Információintegráció

Integrációs fokozatok: vállalati alkalmazások összekapcsolása, kommunikációszervezés üzleti folyamatok alapján, információintegráció. Az első kettő csak az operatív működést támogatja, a stratégiai döntések támogatásához információintegráció szükséges. Adatmennyiséghez, részletességhez, pontossághoz és elosztottsághoz tartozó problémák kezelése.

Az információintegráció folyamata

Adatgyűjtés, feldolgozás és vizualizáció. Döntések meghozatala és hatásuk visszacsatolása
Előkészítés → feljavított adatok az adattárházba

Döntéstámogató rendszerek

DSS áttekintése

Alternatív elnevezés: üzleti intelligencia. Különféle típusok: szakértői rendszerek, tudáskezelő rendszer, felsővezetői információs rendszer. Alapvető célok: a múltbeli adatok áttekintése, és összefüggéseik áttekintése, adöntéstől függő jövőbeli folyamatok modellezése, intuitív felületek, használhatóság, és egyszerű értelmezhetőség.

DSS rendszerek osztályozása

Bemenő adatok (kommunikációalapú, adatalapú, dokumentumalapú, tudásalapú, modellalapú döntéstámogatás), felhasználóval való kapcsolattartás (aktív, passzív, kooperáló), rendszer mérete(teljes vállalatot lefedő, egygépes) alapján

DSS rendszerek eszközei

Adatok begyűjtése: benchmarking és adattárház-ápítés. Feldolgozás: statisztikai jellegű elemzés, adatbányászat, analitika. Információmegjelenítés
Benchmarking: teljesítményfelmérés → vállalat működéséről mérhető információkat szerezni, ez alapján jobban működtetni.

Adattárház-építés

Az operatív adatok felhasználásának problémái

Az adatok heterogén adatforrásokban találhatók(elérésük, beszerzésük külön technológiát igényel), ezek egymással nincsenek összhangban. Az adatok elérésének teljesítményproblémái. Az operatív és a stratégiamegközelítés eltérő adatfeldolgozási módszert igényel. Az operatív rendszerek hibás adatait és hiányosságait kezelni kell.

Az üzleti intelligencia

Üzleti intelligencia: operatív és döntéstámogató rendszerek együttese. A döntéstámogatáshoz az alacsony absztrakciós szintű, heterogén, zajos adathalmazt alkalmasasbb formára alakítjuk, és az adattárházban tároljuk. Teljesítményelőnyök a szétválasztás, a csak olvasható elérés, az előkalkuláció, valamint az optimalizált tárolás és indexelés miatt. Adatok tematikus csoportosítása: függő és független adatpiacok. Adatok hasznosítása: jelentésalapú információfeldolgozás, analitikus feldolgozás, adatbányászat.
Adattárház: az összegzett és megtisztított adatok közös, témaorientált, rendezett és tartós tárolóhelye. Üzleti intelligencia: teljes folyamat az adattárházon át(ábra). Adatpiac: Tranzakciós rendszerekből nyert, összegzett és megtiszttott adatok közös, témaorientált, rendszerezett és tartós tárolóhelye, amely csak egy speciális felhasználói csoport kiszolgálását célozza meg.

Az adattárház-építés menete

- adatkinyerés
- adattisztítás
- átalakítás

- betöltés
- rendszeres frissítés

Adattárház-építés támogatása a relációs adatmodellben

Multidimenziós adatmodell

Tényadat: elemzések alapját képező adatok. Dimenzióadatok: tényadatokat leíró dimenziók, hierarchiába rendezve tároljuk. Multidimenzionális adatmodell leképezése relációs adatmodellre: csillagséma, hóehelyséma. Csillagséma: tényadatok központi táblába, amelyhez egy-egy idegen kulcson keresztül, csillagszerűen kapcsolódnak a dimenzióelemeket tartalmazó táblák. Hópehelyséma: az egyes dimenziók hierarchiaszintjei külön táblákba kerülnek, és közülük csak a legkisebb szinten levő kapcsolódik közvetlenül a tényadatokhoz.

Relációs adatok indexelése adattárházakban

Adatelemzés hatékonyaságának javítása indexeléssel: bitmapindexelés, összekapcsoló indexelés. Bitmapindexelés: szűk értékkészlettel rendelkező attribútumok; az indexben az attribútum értékkészletének minden eleméhez egy-egy bit kerül, amelyek közül minden rekordban csak az egyik, az attribútum értékének megfelelő bit kerül bekapcsolt állapotba. (Első negyedév: 1000, negyedik: 0001). Összekapcsoló indexelés: tény- és dimenziótáblák közözötti gyors kapcsolatteremtést segíti; A szokásos kulcsok úgy segítik a keresést, hogy az indexelt attribútum által felvett értékéhez egy elsődleges kulcsokból álló listát rendelnek(vagyis felsorolják, hogy melyik rekordokban található az adott attribútumérték)

