Power BI - Datové modelování

Datové modelování

- Základní proces analýzy dat
- Pomocí diagramu se vizuálně vyjádří vztah mezi jednotlivými daty, která jsou pro analýzu shromažďována.
- Správný datový model zjednodušuje práci s daty a umožňuje dosáhnout správných výsledků
- Zlepšuje se performance reportu
- Je možné se vyvarovat složitým DAX funkcím, které by mohly report zpomalovat nebo vracet nejednoznačné výsledky
- Model je udržitelný a snadno rozšiřitelný o další data (tabulky)
- Často jde o úplně první krok v analýze a stačí nám k němu tužka a papír
- Datový model musí být jednoznačný

Vazby

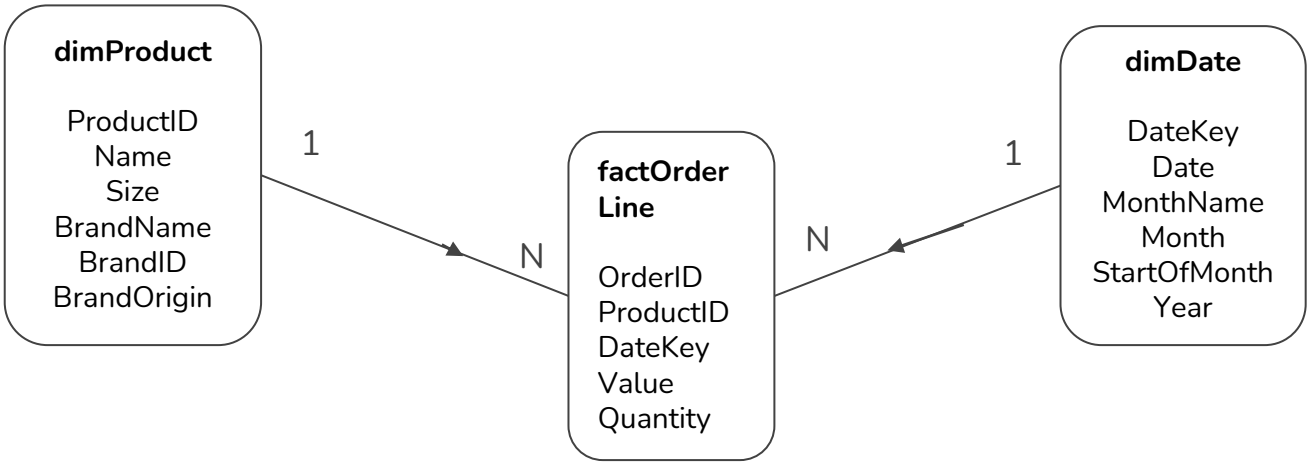
- Rozlišujeme dva typy tabulek
  dimenzní (číselníky), vždy by měly mít unikátní identifikátor pro každý řádek
  faktové
- Typy vazeb
  1:1 (one to one) každá položka je v tabulce právě jednou (produkt, ceník produktů)
  1:N / **1:*** (one to many) každá položka může mít v tabulce více výskytů (produkt, ceník produktů v jednotlivých letech)
  N:N / ***:*** (many to many) Jedna nebo více položek může mít v tabulce více výskytů (tabulka knih v knihovně s autory, která bude mít vazbu na tabulku autorů a jejich knih - autor má více výskytů, protože napsal více knih a zároveň jsou knihy, které může napsat více autorů)
- Směr vazeb
  Jednostranná Jedna tabulka filtruje jednotlivé záznamy v druhé tabulce (typicky dimenzní tabulka filtruje dané záznamy z factové tabulky, druhá tabulka pak ale už nefiltruje tabulku první
  Oboustranná obě tabulky se filtrují navzájem, filtrování vždy probíhá pomocí sloupců přes které je vazba vytvořená

Denormalizace a Normalizace

- Denormalizace i Normalizace jsou postupy, které se používají při vytváření různých modelů
- **Denormalizace -** přidávání redundantních (opakujících se dat) do tabulek v modelu pro účely rychlejšího a jasnějšího prohlížení dat
  (v databázích může jít o materializované tabulky, nebo pohledy (views), které usnadňují analytickou práci tak, aby nebylo nutné na sebe pokaždé napojovat různé tabulky).
- Denormalizovaný model by tedy mohl vypadat jako jedna tabulka, to má ale určité nevýhody
- **Normalizace** je pak proces, kdy naopak tvoříme malé tabulky, snižujeme redundaci dat a zlepšujeme jejich integritu a udržitelnost. Využíváme především pokud máme velké objemy faktových dat.
- Většinou používáme oboje a záleží na jednotlivých případech

