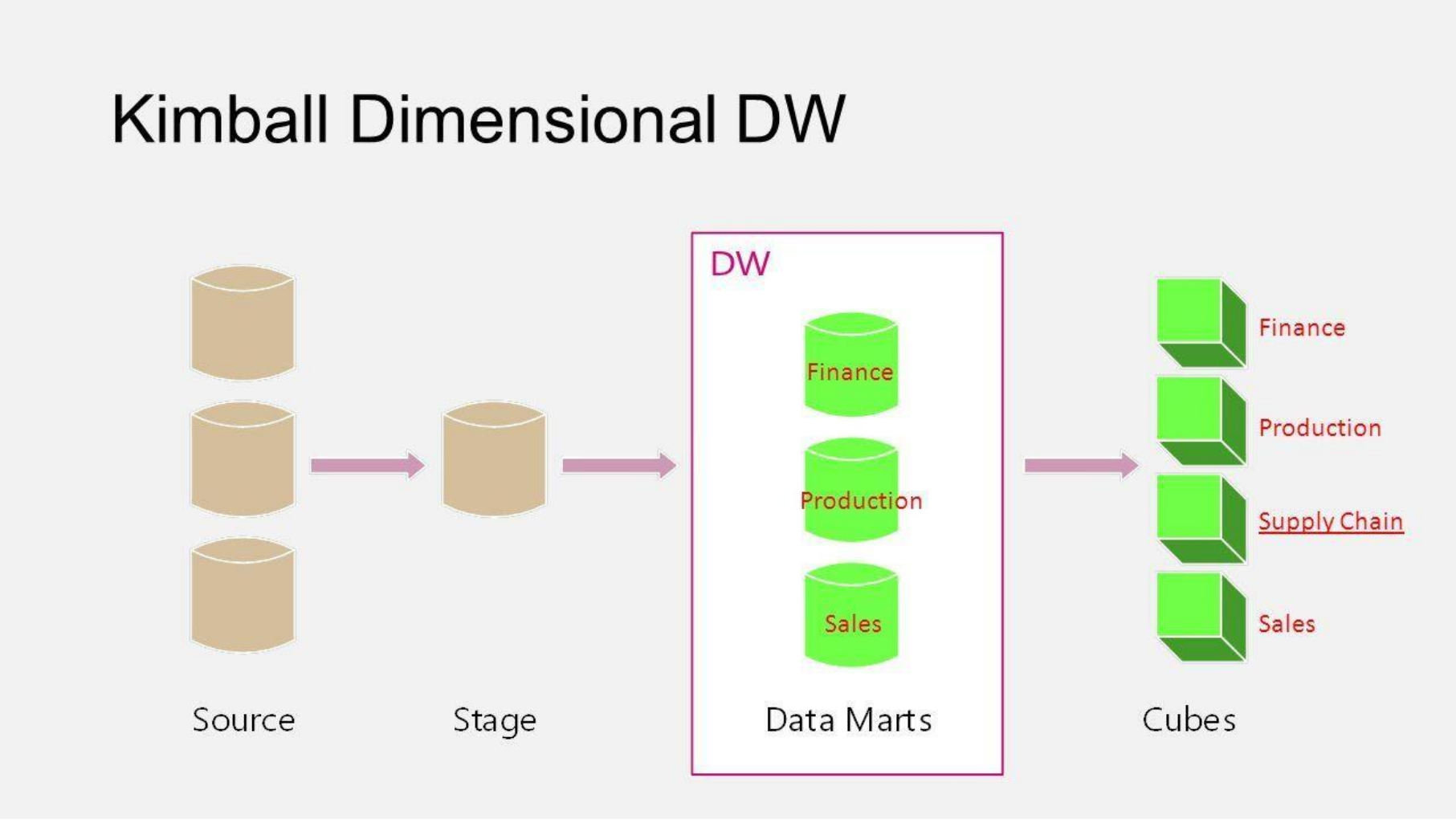
- Big Data rendszerekre jellemző
- Központosított adatkezelés
- Az adattárházaknál rugalmasabb struktúra
- Strukturált és nem strukturált adatokat is tartalmazhatnak

