

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN



Báo cáo môn học
Hệ thống thông tin phục vụ
trí tuệ kinh doanh

Đồ án thực hành

LỚP: CQ2021/1

Giảng viên hướng dẫn:

Hồ Thị Hoàng Vy

Tiết Gia Hồng

Nguyễn Ngọc Minh Châu

Sinh viên thực hiện:

MSSV	Tên thành viên
20120133	Phạm Lê Hoài Minh
21120037	Mã Thùy Anh
21120199	Trần Quốc Anh
1712876	Nguyễn Phạm Anh Tuấn

Mục lục

I. Tổng quan.....	2
1. Mô tả báo cáo	2
2. Phân công và mức độ hoàn thành	2
II. Nội dung báo cáo.....	3
1. Mô tả nguồn dữ liệu	3
2. Phân tích kho dữ liệu.....	4
3. Xác định các transformation	8
4. Phân tích qui trình ETL	9
5. Phân tích qui trình NDS vào DDS	19
6. OLAP	30
7. MDX	34
8. Dashboard.....	43
9. Data mining	43
10. Kết luận	49
III.Tài Liệu tham khảo.....	50

I. Tổng quan

1. Mô tả báo cáo

- Báo cáo trình bày các nội dung đã hoàn thành trong đồ án thực hành ở giai đoạn 2:

ETL source >> Stage >> NDS >> DDS

- Link Github: <https://github.com/Quanhf2704/HTTPVTTKD-01>

2. Phân công và mức độ hoàn thành

Công việc	Người thực hiện	Mức độ đóng góp
Quay video giải thích ETL Phân tích NDS Tạo cơ sở dữ liệu Phân tích DDS Chỉnh sửa, bổ sung qui trình nạp dữ liệu vào DDS Tạo khối OLAP Viết script MDX Làm dashboard	Phạm Lê Hoài Minh	25%
Xác định các transformation Tạo cơ sở dữ liệu Phân tích DDS Nạp dữ liệu vào DDS Viết script MDX Làm dashboard Quay video giải thích MDX và dashboard	Mã Thùy Anh	25%
Quay video giải thích ETL Kết nối ETL Tạo cơ sở dữ liệu Phân tích DDS Chỉnh sửa, bổ sung qui trình nạp dữ liệu vào DDS Làm dashboard	Trần Quốc Anh	25%
Quay video giải thích ETL Tạo cơ sở dữ liệu Phân tích DDS Làm dashboard Thực hiện data mining	Nguyễn Phạm Anh Tuân	25%

II. Nội dung báo cáo

1. Mô tả nguồn dữ liệu

Bộ dữ liệu mô tả chi tiết chất lượng không khí tại các quận trên 10 tiểu bang Mỹ được tổng hợp bởi the Environmental Protection Agency's (EPA) Daily Summary AQI trong khoảng thời gian từ 2021 đến 2023.

a) Source 1: 10_state_aqi.csv

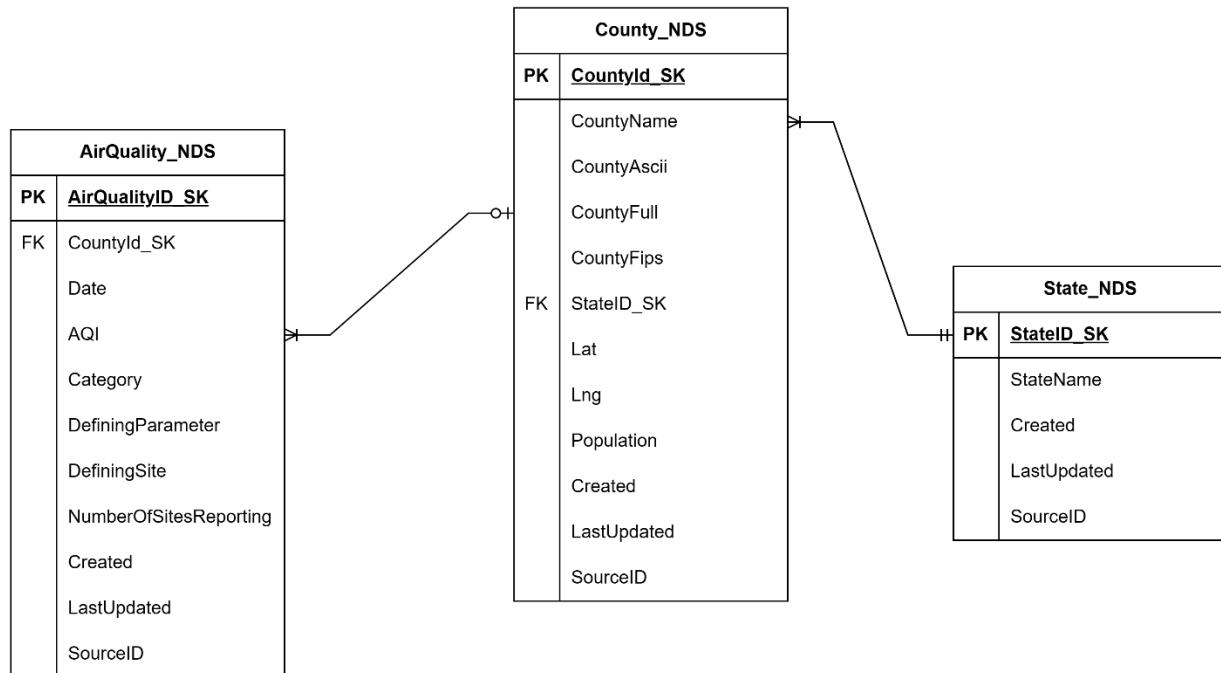
State Code	Mã tiểu bang
State Name	Tên tiểu bang
County Code	Mã quận
County Name	Tên quận
Date	Ngày ghi lại AQI
AQI	Chỉ số không khí đo được
Category	Phân loại mức độ ảnh hưởng đến sức khỏe
Defining Parameter	Thành phần quyết định chỉ số AQI
Defining Site	Trang web quyết định chỉ số AQI
Number of Sites Reporting	Số trang web báo cáo dữ liệu tương ứng với mỗi dòng dữ liệu
Created	Thời gian tạo dữ liệu
Last Updated	Thời gian cập nhật dữ liệu

b) Source 2: (2B)uscounties.csv

County	Tên quận
County_ascii	Tên quận (ascii)
County_full	Tên đầy đủ của từng quận
County_fips	Mã fips của mỗi quận
State_id	Mã tiểu bang
State_name	Tên tiểu bang
Lat	Vĩ độ của quận
Lng	Kinh độ của quận
Population	Mật độ dân số tại mỗi quận

2. Phân tích kho dữ liệu

a) NDS



- Bảng AirQuality_NDS

STT	Thuộc Tính	Kiểu giá trị	Mô tả
1	AiQuanlityID_SK	INT	Khóa chính, tự động tăng dần và không chấp nhận giá trị NULL.
2	CoutyCode_SK	INT	Khóa ngoại, mã định danh mỗi quận
3	Date	DATETIME	Ngày ghi lại AQI
4	AQI	INT	Chỉ số đo mức độ không khí
5	Categogy	NVARCHAR(255)	Phân loại mức độ ảnh hưởng đến sức khỏe
6	DefiningParameter	NVARCHAR(255)	Thành phần quyết định chỉ số AQI
7	DefiningSite	NVARCHAR(255)	Trang web quyết định chỉ số AQI
8	NumberofSitesReporting	INT	Số trang web báo cáo dữ liệu tương ứng với mỗi dòng dữ liệu
9	Created	DATETIME	Ngày tạo
10	LastUpdated	DATETIME	Ngày cập nhật

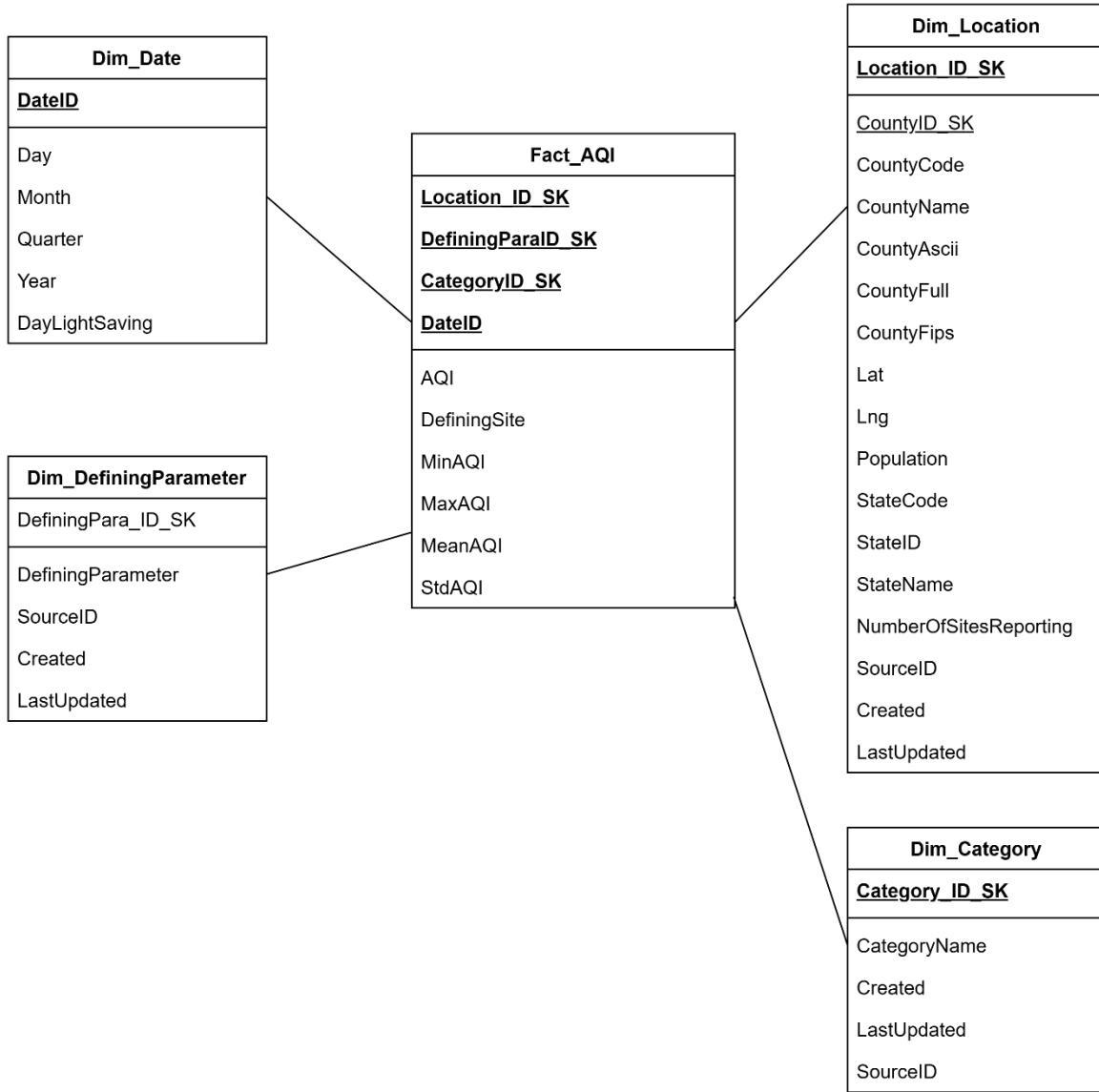
- Bảng County_NDS

STT	Thuộc Tính	Kiểu giá trị	Mô tả
1	CountyCodeID_SK	INT	Khóa chính, không chấp nhận giá trị NULL.
2	CoutyName	NVARCHAR(255)	Tên quận
3	CoutyAscii	NVARCHAR(255)	Tên quận (ascii)
4	CoutyFull	NVARCHAR(255)	Tên đầy đủ quận
5	Coutyfips	INT	Mã fips của mỗi quận
6	StateID	NVARCHAR(2)	Khóa ngoại, mã tiểu bang
7	Lat	NVARCHAR(255)	Vĩ độ của quận
8	Lng	NVARCHAR(255)	Kinh độ của quận
9	Population	FLOAT	Mật độ dân số tại mỗi quận
10	Created	DATETIME	Ngày tạo
11	LastUpdated	DATETIME	Ngày cập nhật

- Bảng State_NDS

STT	Thuộc Tính	Kiểu giá trị	Mô tả
1	StateCodeID_SK	INT	Khóa chính, không chấp nhận giá trị NULL.
2	StateName	NVARCHAR(255)	Tên tiểu bang
3	Created	DATETIME	Ngày tạo
4	LastUpdated	DATETIME	Ngày cập nhật

b) DDS



- Bảng Dim_Date

STT	Thuộc Tính	Kiểu giá trị	Mô tả
1	DateID	INT	Khóa chính, không chấp nhận giá trị NULL.
2	Day	INT	Ngày
3	Month	INT	Tháng
4	Quarter	INT	Quí
5	Year	INT	Năm
6	DayLightSaving	BIT	Quy ước giờ mùa hè

- Bảng Dim_DefiningParameter

STT	Thuộc Tính	Kiểu giá trị	Mô tả
1	DefiningPara_ID_SK	INT	Khóa chính, không chấp nhận giá trị NULL.
2	DefiningParameter	NVARCHAR(50)	Thành phần quyết định chỉ số AQI
3	SourceID	INT	ID nguồn
4	Created	DATETIME	Ngày tạo
5	LastUpdated	DATETIME	Ngày cập nhật

- Bảng Dim_Location

STT	Thuộc Tính	Kiểu giá trị	Mô tả
1	Location_ID_SK	INT	Khóa chính, không chấp nhận giá trị NULL.
2	CountyID_SK	INT	Khóa ngoại, mã quận
3	CountyCode	INT	Mã quận
4	CountyName	NVARCHAR(50)	Tên quận
5	CountyAscii	NVARCHAR(50)	Tên quận (ascii)
6	CountyFull	NVARCHAR(50)	Tên đầy đủ quận
7	CountyFips	NVARCHAR(10)	Mã fips của mỗi quận
8	Lat	NVARCHAR(10)	Vĩ độ của quận
9	Lng	NVARCHAR(15)	Kinh độ của quận
10	Population	NVARCHAR(10)	Mật độ dân số tại mỗi quận
11	StateCode	INT	Mã bang (theo chữ cái)
12	StateID	NVARCHAR(5)	Mã bang (theo số)
13	StateName	NVARCHAR(50)	Tên bang
14	NumberOfSitesReporting	INT	Số trang báo cáo
15	SourceID	INT	ID nguồn
16	Created	DATETIME	Ngày tạo
17	LastUpdated	DATETIME	Ngày cập nhật

- Bảng Dim_Category

STT	Thuộc Tính	Kiểu giá trị	Mô tả
1	Category_ID_SK	INT	Khóa chính, không chấp nhận giá trị NULL.
2	CategoryName	INT	Tên phân loại
3	SourceID	INT	ID nguồn
4	Created	DATETIME	Ngày tạo
5	LastUpdated	DATETIME	Ngày cập nhật

- Bảng Fact_AQI

STT	Thuộc Tính	Kiểu giá trị	Mô tả
1	Location_ID_SK	INT	Khóa chính, không chấp nhận giá trị NULL.
2	DefiningPara_ID_SK	INT	Khóa chính, không chấp nhận giá trị NULL.
3	CategoryID_SK	INT	Khóa chính, không chấp nhận giá trị NULL.
4	DateID	INT	Khóa chính, không chấp nhận giá trị NULL.
5	AQI	INT	Giá trị chỉ số AQI
6	DefiningSite	NVARCHAR(50)	Vị trí ghi nhận
7	MinAQI	INT	Giá trị AQI nhỏ nhất
8	MaxAQI	INT	Giá trị AQI lớn nhất
9	MeanAQI	FLOAT	Giá trị AQI trung bình
10	StdAQI	FLOAT	Phương sai AQI

3. Xác định các transformation

a) Tạo chiêu dữ liệu thời gian

Dựa vào những yêu cầu liên quan đến phân tích, so sánh về thay đổi chỉ số AQI theo từng quý, từng năm và thống kê số ngày trung bình mà chỉ số AQI được đánh giá theo một mức nào đó (Good, Moderate,...) thì việc ta tạo một chiêu dữ liệu thời gian trong cơ sở dữ liệu sẽ giúp quá trình phân tích dễ dàng hơn. Nhờ tạo chiêu dữ liệu thời gian, ta sẽ trình bày được thời gian theo một chuẩn thống nhất giữa các nguồn dữ liệu giúp cho truy vấn nhanh hơn và hỗ trợ các phân tích, so sánh tốt hơn.

Trong báo cáo này, chúng ta sẽ xây dựng chiêu dữ liệu thời gian theo phân cấp như sau:

Year > Quarter > Month > Day

b) Tạo chiêu dữ liệu địa lý

Đối với những yêu cầu liên quan đến phân tích tình hình theo khu vực địa lý, chẳng hạn như thống kê chỉ số theo từng tiểu bang hoặc từng quận, ta nhận thấy rằng việc tạo chiêu dữ liệu địa lý rất quan trọng bởi vì một tiểu bang bao gồm rất nhiều quận, nên nếu ta không tạo chiêu dữ liệu địa lý trong cơ sở dữ

liệu thì khi truy vấn ta sẽ phải mất thời gian kết bảng và dò tên quận, tên tiểu bang.

