Tehtävä 1: Massadata (”Big data”) ja sen käytön haasteet

Oikeita vastauksia tehtävän kysymyksiin on useita ja tässä annetaan vain esimerkkivastauksia, jonka tyyppiset voidaan katsoa ratkaisuiksi. Arvostelussa huomioidaan erityisesti esimerkkien monipuolisuus ja niiden perustelut.

Kysymys 1 mallivastaus:

Massadataa luonnehtivia piirteitä ovat volyymi, vauhti, variaatio, oikeellisuus, oleellisuus ja arvo.

Volyymista esimerkkinä voidaan antaa kauppaketjujen kassapäätteisiin kertyvä data siitä, mitä tuotteita asiakkaat ostavat. Tätä dataa syntyy suurissa kauppaketjuissa päivittäin suuria määriä. Yksinkertaisimmillaan sitä voidaan käyttää esimerkiksi kaupan varastotilannetiedon ylläpitämiseen, mutta sitä analysoimalla voidaan myös yrittää ymmärtää asiakkaiden ostokäyttäytymistä: mitä he ostavat, mihin aikaan vuorokaudesta/vuodesta, kuinka paljon, minkä muiden ostosten yhteydessä jne. ja tätä tietoa voidaan hyödyntää esimerkiksi mainostamisessa.

Vauhdista esimerkkinä voidaan antaa Twitter-viestit, joita syntyy ympäri maailmaa satoja miljoonia päivittäin. Vaikka kukin viesti on vain 140 merkkiä pitkä, niistä päivittäin syntyvän datan määrä on huikea ja sen tallettaminen vaatii suuria määriä levytilaa. Palvelun käyttäjien ja siinä lähetettyjen viestien määrä kasvaa jatkuvasti. Twitter-viestejä analysoimalla voidaan saada selville hyvin monenlaisia asioita alkaen maailmanlaajuisten yhteiskunnallisten mielipiteiden muutoksista ja päätyen esimerkiksi viihteessä sillä hetkellä pinnalla oleviin trendeihin.

Variaatiosta esimerkkinä voidaan antaa yhteisöpalvelu Facebookissa syntyvä data. Data voi olla luonteeltaan esimerkiksi käyttäjien tekstimuotoisia päivityksiä eri kielillä, kuvia, linkkejä tai paikkadataa. Facebookin omistava yhtiö voi hyödyntää tätä dataa esimerkiksi myymällä mainostilaa sopivien keskustelunaiheiden yhteydessä, mutta tekstimuotoisen, monilla eri kielillä tuotetun datan analysointi kulttuurierot huomioon ottaen on haastavaa.

Oikeellisuudesta esimerkkinä voidaan antaa neuvolassa pienten vauvojen punnituksessa vuosikymmenien kuluessa syntyvä painotieto. Yksittäisen vauvan kohdalla data on suhteellisen helppo tulkita tiedoksi: vauva esimerkiksi on kasvanut normaalia vauhtia tai kasvu on jostain syystä hidastunut, johon on aiheellista etsiä syytä. Pienen vauvan painossa kuitenkin pienetkin painon nousut ja laskut ovat merkityksellisiä ja vaa’assa oleva heitto suuntaan tai toiseen tai vauvan juuri syömä ateria voivat vaikuttaa punnitustulokseen. Yksittäisen vauvan kohdalla voidaan tarkkailla muitakin merkkejä kuin painoa (esimerkiksi kuinka virkeä vauva on), mutta mikäli halutaan ymmärtää vuosien saatossa kaikkien suomalaisten vauvojen tai kaikkien eurooppalaisten vauvojen painon kehitystä ja syitä siihen, täsmälleen oikea paino olisi hyödyllistä tietää. Tähän ei kuitenkaan koskaan voida päästä, joten datan tulkinnassa on ymmärrettävä virhemarginaalin tarpeellisuus.

Oleellisuudesta esimerkkinä voidaan antaa linja-autojen sijaintidata. Joillakin paikkakunnilla linja-autoissa on GPS-paikannin, joka kertoo reaaliaikaisesti missä linja-auto kulkee. Kun tämä data sijoitetaan kartalle ja yhdistetään linja-autoaikatauluihin, matkustajat voivat siitä nähdä kuinka kauan menee, että linja-auto saapuu pysäkille, tai onko linja-auto jo ohittanut pysäkin. Tämän datan hyödyllisyys ainakin matkustajan kannalta on siis sidottu ajanhetkeen.