A kép forrása: https://www.youtube.com/watch?v=a3nezIjxI-E

Adattavak vs. adattárházak

Data lake	Data warehouse
Az adatok tárolásának célja előre nem definiált	Előre definiált tárolási cél
Az adatok nyers formában tárolódnak	Az adatok lekérdezésre alkalmas formában tárolódnak
Adattudósok, adatelemzők használják	Üzleti felhasználók használják
Feltörekvő technológia	Kidolgozott technológia
NoSQL lekérdezések	SQL lekérdezések
Gyors válaszidő	Lassú válaszidő
Alacsony költségű tárolás	Magas költségű tárolás

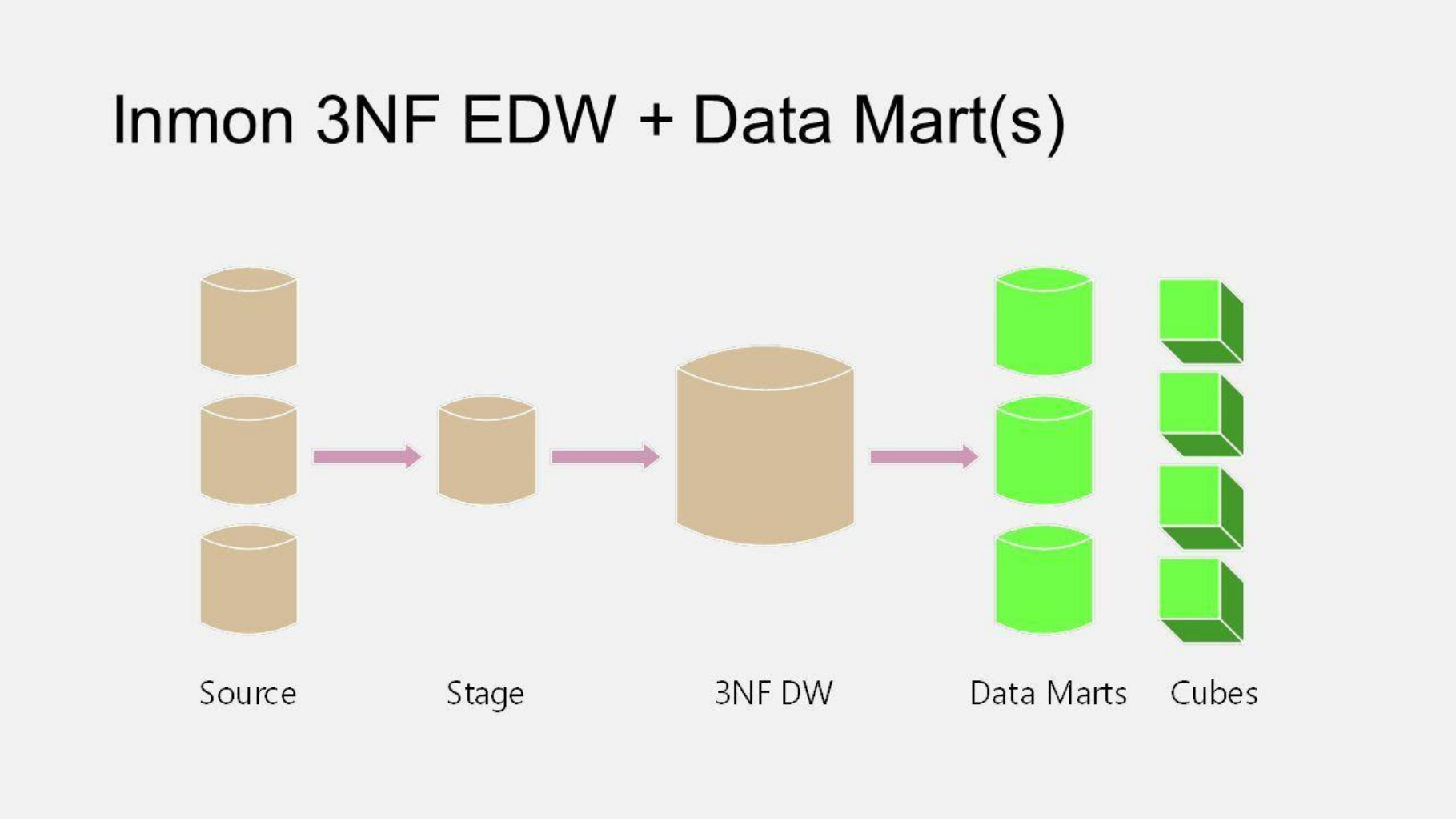
Kimball adattárház modellje – alulról fel megközelítés



