## i Framside / Front page

Institutt for datateknologi og informatikk

Eksamensoppgave i TDT4305 Big Data Architecture

Eksamensdato: 23. mai 2020

**Eksamenstid (fra-til):** 09:00 – 13:00

Hjelpemiddelkode/Tillatte hjelpemidler: A / Alle hjelpemidler tillatt

Faglig kontakt under eksamen:

**Tlf.**: 99 02 76 56

Teknisk hjelp under eksamen: NTNU Orakel

**Tlf:** 73 59 16 00

#### **ANNEN INFORMASJON:**

Gjør dine egne antagelser og presiser i besvarelsen hvilke forutsetninger du har lagt til grunn i tolkning/avgrensing av oppgaven. Faglig kontaktperson skal kun kontaktes dersom det er direkte feil eller mangler i oppgavesettet.

**Lagring**: Besvarelsen din i Inspera Assessment lagres automatisk. Jobber du i andre programmer – husk å lagre underveis.

**Juks/plagiat**: Eksamen skal være et individuelt, selvstendig arbeid. Det er tillatt å bruke hjelpemidler. Alle besvarelser blir kontrollert for plagiat. <u>Du kan lese mer om juks og plagiering på eksamen her</u>.

Kildehenvisninger: Alle oppgaver skal besvares "med egne ord" for å vise forståelse.

**Varslinger**: Hvis det oppstår behov for å gi beskjeder til kandidatene underveis i eksamen (f.eks. ved feil i oppgavesettet), vil dette bli gjort via varslinger i Inspera. Et varsel vil dukke opp som en dialogboks på skjermen i Inspera. Du kan finne igjen varselet ved å klikke på bjella øverst i høyre hjørne på skjermen. Det vil i tillegg bli sendt SMS til alle kandidater for å sikre at ingen går glipp av viktig informasjon. Ha mobiltelefonen din tilgjengelig.

**Vekting av oppgavene**: Som vist i oppgavesettet. Alle deloppgaver innenfor en oppgave teller likt.

#### OM LEVERING:

Besvarelsen din leveres automatisk når eksamenstida er ute og prøven stenger, forutsatt at minst én oppgave er besvart. Dette skjer selv om du ikke har klikket «Lever og gå tilbake til Dashboard» på siste side i oppgavesettet. Du kan gjenåpne og redigere besvarelsen din så lenge prøven er åpen. Dersom ingen oppgaver er besvart ved prøveslutt, blir ikke besvarelsen din levert.

**Trekk fra eksamen:** Ønsker du å levere blankt/trekke deg, gå til hamburgermenyen i øvre høyre hjørne og velg «Lever blankt». Dette kan <u>ikke</u> angres selv om prøven fremdeles er åpen.

Tilgang til besvarelse: Du finner besvarelsen din i Arkiv etter at sluttida for eksamen er passert

