NoSQL-L3

**Fråga 1:**

Företaget ”Vi följer dig” vill flytta in delar av data som tidigare endast har funnits i text-filer i en databas. Användningsområdet är främst att samla in data för att sedan sälja vidare information till andra företag som ber om den för en specifik person. Budget för flytten av data är minimal.  
  
Givet följande exempel datamängd, rekommendera en databas för företaget. **Motivera.**  
  
Datapunkt1:  
Namn: Gösta Knutsson  
Ålder: 45  
Körkort: sant  
Barn: ja  
Antal barn: 2  
Vanliga köp: Mjölk 3, Kaffe 7, Potatis 2, Bad skum 1, Groddar ibland, Pan pizza 6, Tomat 3, Ost 1, Lök 10  
Internetvanor: Aftonbladet, Hemnet, Youtube, uu,  
Favoritfärg: Grön,  
Ekonomi: 450.000  
Bil: ja  
Semester: I regel  
----

Datapunkt2:  
Namn: Bodil Arnesson  
Ålder: 34  
Körkort: sant  
Barn: nej  
Sambo: ja  
Intressen: Bio, Hagelskjutning  
Inkomst: runt 500000kr  
Cykel: ja  
Hund: ja  
A-skatt: ja

Svar: Enligt ovanstående datamängden, rekommenderar jag varmt att företaget använder databasen NoSQL. Anledningen till den databasen är att en NoSQL-databas är en hierarki som ett filmappssystem. Olika data typer har sparat in i olika datapunkter. Data i den databasen är ostrukturerad, det vill säga att data är icke relationell. Som vi se att datatyper som finns in i datapunkt 1 inte kan relatera till datapunkt 2. Alla data typer brister på struktur vilket gör att de snabbt kan hanteras större mängder data. Den största fördelen är att NoSQL bättre utvecklar databasen genom att insatta olika datatyper.

**Fråga 2:**

Givet databas DB2-1 och samlingen fire-incidents besvara frågorna:

1. Ange de 3 vanligaste förekommande utryckningstyperna ”majorcategory” i ordning flest🡪minst. Inkludera query för hur svaret nåtts i MongoDB

db.getCollection("fire-incidents").aggregate([{$group:{\_id:{"majorcategy":"$majorcategory"},count:{$sum:1}}},{$sort:{count:-1}},{$limit:3}])

Svar:

|  |
| --- |
| { \_id: { majorcategy: **'EMS'** }, count: 27930 } |
| { \_id: { majorcategy: **'FALSEALARM'** }, count: 7750 } |

|  |
| --- |
| { \_id: { majorcategy: **'GOODINTENT'** }, count: 4935 } |

1. Vilken station har mest att göra? Svar på frågan samt redogörelse för hur du nådde svaret skall inkluderas *(observera att du inte behöver använda mongo om du inte vill)*.

db.getCollection("fire-incidents").aggregate([{$match: { majorcategory: 'GOODINTENT'}},{$group: {\_id: '$station',count:{$sum:1}}}, {$sort:{count:-1}},{$limit:3}])

Svar: Station “003” har mest att göra när condition är majorcategory: “GOODINTENT” in i DB2-1,

|  |
| --- |
| { \_id: **'003'**, count: 793 } |
| { \_id: **'001'**, count: 732 } |

|  |
| --- |
| { \_id: **'004'**, count: 593 } |

**Fråga 3:**

1. Dra ett sträck som parar ihop ”datalagrings sätt” med rätt strategi:

**Data Warehouse ETL**

**Data Lake ELT**

1. **(VG-bedöms)** Vad står förkortningen ETL för och vad innebär det i sammanhanget vi befinner oss i?

Svar: ETL är en förkortning för ”Extract, Transform and Load”. Det handlar om en process där data extraheras från olika typer datakällor och omvandlas till ett lämpligt format för lagring och framtida referensändamål. På det första steget av ETL-processen importeras och konsolideras både strukturerad och ostrukturerade data i databasen. Vidare läggs en uppsättning regler exempelvis rensning, standardisering, de-duplicering, verifiering, sortering och andra uppgifter till de data som används för att säkerställa datakvalitet och tillgänglighet på det andra steget av ETL-processen. På det sista steget av ETL-processen är att ladda den nyligen transformerade data till ett nytt mål full laddning(samtidigt) eller inkrementell laddning (steg för steg).

**Fråga 4 (VG-bedöms):**

Att analysera data görs oftast i steg, nämn 3 vanligt förkommande steg och beskriv vad varje steg innefattar:

Data Discovery: Det är det första steget för Data Wrangling för att bekanta sig data och förstå mönster med data. Innan vi går djupare på dataanalys måste vi bättre förstå vad som finns i vår data, vilket kommer att informera om hur vi vill analysera data. Exempelvis att hur vi bråkar med kunddata för att vara informerad om var de finns, vad de köpte eller vilka kampanjer de fick.

Structuring: Det är en process för att organisera alla data, se till alla data är strukturerad för nästa steg. Rådata är vanligtvis ostrukturerad och de är inte användbart för slutprodukten därför sådan process är viktigt för enklare beräkning och analys i de senare stegen.

Validating: Det är en process för att dubbel kolla data så att det inte står något fel i data. Använd upprepade sekvenser av valideringsregler som exempelvis att bekräfta fältens noggrannhet via kontroll av data för att säkerställa datakonsistens samt kvalitet och säkerhet.

