NoSQL L3

Del 1 (*Big Data*)

Genomför och besvara följande frågor:

Data Warehouse	ETL
Data Lake	ELT

1. Parar ihop ”datalagrings sätt” med rätt strategi (ETL, ELT):
   (placera in ETL och ELT på passande rad)

2. Vad står förkortningen ETL för och vad innebär det i sammanhanget vi befinner oss i?
ETL står för Extract, Transform, Load. Det är en process för att överföra data från en eller flera källor till en data warehouse.

3. I ett data analys scenario så kan så mycket som 80% av tiden läggas på steget ”data wrangling”. Presentera vanliga uppgifter som kan ingå i nämnt steg, inkludera exempel data där det kan vara passande för att förmedla vinsten med att genomföra aktuell syssla.

Data rensning och data transformation. Data rensning innebär att man tar bort dubletter, felaktiga eller onödiga data. Ett exempel på data där detta kan vara passande kan vara kunddata då man inte vill ha dubletter och korrekt kunduppgifter. Data transformation innebär att man omvandlar data från en form till en annan. Ett exempel kan vara transaktionsdata där man skapar rapporter som visar försäljning.

4. Begreppet Big Data brukar bytas ner och förklaras med hjälp av ord som börjar på bokstaven V. Vanligt är att man lyfter 3-7st "V’n". Presentera minst tre av dessa ord på V med fullständigt namn samt förklarande text.

   Volume är att det är enorma mängder data som genereras dagligen i Big Data.
   Velocity är hastigheten på dataflödena i Big Data.
   Variety är de olika typer av data som genereras. Data kan vara strukturerade, ostrukturerade eller semi-strukturerade.

5. Läs följande artikel:
   https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie *(Nej, du behöver sedan inte göra något mer men visst var det kul läsning!)*

6. Din uppgift är att lokalisera en forskningsartikel (tips: https://scholar.google.com) där något av följande begrepp tas upp:

   Klassificering
   Data wrangling
   Data mining
   Data lake
   Data warehouse
   Big data
   Machine learning

“NoSQL vs Relations-DB”

(Det räcker alltså att ett begrepp förekommer i artikeln du väljer, det går även om artikeln tar upp angränsande begrepp så länge vi befinner oss inom samma kontext).

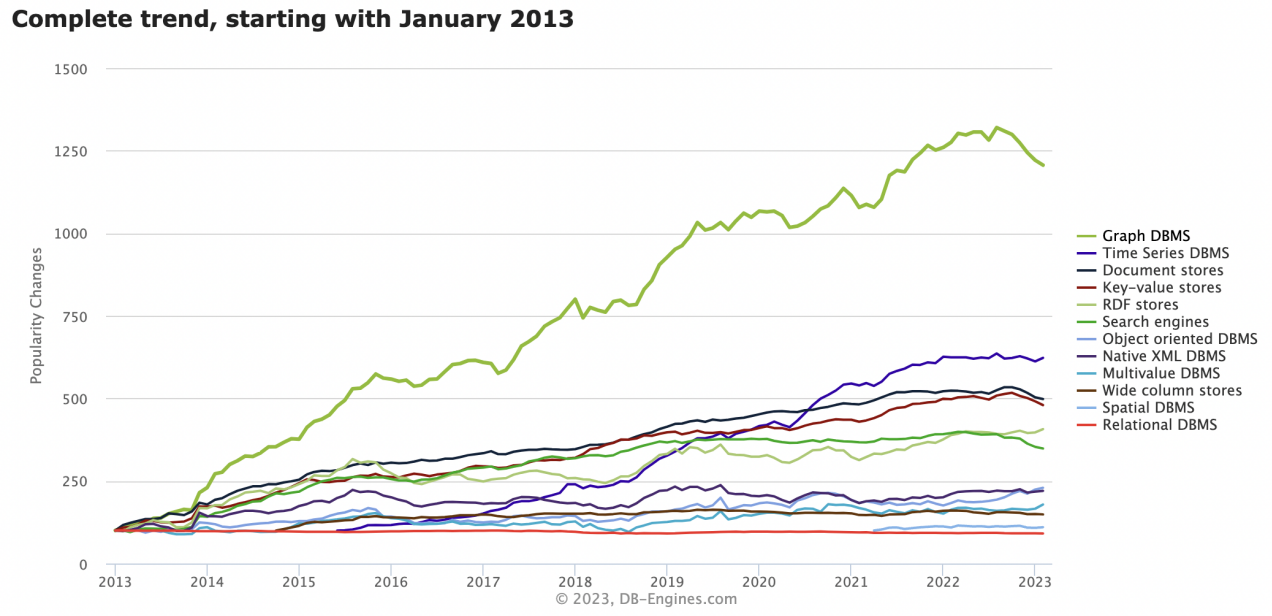
Du skriver sedan en kortare summering över vad som behandlas i artikeln samt vad du fan mest intressant i artikeln du valde. Inkludera länk till vald artikel.