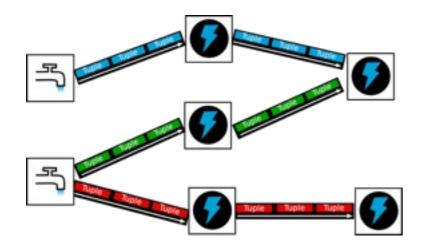
## <sup>1</sup> Oppgave 1

### Oppgave 1 – Systemer for datastrømmer (15%)

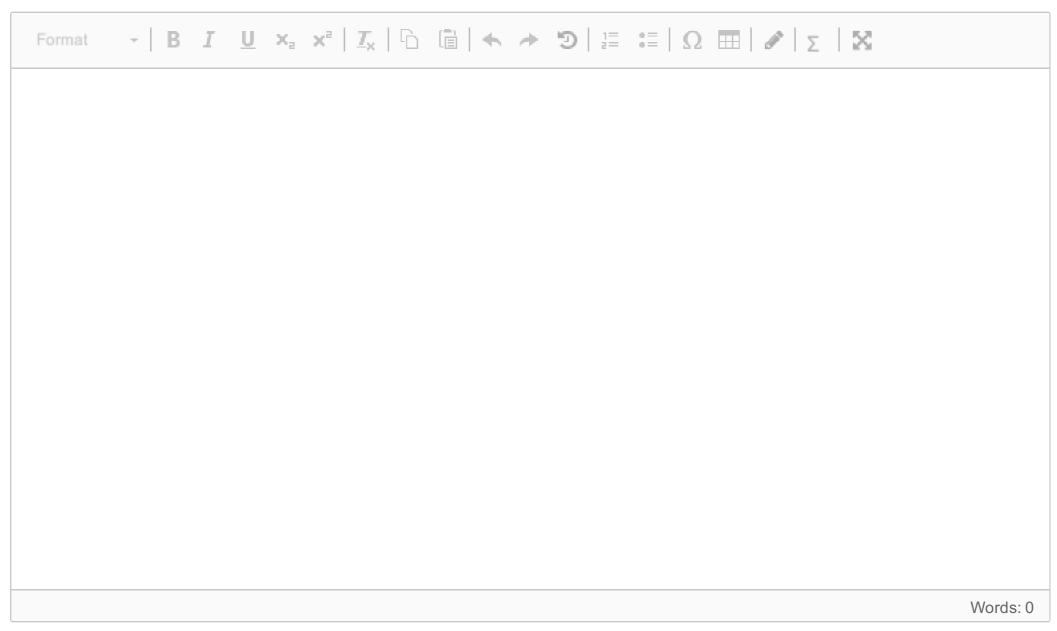
- 1. Drøft prinsippene bak håndteringen av datastrømmer i AsterixDB. Hvordan skiller dette seg fra Spark og Storm?
- 2. Vi skiller mellom tre forskjellige meldingsleveringsgarantier (Delivery semantics). Forklar hva disse tre garantiene betyr i praksis. Lag en tabell som forklarer hvilke garanti(er) som henholdsvis AsterixDB, Spark og Storm klarer å oppfylle. Til dette skal du bruke følgende mal:

	Garanti	Eksempel
AsterixDB		
Spark		
Storm		

3. Forklar følgende figur av et Storm system:



#### Skriv ditt svar her...



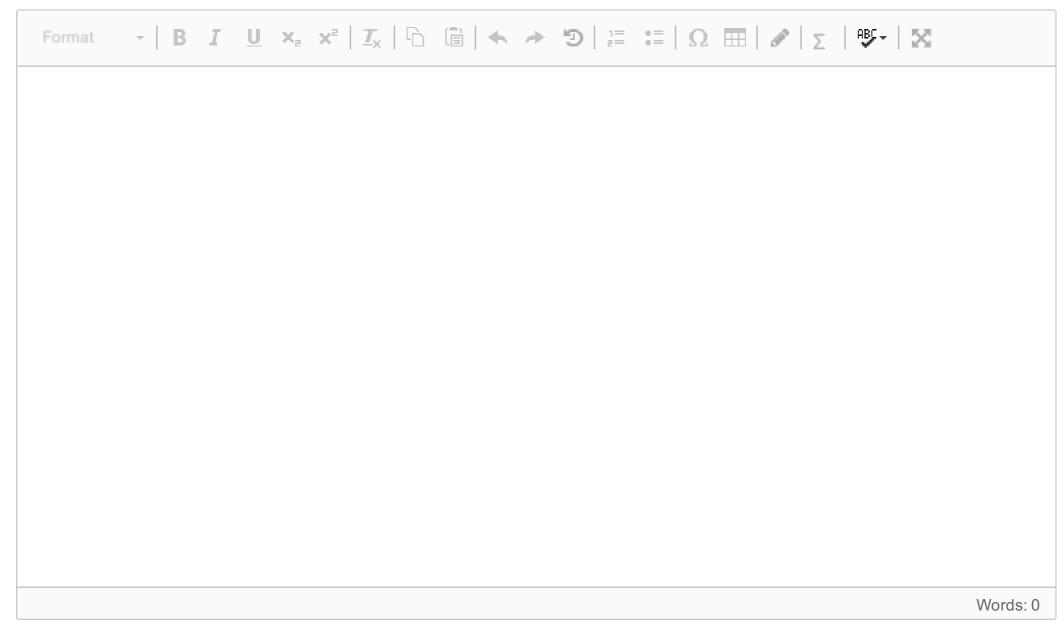
## <sup>2</sup> Oppgave 2

### Oppgave 2 – Håndtering av datastrøm (15%)

REI Coop ønsker å få oversikt over kundene sine kjøpsvaner. De bestemmer seg derfor å lage et system som registrerer hvor ofte en spesifikk kunde er innom nettsidene deres og ser på deres produkt. Basert på dette skal du svare på følgende spørsmål. Gjør de antakelsene du finner nødvendige.

- 1. Drøft hvordan du ville løse denne oppgaven ved hjelp av stående spørring (standing queries).
- 2. Kan denne oppgaven løses ved hjelp av «bit counting»? Forklar.
- 3. Anta at du skal bruke buckets til å løse oppgaven. Hvordan fungerer buckets? Hvordan ville du gå frem for å løse oppgaven?

#### Skriv ditt svar her...



Maks poeng: 15

# <sup>3</sup> Oppgave 3

## Oppgave 3 – Anbefalingssystemer (20%)

- 1. Drøft når du ville ha valgt å bruke «content-based recommendation» og når du ville ha valgt å bruke «collaborative filtering» som basis for anbefalingsmetode i et anbefalingssystem. Ta med fordeler og ulemper med begge som en del av forklaringen din.
- 2. Tabellen nedenfor viser hvordan 5 av deres brukere (U1 U5) har «rated» filmene (M1 M5) de ser på med verdi fra 0 til 10. Fyll ut tabellen for å finne ut hvilke av de andre brukerne som er mest lik bruker 4 (dvs. U4). For å finne ut denne likheten skal du bruke **Pearson correlation coefficient.**

User-ID /Movie-ID	M1	M2	М3	M4	M5	pearson(Ui, U4) (user-user)
U1	7	6	7	4	5	
U2	6	7	0	4	3	
U3	0	3	3	1	1	
U4	1	2	2	3	3	

U5	1	0	1	2	3	

Pearson correlation coefficient-formelen du skal bruke er gitt ved:

$$sim(x,y) = \frac{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x}) (r_{ys} - \overline{r_y})}{\sqrt{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})^2} \sqrt{\sum_{s \in S_{xy}} (r_{ys} - \overline{r_y})^2}}$$

Her er  $r_{xs}$  ratingsverdi gitt av bruker x på produkt s og  $\overline{r_x}$  snittverdi av bruker x sine ratings.  $S_{xy}$  er sett av alle produkter som både bruker x og bruker y har gitt en rating.

### Skriv ditt svar her...

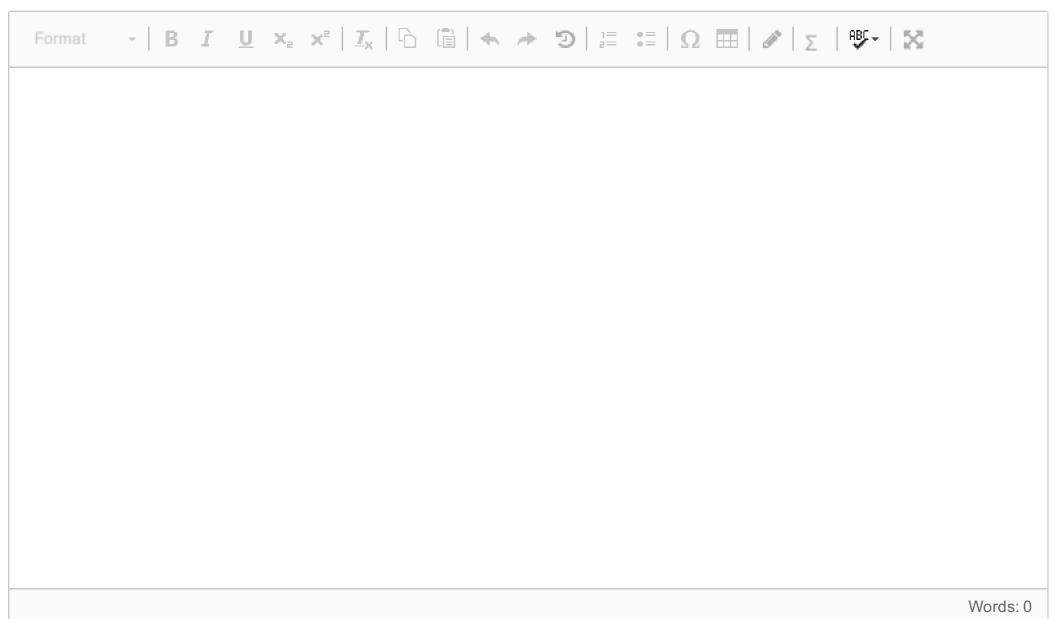
Format	-   B	I <u>U</u>	Xa	$\mathbf{x}^{z} \mid \underline{\mathbf{I}}_{x} \mid \widehat{\Box}$	9   <u>1</u> =	:≣   Ω	⊞   <i>₽</i>   ∑	ABC →   🏡	
									Words: 0
									Words: 0