Kimball adattárház modellje – alulról felfelé megközelítés

- A fókusz az adatpiacokon van
- Az adatpiacok tartalmaznak elemi és összegzett adatokat is
- Az adatpiacok csillag szerkezetűek
- Az architektúra legfontosabb részei a stage terület és az adatpiacok

Inmon adattárház modellje – fentről le megközelítés



Inmon adattárház modellje – fentről le megközelítés

- A fókusz az adattárházon (DW) van
- Az adattárház elemi adatokat tartalmaz normalizált formában
- Az adatpiacok összegzett adatokat tárolnak téma specifikus, dimenzionális modellben
- Az architektúra fontosabb rétegei a staging area, a DW, és az adatpiacok
- A felhasználó lekérdezhetnek akár az adatpiacokból, akár az adattárházból is

Kimball vs. Inmon

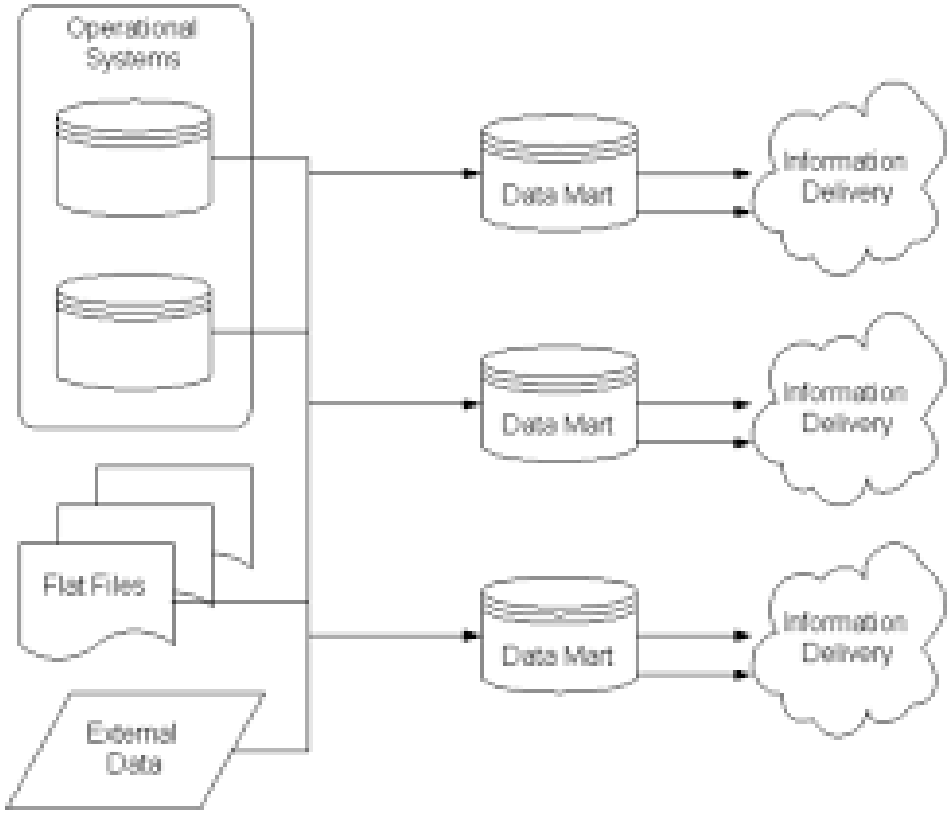
Kimball modellje a megfelelő, ha

- a felhasználók IT területről kerülnek ki
- inkább taktikai döntések szükségesek
- a forrásrendszerek viszonylag stabilak
- minél előbbi eredményt szeretnénk elérni, kis kezdeti befektetéssel és csapattal
- az adatok különálló üzleti területekről jönnek
- a változások köre limitált

Inmon modellje a megfelelő, ha

- a felhasználók nem IT szakemberek
- stratégiai döntések vannak túlsúlyban
- a forrásrendszerek gyakran változnak
- több idő, pénz és nagyobb létszámú csapat áll rendelkezésre
- vállalati szintű adatintegráció szükséges
- a változások köre bővülhet.

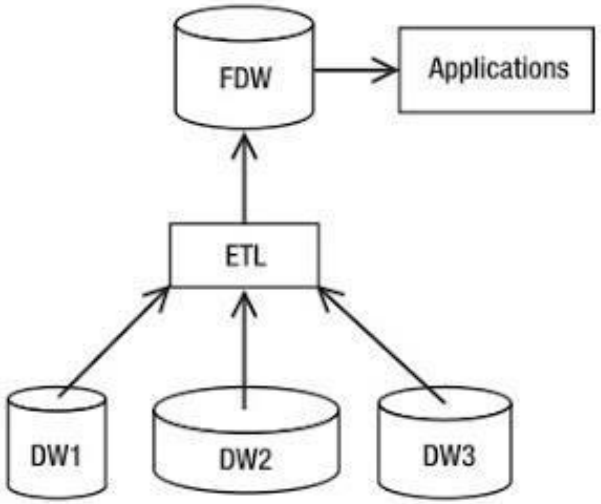
Alternatív adattárház architektúrák
Példa 1: független adatpiacok (independent data marts)



- Nincs központi adattárház
- Nincs központi jogosultság rendszer
- Az adatok közvetlenül a forrásrendszerekből érkeznek
- Biztonságosabb
- Kisebb méretű rendszereknél alkalmazható

A kép forrása: https://www.researchgate.net/figure/Data-flow-when-using-independent-data-marts_fig6_34267502

Alternatív adattárház architektúrák
Példa 2: egyesített adattárház (federated data warehouse)