Như vậy, trong cơ sở dữ liệu, ta sẽ xây dựng chiều dữ liệu địa lí theo phân cấp sau:

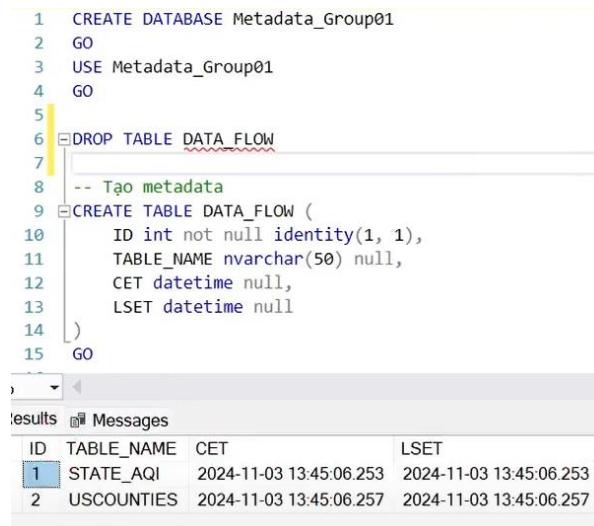
State > County

4. Phân tích qui trình ETL

a) ETL Source đến Stage

Đầu tiên, chúng ta sẽ tạo metadata và các bảng cần thiết trong SQL Server để khi thiết lập kết nối ETL trong Visual Studio, chúng ta sẽ có sẵn được một cơ sở dữ liệu và các bảng để nạp dữ liệu.

- Tạo metadata



```
1 CREATE DATABASE Metadata_Group01
2 GO
3 USE Metadata_Group01
4 GO
5
6 DROP TABLE DATA_FLOW
7
8 -- Tạo metadata
9 CREATE TABLE DATA_FLOW (
10     ID int not null identity(1, 1),
11     TABLE_NAME nvarchar(50) null,
12     CET datetime null,
13     LSET datetime null
14 )
15 GO
```

The screenshot shows the SQL Server Management Studio interface. On the left, the code editor displays the creation of a database 'Metadata_Group01' and a table 'DATA_FLOW' with specific columns and constraints. On the right, the 'Results' tab shows the output of the execution, displaying two rows of data in a table format:

ID	TABLE_NAME	CET	LSET
1	STATE_AQI	2024-11-03 13:45:06.253	2024-11-03 13:45:06.253
2	USCOUNTIES	2024-11-03 13:45:06.257	2024-11-03 13:45:06.257

- Tạo sơ sở dữ liệu stage

```
1 CREATE DATABASE Stage_Group01
2 GO
3 USE Stage_Group01
4 GO
5
6 DROP TABLE STATE_AQI_STAGE
7 DROP TABLE USCOUNTIES_STAGE
8
9 -- Tạo database lưu Stage
10 CREATE TABLE STATE_AQI_STAGE (
11     ID int not null identity(1, 1),
12     state_name varchar(50) null,
13     county_name varchar(50) null,
14     state_code int null,
15     county_code int null,
16     . . .
17 )
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
259
260
261
262
263
264
265
266
267
268
269
269
270
271
272
273
274
275
276
277
278
279
279
280
281
282
283
284
285
286
287
287
288
289
289
290
291
292
293
294
295
296
297
297
298
299
299
300
301
302
303
304
305
306
307
308
309
309
310
311
312
313
314
315
316
317
318
319
319
320
321
322
323
324
325
326
327
328
329
329
330
331
332
333
334
335
336
337
338
339
339
340
341
342
343
344
345
346
347
348
349
349
350
351
352
353
354
355
356
357
358
359
359
360
361
362
363
364
365
366
367
368
369
369
370
371
372
373
374
375
376
377
378
379
379
380
381
382
383
384
385
386
387
388
389
389
390
391
392
393
394
395
396
397
398
399
399
400
401
402
403
404
405
406
407
408
409
409
410
411
412
413
414
415
416
417
418
419
419
420
421
422
423
424
425
426
427
428
429
429
430
431
432
433
434
435
436
437
438
439
439
440
441
442
443
444
445
446
447
448
449
449
450
451
452
453
454
455
456
457
458
459
459
460
461
462
463
464
465
466
467
468
469
469
470
471
472
473
474
475
476
477
478
479
479
480
481
482
483
484
485
486
487
488
489
489
490
491
492
493
494
495
496
497
497
498
499
499
500
501
502
503
504
505
506
507
508
509
509
510
511
512
513
514
515
516
517
518
519
519
520
521
522
523
524
525
526
527
528
529
529
530
531
532
533
534
535
536
537
538
539
539
540
541
542
543
544
545
546
547
548
549
549
550
551
552
553
554
555
556
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
597
598
599
599
600
601
602
603
604
605
606
607
608
609
609
610
611
612
613
614
615
616
617
618
619
619
620
621
622
623
624
625
626
627
628
629
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
697
698
699
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
788
788
789
789
790
791
792
793
794
795
796
797
797
798
799
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
879
879
880
881
882
883
884
885
886
887
888
888
889
889
890
891
892
893
894
895
896
897
897
898
899
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
918
919
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
948
949
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
968
969
969
970
971
972
973
974
975
976
977
978
979
979
980
981
982
983
984
985
986
987
987
988
989
989
990
991
992
993
994
995
995
996
997
997
998
999
999
1000
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1088
1089
1089
1090
1091
1092
1093
1094
1095
1096
1096
1097
1097
1098
1099
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1148
1149
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1178
1179
1179
1180
1181
1182
1183
1184
1185
1186
1187
1187
1188
1188
1189
1189
1190
1191
1192
1193
1194
1195
1195
1196
1196
1197
1197
1198
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1278
1279
1279
1280
1281
1282
1283
1284
1285
1286
1287
1287
1288
1288
1289
1289
1290
1291
1292
1293
1294
1295
1295
1296
1296
1297
1297
1298
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1378
1379
1379
1380
1381
1382
1383
1384
1385
1386
1387
1387
1388
1388
1389
1389
1390
1391
1392
1393
1394
1395
1395
1396
1396
1397
1397
1398
1398
1399
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1418
1419
1419
1420
1421
1422
1423
1424
1425
1426
1427
1427
1428
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1436
1437
1437
1438
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1477
1478
1478
1479
1479
1480
1481
1482
1483
1484
1485
1486
1487
1487
1488
1488
1489
1489
1490
1491
1492
1493
1494
1495
1495
1496
1496
1497
1497
1498
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1518
1519
1519
1520
1521
1522
1523
1524
1525
1526
1527
1527
1528
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1536
1537
1537
1538
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1547
1548
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1577
1578
1578
1579
1579
1580
1581
1582
1583
1584
1585
1586
1587
1587
1588
1588
1589
1589
1590
1591
1592
1593
1594
1595
1595
1596
1596
1597
1597
1598
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1617
1618
1618
1619
1619
1620
1621
1622
1623
1624
1625
1626
1627
1627
1628
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1636
1637
1637
1638
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1647
1648
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1677
1678
1678
1679
1679
1680
1681
1682
1683
1684
1685
1686
1687
1687
1688
1688
1689
1689
1690
1691
1692
1693
1694
1695
1695
1696
1696
1697
1697
1698
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1717
1718
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1727
1728
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1736
1737
1737
1738
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1747
1748
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1777
1778
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1787
1788
1788
1789
1789
1790
1791
1792
1793
1794
1795
1795
1796
1796
1797
1797
1798
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1817
1818
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1827
1828
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1836
1837
1837
1838
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1847
1848
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1877
1878
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1887
1888
1888
1889
1889
1890
1891
1892
1893
1894
1895
1895
1896
1896
1897
1897
1898
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1917
1918
1918
1919
1919
1920
1921
1922
1923
1924
1925
1926
1927
1927
1928
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1936
1937
1937
1938
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1947
1948
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1977
1978
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1987
1988
1988
1989
1989
1990
1991
1992
1993
1994
1995
1995
1996
1996
1997
1997
1998
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2009
2010
2011
2012
2013
2014
2015
2016
2017
2017
2018
2018
2019
2019
2020
2021
2022
2023
2024
2025
2026
2027
2027
2028
2028
2029
2029
2030
2031
2032
2033
2034
2035
2036
2036
2037
2037
2038
2038
2039
2039
2040
2041
2042
2043
2044
2045
2046
2047
2047
2048
2048
2049
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2068
2069
2069
2070
2071
2072
2073
2074
2075
2076
2077
2077
2078
2078
2079
2079
2080
2081
2082
2083
2084
2085
2086
2087
2087
2088
2088
2089
2089
2090
2091
2092
2093
2094
2095
2095
2096
2096
2097
2097
2098
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2117
2118
2118
2119
2119
2120
2121
2122
2123
2124
2125
2126
2127
2127
2128
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2136
2137
2137
2138
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2147
2148
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2168
2169
2169
2170
2171
2172
2173
2174
2175
2176
2177
2177
2178
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2187
2188
2188
2189
2189
2190
2191
2192
2193
2194
2195
2195
2196
2196
2197
2197
2198
2198
2199
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2217
2218
2218
2219
2219
2220
2221
2222
2223
2224
2225
2226
2227
2227
2228
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2236
2237
2237
2238
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2247
2248
2248
2249
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2268
2269
2269
2270
2271
2272
2273
2274
2275
2276
2277
2277
2278
2278
2279
2279
2280
2281
2282
2283
2284
2285
22
```

- Tạo cơ sở dữ liệu NDS

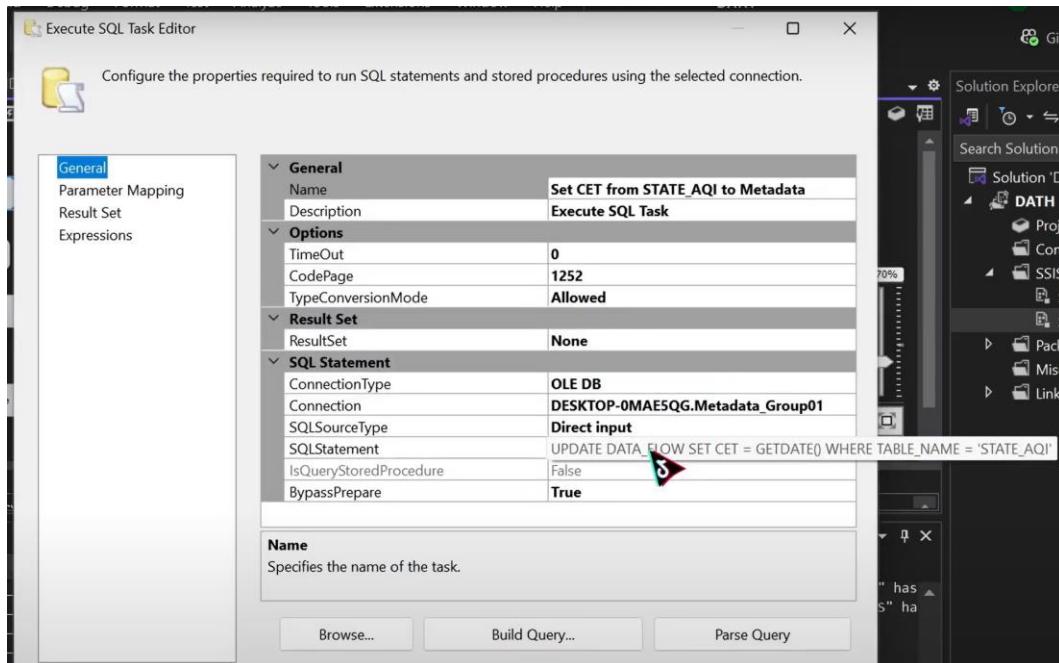
```
13 state_id nvarchar(5) not null,
14 state_name nvarchar(50) null,
15 created datetime null,
16 last_updated datetime null,
17 source_id int null
18 )
19 GO
20
21 CREATE TABLE COUNTY_NDS (
22 county_id_sk int PRIMARY KEY identity(1, 1),
23 county_code int not null,
24 county_name nvarchar(50) null,
25 county_ascii nvarchar(50) null,
26 county_full nvarchar(50) null,
27 county_fips nvarchar(10) null,
```

Tiếp theo, ta bắt đầu xây dựng qui trình ETL trong Visual Studio như sau:

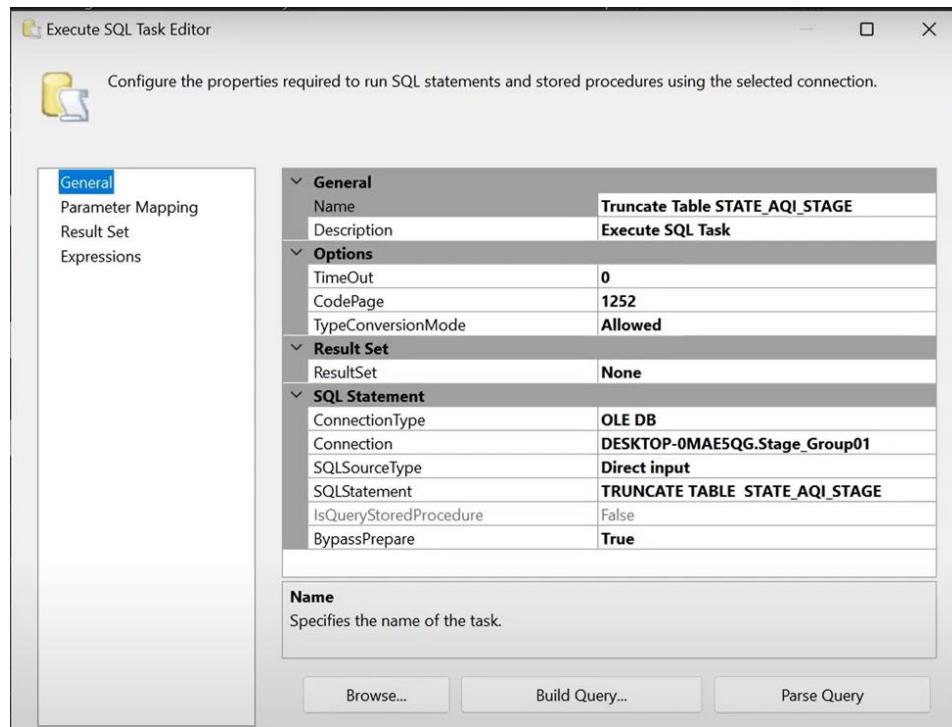


Trong đó:

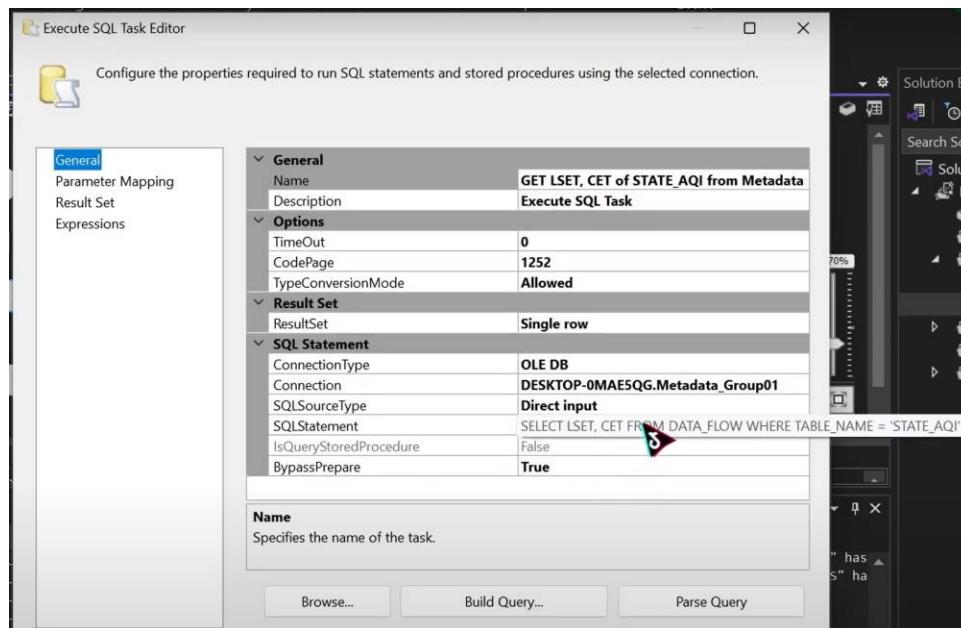
- Component Execute SQL Task mang tên “Set CET from _____ to Metadata” dùng để gán CET giá trị của thời điểm bắt đầu chạy qui trình.