Kysymys 2 mallivastaus:

a) Autojen reaaliaikaisella GPS-seurannalla voitaisiin tunnistaa liikenteen pullonkaulakohdat, esimerkiksi tieosuudet joilla on liian paljon liikennettä tien kapasiteettiin nähden jolloin liikenne ruuhkautuu ja hidastuu. Seurannalla voitaisiin myös tukea joukkoliikenteen suunnittelua tunnistamalla useimmin käytettyjä autoreittejä työmatkaliikenteessä ja siten voitaisiin uudet joukkoliikennepalvelut kohdistaa näille reiteille
b) Autojen reaaliaikaisen GPS-seurannan avulla kaikkien autoilijoiden liikkeitä voitaisiin seurata tarkasti, mikä johtaisi yksityisyyden katoamiseen. Myös datan väärinkäyttö voisi olla ongelma, sillä esimerkiksi keltainen lehdistö haluaisi varmasti mielellään seurata julkisuuden henkilöiden autojen liikkeitä.
c) Autojen reaaliaikaisen GPS-seurannan teknisiä haasteita ovat esimerkiksi paikannuksen epävarmuus sekä siirrettävän datan suuri määrä. Paikannuksessa voi syntyä epävarmuuksia satelliittisignaalien heikkenemisen vuoksi maanalaisissa tunneleissa tai pysäköintihalleissa, eikä paikannuksessa saa olla suuria heittoja heikosta signaalista huolimatta. Autojen suuresta määrästä johtuen seurannasta kerätään niin suuri datamäärä, että sen turvallinen säilyttäminen verotuksen vaatiman ajan voi tulla ongelmaksi.
d) Suuresta datamäärästä pitää pystyä jalostamaan tieto yksittäisen ajoneuvon liikkeistä ja ajomäärästä verotusta ja mahdollisia valituksia varten. Epätarkan signaalin ja signaalin katoamisen aiheuttamat katkokset tallennetussa ajoreitissä pitää pystyä käsittelemään verotettavan kannalta edullisimmalla tavalla.

Kysymys 3 mallivastaus:

1. Datan yksikköä ei ole tallennettu, vain arvo. Esimerkiksi dataan ei ole tallennettu mitä lämpötilan arvo 35 tarkoittaa, ja tämä pitää selvittää muualta kuin itse datasta, esimerkiksi kysymällä dataa hallinnoivilta ihmisiltä tai muualta dokumentaatiosta. Lämpötila voi olla joskus Celsius-, joskus Fahrenheit-asteikolla. Luonnollisesti yksikkö pitäisi lisätä datan yhteyteen tallenteisiin.
2. Valuuttatiedot esimerkkinä ajan suhteen muuttuvista arvoista: on ensinnäkin tiedettävä rahayksikkö, jossa numeerinen valuuttadata on tallennettu. Lisäksi, jos tarvitaan konversioita toiseen valuuttaan, on tiedettävä aika, jolloin valuuttatieto on tallennettu, koska vaihtokertoimet riippuvat ajasta. Tällöin esim. puhelimen hinnasta on tiedettävä hinnan 147 lisäksi valuutta euro, sekä ajankohta 15.6.2014 ja lisäksi paikka, Oulu, Suomi. Vasta näillä tiedoilla varustettuna hintatieto on vertailtavissa ajan ja paikan suhteen muihin tietoihin.
3. Jatkuvasti kerättävä data vanhenee tai on hyödytöntä, jos keräämistä ei jatketa. Esimerkkinä trendien muutokset tweettien kautta seurattuna: on jatkuvasti kerättävä Twitterin kautta dataa, jos halutaan seurata mitkä ilmiöt, brändit, ihmiset tai tapahtumat kiinnostavat ihmisiä, ja miten kiinnostuksen kohteet muuttuvat. On siis huolehdittava datan päivityksestä ja uuden datan tuomisesta mukaan analyysiin sitä mukaa, kun sitä syntyy. Tallennuskapasiteetti saattaa muodostua ongelmaksi, samoin tiedon arkistointi. Analyysistä tulee raskasta. Tarvitaan tehokkaita tietokantoja ja analyysisovelluksia.
4. Aineiston taustat voivat olla erilaiset, esim. jossain aineistossa data voi käsittää vain tietyn populaation tai osan siitä, kun taas toisissa aineistoissa data voi koskettaa ihan erilaista populaatiota. Näin datan analyysissä voidaan tehdä virheitä, jos ei ymmärretä, ettei data ole yhteismitallista populaatioiden erojen takia. Esimerkiksi jos tutkitaan datan avulla puhelimen käyttäjien tottumuksia maailmalla, eikä ymmärretä, että kehittyvissä maissa puhelin ei ole aina henkilökohtainen viestintäväline, vaan puhelimen omistaja saattaa vuokrata puhelintaan muiden käyttöön jatkuvasti. Näin puhelimen käytöstä kertyvä data ei kerro yksilön käyttäytymisestä, toisin kuin rikkaissa maissa. Datan yhteyteen pitää siis jollain tavoin metatietona kirjata millaisesta populaatiosta se on kerätty, jotta aineistot olisivat vertailukelpoisia, tai vaihtoehtoisesti se pitää jotenkin muuten selvittää, kun dataa analysoidaan. Dataa analysoitaessa nämä metatiedot tietysti pitää ottaa huomioon.