Az első két hét eredménye

* Minta applikáció készítése

A munkafolyamat során a következő linket használtam:

<https://help.sap.com/viewer/4505d0bdaf4948449b7f7379d24d0f0d/2.0.03/en-US/4f0a1567616541cbb7bd338fb1b64c35.html>

Ez egy egyszerű TinyWorld alkalmazás, ami ebben a formában még kicsi és használhatatlan, de betanulásnak szerintem megfelelő. Ehhez hozzáadtam egy adatbázis, egy NodeJS és egy HTML5 modult a leírásnak megfelelően, ezek fordultak, és sikeresen futtattam is őket. A következő link ennek kibővítésére vonatkozik, amit megpróbáltam megcsinálni, de a kalkulációs nézetnél elakadt a dolog, mert nem volt semmilyen arra utaló jel, hogy ki lehetett volna javítani a “no measure defined” típusú hibát.

* A kapott pivot alkalmazás adatbázisának kialakítása

A normális szerver kialakulása után az adattáblát először importáltam a HANA rendszerbe, ahol az adatbázis számos hibájára bukkantam rá. Rögtön felfedeztem két olyan sort, amely nem felelt meg a kialakított sémának és formátumnak, ezeket azonnal töröltem. Ezek közül a legfontosabbak a társaságok neveinek különböző módon történő gépelése (akár direkt más módon leírva, akár elgépelve) és a címek nagyfokú inkonzisztenciája. Ettől függetlenül az adattábla egyértelmű kód-érték mezői nagyjából megfelelőek voltak, itt csak egy-egy módosítást voltam kénytelen végrehajtani az adattáblában (például az azonos településkódhoz tartozó Dunavecse és DunavecsDunavecse egységesítve lett Dunavecse néven). A társaságok rövid és hosszú neveinek elméletben bijektívnek kellene lennie, a táblában ez a tulajdonság sajnálatos módon nagyon sok adatnál nem állt fenn, ezért a táblát az újraimportálás előtt több ízben is módosítottam. A rövid neveknél egységesen eltávolítottam a kezdő újsor karaktert, a hosszan kiírt gazdasági formára utaló részt pedig egységesen felváltottam annak rövidítésével, eltávolítottam a dupla szóközöket és pontokat, a cég állapotára vonatkozó részt eltávolítottam, a név végén lévő felesleges szóközöket kitöröltem, továbbá bizonyos adatokon kézi beavatkozást is végeztem. Hasonló módosításokat végeztem a hosszú neveknél is, de ott a rövidítések helyett a teljes gazdasági formát írtam ki, ahol kellett. Ezeken felül mindkét mező értékeit nagybetűssé alakítottam, hogy az átformázást egyszerűbben tudjam végrehajtani.

Ezután a nagy táblát újraimportáltam, de a kódtáblákat módosítani már nem kellett. A rövid és hosszú neveket az adószám segítségével bijektívvé tettem, méghozzá úgy, hogy az azonos rövid név-adószám párhoz tartozó hosszú nevekből egyet tartottam meg, a többit ezzel írtam felül, hasonlóan a hosszú név-adószám párokkal is ugyanezt tettem. A címekből is csak egyet tartottam meg az egyes cégekhez, ugyanis ezek formátuma nagyon inkonzisztens volt az adatbázist tekintve. Emiatt kénytelen voltam törölni azokat a sorokat, ahol a település és a címadat nem volt kompatibilis egymással, ezzel nagyjából harminc rekordot vesztettem. Végül kiszedtem egy külön táblába a társaságok rövid és hosszú neveit, az adószámot és a címet, majd létrehoztam a kapcsolótáblát, a jegyzett tőke ebbe lett elhelyezve, mint egyetlen nem kulcs adat. Végül ezekből a táblákból előállítottam a teljes tábla résztáblákból előálló változatát, az előző módosítások végrehajtása miatt kb. 2% adatvesztéssel. Az eredmény tíz kódtábla, egy névtábla, egy kapcsolótábla és az ezekből előálló teljes tábla.

Ezek után próbáltam újragondolni a tábla normalizálását főként az elveszett rekordok, illetve a név és címadatok miatt. Két kisebb javítási lehetőséget felfedeztem, miszerint a település egyértelműen meghatározza a megyét és az országot, a megye egyértelműen meghatározza a régiót. Ez mínusz három oszlop a kapcsolótáblában, ráadásul így összességében kevesebb helyen tárolunk ugyanannyi adatot. Hasonló módon a nemzetgazdasági szakágazat egyértelműen meghatározza az ágazatot, illetve az ágazat az ágat, ez további mínusz kettő oszlop a kapcsolótáblában. A név és címadatoknál azonban nem ilyen egyszerű a helyzet, szerény véleményem szerint a társaságok rövid és hosszú nevének továbbra is bijektívnek kellene lennie, továbbá a rövid név, az adószám és az adatszolgáltatás éve egyértelműen meghatározza a címet és a jegyzett tőkét is. Ennek kialakítása azonban nem sikerülhetett, ugyanis a kialakításhoz komolyabb adattisztítást kellett volna kialakítani, mint amire képes lettem volna.

* Egyéb feladatok a pivot alkalmazásban
  + Az alkalmazás geokódot előállító részében többszöri alkalommal nyújtottam kisebb segítséget.
  + A geokódot előállító JAVA alkalmazás JSON formátumban várja az adattömeget, ami során felmerült egy probléma, miszerint az általunk ismert konvertálóprogramokkal csak a táblázat első sora alakítható át JSON-be. Így igény merült fel arra, hogy a konvertálóprogramot mi magunk írjuk meg. Több órányi munka után sikerült létrehoznom egy C++ programot, amellyel azt tudjuk elérni, hogy a táblázat txt-be történő exportálása után abból valid JSON-t tudunk létrehozni. Ekkor egy újabb probléma merült fel, miszerint a JAVA alkalmazás már az első rekord minden mezőjénél null értékekként olvassa be az adatokat, sajnos ezzel már nem tudtunk mit kezdeni.
  + A HANA rendszerben erre a táblázatra sikerült létrehozni egy kalkulációs nézetet, és megismerkedtem ennek alapjaival. Ugyanakkor itt is rábukkantam egy-két érdekes anomáliára. Számos esetben a rendszer rosszkor használja a distinct kulcsszót, amely miatt több megjelenítés is invalid adatokon alapul.