A federated data warehouse from several data warehouses

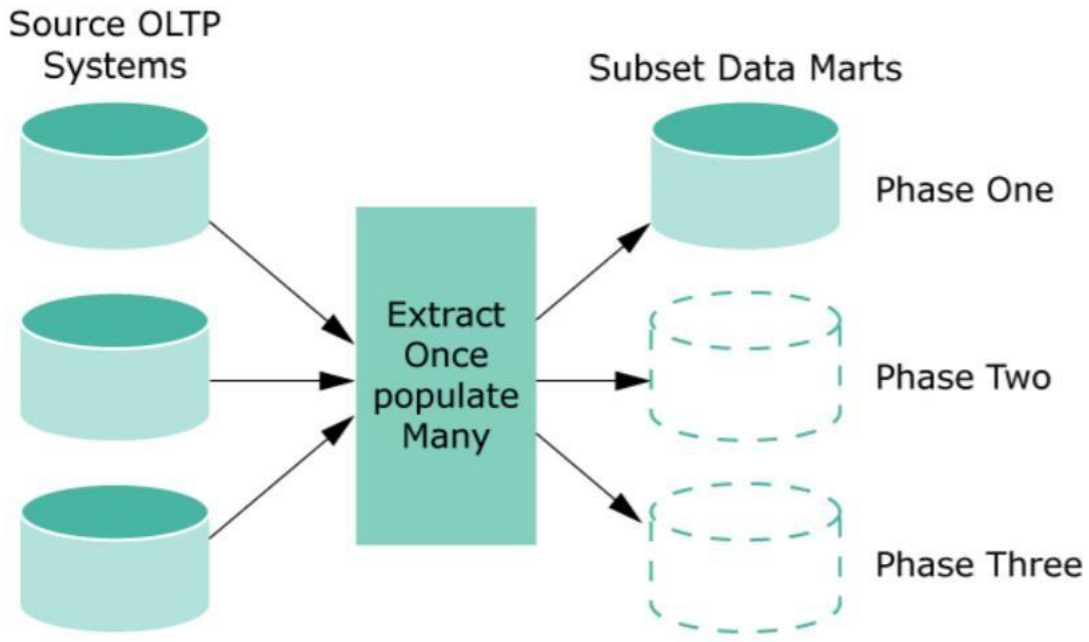
Több adattárházat egyesít

- Pl: nagyobb cég, telephelyek több országban, mindegyiknek saját adattárház
- Közös üzleti szabályok, egyszerű lekérdezés
- Nehéz technikai megvalósítás

A kép forrása: http://blog.umy.ac.id/yusufha/2016/09/21/2-data-warehouse-architecture/

Alternatív adattárház architektúrák

Példa 3: Inkrementálisan felépített adatpiacok (incremental architected data marts)



- Az adatpiacok több fázisban jönnek létre
- Egyszeri adatkinyerés, többszöri felhasználás

A kép forrása:
http://www.just.edu.jo/~mzali/courses/Fall14/Cis330/handouts/Successful%20Data%20Warehouses.pdf

