

2. forduló



A kategória támogatója: DXC Technology

# Ismertető a feladatlaphoz

#### Kérjük, hogy a feladatlap indítása előtt mindenképp olvasd el az alábbi útmutatót:

Helyezéseket a 4. forduló után mutatunk, százalékos formában: adott kategóriában a TOP 20-40-60%-hoz tartozol.

A feltűnően rövid idő alatt megoldott feladatlapok kizárást vonnak maguk után, bármilyen más gyanús esetben fenntartjuk a jogot a forduló érvénytelenítésére!

Jó versenyzést kívánunk! Ez most kicsit nehezebb lesz. :)



Ebben a fordulóban elsősorban Apache Kafka és streaming-el kapcsolatos kérdésekkel találkozhatsz.

# 1. feladat 1 pont

Mi azonosít egyértelműen egy üzenetet Kafkában?

Válasz
Topic-Partíció-Offset
Broker-Id
Topic-Message_Id
Broker-Topic-Offset
2. feladat 1 pont
Melyik mód a legelfogadottabb Kafka üzenetek struktúrájának tárolására?
Válasz
Annyira sokszínűek lehetnek a Kafka üzenetek, hogy nem érdemes a struktúrát tárolni
Egy schema registryben topiconként AVRO (.avsc) formátumban
Egy schema registryben topiconként Parquet (.parq) formátumban
Egy schema registryben particionként Kafka schema definition (.ksd) formátumban
3. feladat 1 pont
Melyik Kafka komponens felelős, hogy a consumer offset-eket kezelje és biztosítsa az üzenetek kézbesítését?
Válasz
Broker
Producer
ZooKeeper
Consumer Group Coordinator

# 4. feladat 1 pont

Az Apache Kafka melyik komponense felelős az adatok tárolásáért és tartósságáért (durability)?

Válasz
Producer
Consumer
Broker
ZooKeeper
5. feladat 4 pont
Az alábbi állítások közül melyek hamisak?
Válaszok
A Kafka támogatja mind az aszinkron, mind a szinkron üzenetküldést.
A Coordinator nem felelős az üzenetek fogyasztásáért.
A Producer csak aszinkron üzenetküldést támogat.
A Kafka Broker felelős az üzenetek tárolásáért és továbbításáért a témák között.
A ZooKeeper felelős a Kafka Broker-ek konfigurálásáért és koordinálásáért.
A Coordinator a Kafka Consumer csoportok koordinálásáért felelős, de nem az üzenetek fogyasztásáért.
A ZooKeeper elengedhetetlen a Kafka rendszer működéséhez.
6. feladat 1 pont
Mi a funkciója a Kafka Connectnek?
Válaszok
Kafka brokerek menedzselése
Komplex stream processing megvalósítása
Integráció külső adatforrásokkal és célrendszerekkel
Producer és Consumer konfigurációk szinkronizálása

#### 7. feladat 1 pont

Tegyük fel, hogy a lenti JSON struktúrát betöltötted egy Python változóba, amit "data"-nak hívunk. Hogyan éred el "Project B" leírását, amin Petra dolgozik?

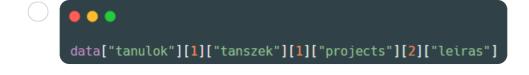
```
• • •
   "Szervezet": "Egyetem", "tanulok": [
         "id": 1,
"nev": "Petra",
                      "nev": "Project A",
"leiras": "Developing new software"
                      "nev": "Campaign X",
"leiras": "Product promotion"
         "id": 2,
"nev": "Barni",
"tanszek": [
                "nev": "Engineering", "projects": [
                      "nev": "Project C",
"leiras": "Network infrastructure"
```

#### Válasz

```
data["tanulok"][0]["tanszek"][0]["projects"][1]["leiras"]

data["tanulok"][0]["projects"][0]["leiras"]

data["tanulok"][1]["tanszek"][0]["projects"][0]["leiras"]
```



# 8. feladat 1 pont

Egy új, valós idejű (real-time) adattárházat építesz a cégednek, amihez BigQuery streaming insert-et fogsz használni. Nem garantált, hogy minden adat egyszer és csak egyszer fog betöltődni, de minden sor el van látva egy egyedi ID-val és időbélyeggel. Biztosítanod kell, hogy a duplikált sorok eliminálva legyenek az interaktív query eredményekből. Melyik query típust kell használnod az alábbiak közül ?

V	á	la	SZ
w	$\sim$	$\sim$	~

ORDER BY időbélyeg szerint, és LIMIT 1
GROUP BY használata az egyedi azonosítón és időbélyegen, majd SUM az értékeken
A LAG ablak funkció használata PARTITION BY az egyedi azonosítón, valamint WHERE LAG IS NOT NULL
ROW_NUMBER használata PARTITION BY az egyedi azonosítón, valamint WHERE ROW_NUMBER = 1

# 9. feladat 1 pont

Spark streaming "csúszó ablakos" feldolgozásnál melyik két paramétert kell definiálnunk?

### Válasz

Ablak méret, csúszás intervalluma
Állapot hossza, ablak méret
Állapot hossza, csúszás intervalluma
A fentiek közül egyik sem.