Mi a IBM QualityStage alkalmazás fő célja?

Az alkalmazás fő célja sokféle adatintegrációs (-elemzési, -tisztítási) eszköz integrálása egy egységes tervező és futtató rendszerben, közös metaadat rétegre építve. Az egységes keretrendszer és a megosztott információleírások segítségével az adattisztítási folyamat korábban különálló lépései egy közös rendszerbe integrálhatók. Az egyes lépések eredményei egymás számára azonnal láthatóvá válnak, valamint egységes szolgáltatásokkal (adatelérés, párhuzamos futtatás, naplózás, megjelenítés, stb.) dolgozhatunk a különböző lépésekben.

A rendszer másik célja, hogy a korábban különálló módon létező, különféle felületekkel ellátott (gyakran különböző gyártók által előállított) eszközöket egységes felhasználó felületbe foglalja, ezáltal lehetővé téve, hogy a fejlesztők könnyen mozoghassanak az eszközök között. Ez az egységes felület háromféle aktivitás köré szerveződik: elemzés, fejlesztés és adminisztráció.

Az elemzés alapvető célja a rendelkezésükre álló adatok mélyebb megismerése annak érdekében, hogy a feldolgozási, tisztítási folyamatukat megtervezhessük. A fejlesztőeszköz segítségével egy grafikus környezetben építhetjük fel az adattisztítási folyamatot. Az adminisztratív interfész segítségével alapvetően a munkáink futását állíthatjuk be illetve követhetjük nyomon.

Mi az adattisztítás célja? Mondjon példát rá!

Illeszkedve a forrás és alkalmazói rendszerekhez, biztosítsa az adatforrásokból származó adatok egységesítését, megfelelő minőségi szintre hozását az azokat felhasználó rendszerek számára.

Az IBM WebSphere QualityStage egy olyan adattisztító rendszer, amely abban segíti az üzleti rendszerek alkalmazóit és fejlesztőit, hogy egységesítsék, megtisztítsák és megfelelő minőségi szintre hozzák a rendszerükben tárolt adatokat. A rendszer képes a forrásrendszerek adatainak vizsgálatára (investigate), egységesítésére (standardize), illesztésére (match) és az illeszkedő adatok összeolvasztására (survivorship).

példa:

távközlési szolgáltató

Milyen főbb kliens alkalmazásokból áll az IBM QualityStage?

A teljes rendszer az alábbi három összetevőből áll: kliensek, futtatógépek, és megosztott adathalmaz.

Kliensek:

- **Administrator kliens:** Az adminisztrátor segítségével általános szerver beállításokat adhatunk meg, létrehozhatunk és törölhetünk projekteket, valamint beállíthatjuk azok tulajdonságait.
- **Director kliens:** Segítségével ellenőrizhetjük és futtathatjuk a munkákat, követhetjük a működésüket, statisztikákat gyűjthetünk.
- **Designer kliens:** Segítségével egy grafikus felületen állíthatjuk össze a munkákat.

Mik az adattisztítás lépései?

A minőségi adatok biztosítása négy lépésben történik:

- **Vizsgálat (investigate):** a beérkező információ teljes megértése
- **Standardizálás:** az információ teljes tisztítása
- **Illesztés (matching):** az összetartozó információk összekapcsolása
- **Túlélés (survivorship):** az összefüggő információk legjobb egységesített nézetének előállítása

Mi történik a vizsgálati lépésben?

(Investigate stage)

Az adatok megértése fontos eleme azok tisztításának. Az Information Analyzer vagy a vizsgálati lépés közvetlen adatokkal szolgálhat a további lépések számára az adatok minőségéről.

A vizsgálati lépések kimenete (minta riport) megjeleníti azon adatok számát illetve százalékos arányát, amelyek az adott mintára illeszkednek, valamint adatmintákat is kapunk az elemzésekből.

Mi történik a standardizálási lépésben?

A vizsgálati lépésben feltárt jellemzők alapján előre gyártott szabályok segítségével átformázhatjuk a bejövő adatokat. Ez a lépés mintaillesztéssel és kimeneti formázással valósul meg.

Példa bejövő „piszkos” adatokra:
Ebes, Fő u. 2
4330 Ebes, Kertsor
Malompark
Nagyvigad, Kapa u. 12.
Átalakításuk után egységes irányítószám:város:közterület:házszám alakot kapnak.

Mi az illesztési lépés feladata?

Az adatillesztés megkeresi az adatforrásokból származó rekordok között azokat, amelyek ugyanarra az entitásra (emberre, címre, szervezetre, helyre, termékre, stb.) vonatkoznak. Ez az illesztés még akkor is megvalósítható, ha nincs előre meghatározott kulcs, amely az egyértelmű összerendelést leírná. Az adatokat bármilyen kapcsolat mentén összekapcsolhatjuk, mint például egy közös ember, cég, hely, időpont, stb.

Az adatillesztés a következő képességekkel rendelkezik:

- Duplikátumok azonosítása egy vagy több adatforrásban.
- Egységesített nézetek készítése az üzleti szabályoknak megfelelően.
- Az azonos fizikai helyhez kötött egyedek (emberek, üzleti entitások) felismerése és kezelése (householding).
- Illesztési csoportok készítésének lehetősége előre meghatározott közös kulcsok esetében, illetve azok hiányában.
- Létező adatok gazdagítása külső forrásokból származó illeszkedő új attribútumokkal.

Az illesztés egy kétlépéses folyamat: először blokkokra vágjuk az adatokat, majd ezeket a blokkokat illesztjük egymással. A blokkosítás az adatok között olyan részhalmazokat határoz meg, amelyek hatékonyan illeszthetők. Az illesztés lehet referencia meghatározás vagy duplikátumok jelölése. Az előbbi a rekordok közötti kapcsolatokat térképezi fel, míg az utóbbi a hasonló (potenciálisan törlésre jelölhető) rekordokat csoportosítja egy adott adatforrásban.

Mi történik a túlélés lépésben?

(Survive stage)

Ez a lépés összevonja a duplikátumokat és elkészíti a legjobb közös reprezentációját az illeszkedő adatoknak.

Lépései:

- Egy adott rekordban hiányzó adatok pótlása ugyanannak az entitásnak más rekordokból származó adataival.
- Hiányzó adatok pótlása olyan rekordokból származó értékekkel, amelyeket az illesztés ugyanabba a csoportba tartozónak jelölt.
- A meglevő adatok gazdagítása külső forrásokból.