# Prova laboratorio Databricks

## La base dati

I dati sono relativi al Contact Center della sociatà X. Il Contact Center gestisce sia telefonate che schede contatto (OrigineCaso) di soggetti residenti in diverse regioni d'Italia (e con un diverso livello di istruzione).

I casi gestiti, sia telefonate che schede contatto, sono classificate da un sistema automatico e lavorate da diversi soggetti. Più è complicata la richiesta, più è elevato il livello di escalation del caso.

I file disponibili sono 4 e sono in formato csv:

* Calendario
* Casi
* Contatti
* DecodificaOrigineCaso

I file riportano la seguente struttura.

## Calendario

* KEY\_DATA data nel formato AAAA|MM|GG
* DayNum numero che identifica il giorno della settimana

## Casi

* customerid identificativo univoco del contatto (si relaziona con contactid della Dataset Contatti)
* LivelloEscalation identifica il livello di escalation di un determinato caso
* KEY\_DATA data nel formato AAAA|MM|GG
* OraArrivo fascia oraria di arrivo del caso
* NumeroTicket identificativo univoco del caso
* DataArrivo data di arrivo del caso
* OrigineCaso codice identificativo dell'origine del caso (si relaziona con DecodificaOrigineCaso)
* ClassificazioneTicket classificazione del caso

## Contatti

* contactid identificativo univoco del contatto (si relaziona con customerid della Dataset Casi)
* EducationCodeValue livello di istruzione del contatto
* GenderCodeValue genere del contatto
* RegioneDiResidenza regione del contatto
* StatoOccupazionale stato occupazionale del contatto

## DecodificaOrigineCaso

* ID identificativo dell'origine del caso
* Valore decodifica dell'origine del caso

# Domande

Rispondi alle domande e allega anche gli screen shot del risultato ottenuto sul notebook. Entrambi gli elementi serviranno a rispondere alla domanda.

PS: assicurati che nello screen shot sia ben visibile il nome dell’account e del cluster usato per eseguire lo script.

## Parte I

1. **Creare** un **cluster** utilizzando il proprio nome e cognome, del tipo Cluster\_Nome\_Cognome
2. Inserire LUMSA nel percorso del FileStore di Databricks e **caricare** i **file** all’interno del **DBFS**
3. **Creare** un **Notebook** sull’area personale (no shared) e rinominarlo con nome e cognome (es. nome\_cognome)

## Parte II

1. Creare i DataFrame (DF) relativi ai CSV caricati nel DBFS (rinomina i file con lo stesso nome dei file csv)
2. Creare le **Temporary Table** partendo dai DF del punto 4
3. Attraverso **.count()** applicato al DF, contare il numero di righe dei diversi DF
4. Rinominare i giorni della settimana all’interno del DF Calendario, creando un **nuovo** **DF** rinominato **Calendario\_DAY**, in cui il nuovo campo è chiamato **day\_of\_week** ed il mapping segue il seguente criterio:
   1. 0 = lunedì
   2. 1 = martedì
   3. 2 = mercoledì
   4. 3 = giovedì
   5. 4 = venerdì
   6. 5 = sabato
   7. 6 = domanica

Utilizzando la funzione .count() dell’esercizio precedente, qual è il numero totale di ticket presenti nella nostra base dati? Verifica che la risposta sia corretta utilizzando anche SQL.

## Parte III

1. Unisci le diverse tabelle in un unico DF tenendo conto delle relazioni di join. Crea una temporary table in cui la tabella principale è quella dei Casi. Analizza con l’apposita funzione la qualità dei campi a disposizione (ricorda: esiste funzione per ispezionare i valuti non-nulli)

Attraverso l’aiuto di **SQL**, rispondi alle seguenti domande.

1. Considerando che ogni ticket nella base dati ha un id di riferimento (**NumeroTicket**), attraverso il linguaggio **SQL** conta i ticket (chiamate la nuova variabile **total\_ticket**), raggruppandoli per giorno della settimana (**day\_of\_week**) e ordinandoli in modo decrescente (rispetto alla variabile **total\_ticket**).

Produci un **bar chart** così definito:

- **asse x** il giorno della settimana ordinato in modo decrescente

- **asse y** il numero di ticket totali

Ora rispondete alle seguenti domande:

* 1. Qual è il giorno con il maggior numero di ticket?
  2. Qual è il giorno con il minor numero di ticket?
  3. Quantifica il numero di ticket per i 2 giorni sopra indicati

1. Sulla base di quanto indicato nella domanda precedente, raggruppa ora gli ordini per ora in cui questo è stato effettuato (**OraArrivo**). Mostra l’andamento in una **trend line** rappresentando l’ora in modo crescente e commenta quanto emerge dalla visualizzazione (es. ore in cui vi sono più ticket, ore dove si fanno meno chiamate, etc).
2. Identifica quali sono le prime 5 regioni che hanno più chiamate e mostrale in un grafico (utilizza lo strumento grafico più opportuno fra quelli proposti da Databricks). Raggruppa le altre regioni diverse dalle prime 5 in una categoria residuale definita “Altre”. Prova a rappresentare i valori anche in termini percentuali.

## Parte IV

Aiuta il Manager del Contact Center a rispondere alle seguenti Business Question:

1. Quanti sono i giorni di attività del Contact Center durante tutto il periodo considerato? (attraverso il dataset Casi)
2. Quali sono i 3 giorni (es. lunedì xx/11/2020) con il maggior numero assoluto di casi, cioè i giorni in cui si verificano dei picchi di lavorazione?
3. Quale fascia oraria mediamente riceve più telefonate?
4. Qual è la percentuale di casi per ogni regione?
5. Qual è la percentuale per genere, livello di istruzione e stato occupazionale declinata per telefonate e schede contatto? (tipologia di origine caso -> è possibile filtrare via i valori nulli per ogni per genere quando la percentuale da calcolare è genere, e così via)

**Esportare** il Notebook in formato **IPython Notebook**.