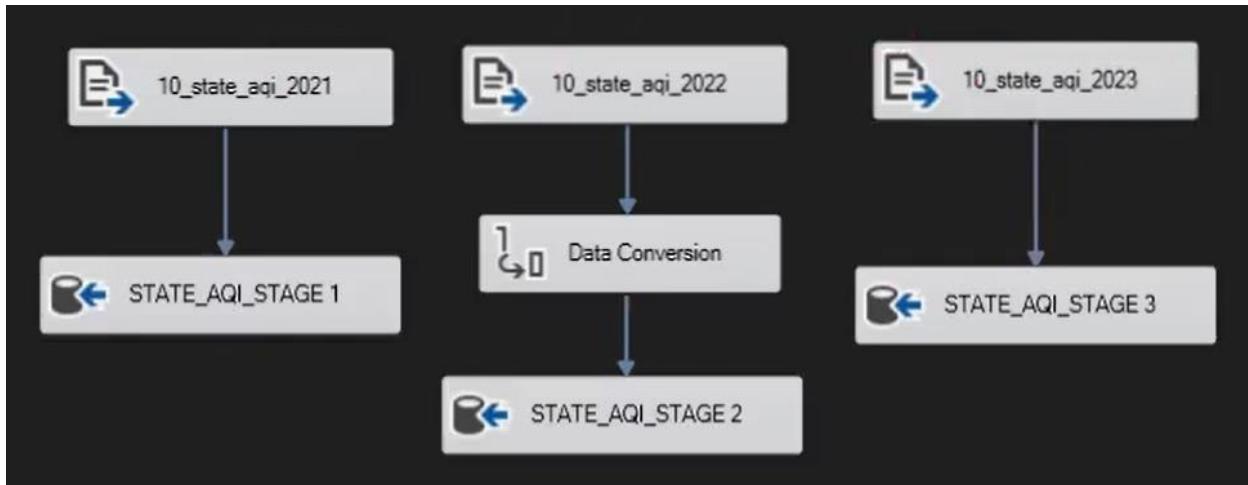
- Component Execute SQL Task mang tên “Truncate Table” dùng để xóa bảng có sẵn trong cơ sở dữ liệu SQL Server đang được kết nối.



- Component Execute SQL Task mang tên “GET LSET, CET” dùng để lấy giá trị LSET, CET trong bảng Stage.



- Khi nhấp vào component Data Flow Task “Import data from _____”, ta xây dựng các chiêu nạp dữ liệu từ flat file (file .csv) vào bảng trong cơ sở dữ liệu. Dưới đây là kết quả sau khi nhấp. Ngoài ra ta còn thấy trong sơ đồ này còn có một component khác là Data Conversion được dùng để biến đổi kiểu dữ liệu khi kiểu dữ liệu trong file .csv không phù hợp với cơ sở dữ liệu.



1 Data Conversion Transformation Editor

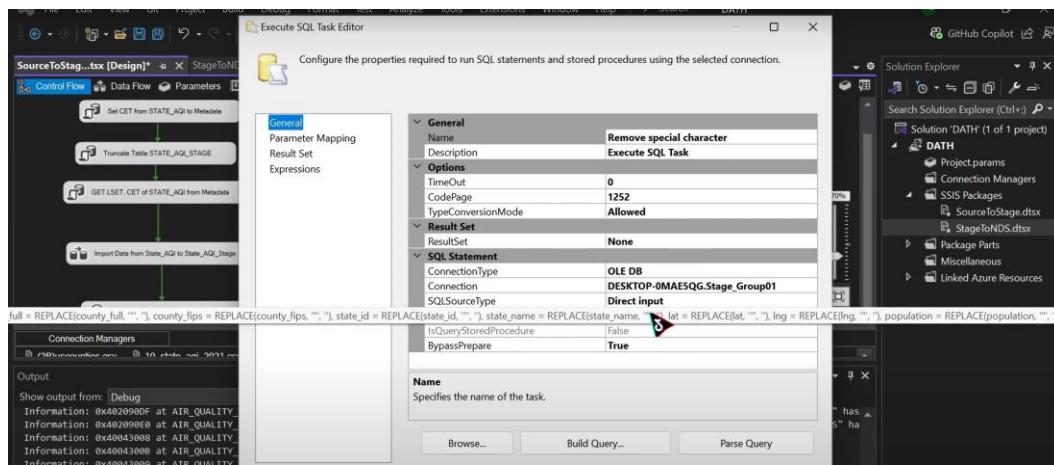
Configure the properties used to convert the data type of an input column to a different data type. Depending on the data type to which the column is converted, set the length, precision, scale, and code page of the column.

Input Column	Output Alias	Data Type	Length	Precision	Scale	Code Page
State Name	Copy of State Name	string [DT_STR]	50			1252 (ANSI)
county Name	Copy of county Name	string [DT_STR]	50			1252 (ANSI)
Category	Copy of Category	string [DT_STR]	50			1252 (ANSI)
Defining Parameter	Copy of Defining Par...	string [DT_STR]	50			1252 (ANSI)
Defining Site	Copy of Defining Site	string [DT_STR]	50			1252 (ANSI)

Available Input Columns:

- Name
- State Name**
- county Name
- State Code
- County Code
- Date
- AQI
- Category

- Component Execute SQL Task mang tên “Remove special character” được dùng để xóa các kí tự đặc biệt có trong dữ liệu từ nguồn.



Sau khi hoàn thành xong xây dựng qui trình từ ETL Source vào Stage, ta thu được kết quả như sau:

```

28
29     INSERT INTO DATA_FLOW (TABLE_NAME, CET, LSET) VALUES ('STATE_AQI', NULL, NULL);
30     INSERT INTO DATA_FLOW (TABLE_NAME, CET, LSET) VALUES ('USCOUNTIES', NULL, NULL);
31     GO
32
33     SELECT * FROM DATA_FLOW
34     GO

```

ID	TABLE_NAME	CET	LSET
1	STATE_AQI	2024-11-03 13:51:42.417	2024-11-03 13:52:43.850
2	USCOUNTIES	2024-11-03 13:51:42.407	2024-11-03 13:52:11.013

```

45     SELECT DISTINCT COUNT(*) FROM STATE_AQI_STAGE
46     SELECT DISTINCT COUNT(*) FROM USCOUNTIES_STAGE
47     GO

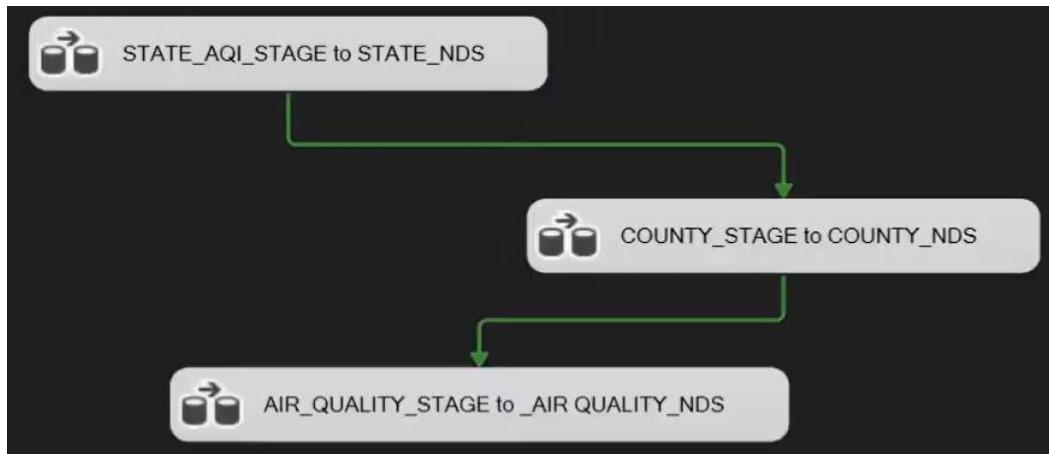
```

(No column name)
194971

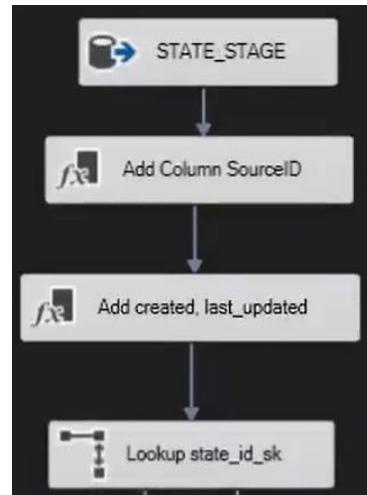
(No column name)
3144

b) Stage đến NDS

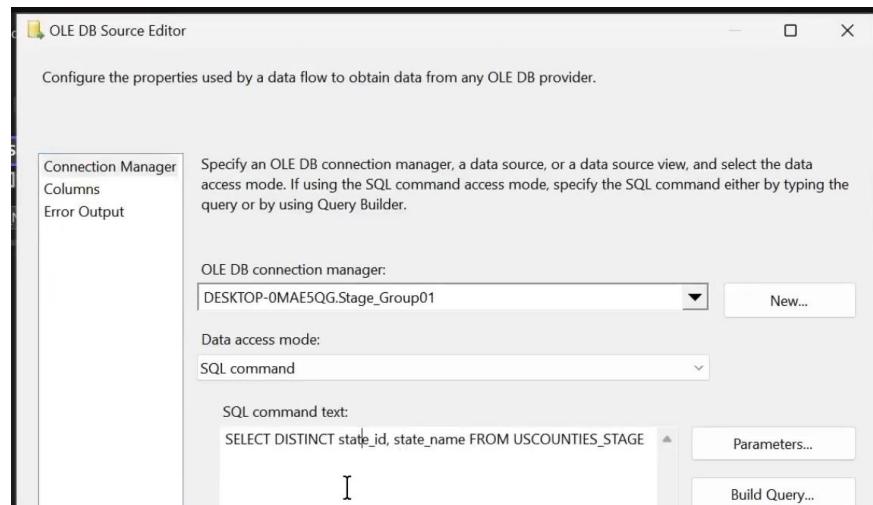
Tại bước này, ta lập các component Data Flow Task như sau để nạp dữ liệu theo thứ tự STATE > COUNTY > AIR_QUALITY



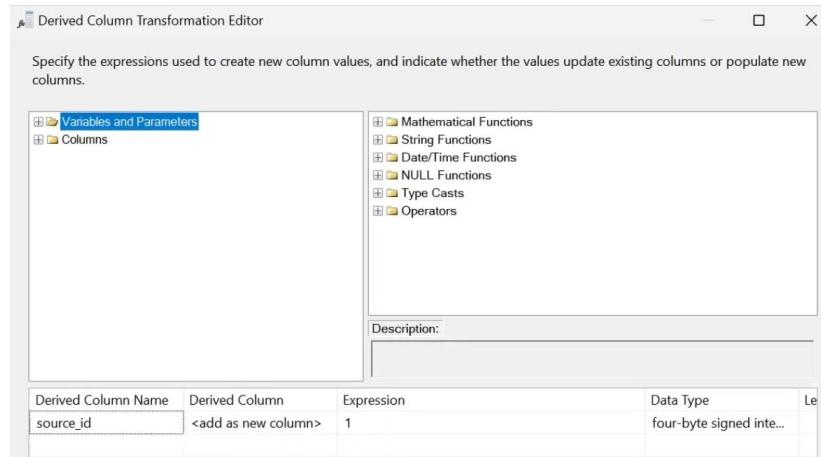
- Component Data Flow Task tên “STATE_AQI_STAGE to STATE_NDS”:



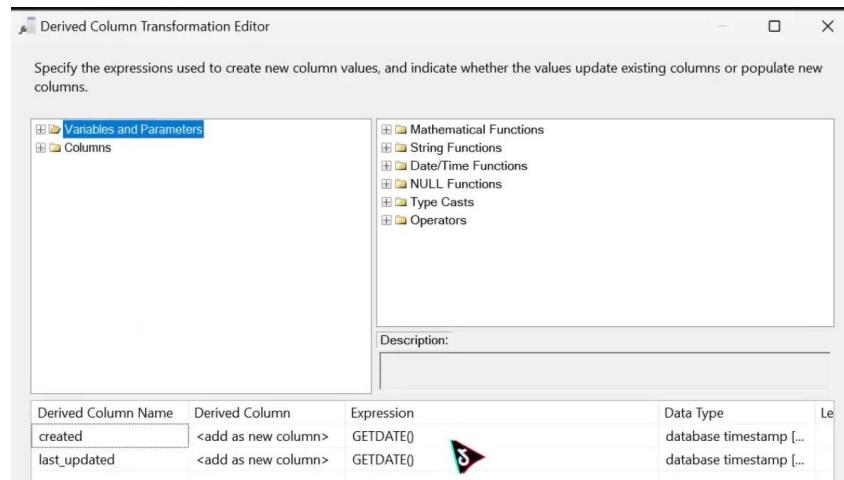
- o Để thiết lập qui trình nạp dữ liệu cho bảng STATE_NDS, ta thiết lập nguồn dữ liệu như sau để lấy dữ liệu từ bảng Stage vào bảng NDS.



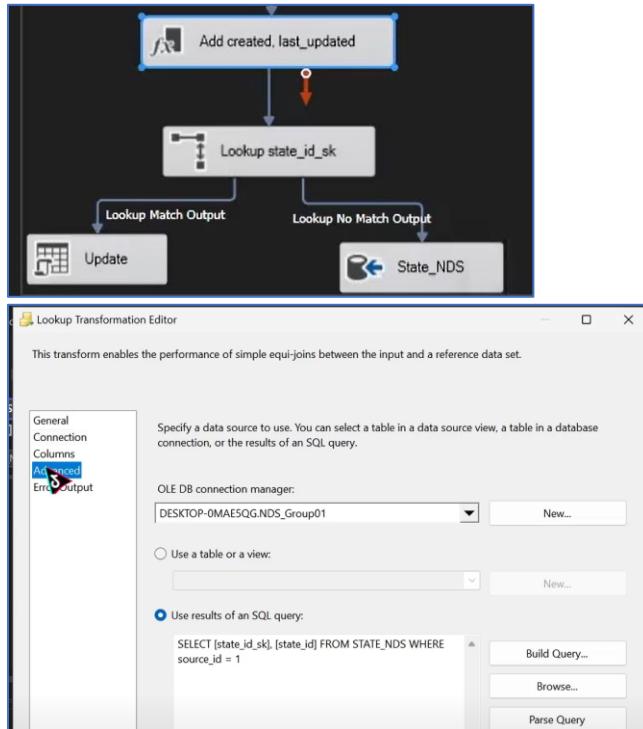
- Tại component Drived Column tên “Add Column SourceID”, ta thêm vào một cột tên “source_id”.



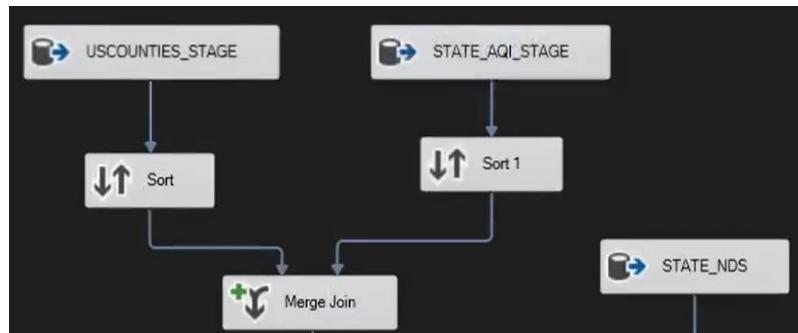
- Tương tự tại component “Add created, last_updated”, ta thêm vào cột “created” và “last_updated”.



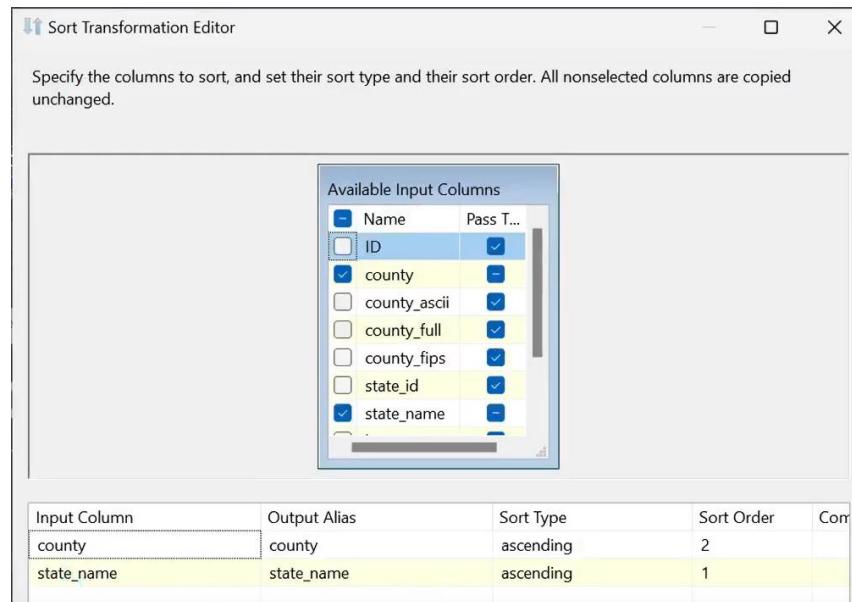
- Tại component Lookup tên “Lookup state_id_sk”, chúng ta thiết lập câu lệnh để có thể cập nhật đúng thông tin bản ghi hoặc thêm vào dòng mới trong bảng.



