Cognome e Nome _____

Matricola _____

Big data per l'IA
A.A. 2026/2027

Esonero del 13 novembre 2026

Esercizio 1 [8 punti]
Si consideri un marketplace (simile a eBay).
Siano date in input:

- Delle quadruple **(id_acquirente, id_articolo, quantità, data)** che rappresentano degli acquisti effettuati da **id_acquirente** (che varia tra 0 e k-1) aventi come oggetto **quantità** pezzi dell'articolo **id_articolo** (che varia tra 0 e m-1) in data **data**.
- Delle triple **(id_articolo, prezzo_unitario, id_venditore)** che danno le seguenti informazioni sull'articolo **id_articolo**: ha prezzo unitario **prezzo_unitario** ed è venduto dal venditore **id_venditore** (che varia tra 0 e n-1).

Si noti che ogni articolo è associato ad un unico venditore.
Si assuma che ogni articolo è coinvolto in un numero costante di acquisti.

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map e Reduce* per dare in output **un venditore che ha venduto di più in termini di importo totale (che ha cioè il ricavo totale più alto)**, utilizzando spazio locale $O(n^{1/2})$. Si discuta il *reducer size* di tutti gli step individuati.

Suggerimento: calcolare prima, per ogni articolo, il numero di pezzi venduti ottenendo così le quadruple *(id_articolo, id_venditore, prezzo_unitario, pezzi_venduti);* calcolare poi, per ogni venditore, l'importo totale ricavato dalla vendita e procedere infine ad individuare un venditore con il ricavo totale massimo.

[Esercizio alternativo, valutazione massima 5 punti]: Si assuma di partire dalle quadruple *(id_articolo, id_venditore, prezzo_unitario, pezzi_venduti)* date direttamente in input.

[Esercizio alternativo, valutazione massima 3 punti]: Si assuma di partire dalle coppie *(id_venditore, totale_ricavo)* date direttamente in input.

Esercizio 2 (9 punti)
[5 punti] Si consideri la variante multi-hash dell'algoritmo *Park Chen Yu* **con tre tabelle hash** nell'ambito del problema degli *itemset frequenti*, e se ne descriva il funzionamento in modo dettagliato, fornendo lo pseudocodice e descrivendo come viene utilizzata la memoria primaria a disposizione.
[4 punti] Si dica inoltre, in modo **sintetico e puntuale, giustificando la risposta,** cosa si può affermare in termini di **falsi positivi** per questa variante e una data istanza che, sulla macchina in uso con una certa memoria primaria a disposizione, laddove si utilizzasse il PCY classico con una tabella hash di dimensione **h** che sfrutta tutta la memoria a disposizione, farebbe sì che al più il 5% dei contatori cumulativi raggiungono la soglia **s** di supporto (perché la somma di tutti i contatori è minore o uguale di $\frac{s \cdot h}{20}$).

[Esercizio alternativo, valutazione massima 3+3=6 punti]: Svolgere l'esercizio relativamente variante multi-hash dell'algoritmo *Park Chen Yu* **con due tabelle hash**

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **90** minuti.
- Scrivere **subito** nome, cognome, matricola su OGNI FOGLIO (**compreso questo**).
- Durante la prova scritta **non** è possibile abbandonare l'aula.
- Non è ammesso **per nessun motivo** comunicare in qualsiasi modo con altre persone
- Non è possibile consultare appunti, libri e dispense.
- Qualsiasi strumento elettronico di calcolo o comunicazione (telefoni cellulari, calcolatrici, palmari, computer, etc…) deve essere **completamente disattivato** e **depositato in vista sulla cattedra**
- Mettere in vista sul banco un valido documento di identità.