Artikeln *Data mining: an overview from a database perspective,* skriven av Ming-Syan Chen, Jiawei Han och P.S. Yu, diskuterar “explosive growth” av data och databaser pågrund av tillgänglighet av kraftfulla men även prisvärda databas system, vilket har lett till ett behov av nya verktyg som man kan användas för att omvandla data till användbar information. Data mining är en process som handlar om att analysera stora mängder data för att hitta mönster och få en förståelse för hur datan kan användas.

Del 2 (*Neo4j*)

I en värld med 1000-tals olika databaser så ställs vi ständigt i situationer där vi behöver göra val, vilken passar oss bäst just nu? Valet är komplext och många parametrar behöver förhållas till. För att överhuvudtaget kunna börja så är ett en god idé att göra en inledande botanisering samt testskott med nya databaser då och då för att se vad de har att erbjuda.

En tydlig trend de senaste åren (*observera att trend här är inte samma sak som marknadsandelar*) har varit Graph-databasen *(alt senare någon subvariant, RDF store’s).*

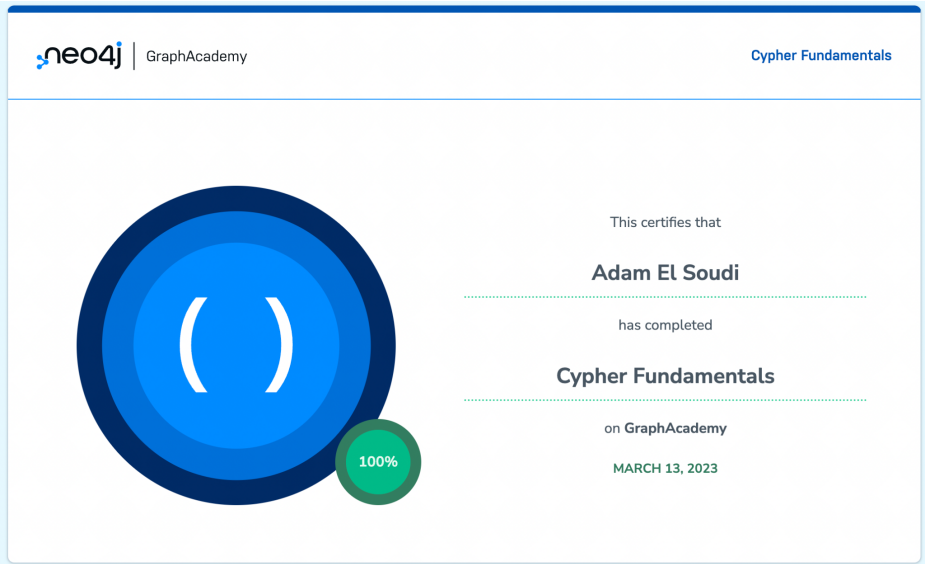


Din uppgift är således att göra ett testskott med Neo4j som är en databas av nämnd familj.

Att göra:

1. Genomgå kursen "Neo4j Fundamentals" som hittas här: https://graphacademy.neo4j.com/

2. Genomgå kursen "Cypher Fundamentals" som hittas här: https://graphacademy.neo4j.com/



3. Logga in i https://sandbox.neo4j.com/ och genomför *(starta upp och gå igenom tillhörande guid)* minst två "project" för att titta närmare på användarscenarion *(tipsar om Crime Investigation projektet som jag finner intressant!)*.

4. Ge en kortare presentation av dina inledande tankar kring Neo4j givet dessa övningar du precis gjort. Besvara i texten om du tyckte att kurserna och guiderna som Neo4j gett dig var tydliga och på lagom nivå att ta till sig som en nybörjare. Presenter även vilka två "sandbox projekt" du genomförde och vad de handlade om.

Neo4j var häftigt att få jobba med. Jag tyckte det var kul att få lära sig cypher och jobba lite med noder och sånt i en grafisk databas. Det hela var rätt enkelt att lära sig och ibland lite för basic så att man tappade fokus men det är väl bra för nybörjare. Jag valde att jobba med projekten “Cyber Security” och “Crime Investigation” i Neo4J sandbox. Det var lite svårt att faktiskt se vad som hände i databaserna i början men med lite cypher så kunde jag se lite mer. Men gennerellt var kursen bra för nybörjare men jag tyckte att sandbox projekten var lite svårt att komma igång med i början. Viste inte riktigt vad jag ens skulle göra efter att jag valde projekten.