











A kategória támogatója: DXC Technology

## Ismertető a feladathoz

Ebben a fordulóban Streaming témájú kérdésekkel találkozhatsz.

NEM lesz szükséged Google Cloud accountra, vagy azon történő munkára a feladatok megoldásához.

Egyes válaszlehetőségeknél "Option1", "Option2" stb. megjelöléssel találkozhatsz, ez szövegileg <u>sosem része</u> az adott válasznak, csupán a válaszok későbbi összekapcsolódását biztosítja a magyarázatokkal.

Felhasznált idő: 00:00/34:00 Elért pontszám: 0/9

# 1. feladat 0/1 pont

Egy új, valós idejű (real-time) adattárházat építesz a cégednek, amihez BigQuery streaming insert- et fogsz használni. Nem garantált, hogy minden adat csak 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?

## Válasz

Option1: ORDER BY időbélyeg szerint, és LIMIT 1

Option2: GROUP BY használata az egyedi aznosítón és időbélyegen, majd SUM az értékeken

Option3: A LAG ablak funkció használata PARTITION BY az egyedi azonosítón, valamint WHERE LAG IS NOT NULL

Option4: ROW\_NUMBER használata PARTITION BY az egyedi azonosítón, valamint WHERE ROW\_NUMBER = 1
 Ez a válasz helyes, de nem jelölted meg.

## Magyarázat

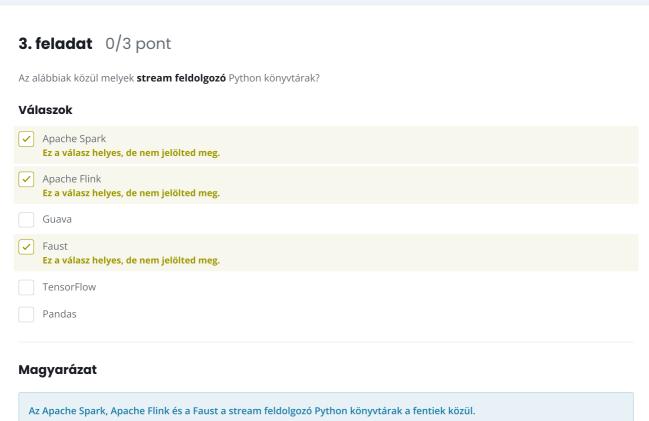
Az Option4 helyes, mert csak 1 sort fog választani a duplikátumok közül.

Az Option1 helytelen, mert csak egyetlen sort fog visszaadni.

Az Option2 helytelen, mert nem egyet választ a duplikáltak közül, hanem aggregálja azokat.

Az Option3 inkorrekt, mert azok a sorok, amik nem duplikáltak, exkludálva lesznek.

2. feladat 0/1 pont
10 000 új IoT eszközt telepítesz világszerte a raktáraidban, hőmérsékleti adatok gyűjtésére. Valahol fel kell dolgoznod, tárolnod és elemezned kell ezt a nagymennyiségű adatot, valós időben, Google Cloud platformot használva. <b>Az alábbiak közül mit fogsz tenni?</b>
Válasz
Option1: Beküldöd az adatokat Google Cloud Datastore-ba, aztán exportálod BigQuery-be.
<ul> <li>Option2: Becsatornázod az adatokat Googe Cloud Pub/Sub-ba, innen tovább streameled Google Cloud Dataflow-ba, végül Google BigQuery-ben letárolod azokat.</li> <li>Ez a válasz helyes, de nem jelölted meg.</li> </ul>
Option3: Beküldöd az adatokat Cloud Storage-be, és indítasz egy akkora Apache Hadoop klasztert, amit az adatok Google Cloud Dataprocban való elemzése megkövetel
Option4: Kötegekben (batch) exportálod az adatokat Google Cloud Storage-ba, indítasz egy Google Cloud SQL példányt ahova átimportálód az adatokat a Google Cloud Storage-ből, majd itt futtatod a szükséges elemzéseket
Magyarázat
Az Option2 az egyetlen valós idejű feldolgozásra alkalmas opció.
3. feladat 0/3 pont



# 4. feladat 0/1 pont

Spark streaming csúszó ablakos feldolgozásnál melyik két paramétert kell definiálnunk?

