:ETL

בחרנו לקרוא נתונים על שתי מדינות שהאקלים בהן שונה על מנת לקבל תמונת מצב רחבה יותר על כמות המשקעים העולמית. המדינות שבחרנו הן קנדה וטנזניה, אחת קרובה לקו המשווה ובעלת מזג אוויר הפכפך והשנייה רחוקה וקרה יחסית ולכן ציפינו שתהיה בה כמות משקעים גדולה.

בחרנו להוציא שתי טבלאות:

County_year_final(country,year,total_sum,total_count,Avg_PRCP)

Country – סימול המדינה (2 תווים המייצגים מדינה)

Year – שנה

Total_sum – סה״כ המשקעים לאורך השנה

Total_count – מס׳ המדידות הקיימות בשנה

Avg_PRCP – ממוצע המשקעים השנתי ליום

בחרנו להוציא את הנתונים האלה לטובת שאלת המחקר שלנו – האם קיים שינוי בכמות המשקעים הממוצעת לאורך השנים והאם קיים הבדל בתנודתיות בין מדינות שונות כתלות במרחק מקו המשווה.
הפרמטרים Total_sum, Total_sum הוצאו על מנת להוציא את ממוצע המשקעים השנתי לכל מדינה ושנה תוך כדי הזרמת הנתונים.

Station_year_final(country,StationId,year,latitude,longitude,elevation,total_sum,total_count,avg_PRCP)

בחרנו להוציא את הנתונים האלה לצורך ביצוע אגרגציה על elevation לטובת שאלת החקר שלנו האם הגובה משפיע על כמות המשקעים השנתית הממוצעת ליום ואת שאר המשתנים לטובת החלק של למידת המכונה – בחרנו לבדוק האם ניתן לחזות את ממוצע המשקעים השנתי ליום לפי מיקום(קווי אורך, קווי רוחב וגובה) ולפי שנה ולכן בחרנו לא לבצע את החלוקה לקטגוריות של הגובה כבר בתהליך ה-ETL במחשבה שהאגרגציה לפי קטגוריות הגובה תהיה פשוטה יותר על הטבלה המתקבלת בסוף התהליך.

על מנת לייעל את תהליך ההזרמה השתמשנו בפרגמנטציה אופקית לפי מדינה, סימול תחנה ושנה,

מה שעוזר לביצוע האגרגציות בשלב הטיפול בכל באצ׳.

מפאת חוסר הזמן, לא הצלחנו להתמודד עם עצירת הquery עד סיום הזרמת הנתונים. לכן, בחרנו להשתמש בבאצ׳ים בגודל מליון ותחזקנו מונה שיסיים את פעולת ההזרמה לאחר 100 איטרציות, כלומר לאחר שקראנו סה״כ 100 מיליון רשומות כנדרש.