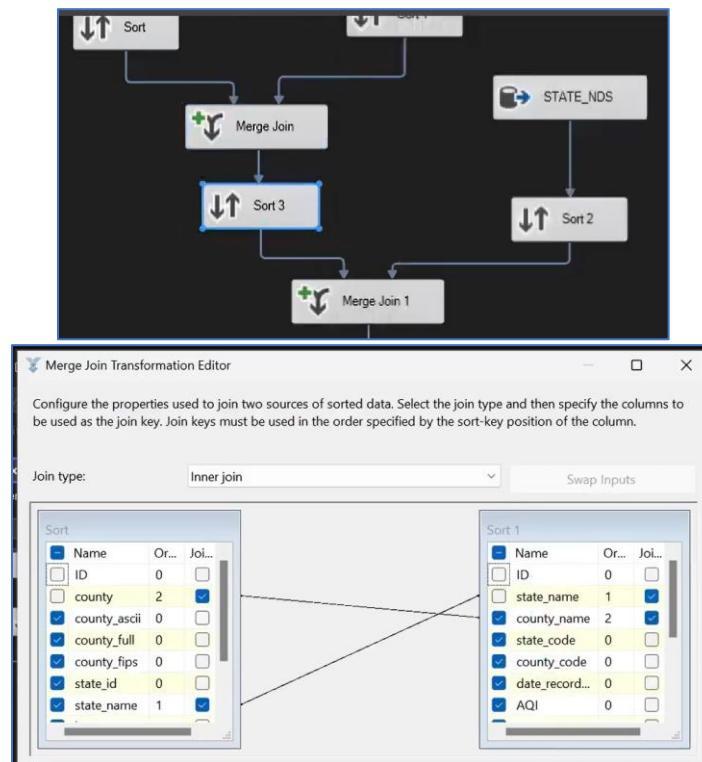
- Sau khi thiết lập Lokup theo state_id_sk và state_id, nếu giá trị cả hai trường này trong nguồn không trùng với bản ghi nào trong bảng NDS thì bảng NDS sẽ tự thêm một dòng mới, còn ngược lại, thì bảng NDS sẽ cập nhật lại giá trị của bản ghi có state_id_sk và state_id trùng với câu lệnh.
- Component Data Flow Task tên “COUNTY_STAGE to COUNTY_NDS”:



- Tại đây, ta thấy trong sơ đồ này có sử dụng component Sort để sắp xếp dữ liệu theo giá trị cột state_name và county.



- Tại component Merge Join tên “Merge Join”, ta ghép hai bảng USCOUNTIES_STAGE và STATE_AQI_STAGE dựa vào county_name và state_name. Tương tự với component “Merge Join 1”, ta cũng ghép bảng STATE_NDS với bảng vừa được ghép tại “Merge Join” dựa vào state_id và state_name.



- Phân cõi lại của qui trình này sẽ tương tự như Data Flow Task tên “STATE_AQI_STAGE to STATE_NDS”.

Kết quả sau khi chạy qui trình:

```

54
55 SELECT* FROM AIR_QUALITY_NDS
56 SELECT* FROM COUNTY_NDS
57 SELECT* FROM STATE_NDS
58 GO
59
60 SELECT DISTINCT COUNT(*) FROM AIR_QUALITY_NDS
61 SELECT DISTINCT COUNT(*) FROM COUNTY_NDS
62 SELECT DISTINCT COUNT(*) FROM STATE_NDS
63 GO

```

air_quality_id_sk	county_id_sk	date_recorded	aqi	category	defining_parameter	defining_site	number_of_sites_reporting	created	last_updated
1	3	2022-10-31 00:00:00.000	26	Good	Ozone	18-003-0002	1	2024-11-03 14:05:55.030	2024-11-03
2	75	2023-04-16 00:00:00.000	56	Good	Ozone	06-075-0005	1	2024-11-03 14:05:55.030	2024-11-03
3	7	2021-10-02 00:00:00.000	45	Good	Ozone	04-007-0010	4	2024-11-03 14:05:55.030	2024-11-03
4	255	2021-04-02 00:00:00.000	7	Good	NO2	48-255-1070	1	2024-11-03 14:05:55.030	2024-11-03
5	157	2022-09-22 00:00:00.000	39	Good	Ozone	17-157-0001	1	2024-11-03 14:05:55.030	2024-11-03
6	41	2021-03-20 00:00:00.000	37	Good	Ozone	06-041-0001	2	2024-11-03 14:05:55.030	2024-11-03

county_id_sk	county_code	county_name	county_ascii	county_full	county_tips	state_id_sk	lat	lng	population	created	last_updated	source_
8	119	Sumter	Sumter	Sumter County	01119	2	32.5911	-88.1988	12196	2024-11-03 14:04:30.737	2024-11-03 14:04:30.737	1
9	119	Sumter	Sumter	Sumter County	01119	2	32.5911	-88.1988	12196	2024-11-03 14:04:30.737	2024-11-03 14:04:30.737	1
10	119	Sumter	Sumter	Sumter County	01119	2	32.5911	-88.1988	12196	2024-11-03 14:04:30.737	2024-11-03 14:04:30.737	1
11	119	Sumter	Sumter	Sumter County	01119	2	32.5911	-88.1988	12196	2024-11-03 14:04:30.737	2024-11-03 14:04:30.737	1
12	119	Sumter	Sumter	Sumter County	01119	2	32.5911	-88.1988	12196	2024-11-03 14:04:30.737	2024-11-03 14:04:30.737	1

state_id_sk	state_id	state_name	created	last_updated	source_id
1	AK	Alaska	2024-11-03 14:02:24.773	2024-11-03 14:02:24.773	1
2	AL	Alabama	2024-11-03 14:02:24.773	2024-11-03 14:02:24.773	1
3	AR	Arkansas	2024-11-03 14:02:24.773	2024-11-03 14:02:24.773	1
4	AZ	Arizona	2024-11-03 14:02:24.773	2024-11-03 14:02:24.773	1
5	CA	California	2024-11-03 14:02:24.773	2024-11-03 14:02:24.773	1

5. Phân tích qui trình NDS vào DDS

a) Các nhu cầu cần phân tích

i. Report the min and max of AQI value for each State during each quarter of years.

- Sự kiện: Chỉ số AQI được ghi nhận cho từng bang trong mỗi quý.
- Bối cảnh:
 - Ở đâu: Các khu vực trong từng bang
 - Cái gì: Chỉ số AQI
 - Khi nào: Trong mỗi quý của các năm
- Đo lường:
 - Giá trị AQI tối thiểu và tối đa - Tìm giá trị nhỏ nhất và lớn nhất của AQI cho từng bang trong từng quý.

ii. Report the mean and the standard deviation of AQI value for each State during each quarter of years.

- Sự kiện: Tính toán các chỉ số AQI trung bình và độ lệch chuẩn cho từng bang theo từng quý.
- Bối cảnh:
 - Ở đâu: Các khu vực trong từng bang
 - Cái gì: Chỉ số AQI

- Khi nào: Trong mỗi quý của các năm
 - Đo lường:
 - Giá trị trung bình và độ lệch chuẩn của AQI: Tính giá trị trung bình (mean) và độ lệch chuẩn (standard deviation) cho từng bang trong từng quý.
- iii. **Report the number of days, and the mean AQI value where the air quality is rated as "Very Unhealthy" or worse for each State and County.**
- Sự kiện: AQI được ghi nhận khi chất lượng không khí được đánh giá là "Very Unhealthy" hoặc tệ hơn tại từng quận thuộc các bang.
 - Bối cảnh:
 - Ở đâu: Các quận trong từng bang
 - Cái gì: Chỉ số AQI
 - Khi nào: Mỗi ngày
 - Đo lường:
 - Số ngày AQI được xếp loại "Very Unhealthy" hoặc tệ hơn: Đếm số ngày AQI vượt ngưỡng của mức này.
 - Giá trị AQI trung bình trong những ngày này: Tính giá trị trung bình của AQI trong các ngày có chất lượng không khí là "Very Unhealthy" hoặc tệ hơn.
- iv. **For the four following states: Hawaii, Alaska, Illinois and Delaware, count the number of days in each air quality Category (Good, Moderate,etc.) by County.**
- Sự kiện: Khi một quận thuộc một quận trong bốn tiểu bang Hawaii, Alaska, Illinois, Delaware được đo lường chỉ số AQI trong ngày.
 - Bối cảnh:
 - Ở đâu: Mỗi quận trong các tiểu bang Hawaii, Alaska, Illinois, Delaware
 - Cái gì: Chỉ số AQI
 - Khi nào: Mỗi ngày
 - Đo lường:
 - Số lượng ngày theo từng loại chất lượng không khí (Air Quality Category)
- v. **For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the mean AQI value by quarters.**
- Sự kiện: Khi một quận thuộc một quận trong bốn tiểu bang Hawaii, Alaska, Illinois, Delaware được đo lường chỉ số AQI trong ngày.
 - Bối cảnh:
 - Ở đâu: Mỗi quận trong các tiểu bang Hawaii, Alaska, Illinois, Delaware

- Cái gì: Chỉ số AQI
 - Khi nào: Mỗi quý
 - Đo lường:
 - Chỉ số AQI trung bình của từng bang theo từng quý trong năm
- vi. Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California.**
- Sự kiện: Khi một quận thuộc một tiểu bang (Hawaii, Alaska, Illinois, California) được đo lường chỉ số AQI trong ngày.
 - Bối cảnh:
 - Ở đâu: Mỗi quận trong các tiểu bang Hawaii, Alaska, Illinois, California
 - Cái gì: Chỉ số AQI
 - Khi nào: Trong ngày
 - Đo lường:
 - Chỉ số AQI trung bình của từng bang theo từng năm
- vii. Report the mean, the standard deviation, min and max of AQI value group by State and County during each quarter of the year.**
- Sự kiện: Khi một quận thuộc một tiểu bang được đo lường chỉ số AQI trung bình trong một quý.
 - Bối cảnh:
 - Ở đâu: Mỗi quận thuộc một tiểu bang
 - Cái gì: Chỉ số AQI
 - Khi nào: Mỗi quý
 - Đo lường:
 - Chỉ số AQI trung bình, phương sai, nhỏ nhất và lớn nhất theo từng quận và từng tiểu bang trong mỗi quý của từng năm
- viii. Create a new attribute, DayLightSaving, in a suitable table. DayLightSaving may have two values:**
- True: Between March 12, 2023, and November 5, 2023**
- False: Otherwise**
- Report the mean AQI value by State, Category, DayLightSaving over years.**
- Sự kiện: Khi một quận thuộc một tiểu bang được đo lường chỉ số AQI trong ngày Day Light Saving.
 - Bối cảnh:
 - Ở đâu: Một quận thuộc một tiểu bang
 - Cái gì: Chỉ số AQI
 - Khi nào: Mỗi ngày Day Light Saving
 - Đo lường:
 - Chỉ số AQI trung bình của một tiểu bang trong ngày Day Light Saving

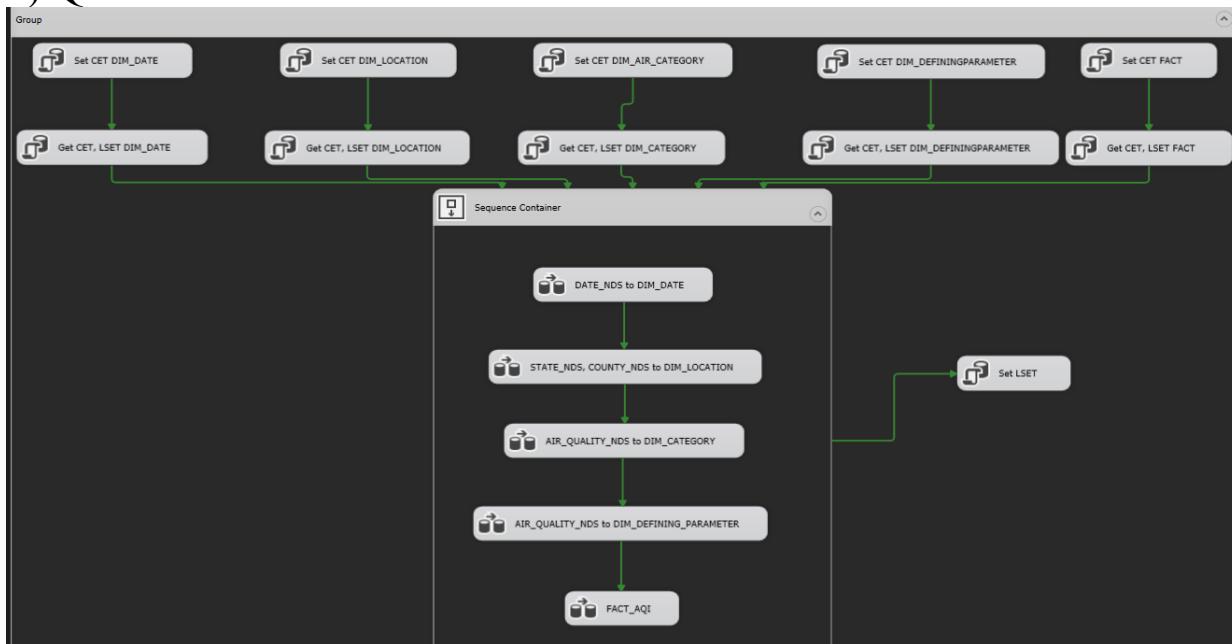
ix. Count the number of days by State, Category in each month.

- Sự kiện: Khi một tiểu bang được đếm số lượng ngày theo đánh giá chất lượng không khí.
- Bối cảnh:
 - Ở đâu: Một tiểu bang
 - Cái gì: Số lượng ngày
 - Khi nào: Trong tháng
- Đo lường:
 - Số lượng ngày trong tháng được đo lường chỉ số AQI

x. Report the number of days by Category and Defining Parameter.

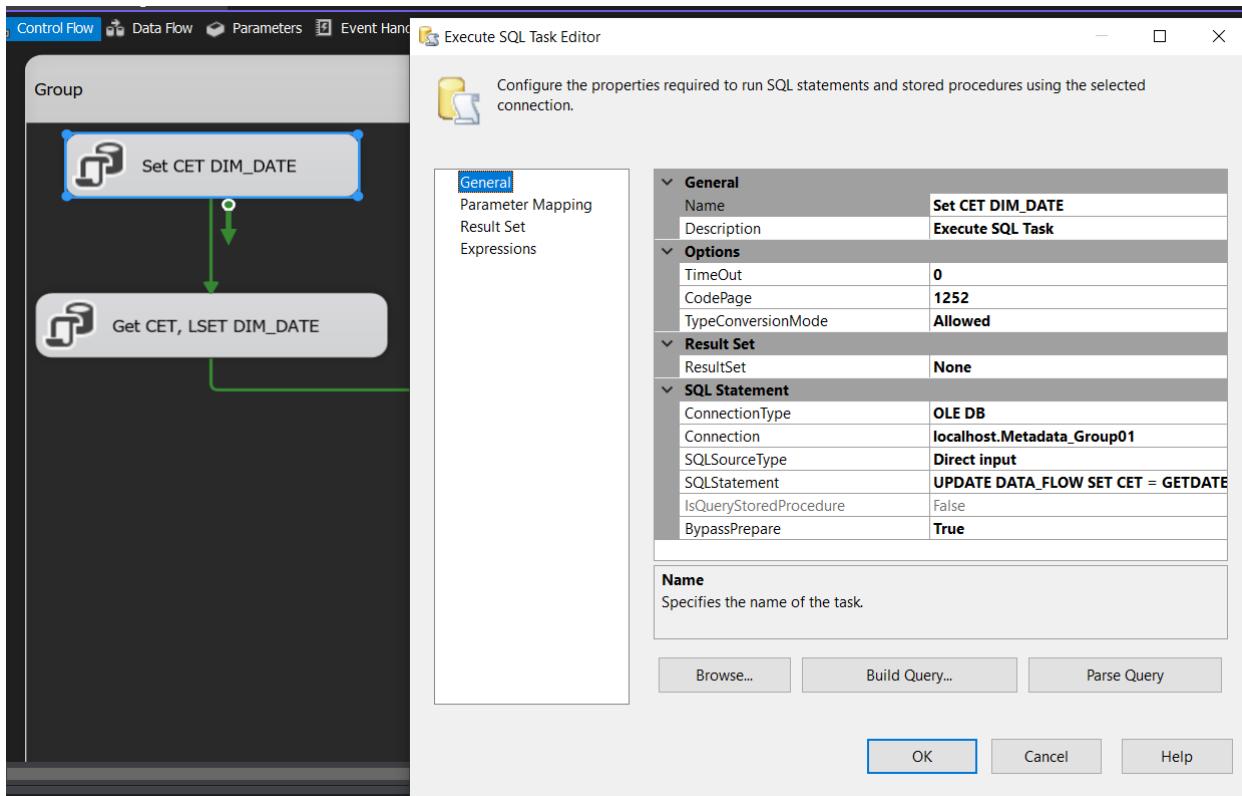
- Sự kiện: Khi dữ liệu được đếm số lượng ngày theo đánh giá chất lượng không khí và các thông số.
- Bối cảnh:
 - Ở đâu: Một tiểu quận
 - Cái gì: Số lượng ngày
 - Khi nào: Trong năm
- Đo lường:
 - Số lượng ngày trong năm được đánh giá theo phân loại không khí (Category) và thông số (Defining Parameter)