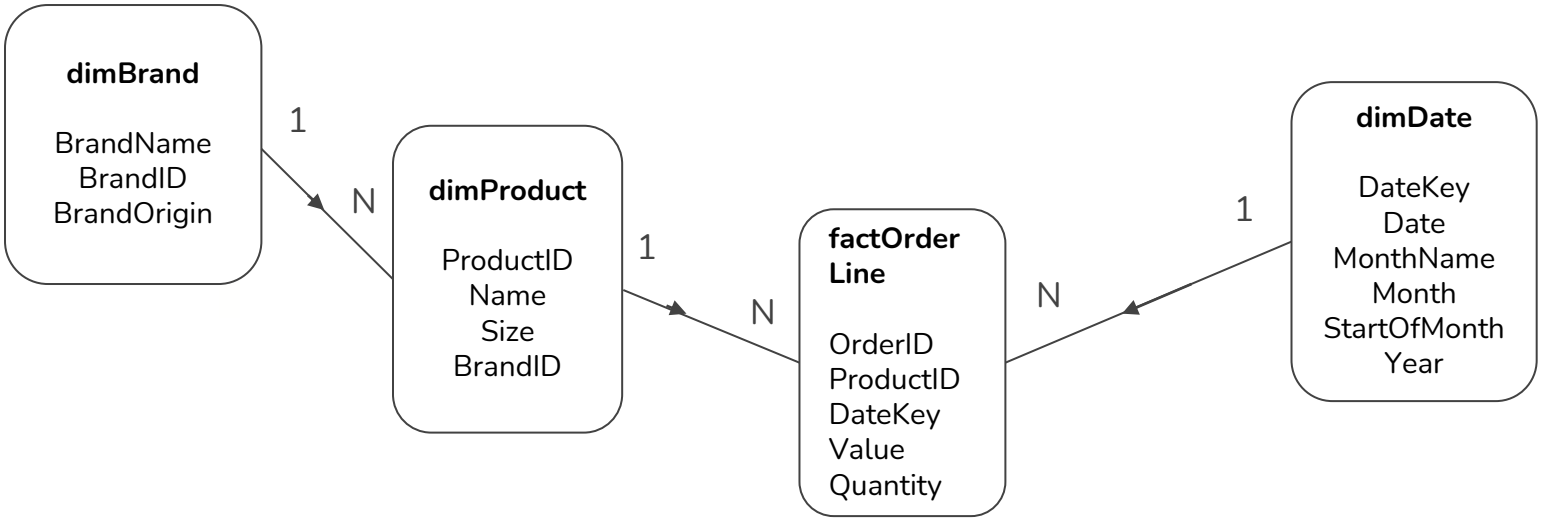
Typy datových modelů - Star Schema

- jde o best practice model, který zajišťuje nejoptimálnější řešení. Pokud to je možné, chceme vždy vytvořit star schema. Model se snadno udržuje a rozšiřuje. Je to model na který bylo Power BI optimalizováno



Typy datových modelů - Snowflake Schema

- Snowflake vzniká normalizací star schema, pokud není nezbytně nutné, tak je vždy na uvážení, jestli je snowflake schéma potřeba
- Kdy je vhodné vytvářet nové dimenze?



Typy Datových modelů

- **Více star schémat - často máme více faktových tabulek a více různých modelů v jednom velkém datovém modelu.**

