Akademia Górniczo-Hutnicza
Wydział Elektrotechniki, Automatyki, Informatyki i Inżynieri Biomedycznej
Kierunek Informatyka

Hurtownie danych

Projekt hurtowni danych

...

Patryk Gałczyński

Kraków, 8 maja 2017

1 Analiza wymagań systemu:

Ponieważ z założenia hurtownia danych ma wspomagać podejmowanie decyzji biznesowych, oraz ułatwiać analizy w tym wypadku sprzedażowe, zdecydowano się na **schemat hybrydowy**, aby zapewnić dobry kompromis pomiędzy wydajnością a łatwością tworzenia zapytań przez analityków biznesowych. Jest to o tyle wygodne rozwiązanie, iż w trakcie dewelopowania złożonych aplikacji klienckich, analitycy będą już mogli korzystać z dostępu do ustrukturyzowanych danych.

2 Źródła danych - input datasources

W tej hurtowni danych, **na wczesnym etapie rozwoju**, głównym źródłem danych mają być **arkusze xls** dostarczane co pewien okres. Aby dostosować to źródło danych, do schematu tej hurtownii powstanie specjalny **driver etl**.

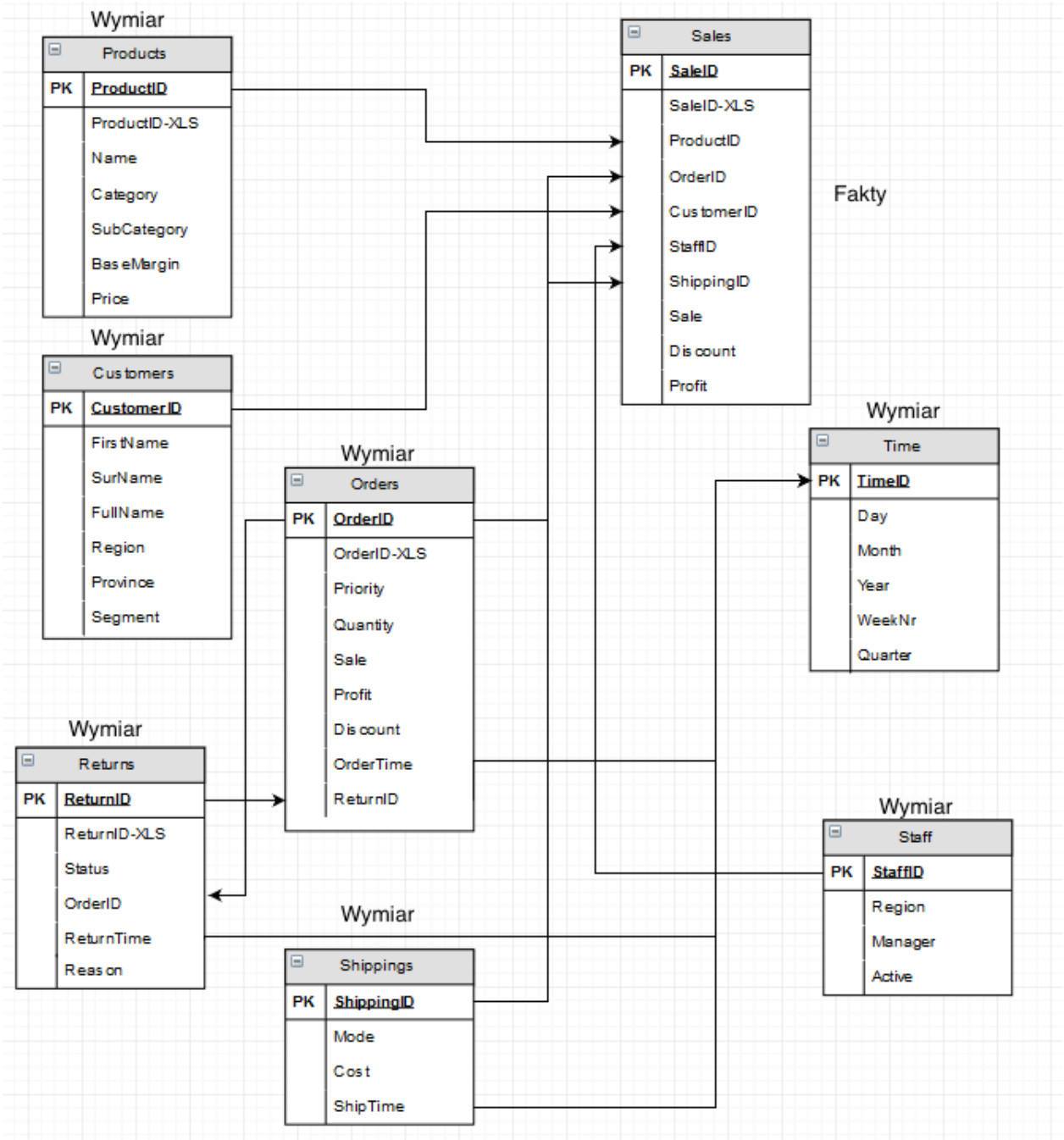
	A	B	C	
1	Row ID	Order ID	Order Date	Order Priority
2	1	3	10/13/2010	Low
3	49	293	10/1/2012	High
4	50	293	10/1/2012	High
5	80	483	7/10/2011	High
6	85	515	8/28/2010	Not Specified
7	86	515	8/28/2010	Not Specified
8	97	613	6/17/2011	High
9	98	613	6/17/2011	High
10	103	643	3/24/2011	High
11	107	678	2/26/2010	Low
12	127	807	11/23/2010	Medium
13	128	807	11/23/2010	Medium
14	134	868	6/8/2012	Not Specified
15	135	868	6/8/2012	Not Specified
16	149	933	8/4/2012	Not Specified
17	160	995	5/30/2011	Medium