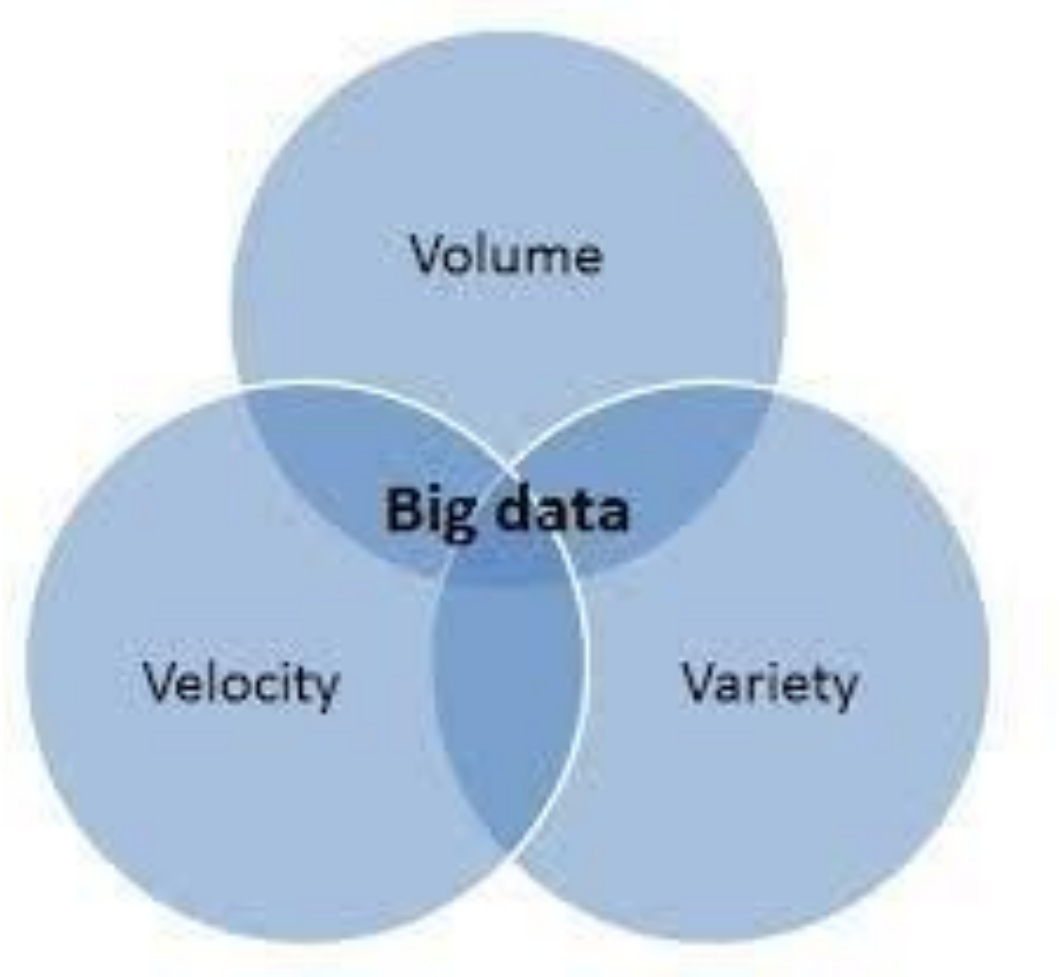
Adattárházak – új kihívások

- Növekvő adatmennyiség
- Adatfrissítés szinte azonnal
- Gyakori változások az adatmodellben
- Lekérdezések futtatása a memóriában