# 4 Oppgave 4

Oppgave 4 – Hadoop og Spark – 15 %

- 1. Forklar hvordan skriving til fil foregår i HDFS.
- 2. Forklar konseptene "narrow dependency" og "wide dependency", og hvordan disse påvirker ytelse.

### Skriv ditt svar her...

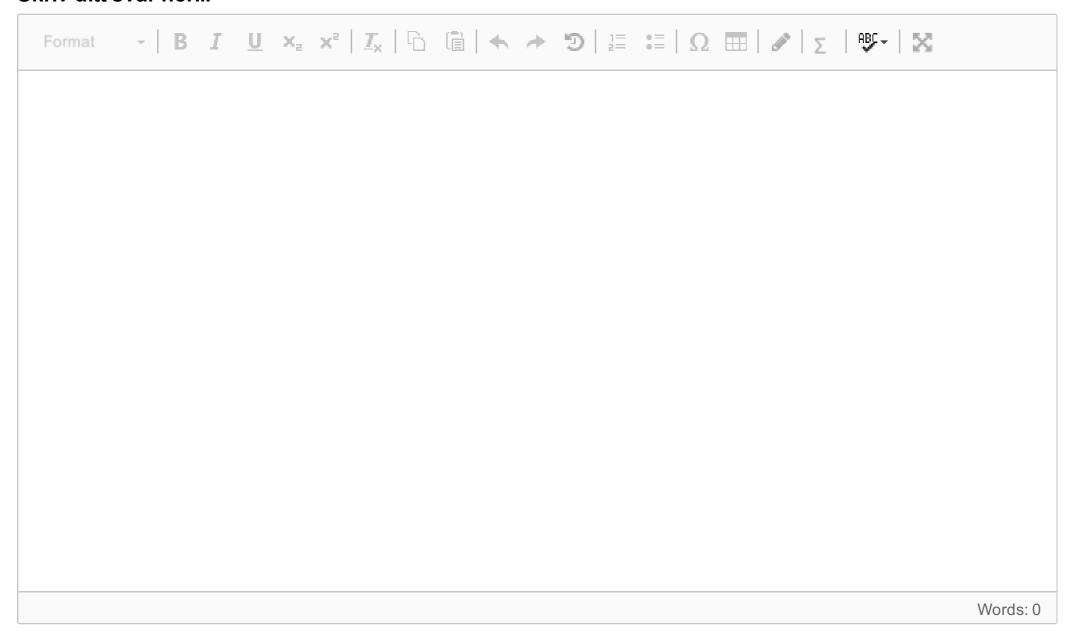


# <sup>5</sup> Oppgave 5

Oppgave 5 - MinHashing - 20 %

Forklar "shingles", "MinHashing", og LSH.

### Skriv ditt svar her...



# <sup>6</sup> Oppgave 6

Oppgave 6 – Adwords – 15 %

Forklar "Balance-"algoritmen (i kontekst av Adwords).

### Skriv ditt svar her...

