

CLOUD BI

5. forduló



A kategória támogatója: DXC Technology

Ismertető a feladathoz

Ebben a fordulóban Streaming témájú kérdésekkel találkozhatasz.

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

Egyes válaszlehetőségeknél "Option1", "Option2" stb. megjelöléssel találkozhatasz, ez szövegileg sosem része 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) adattárházát 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 azonosítón 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. **Az alábbiak közül mit fogsz tenni?**

Válasz

- ☐ Option1: Beküldöd az adatokat Google Cloud Datastore-ba, aztán exportárod BigQuery-be.
- ☒ Option2: Becsatornázod az adatokat Google Cloud Pub/Sub-ba, innen tovább streameled Google Cloud Dataflow-ba, végül Google BigQuery-ben letárolod azokat.
Ez a válasz helyes, de nem jelölted meg.
- ☐ 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) exportárod az adatokat Google Cloud Storage-ba, indítasz egy Google Cloud SQL példányt ahova átimportárod az adatokat a Google Cloud Storage-ból, majd itt futtatod a szükséges elemzéseket

Magyarázat

Az Option2 az egyetlen valós idejű feldolgozásra alkalmas opció.

3. feladat 0/3 pont

Az alábbiak közül melyek **stream feldolgozó** Python könyvtárak?

Válaszok

- ☒ Apache Spark
Ez a válasz helyes, de nem jelölted meg.
- ☒ Apache Flink
Ez a válasz helyes, de nem jelölted meg.
- ☐ Guava
- ☒ Faust
Ez a válasz helyes, de nem jelölted meg.
- ☐ TensorFlow
- ☐ Pandas

Magyarázat

Az Apache Spark, Apache Flink és a Faust a stream feldolgozó Python könyvtárak a fentiek közül.

4. feladat 0/1 pont

Spark streaming csúszó ablakos feldolgozásnál melyik két paramétert kell definiálnunk?

Válasz

- ☒ Option1: Ablak méret, csúszás intervalluma
Ez a válasz helyes, de nem jelölted meg.

- ☐ Option2: **Állapot hossza, ablak méret**
- ☐ Option3: **Állapot hossza, csúszás intervalluma**
- ☐ Option4: **A fentiek közül egyik sem**

Magyarázat

A helyes válasz az Option1.

Ablak méret: az ablak méretét definiálja (pl. 3 napos ablak); csúszás intervalluma: milyen időközönként történik az ablakos feldolgozás (pl. 2 naponta)

5. feladat 0/0 pont

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

Válasz

☐

```
df.max( )
```

☒

```
df.agg(max( "id" ) ).show( )
```

Ez a válasz helyes, de nem jelölted meg.



```
df.agg( ).max("id").show( )
```



Fentiek közül egyik sem

Magyarázat

A helyes válasz a `df.agg(max("id")).show()`

6. feladat 0/1 pont

A következő PySpark kódrészletek közül melyik tárolja csak a merevlemezen a dataframe tartalmát?

Válasz



```
df.cache( )
```



```
df.persist( )
```



```
df.persist(StorageLevel.MEMORY_AND_DISK)
```



```
df.persist(StorageLevel.DISK_ONLY)
```

Ez a válasz helyes, de nem jelölted meg.

Magyarázat

A fenti kódrészletek közül a `df.persist(StorageLevel.DISK_ONLY)` kódrészlet tárolja csak a merevlemezen a dataframe tartalmát.

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

☒ **RDD**
Ez a válasz helyes, de nem jelölted meg.

☐ Fentiek közül egyik sem.

Magyarázat

A parallelize metódus egy 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 használnod?**

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 üzenetküldési protokoll. A többi komponens mind adatbázis megoldások.



[Legfontosabb tudnivalók](#)  [Kapcsolat](#)  [Versenyszabályzat](#)  [Adatvédelem](#) 

© 2023 Human Priority Kft.

KÉSZÍTETTE  **cone**

Megjelenés

 Világos 