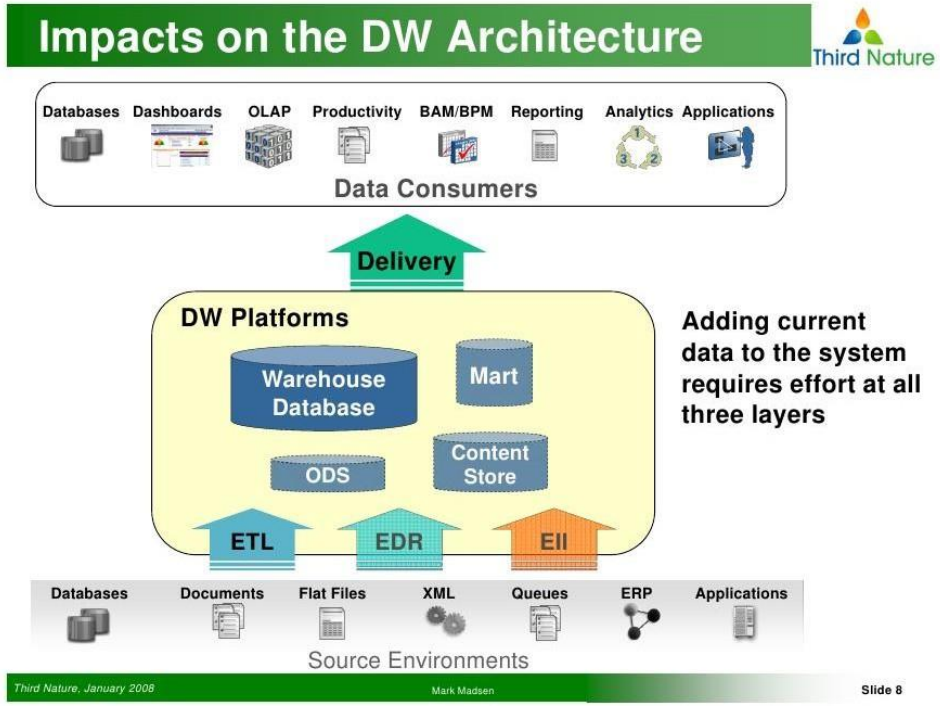
Egyre növekvő adatmennyiség - Big data

Két fő probléma

- ETL-folyamatok
- Számítások



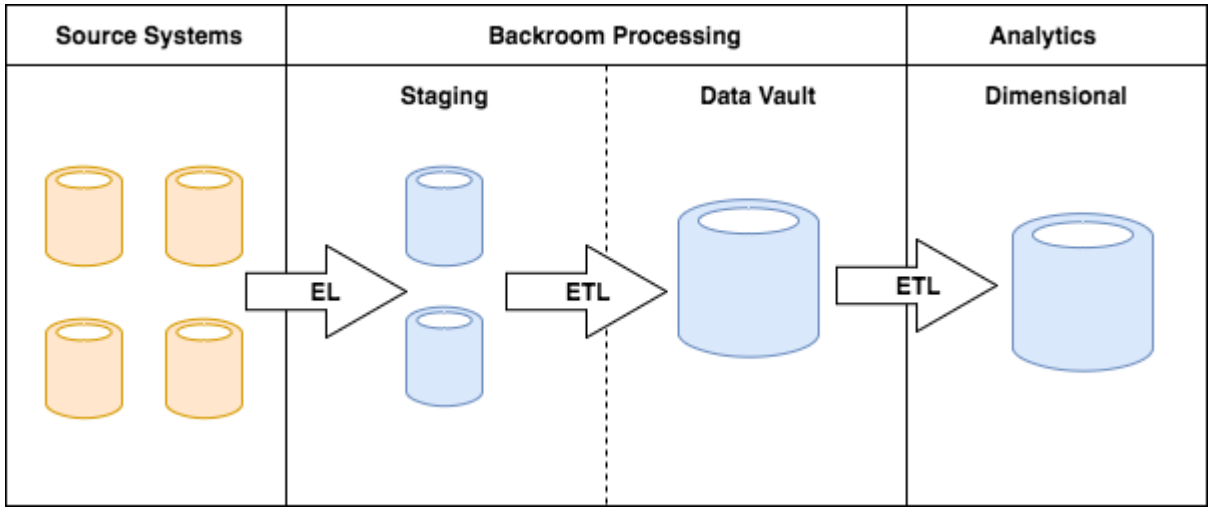
Gyakori adatfrissítés - Valós idejű adattárház



- Közel valós idejű adatok
- Gyors válaszidő
- Nagy számú felhasználói kérés
- Flexibilis, ad-hoc riportolási lehetőség

A kép forrása: https://www.slideshare.net/mrm0/how-real-time-data-changes-the-data-warehouse

Gyakori adatmodell változás - Data vault



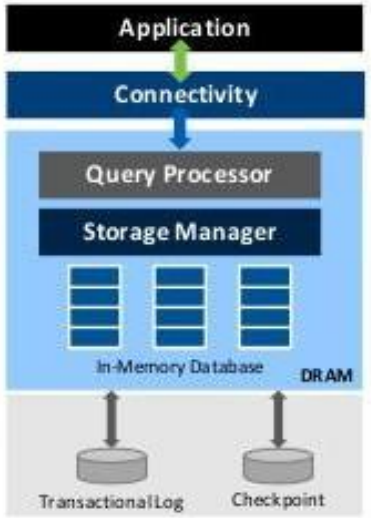
Az adatmodellt metamodell szintjén kezeli

- Speciális fogalmak
  - Üzleti kulcs (hub)
  - Üzleti kulcs tranzakció (link)
  - Üzleti kulcs történet (sat)

In-memory adatbázisok

+ In-Memory Database Technology

- Extremely Fast Transaction processing
  - Entire database resides in computer's memory
  - Powered by special algorithms and data structures that are highly optimized for in-memory computing
  - Hundreds of thousands of transactions per second
- Short and Predictable Response times
  - Optimized for fastest transactional processing with the shortest response times measured in microseconds
  - The improved response times fuel High Throughput.



A kép forrása: https://www.slideshare.net/Altibase/solve-big-data-problem-the-in-memorydatabase-solution-17179969

BUDAPESTI CORVINUS EGYETEM

Az előadás diák egy része, illetve azok alap ötlete dr. Kő Andrea: „Adattárházak áttekintés” előadásából származik

Köszönöm a figyelmet!