# Válasz



Ablak méret: az ablak méretét definiálja (pl. 3 napos ablak); csúszás intervalluma: milyen időközönként történik az ablakos

# 5. feladat 0/0 pont

feldolgozás (pl. 2 naponta)

Az alábbi PySpark kódrészletek közül melyik adja vissza a "df" dataframe "id" oszlopának maximum értékét?

### Válasz





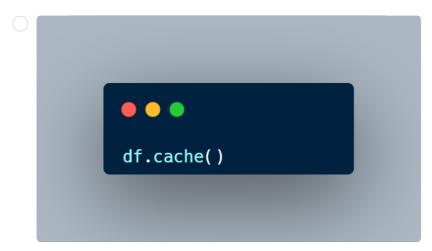
Ez a válasz helyes, de nem jelölted meg.

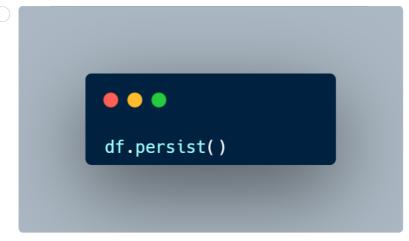


# 6. feladat 0/1 pont

A következő PySpark kódrészletek közül melyik tárolja <u>csak</u> a merevlemezen a dataframe tartalmát?

# Válasz









# Magyarázat

A fenti kódrészletek közül a **df.persist(StorageLevel.DISK\_ONLY)** kódrészlet tárolja <u>csak</u> a merevlemezen a dataframe

# 7. feladat 0/0 pont

Mi lesz a "df" adattípusa az alábbi PySpark kódrészletben?

```
val data = Array(1, 2, 3, 4, 5)
val dt = sc.parallelize(data)
```

## Válasz

array

dataframe

Ez a válasz helyes, de  Fentiek közül egyik	
Fentiek közül egyik	.sem.
agyarázat	
	gy létező collectionból (pl. array) készít egy RDD-t (Resilient Distributed Datasets), így a helyes válasz az
RDD.	

# 8. feladat 0/1 pont

Az alábbiak közül melyik **NEM** igaz a Cloud Pub/Sub-ra?

# Válasz

- Option1: A Pub/Sub egyszerűsíti a rendszereket az által, hogy nem kell minden komponensnek kommunikálnia minden más komponenssel.
- Option2: A Pub/Sub az alkalmazásokat és szolgáltatásokat egy üzenetküldő infrastruktúrán keresztül kapcsolja össze.
- Option3: A Pub/Sub határozatlan ideig eltárolja az üzeneteket, addig amíg azt le nem kérjük. Ez a válasz helyes, de nem jelölted meg.

# Magyarázat

A helyes válasz az Option3: a Pub/Sub maximum 31 napig képes tárolni az üzeneteket az egyes topicokban.

# 9. feladat 0/1 pont

Egy olyan alkalmazást tervezel, amely MQTT (Message Queuing Telemetry Transport) protokoll segítségével fogja küldeni az üzeneteket. **Melyik Google Cloud szolgáltatás kell ehhez használod?** 

#### Válasz

BigQuery

Cloud Pub/Sub

Ez a válasz helyes, de nem jelölted meg.

Cloud Spanner

Bigtable

# Magyarázat

A helyes válasz a Cloud Pub/Sub, ugyanis az MQTT = Message Queuing Telemetry Transport, egy IoT üzenetüldési protokoll. A többi komponens mind adatbázis megoldások.

Legfontosabb tudnivalók 🖸 Kapcsolat 🖸 Versenyszabályzat 🖸 Adatvédelem 🖸

© 2023 Human Priority Kft.

кészíтетте **c⊗ne** 

Megjelenés