Adatelőkészítési technikák

Integráció, adattisztítás, transzformáció és redukció

Alapfeladat	***Cél***	***A megoldáshoz használható módszerek***
Integráció	adatok központi kezelése és egységes elérése, leválasztás az operatív rendszerektől	adattárház-építés
Adattisztítás	a zaj, a redundancia és az inkonzisztencia eltávolítása, hiányzó adatok kezelése	zajszűrés: klaszterezés, osztályozás, regresszióa, szótárak, ismétlődések szűrése, hiánykezelés: törlés, modellezés, klaszterezés, átlagolás
Transzformáció	adatok illesztése a későbbi a későbbi feldolgozási lépésekhez (forma és teljesítmény)	transzformációs táblázatok, áttkódolás, normalizálás, konstrukció (új számított mezők bevezetése
Redukció	adatmennyiség és diverzitás csökkentése az információk megőrzését szem előtt tartva	dimenzió- (attribúzumszám) csökkentés, tömörítés, rekordszámosság-csökkentés klaszterezéssel és modellek alapján, általánosítás, összegzés

Adattisztítás

A hiányzó értékek kezelése

A hiányos rekordok eldobása vagy korlátozott felhasználása. Hiány pótlása: helyettesítés alapértelmezett, átlagos, tipikus vagy legvalószínűbb értékkel. Helyettesítés más hasonló rekordok adatai alapján vagy csoportokba sorolás után a csoport reprezentatív értékével

Zaj-, redundancia- és inkonzisztenciaszűrés

Duplikátumok helettesítése egy elemmel, egymásnak és a kényszereknek ellentmondó rekordok eltávolítása. Szélsőséges értékek kiszűrése. Adatsimítás kosarazással, klaszterezéssel, regresszióval. Szöveges attribútumok zajszűrése klaszterezéssel és szótárakkal

A kényszereket be nem tartó, üzleti logikának ellentmondó adatok eltávolítása.

Adatsimítás: ***kosarazás, klaszterezés, modellillesztés, szövegre szótáralapú is*** ; az adathatlmaz lokális tulajdonsága alapján becsüljük a zajt, majd új, reprezentatívabb értékkel helyettesítjük a zajjal terhelt adatokat. Kosarazás: a zajosnak vélt attribútum szerint rendezzük az adatokat, az eredményt azonos méretű csoportokra bontjuk, ezután az egy kosárba került elemek adott attribútumát egy, a kosárra jellemző értékkel helyettesítjük. Klaszterezés: kosarazás, csak nem rögzítettek a kosarak mérete. Modellillesztés: az adatsorunkat egy vagy több másik változó függvényében elemezzük, és az attribútumértékek helyettísére a szabály alapján leképzett modellt használjuk. Szövegre működik az előbbiek mindegyike, plusz Szótáralapú módszer: explicit módon felsoroljuk a leggyakoribb helyes és hibás formákat.

Az attribútumok transzformálása

Szótárak, algoritmikus transzformációk, attribútum-összevonások és -szétválasztások, normalizálás, diszkretizálás

Normalizálás

Lineáris min-max normalizálás, standard normalizálás

Miért? → Ne torzítsa el a széles értékkészletű dimenzió az adatbányászati elemzést. Lényege, hogy egy fix tartományba transzformálj az adatokat. Lineáris min-max: xi' = (xi - min(X))/(max(X) - min(X)) ; standard normalizálás: noirmál eloszlásos handzsár, "az attribútum empirikus várható értékét nullává változtatja, míg az empirikus szórása 1 lesz

Diszkretizálás

Kvantálás azonos méretű és azonos elemszámot tartalmazó értékintervallumokkal. Hisztogramok

Kvantálás: kategóriákat alakítunk ki, ezekhez alsó és felső határokat rendelünk az elemek besorolásához. (cm helyett alacsony, magas..). Hisztogram: az adattartományok, illetve a hozzá tartozó elemek relatív gyakoriságának szemléltetése.

Adatredukciós eljárások

Tárhely, adatelem- és attribútumszámosság csökkentő eljárások. Memórián kiívüli feldolgozás

Ha túl nagy adathalmazzal kell dolgozni, azt "memórián kívüli" feldolgozásnak hívjuk (nem fér be a memóriába).

A tárhely csökkentését szolgáló eljárások

Duplikátumok törlése, rekordok összegzése. Átkódolás, veszteséges és veszteségmentes tömörítés

A feldolgozandó elemek számosságának csökkentése

Mintavételezés, hasonló elemek csoportosítása és egyesítése, általánosítás

Mintavételezés: az eredeti adathalmazból egy, a memóriában elférő mintát veszünk, majd ezen végezzük el az elemzést.

A feldolgozandó elemek méretének csökkentése

Dimenziók számának csökkentése az attribútumok fontossága és korreláltsága alapján. Faktor- és főkomponens analízis

Faktor-(SVD) és főkomponens(PCA) analízis: meglévő dimenziókból új dimenziókat származtatnak, amelyek korrelálatlansága sokkal nagyobb, mint az eredeti dimenzióké volt.