b) Qui trình NDS vào DDS

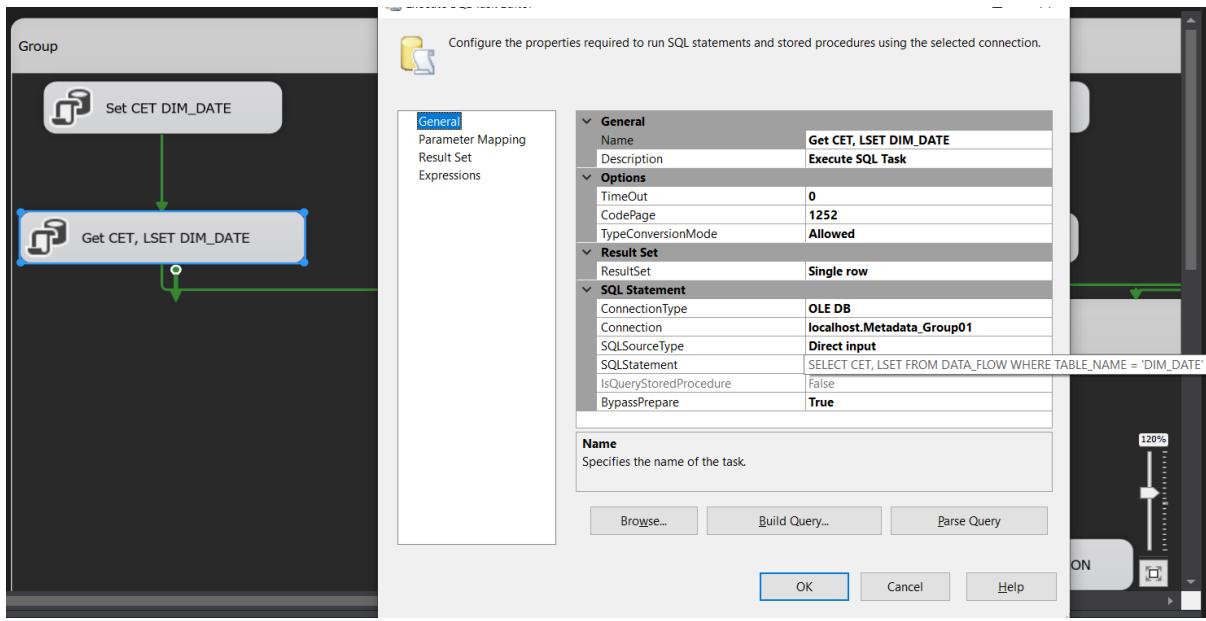


Qui trình thực hiện:

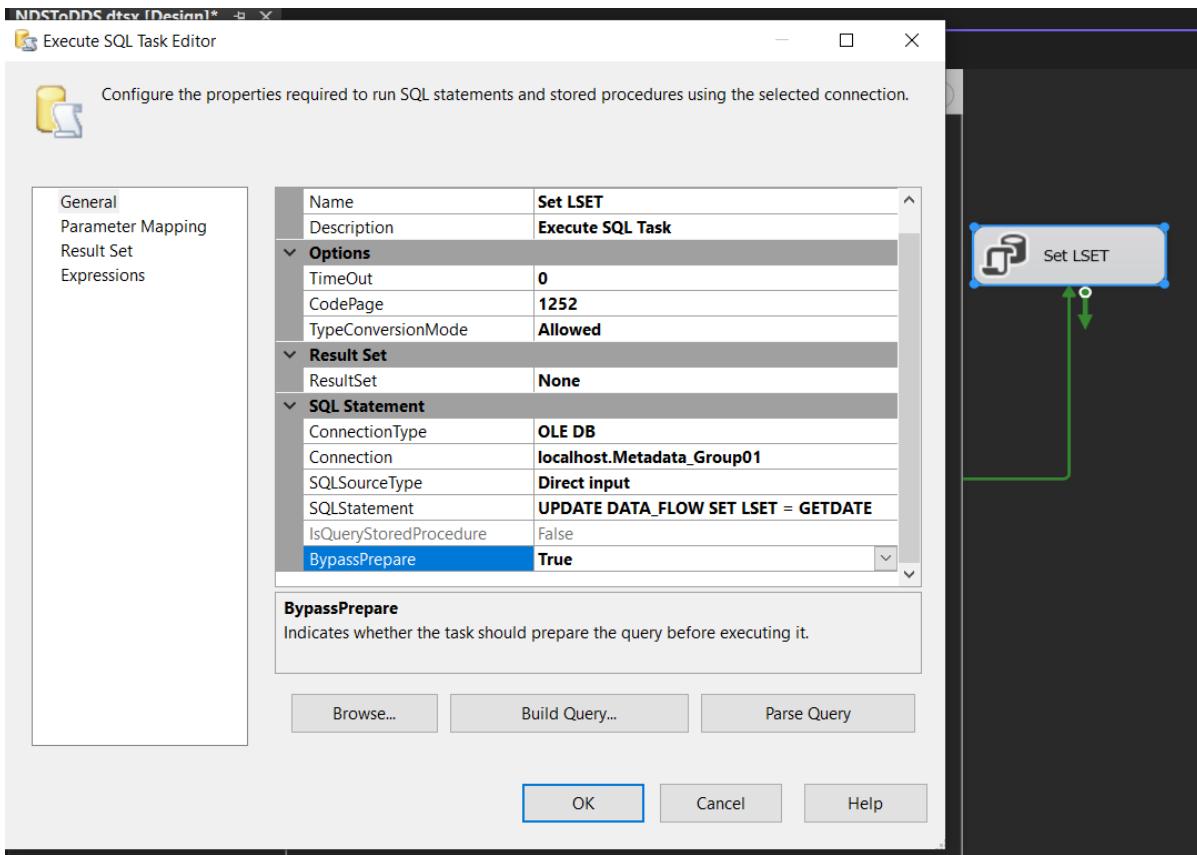
- Thêm DIM_DATE:
 - Cập nhật lại CET bằng GETDATE() cho DIM_DATE trong Metadata:



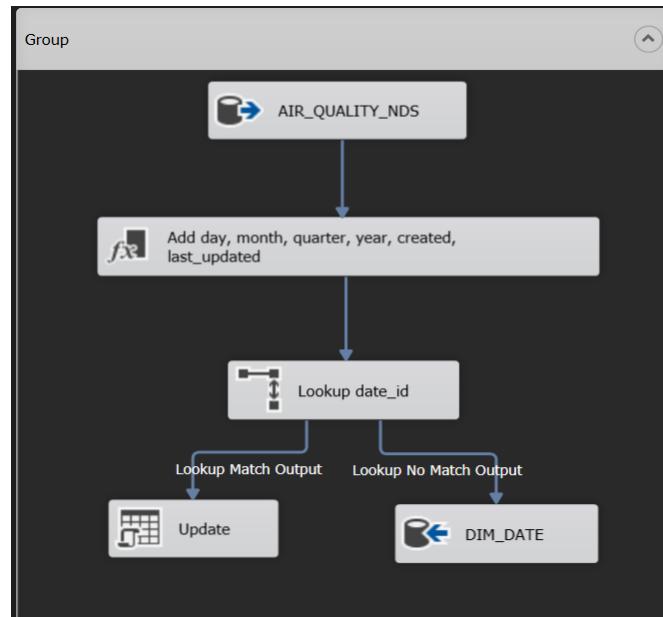
- Select CET, LCET của bảng DIM_DATE trong Metadata



- Sau khi nạp dữ liệu cho các bảng từ NDS vào DDS thì tạo cài đặt lại LSET = GETDATE():



- Ta thực hiện tương tự với bảng DIM_LOCATION, DIM_CATEGORY, DIM_DEFININGPARAMETER, FACT_AQI.
- Nạp dữ liệu từ AIR_QUALITY_NDS vào DIM_DATE:



- Đầu tiên, ta lấy dữ liệu thời gian từ AIR_QUALITY_NDS rồi tạo thêm các cột DAY, MONTH, QUARTER, YEAR,

DAY_LIGHT_SAVING từ cột DAY_RECORDING trong AIR_QUALITY_NDS:

Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

Variables and Parameters

Columns

Mathematical Functions

String Functions

Date/Time Functions

NULL Functions

Type Casts

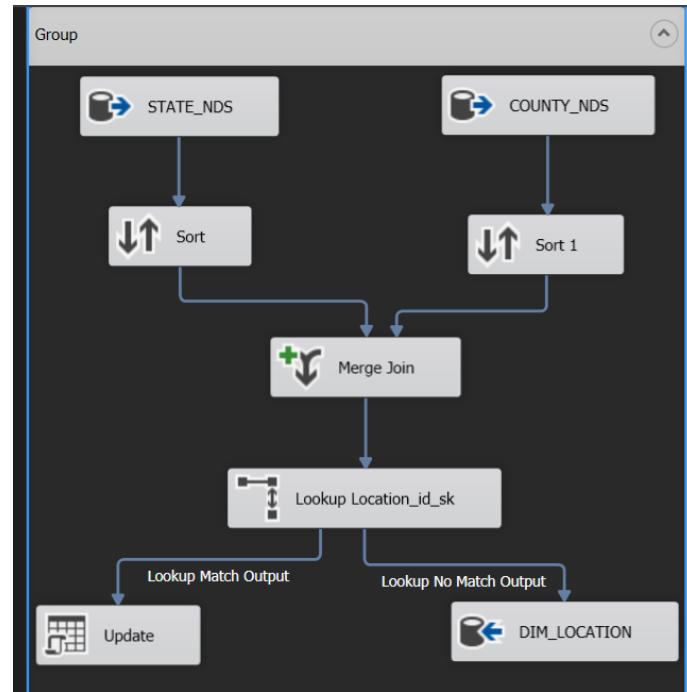
Operators

Description:

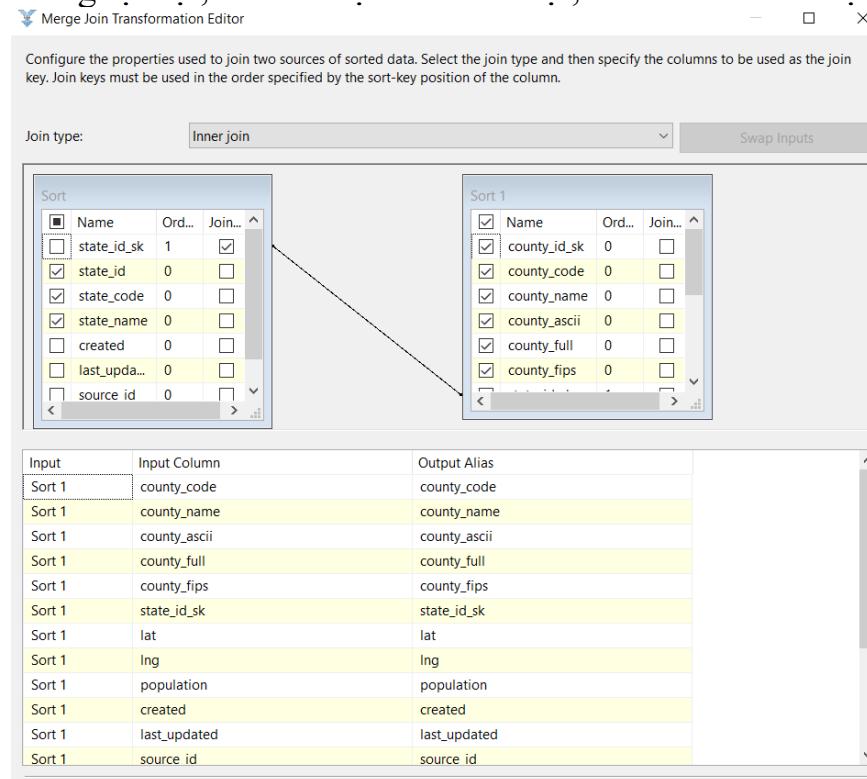
Derived Column Name	Derived Column	Expression	Data Type	Length
day	<add as new column>	DAY(date_recorded)	four-byte signed integer	4
month	<add as new column>	MONTH(date_recorded)	four-byte signed integer	4
quarter	<add as new column>	DATEPART("qq",date_recorded)	four-byte signed integer	4
year	<add as new column>	YEAR(date_recorded)	four-byte signed integer	4
created	<add as new column>	GETDATE()	database timestamp	8
last_updated	<add as new column>	GETDATE()	database timestamp	8
day_light_saving	<add as new column>	(DT_I4)(YEAR(date_recorded) == 2023 && ((M...))	four-byte signed integer	4

- Tiếp theo, ta tạo khối LOOKUP sử dụng date_id để thêm và cập nhật dữ liệu. Nếu dữ liệu tồn tại thì ta cập nhật lại dữ liệu, và ngược lại, nếu dữ liệu chưa tồn tại, ta thêm vào dữ liệu mới.

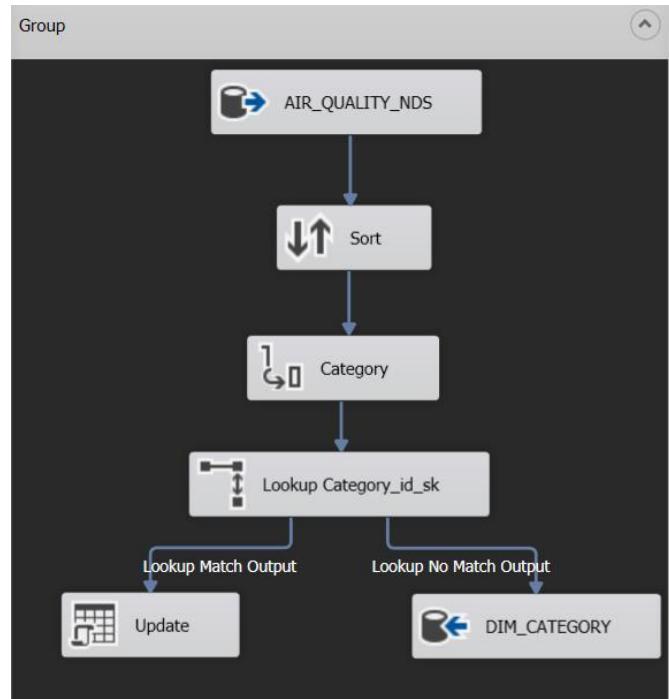
- Nạp dữ liệu từ STATE_NDS, COUNTY_NDS vào DIM_LOCATION:
 - o Ở đây, ta lấy dữ liệu từ STATE_NDS, COUNTY_NDS, sort lại theo state_id_sk và merge join 2 bảng lại.



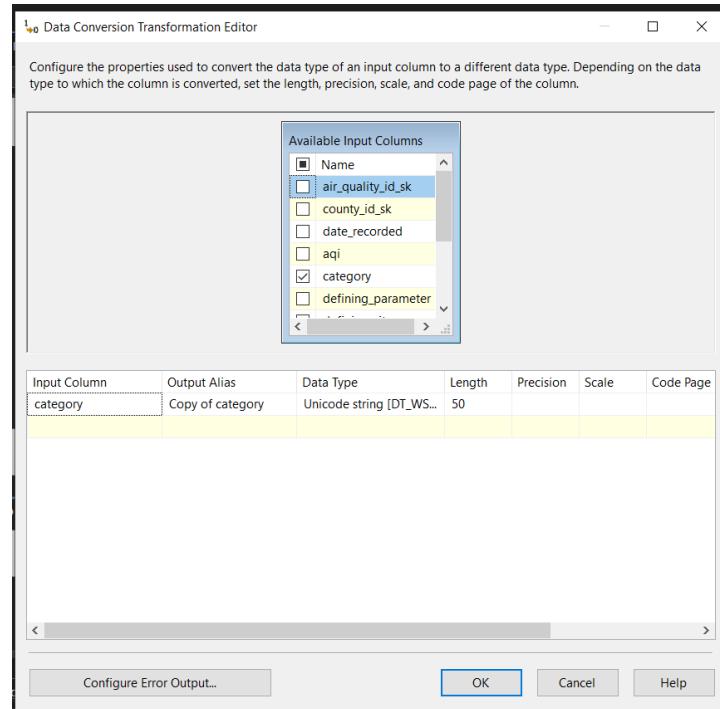
- o Tiếp theo, ta tạo khối LOOKUP sử dụng location_id_sk để thêm và cập nhật dữ liệu. Nếu dữ liệu tồn tại thì ta cập nhật lại dữ liệu, và ngược lại, nếu dữ liệu chưa tồn tại, ta thêm vào dữ liệu mới.



