Cheat sheets

Monday, 13 February, 2023 4:39 PM

No	Objective	Language	Script
1	***Get last month end date***	SQL	select left(dateadd(day, -1, to_date(left(to_char(GETDATE(),'YYYY-MM-DD'),8)+'01', 'YYYY-MM-DD')),10) as lst_mth_date select left(dateadd(day, -1, to_date(left(to_char(DATEADD(Month,-1,GETDATE()),'YYYY-MM-DD'),8)+'01', 'YYYY-MM-DD')),10) as lst_2mth_date

2	***Create new columns based on the values in the column & assign 1***	Pyspark	for p_id in pack_id:
			    df = pp_recharges.withColumn(p_id, when(col('l9_package_id')==p_id, 1).otherwise(0))
			    pp_recharges = df

3	***Group by without assigning the columns name (need to have the list of columns)***	Pyspark	exprs = {x: "max" for x in pack_id} df1 = df.groupBy("account_no").agg(exprs)

4	***Find the possible combinations from the list (need to have the list of elements)***	Pyspark	from itertools import combinations
			sample_list = ['a','b','c']
			list_combinations = list()
			for n in range(len(pack_id)+1):
			    list_combinations += list(combinations(pack_id,2))

5	***Get the group by and the aggregation based on the columns of the dataframe***		def view_dur_pivot(prefix):
			    df_ivpevo_sum = df_ivpevo.filter(col('content_pillar_type') == prefix).groupby('account_no') \\
			    .pivot('month_key') \\
			    .agg(round(sum('duration_in_sec')/60, 2)) \\
			    .na.fill(0)
			    df_ivpevo_sum = df_ivpevo_sum.select([f.col(c).alias(prefix + "_" + "view_dur_" + c) \\
			                    if c not in {'account_no'} else 'account_no' for c in df_ivpevo_sum.columns])
			    return df_ivpevo_sum
			def view_tag_pivot(prefix):
			    df_ivpevo_sum = df_ivpevo.filter(col('content_pillar_type') == prefix).groupby('account_no') \\
			    .pivot('month_key') \\
			    .agg(max('view_cnt')) \\
			    .na.fill(0)
			    df_ivpevo_sum = df_ivpevo_sum.select([f.col(c).alias(prefix + "_" + "view_tag_" + c) \\
			                    if c not in {'account_no'} else 'account_no' for c in df_ivpevo_sum.columns])
			    return df_ivpevo_sum
			def view_sum_pivot(prefix):
			    df_ivpevo_sum = df_ivpevo.filter(col('content_pillar_type') == prefix).groupby('account_no') \\
			    .pivot('month_key') \\
			    .agg(sum('view_cnt')) \\
			    .na.fill(0)
			    df_ivpevo_sum = df_ivpevo_sum.select([f.col(c).alias(prefix + "_" + "view_cnt_" + c) \\
			                    if c not in {'account_no'} else 'account_no' for c in df_ivpevo_sum.columns])
			    return df_ivpevo_sum

6	***Get a list of date range***	Pyspark	months = pd.date_range(start='07-01-2022', periods=6, freq='M').strftime('%Y%m').tolist()

7	***Save the file into s3***	Pyspark	df.repartition(1).write.csv(s3_path + file_name, header="true", mode="overwrite")
			df.write.mode("overwrite").partitionBy("month_key").format('orc').save(s3_path + file_name, header=True)
			df.repartition(1).write.parquet(s3_path+'.parquet')

8	***Get the week from date***	SQL	concat ('WK', lpad (date_part (w, cast ('2022-01-22' as date)), 2, '0')) as week

9	***Pivot***	Pyspark	cols = engage.drop(col('viewer_id')).columns
			pivot = engage.groupBy(cols).agg(count('viewer_id').alias('cust_cnt'))

10	***Assign date to variable***	Pyspark	month_ref = '202204'
			last_3 = (datetime.strptime(month_ref, '%Y%m') - relativedelta(months=6)).strftime('%Y-%m-%d')
			end = (datetime.strptime(month_ref, '%Y%m') - relativedelta(days=1)).strftime('%Y-%m-%d')

11	**Append new column from another df**	Pyspark	def append_dfs(df1,df2):
			        list1 = df1.columns
			        list2 = df2.columns
			        For col in list2:
			                If(col not in list1):
			                        Df1 = df1.withColumn(col, f.lit(None))
			        For col in list1:
			                If(col not in list2):

			Df2 = df2.withColumn(col, f.lit(None)) Return df1.unionByName(df2)