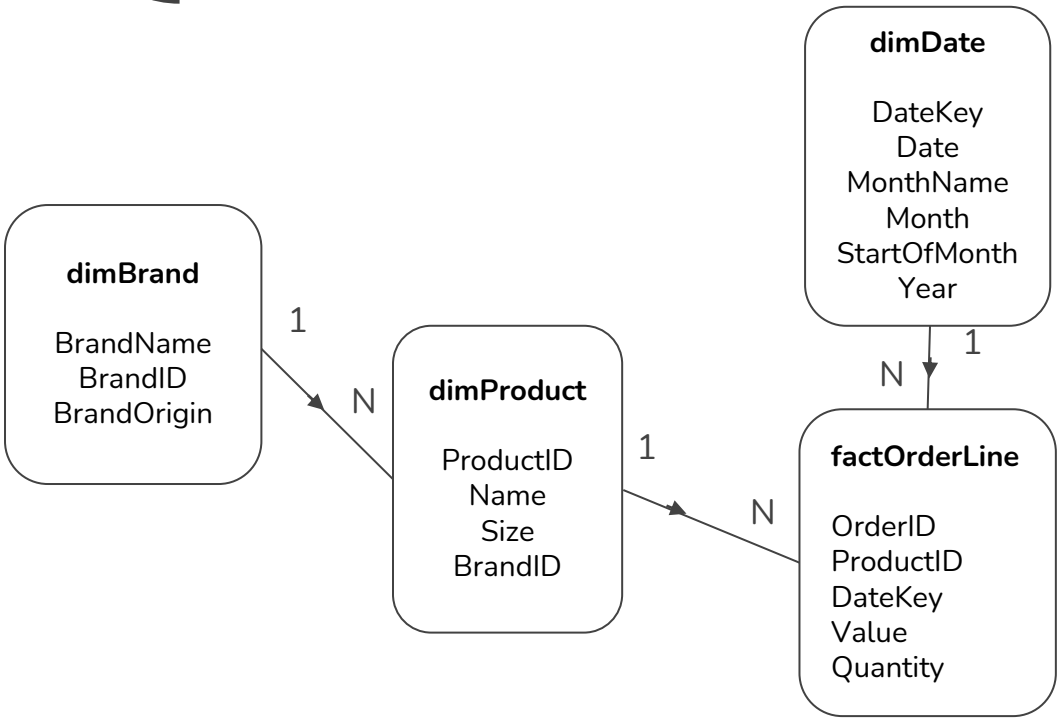
- **Header - Detail datový model**
  Umožňuje při zachování správného detailu pracovat s hodnotami tak, aby se nepřepočítávaly, ale ukazovaly se jen relevantní, (Objednávka s hodnotou dopravy / Objednávka s hodnotou produktů - ke každému produktu zobrazím celkovou cenu dopravy, ale i k objednávce jsem schopen zobrazit celkovou cenu dopravu (nesčítám přes produkty))

- Více vazeb mezi dvěmi tabulkami
  Aktivní a neaktivní vazba
  Vazby se dají vytvářet i dynamicky (pouštět) pomocí DAX funkcí, ale vždy musí existovat

- **Ambiguity** = Mnohoznačný model, více cest jak filtrovat tabulku

- Praktický tip - možnost skrývání sloupců v datovém modelu

Typy datových modelů - Header Detail



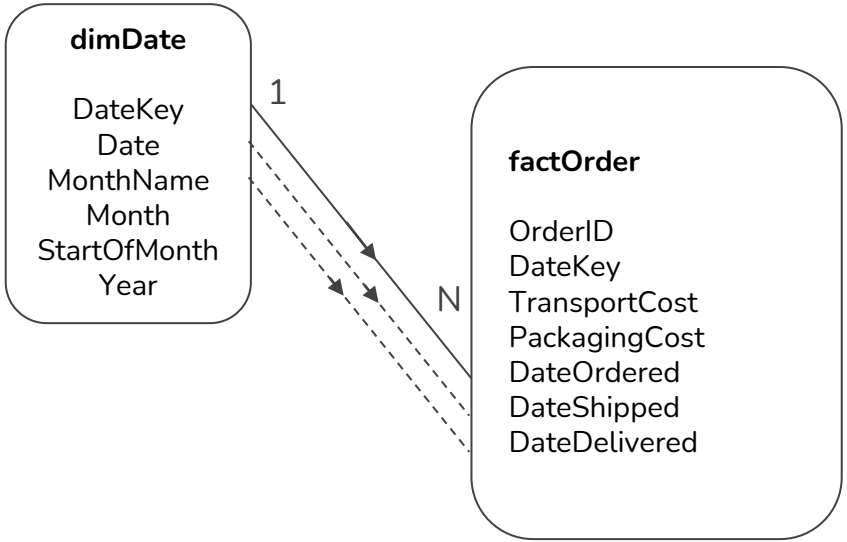
factOrder

OrderID
DateKey
TransportCost
PackagingCost

Oboustranná vazba mezi OrderID

Typy datových modelů – Více vazeb

- neaktivní vazby jdou pouštět pomocí DAX kalkulace - USERELATIONSHIP, umožní napočítat metriky Count of Shipped Orders, Count of Delivered Orders bez nutnosti mít více datumových dimenzí a více filtrů.



Power BI -
ETL a Čištění dat

ETL (Extract Transform Load)

- Jde o proces, kdy data kombinujeme z různých zdrojů, data očišťujeme, upravujeme a organizujeme do jednoho konzistentního datového zdroje (data warehouse, data lake, ...)