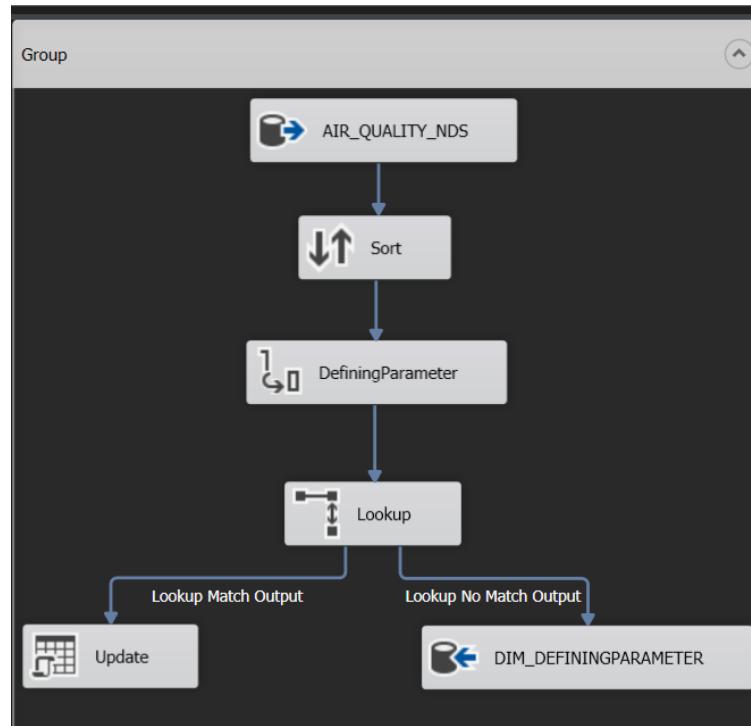
- Nạp dữ liệu từ AIR_QUALITY_NDS vào DIM_CATEGORY:
 - o Ở đây, ta lấy dữ liệu từ AIR_QUALITY_NDS, sort lại theo air_quality_id_sk, và convert kiểu dữ liệu cho category.



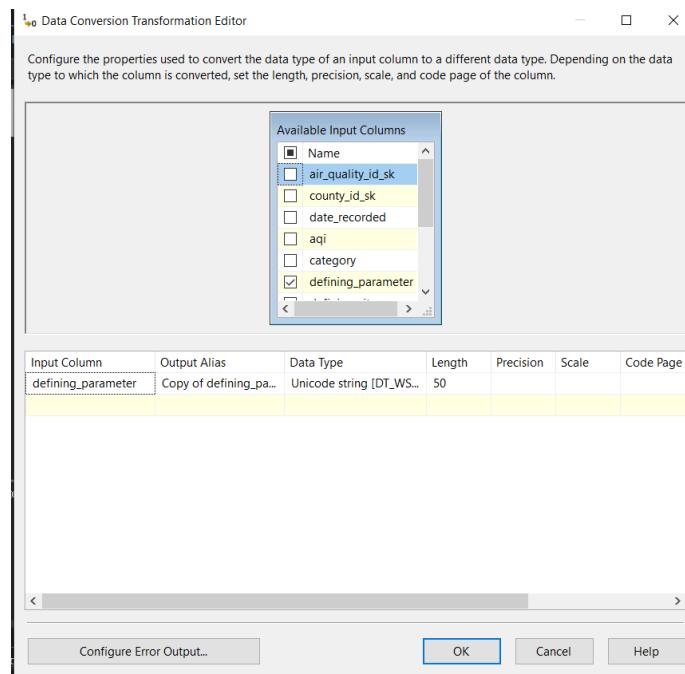
- o Tiếp theo, ta tạo khối LOOKUP sử dụng category_id_sk để thêm và cập nhật dữ liệu. Nếu dữ liệu tồn tại thì ta cập nhật lại dữ liệu, và ngược lại, nếu dữ liệu chưa tồn tại, ta thêm vào dữ liệu mới.



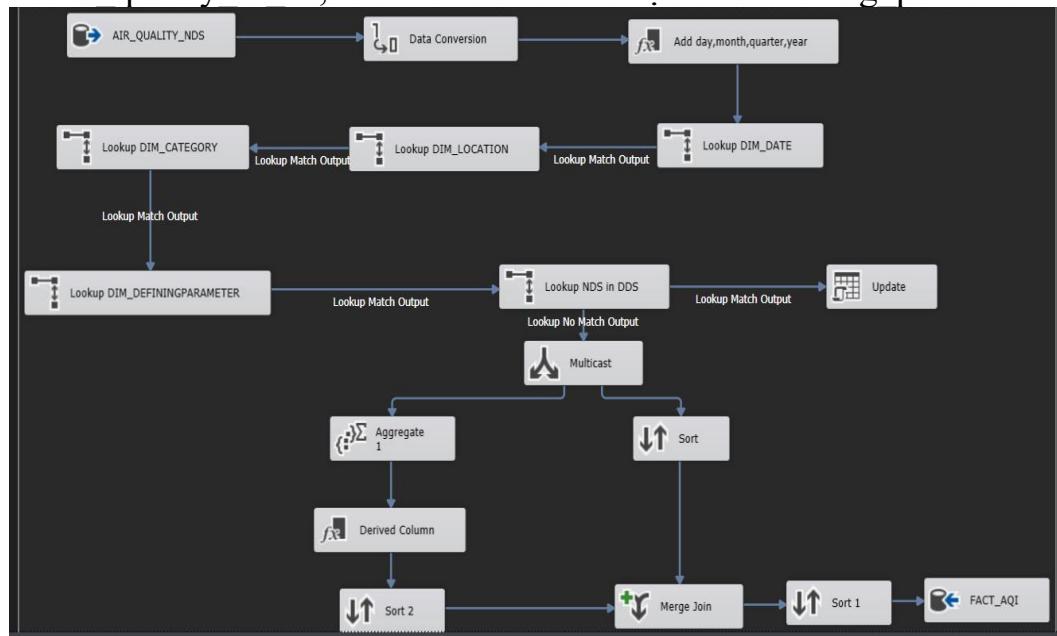
- Nạp dữ liệu từ AIR_QUALITY_NDS vào DIM_DEFINING_PARAMETER:
 - o Ở đây, ta lấy dữ liệu từ AIR_QUALITY_NDS, sort lại theo air_quality_id_sk, và convert kiểu dữ liệu cho defining_parameter.



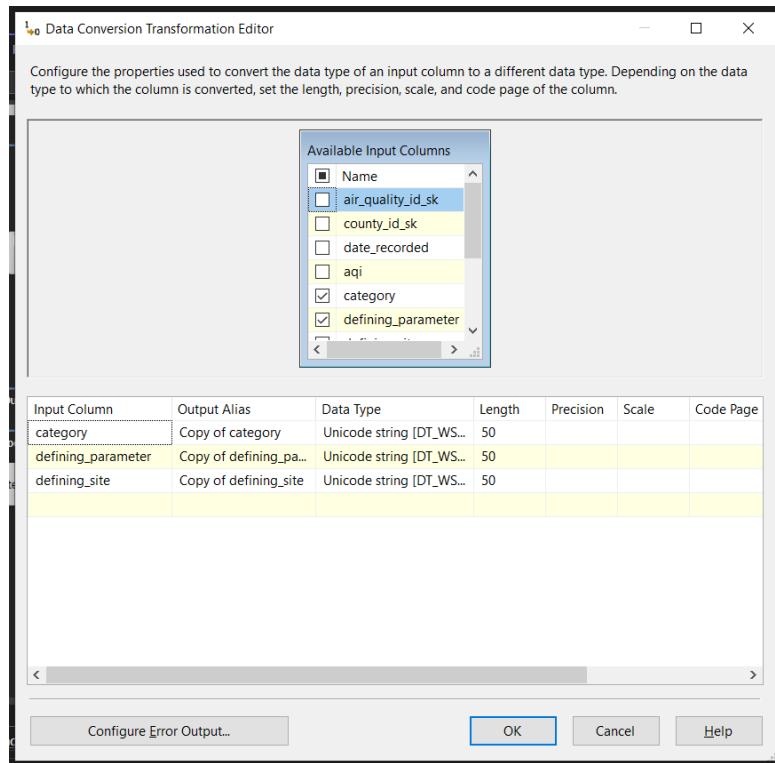
- o Tiếp theo, ta tạo khối LOOKUP sử dụng defining_parameter_id_sk để thêm và cập nhật dữ liệu. Nếu dữ liệu tồn tại thì ta cập nhật lại dữ liệu, và ngược lại, nếu dữ liệu chưa tồn tại, ta thêm vào dữ liệu mới.



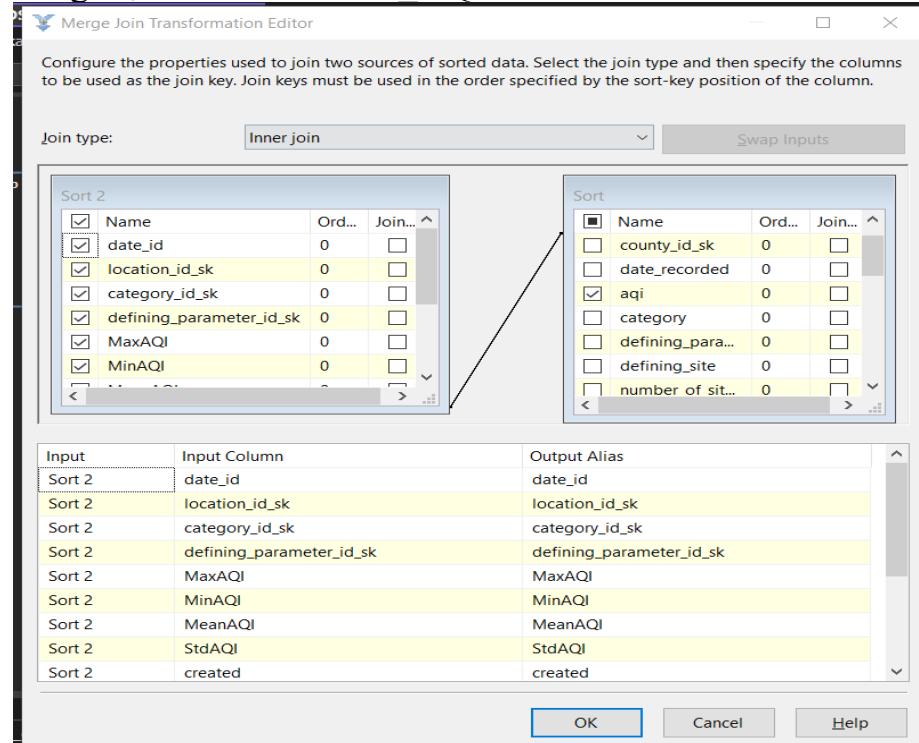
- Nạp dữ liệu từ AIR_QUALITY_NDS vào FACT_AQI:
 - o Ở đây, ta lấy dữ liệu từ AIR_QUALITY_NDS, sort lại theo air quality id sk, và convert kiểu dữ liệu cho defining parameter.



- o Tiếp theo, ta tạo khối LOOKUP lần lượt lookup dữ liệu với bảng DIM_DATE, DIM_LOCATION, DIM_CATEGORY, DIM_DEFINING_PARAMETER để lấy thông tin các trường date_id, category_id_sk, defining_parameter_id_sk, và location_id_sk. Nếu dữ liệu tồn tại thì cập nhật lại dữ liệu, còn nếu dữ liệu chưa tồn tại thì ta chuyển dữ liệu theo hai đường bằng Multicast và ở đường thứ nhất thì tạo khối Aggregate, để nhóm lại các dòng theo date_id, category_id_sk, defining_parameter_id_sk, location_id_sk và tính các giá trị như: MinAQI, MaxAQI, MeanAQI, StdAQI. Sau đó, ta tạo Created, Last_updated để lưu lại ngày tạo, ngày cập nhật.

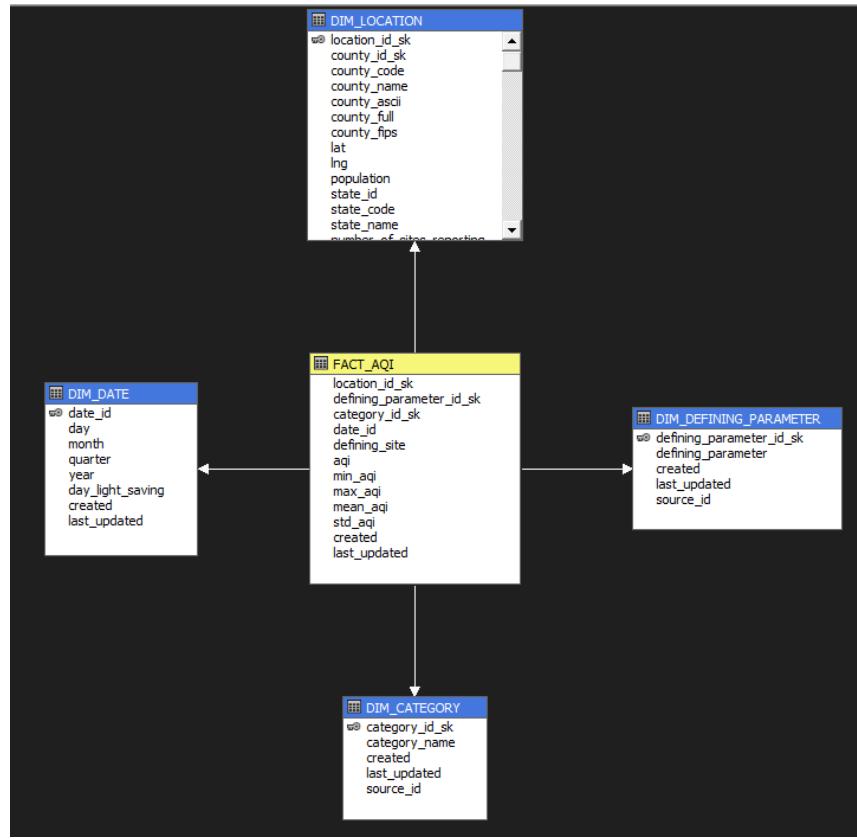


- Và ở đường thứ hai thì chuyển dữ liệu defining_site, aqi sau đó merge lại vào đó vào Fact AQI



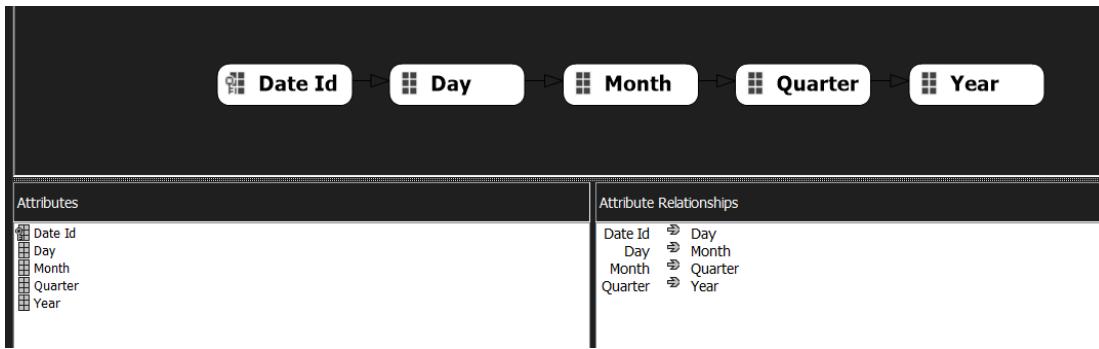
6. OLAP

- OLAP cube:

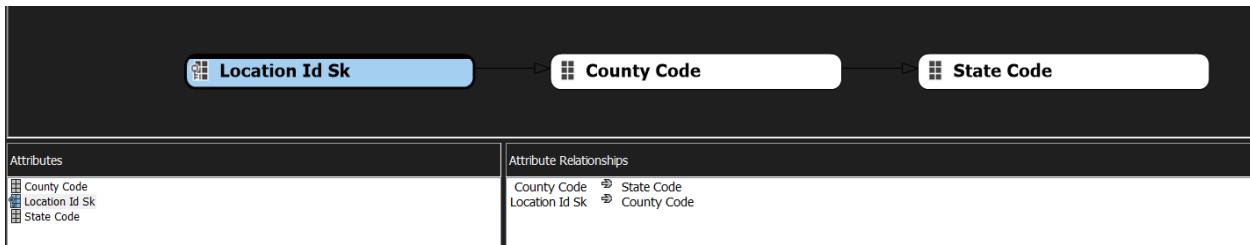
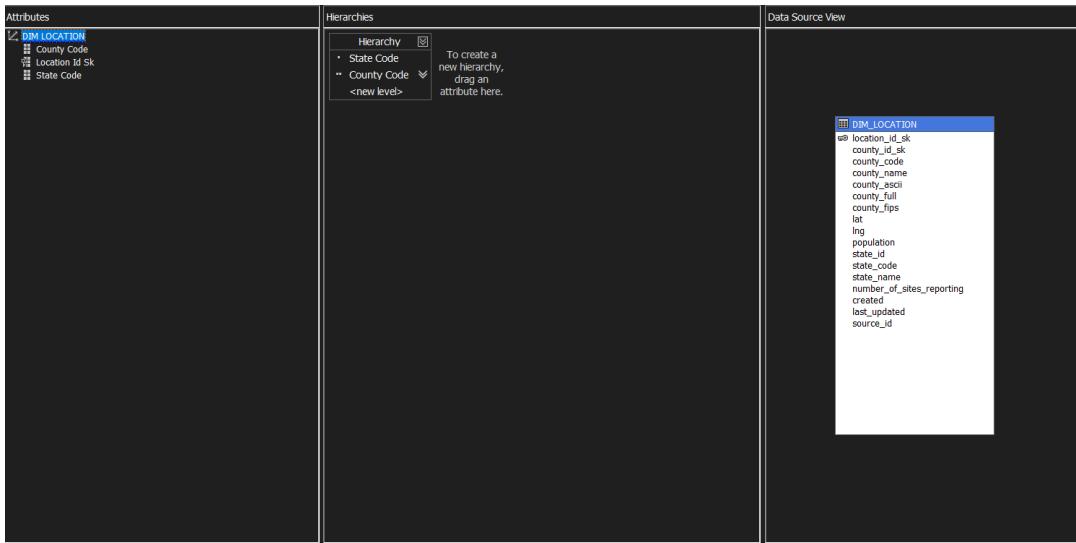