Same pliki xls wydają się być średnio-praktycznym źródłem danych, gdyż pewnie ktoś generuje je ręcznie (sic!), w związku z czym niżej przedstawiam przykładowe, bardziej praktyczne (z mojego punktu widzenia, sensowne ze względu na domniemaną działalność wywnioskowaną z danych z pliku xls) źródła danych:

- system fakturowy jak np. popularny *Comarch ERP Optima*
- dzienny dump danych z systemów sprzedaży online przez api, jak np shopify
- dane o wysyłkach z api popularnych firm przewozowych typu UPS

I wiele innych, z zaznaczeniem, że do każdego źródła trzeba by napisać driver etl.

3 Schemat hurtownii

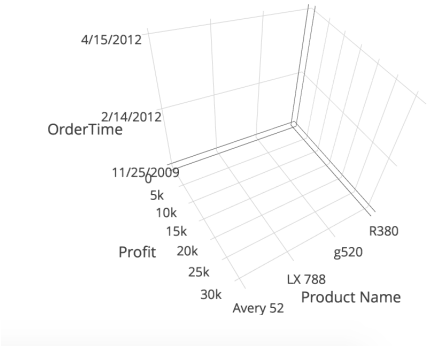


Schemat został zaprojektowany w oparciu o pola w pliku xls, w założeniu dodania nowych źródeł, mógłby on swobodnie ewoluować. Typy pól nie są explicite podane na schemacie, gdyż dostosowanie konkretnych typów zazwyczaj "wychodzi w praniu"przy tworzeniu konkretnej schemy dla konkretnego silnika bazy danych.

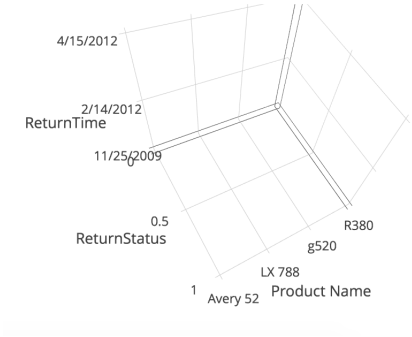
4 Propozycja kostek OLAP

Ponieważ nie byłem w stanie znaleźć dobrego tool'a generującego wizualizację kostki w raz z opisem, poniżej przedstawiam słowny opis przykładowych kostek stworzonych na podstawie powyższego schematu:

1. ProductName, Profit, OrderTime - np. Jakie produkty przynoszą największe zyski w danym kwartale



2. ProductName, Return status, ReturnTime - np. Jakie produkty zwracają się najczęściej w danym okresie



3. ProductName, OrderQuantity, Region - np. Których produktów jest kupowane najwięcej w danym regionie
4. OrderId, StaffId, ReturnStatus - np. Zamówienia obsługiwane przez obsługę z którego regionu są zwracane najczęściej
5. ShippingCost, ShippingTime, Region - Wysyłanie do którego regionu jest najbardziej optymalne, tzn. najmniejszy koszt oraz najkrócej idzie

itd..