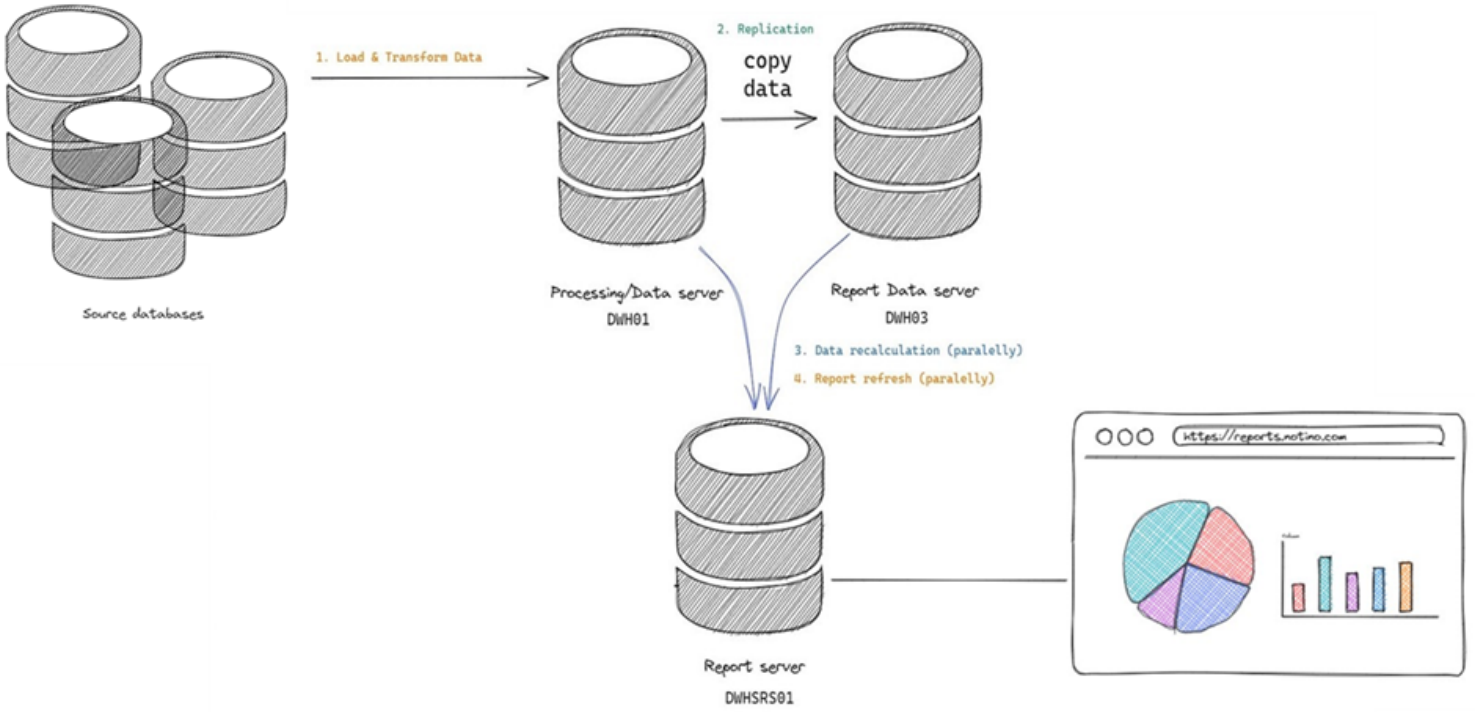
- Většinou ETL proces vzniká na samém začátku a pro účely vizualizací (reportingu) se pracuje s očištěnými daty. Někdy je však potřeba data upravit pro specifické účely vizualizací a nebo se Power BI používá jako prototypovací případně testovací nástroj.

- Úpravu dat je nejlepší provádět co nejblíže zdroji (například pokud opakovaně v různých reportech musíme data upravovat stejným způsobem, dává smysl aby se tato úprava udělala už někde blíže zdroji.)

Typy operací

- Nastavují se správné datové typy
- Vynechání nepotřebných sloupců, duplicit, chybových záznamů
- Měníme a sjednocujeme názvy
- Přidávání sloupců, díky kterým dokážeme data propojit mezi sebou
- kontroluje se datová kvalita

Příklad z praxe



Power Query

- umožňuje automatizaci procesu čištění a transformace dat.
- krok za krokem provádí operace, které následně při refreshi dat opakuje
- Používá jazyk M
- Power Query je i v excelu