- **DIM_DATE:**

<p>Attributes</p> <p><input checked="" type="checkbox"/> DIM_DATE</p> <ul style="list-style-type: none"> Date Id Day Month Quarter Year 	<p>Hierarchies</p> <p>Hierarchy <input checked="" type="checkbox"/></p> <ul style="list-style-type: none"> Year Quarter Month Day <p>To create a new hierarchy, drag an attribute here.</p> <p><new level></p>	<p>Data Source View</p> <p>DIM_DATE</p> <ul style="list-style-type: none"> date_id day month quarter year day_light_saving created last_updated
--	--	--



- DIM_LOCATION:



- DIM_CATEGORY:

Attributes	Hierarchies	Data Source View
<p><input checked="" type="checkbox"/> DIM CATEGORY</p> <ul style="list-style-type: none"> Category Id Sk Category Name 	<p>Hierarchy <input checked="" type="checkbox"/></p> <ul style="list-style-type: none"> Category Name <input checked="" type="checkbox"/> <p>To create a new hierarchy, drag an attribute here.</p>	<p>DIM_CATEGORY</p> <ul style="list-style-type: none"> category_id_sk category_name created last_updated source_id

- DIM_DEFINING_PARAMETER:

Attributes	Hierarchies	Data Source View
<p><input checked="" type="checkbox"/> DIM DEFINING_PARAMETER</p> <ul style="list-style-type: none"> Defining Parameter Defining Parameter Id Sk 	<p>Hierarchy <input checked="" type="checkbox"/></p> <ul style="list-style-type: none"> Defining Parameter <input checked="" type="checkbox"/> <p>To create a new hierarchy, drag an attribute here.</p>	<p>DIM_DEFINING_PARAMETER</p> <ul style="list-style-type: none"> defining_parameter_id_sk defining_parameter created last_updated source_id

7. MDX

i. Report the min and max of AQI value for each State during each quarter of years.

- Max

```

3  SELECT
4      {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON ROWS,
5      NON EMPTY {[DIM LOCATION].[State Name].[State Name]} ON COLUMNS
6
7  FROM
8      [DDS Group01]
9
10 WHERE
11     [Measures].[Max Aqi];
12
13 SELECT
14     {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON ROWS,
15     NON EMPTY {[DIM LOCATION].[State Name].[State Name]} ON COLUMNS

```

- Min

```

13  SELECT
14      {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON ROWS,
15      NON EMPTY {[DIM LOCATION].[State Name].[State Name]} ON COLUMNS
16
17  FROM
18      [DDS Group01]
19
20  WHERE
21      [Measures].[Min AQI];
22

```

- Nhận xét:

+ Nhìn chung các chỉ số min, max AQI của các tiểu bang đều có xu hướng giảm theo thời gian

+ Có tiểu bang California thì luôn có ngày trong mỗi quý đạt tới mốc 500

+ Có vài tiểu bang thường có ngày mà chỉ số min bằng 0 như Alaska, California, Texas

+ Các giá trị min dao động từ 0 – 30, các giá trị max dao động từ 0 – 500. Điều này có thể do có các tiểu bang không có chỉ số đo lường.

ii. Report the mean and the standard deviation of AQI value for each State during each quarter of years.

```
--Cau 2: Report the mean and the standard deviation of AQI value for each State during
--each quarter of years.
WITH MEMBER [Measures].[State AQI] AS
    AVG([DIM LOCATION].[State Name].[State Name],
    [Measures].[Mean Aqi])
SELECT
    NON EMPTY {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON COLUMNS,
    NON EMPTY {[DIM LOCATION].[State Name].[State Name]} ON ROWS
FROM
    [DDS Group01]
WHERE
    [Measures].[State AQI]
---
WITH MEMBER [Measures].[State AQI] AS
    STDEV([DIM LOCATION].[State Name].[State Name],
    [Measures].[Std Aqi])

```

90 %

Messages Results

	2021	2022	2023	2023
	2	3	1	4
Alabama	71648.625	74882.12121212	63812.18181818	65723.0303030303
Alaska	71648.625	74882.12121212	63812.18181818	65723.0303030303
Arizona	71648.625	74882.12121212	63812.18181818	65723.0303030303
Arkansas	71648.625	74882.12121212	63812.18181818	65723.0303030303
California	71648.625	74882.12121212	63812.18181818	65723.0303030303
Colorado	71648.625	74882.12121212	63812.18181818	65723.0303030303
Connecticut	71648.625	74882.12121212	63812.18181818	65723.0303030303
Country Of Mexico	71648.625	74882.12121212	63812.18181818	65723.0303030303
Delaware	71648.625	74882.12121212	63812.18181818	65723.0303030303
District of Columbia	71648.625	74882.12121212	63812.18181818	65723.0303030303

```
WITH MEMBER [Measures].[State AQI] AS
    AVG([DIM LOCATION].[State Name].[State Name],
    [Measures].[Std Aqi])
SELECT
    NON EMPTY {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON COLUMNS,
    NON EMPTY {[DIM LOCATION].[State Name].[State Name]} ON ROWS
FROM
    [DDS Group01]
WHERE
    [Measures].[State AQI]

```

--Cau3: Report the number of days, and the mean AQI value where the air quality is rated
--as "very unhealthy" or worse for each State and County.

SELECT

non empty{

[DIM CATEGORY].[Category Name].&[Very Unhealthy],

[DIM CATEGORY1].[Category1 Name]&[Unhealthy]

90 %

Messages Results

	2021	2022	2023	2023
	2	3	1	4
Alabama	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
Alaska	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
Arizona	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
Arkansas	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
California	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
Colorado	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
Connecticut	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
Country Of Mexico	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
Delaware	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692
District of Columbia	2003.42527913629	2199.2294256904	1680.77994810561	2128.29944032692

- Nhận xét:

+ Nhìn chung các chỉ số mean, std AQI của các tiểu bang đều có xu hướng giảm theo thời gian

+ Các giá trị mean dao động từ 0 – 60, các giá trị max dao động từ 0 – 100. Điều này có thể do có các tiểu bang không có chỉ số đo lường.

iii. Report the number of days, and the mean AQI value where the air quality is rated as "Very Unhealthy" or worse for each State and County.

- Number of days

```
--Cau3: Report the number of days, and the mean AQI value where the air quality is rated
--as "very unhealthy" or worse for each State and County.
SELECT
non empty{
    [DIM CATEGORY].[Category Name].&[Very Unhealthy],
    [DIM CATEGORY].[Category Name].&[Hazardous]
} ON COLUMNS,
non empty{
    [DIM LOCATION].[State Name].[State Name]*
    [DIM LOCATION].[County Name].[County Name]
} ON ROWS
FROM [DDS Group01]
WHERE [Measures].[FACT AQI Count];

--Cau4: For the four following states: Hawaii, Alaska, Illinois and Delaware, count the
--number of days in each air quality Category (Good, Moderate,etc.) by County.
SELECT
    NON EMPTY {
        [DIM CATEGORY].[Category Name].[Category Name]
    } ON COLUMNS,
    NON EMPTY {

```

90 %

Messages Results

		Very Unhealthy	Hazardous
Arizona	Coconino	1	(null)
Arizona	Maricopa	79	(null)
Arizona	Pima	(null)	1
Arizona	Pinal	3	4
California	Bute	3	(null)
California	Colusa	1	(null)
California	El Dorado	1	1
California	Fresno	2	(null)
California	Humboldt	4	(null)
California	Imperial	6	12
California	Inyo	8	14
California	Kern	2	1
California	Los Angeles	5	(null)
California	Mariposa	2	(null)
California	Mesa	1	??

- Nhận xét:

+ Đa phần số lượng ngày ở các tiểu bang ở mức “không tốt cho sức khỏe” và “ ô nhiễm” là không nhiều nên chất lượng không khí ở các tiểu bang đa phần là tốt hay ở mức chấp nhận được

iv. For the four following states: Hawaii, Alaska, Illinois and Delaware, count the number of days in each air quality Category (Good, Moderate,etc.) by County.

```
--Cau4: For the four following states: Hawaii, Alaska, Illinois and Delaware, count the
--number of days in each air quality Category (Good, Moderate,etc.) by County.
SELECT
    NON EMPTY {
        [DIM CATEGORY].[Category Name].[Category Name]
    } ON COLUMNS,
    NON EMPTY {
        {
            [DIM LOCATION].[State Name].[Hawaii],
            [DIM LOCATION].[State Name].[Alaska],
            [DIM LOCATION].[State Name].[Illinois],
            [DIM LOCATION].[State Name].[Delaware]
        } *
        [DIM LOCATION].[County Name].[County Name]
    } ON ROWS
FROM [DDS Group01]
WHERE [Measures].[FACT AQI Count];

--Cau 5: For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the
--mean AQI value by quarters.
WITH MEMBER [Measures].[State AQI] AS
    AVG([DIM LOCATION].[State Name].[State Name].[Measures].[Mean AQI])

```

90 %

Messages Results

		Good	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Hawaii	Hawaii	993	95	(null)	(null)	(null)
Hawaii	Honolulu	1057	37	(null)	1	(null)
Hawaii	Kauai	454	(null)	(null)	(null)	(null)
Hawaii	Maui	1018	8	1	(null)	(null)
Illinois	Champaign	625	410	1	12	1
Illinois	Cook	308	724	7	54	2
Illinois	DuPage	524	551	4	14	2
Illinois	Jersey	664	401	1	24	(null)
Illinois	Jo Daviess	966	83	1	8	(null)
Illinois	Kane	555	278	5	20	(null)
Illinois	Lake	2416	443	6	42	(null)
Illinois	Macon	559	511	1	19	1
Illinois	Macoupin	1001	73	(null)	12	(null)
Illinois	McHenry	577	490	4	21	1

- Nhận xét:

+ Tại bang Hawaii, số lượng ngày có chỉ số AQI ở mức Good, Moderate khá nhiều, và không nhận thấy mức độ tệ hơn, cho thấy được sự trong sạch của hòn đảo này.

+ Tại bang Alaska, số lượng ngày có chỉ số AQI ở mức Good, Moderate khá nhiều, và không nhận thấy mức độ tệ hơn, cho thấy được sự trong sạch tại đây.

+ Tại bang Illinois, số lượng ngày có chỉ số AQI ở mức Good, Moderate khá nhiều. Ngoài ra cũng đo được các bang có vài ngày ở mức ô nhiễm nặng hơn.

+ Tại bang Delaware, số lượng ngày có chỉ số AQI ở mức Good, Moderate khá nhiều. Ngoài ra cũng đo được các bang có vài ngày ở mức ô nhiễm nặng hơn.

v. For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the mean AQI value by quarters.

```
--Cau 5: For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the
--mean AQI value by quarters.

WITH MEMBER [Measures].[State AQI] AS
    AVG([DIM LOCATION].[State Name].[State Name], [Measures].[Mean Aqi])

SELECT
    NON EMPTY {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON COLUMNS,
    NON EMPTY {[DIM LOCATION].[State Name].[Hawaii],
    [DIM LOCATION].[State Name].[Alaska],
    [DIM LOCATION].[State Name].[Illinois],
    [DIM LOCATION].[State Name].[Delaware]} ON ROWS

FROM
    [DDS Group01]

WHERE
    [Measures].[State AQI]

--Cau 6: Design a report to demonstrate the AQI fluctuation trends over the year for the four
--following states: Hawaii, Alaska, Illinois and California.

```

90 %

	2021	2022	2023	2023
	2	3	1	4
Hawaii	71648.625	74882.1212121212	63812.1818181818	65723.0303030303
Alaska	71648.625	74882.1212121212	63812.1818181818	65723.0303030303
Illinois	71648.625	74882.1212121212	63812.1818181818	65723.0303030303
Delaware	71648.625	74882.1212121212	63812.1818181818	65723.0303030303

- Nhận xét:

- + Tại bang Hawaii, chỉ số mean AQI dao động từ 19 đến 26.
 - + Tại bang Hawaii, chỉ số mean AQI dao động từ 20 đến 30.
 - + Tại bang Illinois, chỉ số mean AQI dao động từ 43 đến 52.
 - + Tại bang Hawaii, chỉ số mean AQI dao động từ 43 đến 46.
- => Nhìn chung, các chỉ số mean AQI ở đây đều phản ánh đúng với chỉ số mean AQI trên toàn bang.

vi. Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California.

```
-- Cau 6: Design a report to demonstrate the AQI fluctuation trends over the year for the four
--following states: Hawaii, Alaska, Illinois and California.

WITH MEMBER [Measures].[State AQI] AS
    AVG([DIM LOCATION].[State Name].[State Name],
    [Measures].[Mean Aqi])

SELECT
    NON EMPTY {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON COLUMNS,
    NON EMPTY {[DIM LOCATION].[State Name].[Hawaii],
    [DIM LOCATION].[State Name].[Alaska],
    [DIM LOCATION].[State Name].[Illinois],
    [DIM LOCATION].[State Name].[Delaware]} ON ROWS

FROM
    [DDS Group01]

WHERE
    [Measures].[State AQI]

--Cau 9: Report the mean, the standard deviation, min and max of AQI value group by
--State and County during each quarter of the year.
```

90 %

	2021	2022	2023	2023
Hawaii	2	3	1	4
Alaska	71648.625	74882.1212121212	63812.1818181818	65723.0303030303
Illinois	71648.625	74882.1212121212	63812.1818181818	65723.0303030303
Delaware	71648.625	74882.1212121212	63812.1818181818	65723.0303030303

- Nhận xét: Nhìn chung, các chỉ số AQI ở đây đều phản ánh đúng với xu hướng thay đổi chỉ số mean AQI trên toàn bang (giảm dần theo thời gian).

vii. Report the mean, the standard deviation, min and max of AQI value group by State and County during each quarter of the year.