**Fråga 5 (VG-bedöms):**

1. Beskriv ett tillfälle då du anser att en MySQL databas passar bäst (**motivera**):

Svar: Jag anser att MySQL databas passar bäst på banken eftersom MySQL kan hantera strukturerade data där vi har en relation mellan olika enheter och variabler i data. En relationsdatabas kan lösa problem om olika typer data, dublettdata och ingen organisation genom att skapa separat tabell som är relaterat till med en annan separat tabell. Exempelvis att Kunder (deras namn, telefoner) och Konto (vem som äger, priset, vilka personaler som hjälper till att hantera). Med hjälp av ”nyckel” kan vi länka samma data från två olika tabeller så att vi kan manipulera och kombinera data i olika tabeller efter behov. Det är ett bättre alternativ för applikationer som behöver transaktioner med flera rader eftersom den är säkerhet och stabilitet. MySQL passar bra för bankens komplexa frågor eftersom den har ett standardgränssnitt för hantering av frågor. Dessutom MySQL följer ACID-egenskaper korrekt för att säkerställa säkra och pålitliga databastransaktioner.

1. Beskriv ett tillfälle då du anser att en NoSQL databas passar bäst (**motivera samt nämn** **vilken NoSQL databasfamilj du har förutsatt**):

Svar: Jag anser att NoSQL passar bäst på Webbplatser för sociala nätverk som exempelvis Instagram eller Facebook för att hantera ostrukturerade data eftersom för en applikation som den enorm mängd ostrukturerade data kan vi inte enbart använda SQL-baserade databaser. Sådan applikation har stora krav för skalbarhet och tillgänglighet och dessa krav betjänas av NoSQL-databaser. NoSQL-databaser har väldigt få inbyggda säkerhetsfunktioner för att möjliggöra snabbare dataåtkomst. Dessutom NoSQL är schemafritt och bordsfritt erbjuder det en hög grad av flexibilitet med datamodeller. Det är en billig databas och öppen källkod NoSQL-databaser ger mycket prisvärda lösningar till Webbplatser. NoSQL krävs inte detaljerad databasmodellering och vilket gör att sådan databas sparar tid och ansträngning.

**Fråga 6:**

Din uppgift är att lokalisera två stycken forskningsartiklar (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 per artikel 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 artiklarna du valde. Inkludera länk till valda artiklar.

1. Big Data and Data Mining Technologies Application at Road Transport Logistics

<https://doi.org/10.1016/j.trpro.2022.01.075>

Summering: Artikeln handlar om hur moderna Big Data och Data Mining teknik tillämpar i logistik för att lösa problem av tidsförluster och att bygga nya transportinfrastruktur i moderna logistikprocesser. Med hjälp av Big data och Data Mining för att integrera data om väderförhållanden och att bilda en prediktiv modell för vägtransportsystemet. Vägtrafikolyckor i den prediktiva modellen gör möjligt att uppskatta trängsel av vägbeskrivningar och det är en bra planering på återuppbyggnad och utveckling av logistik.

Anledning till att detta val är att jag bättre kan förstå hur Big Data och Data Mining fungerar i moderntid, inte bara lär mig kunskaper från böcker och lärare. Jag får djupare kunskaper om problem som finns i vårt transportsystem och lösningar på problem. Det väcker dessutom min nyfiken om hur de teknikerna påverkar på andra system exempelvis registreringssystem och bokningssystem i olika branscher.

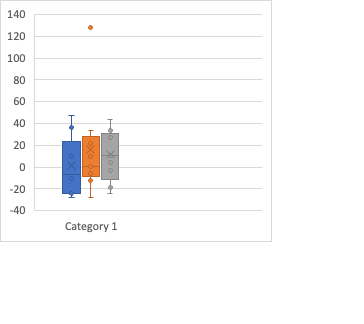
1. Machine learning and pre-medical education

<https://doi.org/10.1016/j.artmed.2022.102313>

Summering: Artikeln tar upp fram att Maskininlärning och artificiell intelligens (AI) är betydelsefull för olika aspekter av medicin och vårdhantering. Artikeln presenterade också att AI-ramverk i maskinlärning har stora potential för att påskynda arbetsflödet för vårdpersonal och vidare förbättrar vården. Men enligt författarna är det stora utmaningar på vårdstudent och vårdpersonal för att behärska principerna för maskininlärning och dessa medicinska tillämpningar. Rekommendation till denna utmaning är att skol och sjukhus erbjuder utbildningsmoduler för maskinlärning och flera introduktionskurser i maskininlärning.

Anledning till detta val är att jag vill gärna lära mig djupare kunskaper om begreppet maskininlärning och att jag är mycket nyfiken om hur maskininlärning tillämpar i vården. Artikeln ger mig en tydlig insikt av hur maskinlärning bidra till en bättre vård, samtidigt presenterar att vilka utmaningar finns i kliniska tillämningar. Jag själv anser att lösningar på utmaningar är mest intressant eftersom det väcker mina nya syner om vårdutbildningar. Jag tror aldrig att kunskaper om maskinlärning är viktigt för vårdstudent och vårdpersonaler tidigare, men artikeln utvecklar mina kunskaper om tillämpningen av maskinlärning inom vårdutbildning.

**Följande frågor behöver endast genomföras av dig som satsar på det högre betyget (VG)  
  
Fråga 7 (VG):**Vilket typ av diagram är det som visas i illustrationen nedan samt för en diskussion gällande användningsområde i samband med dataanalys.

****

**Fråga 8 (VG):**

Givet databas DB2-1 och samlingen fire-incidents besvara frågan:

För varje station, ange vilket skift man bör undvika. Motivera ditt svar samt inkludera eventuella querys och **metoder** du använde för att nå svaret *(observera att du inte behöver använda mongo om du inte vill).*

**Fråga 9 (VG):**

*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.