University of Pisa

Master's degree in

Data Science & Business Informatics

Big Data Analytics

Project Assignment - Part 2

Group 17

Carlo Alberto Carrucciu (533967), Giacomo Lo Dico (600002),
Gian Maria Pandolfi (607268).

Parte 2

In questa seconda parte di progetto venivano principalmente richiesti risultati specifici sul database, attraverso lo sviluppo di query utilizzando il tool **SISS (SQL Server Integration Services)** che ci permette di estrarre, integrare e trasformare i dati. Pertanto tutti i risultati ottenuti dalla varie query sono stati salvati in formato CSV e verranno riportati insieme a questo documento ed ai file del progetto.

- **Task 0:**

Dopo aver eseguito una scansione completa di tutta la tabella **ram_sales** è stato eseguito un *look up* con la tabella **Geography** e subito dopo sono stati filtrati solamente i record registrati nell'area geografica *Germania*. Un secondo *look up* è stato eseguito sulla tabella **ram_product** per ricavare il "*brand name*" utile per il risultato finale della query.

Dopodichè con unica componente di aggregazione è stato fatto un raggruppamento per la regione e per il brand sulla ram e in conclusione, vengono ordinati i brand in base alle vendite totali registrate.

- **Task 1:**

Inizialmente è stata letta la tabella **gpu_sales** sulla quale è stato fatto un *look up* con la tabella **gpu_product** per recuperare il brand delle gpu e con la tabella **time** per risalire al *day of the week*.

A questo punto si può dividere in base al giorno della settimana usando una componente di split, la quale separa i record controllando che siano giorni del fine settimana o meno.

Entrambi i flussi sono stati aggregati sull'attributo **brand** per ottenere le vendite rispettive totali: *weekend sales* e *workday sales*. In seguito, vengono riuniti i flussi con una componente di merge e viene generata una nuova colonna per ottenere il ***ratio*** tra le due misure.

- **Task 2:**

Per poter rispondere alla terza query invece, sono stati programmati tre flussi diversi, uno per ogni prodotto diverso. Sono state poi lette separatamente le tre tabelle **ram_sales**, **gpu_sales** e **cpu_sales**, per tutte e tre è stato fatto un *look up* con la tabella **Geography** per poi aggregare le vendite per continente. Solo a questo punto i tre flussi son stati uniti con due componenti per fare il merge, ottenendo così una tabella che per ogni continente riportava le vendite per ognuno dei tre prodotti selezionati. Adesso per ottenere il risultato desiderato, viene creata una colonna derivata, con l'uso degli operatori ternari, per ogni riga è stato selezionato il valore di interesse che era quello più alto.

Infine viene estratto il CSV con il nome del prodotto corrispondente: **ram, gpu o cpu**.