- Max

```
126  SELECT
127      {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON ROWS,
128      NON EMPTY
129      {[DIM LOCATION].[State Name].[State Name]*
130      [DIM LOCATION].[County Name].[County Name]} ON COLUMNS
131
132  FROM
133      [DDS Group01]
134
135  WHERE
136      [Measures].[Max Aqi];
137
```

155 %

	Alabama	Alabama	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Arizona	Arizona	Arizona	Arizona	Arizona	Arizona	Gila	La Paz	Micopa
	Sumter	Tuscaloosa	Aleutians East	Anchorage	Denali	Fairbanks North Star	Juneau	Kenai Peninsula	Matanuska-Susitna	North Slope	Apache	Cochise	Coconino	Maricopa	Maricopa	Maricopa	Maricopa	Maricopa		
2021	1	(null)	76	76	112	174	73	19	62	8	57	90	84	136	156	239				
2021	2	(null)	132	132	47	189	89	163	163	18	48	105	84	161	187	297				
2021	3	(null)	77	61	74	174	157	64	68	68	53	90	87	180	108	214	237			
2021	4	(null)	46	59	86	76	162	69	62	74	28	49	104	84	115	221	236			
2022	1	7	61	67	77	157	71	115	63	90	104	71	77	161	297	226	46			
2022	2	36	86	68	189	134	73	74	75	53	43	104	77	136	161	239	49			
2022	3	51	132	76	169	145	64	163	71	41	53	87	84	115	187	236	156			
2022	4	(null)	59	81	116	183	157	101	74	63	55	61	93	122	180	234	214			
2023	1	(null)	33	62	76	162	157	73	22	115	24	105	84	84	115	297	239			
2023	2	(null)	39	69	64	160	163	73	60	78	57	90	87	93	161	197	230			

- Min

```

138 --Min
139 SELECT
140 {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON ROWS,
141 NON EMPTY
142 {[DIM LOCATION].[State Name].[State Name]*[DIM LOCATION].[County Name].[County Name]} ON COLUMNS
143
144
145 FROM
146 [DDS Group01]
147
148 WHERE
149 [Measures].[Min Aqi];

```

155 %

Messages Results

		Alabama	Alabama	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Arizona	Arizona	Arizona	Arizona	Arizona	Arizona
		Sumter	Tuscaloosa	Aleutians East	Anchorage	Denali	Fairbanks North Star	Juneau	Kenai Peninsula	Matanuska-Susitna	North Slope	Apache	Cochise	Cochitano	Gila	La Paz	Maricopa	
2021	1	(null)	19	4	7	12	4	0	4	0	2	5	26	33	9	3	6	
2021	2	(null)	(null)	6	7	11	11	2	0	0	2	3	33	35	26	3	46	
2021	3	(null)	4	3	9	11	2	0	1	1	4	21	32	18	33	6		
2021	4	(null)	3	2	4	12	3	1	2	0	1	4	27	9	29	33	5	
2022	1	7	3	15	16	9	1	0	1	1	2	32	35	18	28	9	1	
2022	2	4	3	13	17	11	0	0	0	1	4	32	35	26	3	6	5	
2022	3	4	2	11	22	4	0	0	1	1	6	26	31	27	37	4	4	
2022	4	(null)	4	5	8	8	2	2	2	0	3	2	32	31	30	6	5	
2023	1	(null)	14	2	5	18	8	0	1	0	1	2	33	26	26	25	10	
2023	2	(null)	3	4	5	11	5	0	1	1	1	3	28	26	30	29	4	
2023	3	(null)	7	4	8	12	15	2	4	1	1	2	22	21	21	20	8	

- Mean

```

WITH
MEMBER [Measures].[Avg] AS AVG([DIM LOCATION].[County Name].[County Name], [Measures].[Mean Aqi])
SELECT
NON EMPTY {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON ROWS,
NON EMPTY {[DIM LOCATION].[State Name].[State Name]*[DIM LOCATION].[County Name].[County Name]} ON COLUMNS
FROM
[DDS Group01]
WHERE
[Measures].[Avg];
---
```

WITH

```

MEMBER [Measures].[Std] AS AVG([DIM LOCATION].[County Name].[County Name], [Measures].[Std Aqi])
SELECT
NON EMPTY {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON ROWS,
NON EMPTY {[DIM LOCATION].[State Name].[State Name]*[DIM LOCATION].[County Name].[County Name]} ON COLUMNS

```

90 %

Messages Results

	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama								
	Autauga	Baldwin	Barbour	Bibb	Blount	Bullock	Butler	Calhoun	Chambers	Cherokee	Chilton	Choctaw	Clarke	Clay	Cleburne	Coffee	Colbert	Cone
2021	2	14163	14163	14163	14163	14163	14163	14163	14163	14163	14163	14163	14163	14163	14163	14163	14163	141
2022	3	15250	15250	15250	15250	15250	15250	15250	15250	15250	15250	15250	15250	15250	15250	15250	15250	152
2023	1	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	8836.5	883
2023	4	9591	9591	9591	9591	9591	9591	9591	9591	9591	9591	9591	9591	9591	9591	9591	9591	9591

- Std

```

WITH
MEMBER [Measures].[Std] AS AVG([DIM LOCATION].[County Name].[County Name], [Measures].[Std Aqi])

SELECT
NON EMPTY {[DIM DATE].[Year].[Year]*[DIM DATE].[Quarter].[Quarter]} ON ROWS,
NON EMPTY {[DIM LOCATION].[State Name].[State Name]*[DIM LOCATION].[County Name].[County Name]} ON COLUMNS

FROM
[DDS Group01]

WHERE
[Measures].[Std];

--Cau11: Count the number of days by State, Category in each month.
SELECT
non empty{
    [DIM DATE].[Month].[Month]*
    [DIM DATE].[Year].[Year]
} ON COLUMNS

```

90 %

Messages Results

	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama
	Autauga	Baldwin	Barbour	Bibb	Blount	Bullock	Butler	Calhoun	Chambers	
2021	2	3099.95612872182	3099.95612872182	3099.95612872182	3099.95612872182	3099.95612872182	3099.95612872182	3099.95612872182	3099.95612872182	3099.95612872182
2022	3	3391.28412257068	3391.28412257068	3391.28412257068	3391.28412257068	3391.28412257068	3391.28412257068	3391.28412257068	3391.28412257068	3391.28412257068
2023	1	894.490078200983	894.490078200983	894.490078200983	894.490078200983	894.490078200983	894.490078200983	894.490078200983	894.490078200983	894.490078200983
2023	4	1849.79133958401	1849.79133958401	1849.79133958401	1849.79133958401	1849.79133958401	1849.79133958401	1849.79133958401	1849.79133958401	1849.79133958401

viii. Report the mean AQI value by State, Category, DayLightSaving over years.

- Nhận xét:

- + Trong khoảng thời gian DayLightSaving, chỉ số AQI giảm đi một cách đáng kể.
- + Thông kê cho thấy rằng, các chỉ số như max, min, mean, std thay đổi không đáng kể, tuy nhiên số lượng ngày có mức độ không khí từ Very Unhealthy trở lên giảm đi rõ rệt (có DayLightSaving: từ 50- 100, không có DayLightSaving: từ 200 – 300).

ix. Count the number of days by State, Category in each month

```

--Cau11: Count the number of days by State, Category in each month.
SELECT
non empty{
    [DIM DATE].[Month].[Month]*
    [DIM DATE].[Year].[Year]
} ON COLUMNS,
non empty{
    [DIM LOCATION].[State Name].[State Name]*
    [DIM CATEGORY].[Category Name].[Category Name]
} ON ROWS
FROM [DDS Group01]
WHERE [Measures].[FACT AQI Count];

--Cau12: Report the number of days by Category and Defining Parameter.
SELECT
non empty{
    [DIM CATEGORY].[Category Name].[Category Name]
} ON COLUMNS

```

90 %

Messages Results

	1	10	11	12	2	3	4	5	6	7	8	9	
	2023	2023	2023	2023	2023	2023	2021	2021	2021	2022	2022	2022	
Alabama	Good	58	173	69	70	47	128	131	114	96	111	143	161
Alabama	Moderate	72	67	69	69	70	68	75	95	93	97	94	73
Alabama	Unhealthy	(null)	2	(null)	(null)	(null)							
Alabama	Unhealthy for Sensitive Groups	1	1	(null)	(null)	2	1	(null)	4	14	5	1	2
Arizona	Good	800	805	732	790	764	788	525	420	475	560	654	766
Arizona	Hazardous	(null)	1	(null)	(null)	(null)	(null)	(null)	1	2	1	(null)	
Arizona	Moderate	174	217	247	208	204	249	468	611	481	421	347	235
Arizona	Unhealthy	20	23	(null)	5	(null)	24	3	1	6	12	3	10
Arizona	Unhealthy for Sensitive Groups	56	17	12	29	7	10	23	47	68	83	70	16
Arizona	Very Unhealthy	1	13	(null)	(null)	1	9	28	(null)	8	1	1	21
California	Good	2371	2573	2315	2464	2575	3468	2792	2856	2825	2255	1862	1960
California	Hazardous	4	9	4	1	9	5	10	10	5	4	30	6
California	Moderate	2300	1946	2125	2119	1711	1321	1731	1743	1512	1969	2086	2159

- Nhận xét:

- + Tùy vào mỗi bang có số ngày ở mỗi loại không khí khác nhau nhưng nhìn chung thì trong từng bang thì chất lượng không khí ở mỗi tháng không thay đổi hay chênh lệch nhiều.

x. Report the number of days by Category and Defining Parameter.

```
--Cau12: Report the number of days by Category and Defining Parameter.
SELECT
non empty{
    [DIM CATEGORY].[Category Name].[Category Name]
} ON COLUMNS,
non empty{
    [DIM DEFINING PARAMETER].[Defining Parameter].[Defining Parameter]
} ON ROWS
FROM [DDS Group01]
WHERE [Measures].[FACT AQI Count]
```

90 %

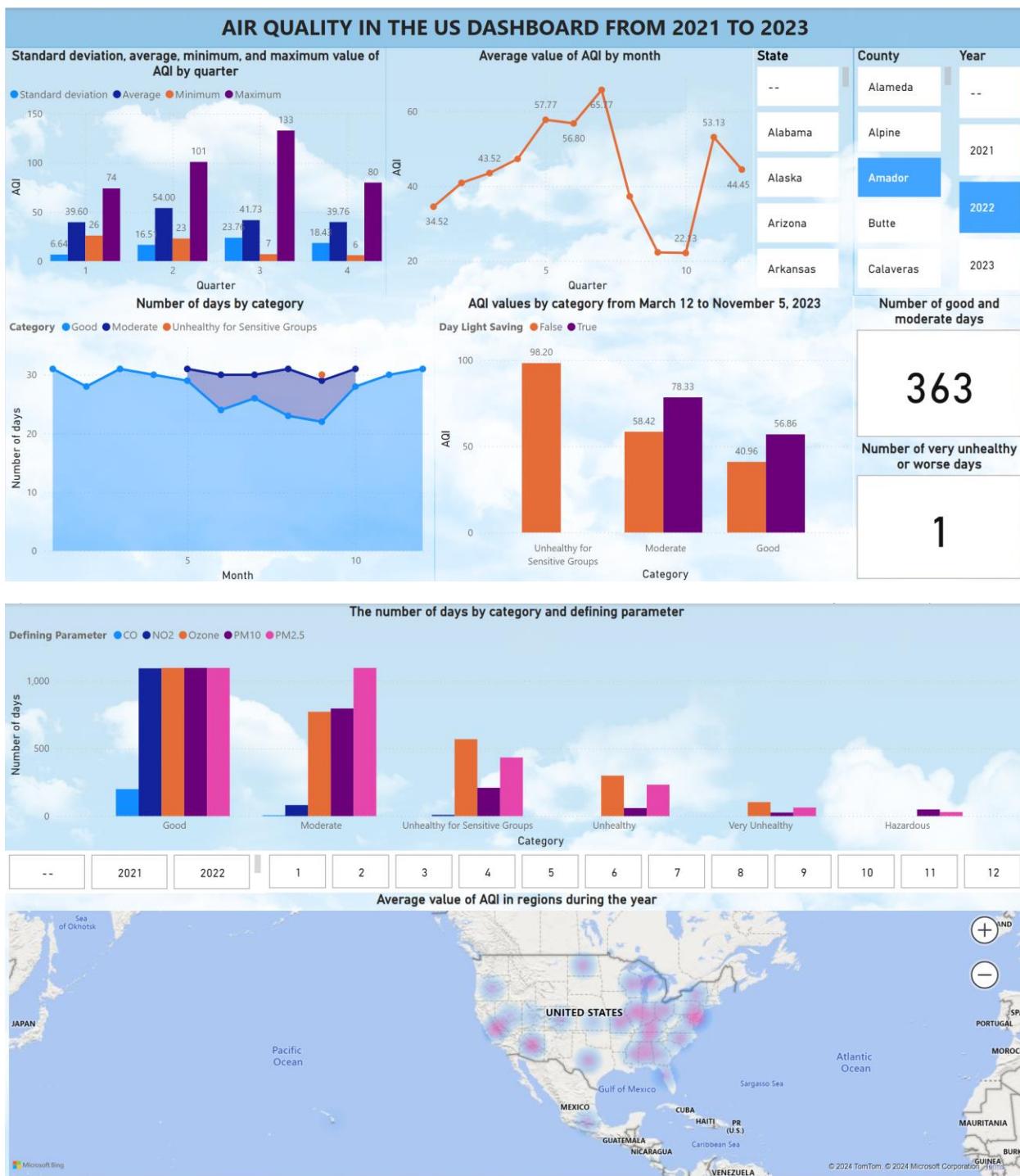
Messages Results

	Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
CO	214	(null)	2	(null)	(null)	(null)
NO2	3566	(null)	89	(null)	9	(null)
Ozone	71367	(null)	18285	580	3176	113
PM10	3815	67	2317	71	263	26
PM2.5	38333	40	42956	484	1112	104

- Nhận xét:

- + Thống kê cho thấy rằng, tác nhân gây ô nhiễm chính trên nước Mỹ là bụi mịn PM2.5 và PM10. Điều này có thể lí giải là do Mỹ có nền công nghiệp phát triển hàng đầu thế giới, hệ thống giao thông vận tải trèsnđa dạng, đô thị hóa cao.
- + Ngoài ra, Ozone cũng là nguyên nhân làm cho không khí ở Mỹ bị ô nhiễm. Điều kiện khí hậu khắc nghiệt, hoạt động công nghiệp... là nguyên nhân chính tạo ra khí thải này.

8. Dashboard



9. Data mining

Yêu cầu: Sử dụng mô hình để dự báo chất lượng không khí trong các giai đoạn tiếp theo như quý tiếp theo (Q1-2024), tháng tiếp theo (01-2024), v.v.

Ở đây em sẽ chọn mô hình **LSTM (Long Short-Term Memory)**

a) Giải thích thuật toán LSTM

i. Tổng quan về LSTM

- LSTM (Long Short-Term Memory) là một kiến trúc mạng nơ-ron hồi tiếp (RNN) đặc biệt, được thiết kế để xử lý và dự đoán dữ liệu chuỗi thời gian.
- Đặc điểm nổi bật:
 - Bộ nhớ dài hạn và ngắn hạn: LSTM sử dụng các "cổng" để kiểm soát dòng chảy thông tin (input gate, forget gate, output gate), cho phép ghi nhớ và quên thông tin một cách có chọn lọc.
 - Khắc phục nhược điểm của RNN truyền thống: LSTM giải quyết vấn đề gradient biến mất hoặc bùng nổ, giúp nó phù hợp với các chuỗi thời gian dài.

ii. Cấu trúc của LSTM

LSTM bao gồm 4 thành phần chính:

1. Cổng đầu vào (Input Gate): Xác định thông tin mới nào nên được lưu trữ trong trạng thái ẩn.
2. Cổng quên (Forget Gate): Quyết định thông tin nào từ trạng thái cũ nên bị loại bỏ.
3. Cổng đầu ra (Output Gate): Xác định thông tin nào từ trạng thái hiện tại nên được đưa ra làm đầu ra.
4. Trạng thái bộ nhớ (Cell State): Lưu trữ thông tin quan trọng qua thời gian.

iii. Quá trình hoạt động

- Bước 1: LSTM nhận dữ liệu đầu vào của chuỗi thời gian (12 giá trị AQI gần nhất).
- Bước 2: Qua các cổng, mô hình chọn lọc thông tin hữu ích và cập nhật trạng thái bộ nhớ.
- Bước 3: Trả về dự đoán AQI cho tháng tiếp theo.

iv. Ứng dụng với dữ liệu AQI

LSTM rất phù hợp để xử lý dữ liệu AQI vì:

- AQI có tính phụ thuộc thời gian: Giá trị AQI trong tương lai chịu ảnh hưởng bởi giá trị quá khứ.

- Dữ liệu AQI thường không tuyến tính và có xu hướng dài hạn, điều này phù hợp với khả năng của LSTM.

b) Lý do chọn LSTM

- Xử lý phi tuyến tính: Dữ liệu AQI không tuân theo một mô hình tuyến tính đơn giản. LSTM có khả năng học các mối quan hệ phức tạp này.
- Phụ thuộc dài hạn: Với chuỗi thời gian dài, LSTM vẫn có thể ghi nhớ các thông tin quan trọng từ quá khứ.
- Hiệu quả thực nghiệm: Các nghiên cứu trước đây cho thấy LSTM hoạt động tốt hơn các mô hình thống kê (như ARIMA) trong dự báo dữ liệu phi tuyến tính.

c) Kết quả dự báo

i. Thông tin dự báo

- **Dữ liệu đầu vào:**
 - 12 tháng gần nhất của AQI (từ tháng 1/2023 đến tháng 12/2023) được chuẩn hóa và đưa vào mô hình.
- **Dữ liệu đầu ra:**
 - Giá trị AQI dự báo cho tháng 1/2024 được tính toán dựa trên thông tin từ 12 tháng trước đó.

ii. Quy trình dự báo

```
# Đọc dữ liệu
file_path = 'NDS_AIR.csv'
data = pd.read_csv(file_path)

# Chuyển đổi cột 'date_recorded' thành kiểu datetime
data['date_recorded'] = pd.to_datetime(data['date_recorded'])
filtered_data = data[['date_recorded', 'aqi']].drop_duplicates().dropna()

# Gộp dữ liệu theo tháng
monthly_data = filtered_data.set_index('date_recorded').resample('ME').mean()

# Chuẩn hóa dữ liệu
scaler = MinMaxScaler(feature_range=(0, 1))
scaled_data = scaler.fit_transform(monthly_data)
```

- Dữ liệu AQI được ghi nhận hàng ngày, sau đó gộp theo trung bình tháng để tạo dữ liệu đầu vào cho mô hình.

Chuẩn hóa dữ liệu:

- Dữ liệu AQI được chuẩn hóa về khoảng [0, 1] bằng MinMaxScaler nhằm tăng hiệu quả huấn luyện mạng nơ-ron.
- Quá trình chuẩn hóa đảm bảo dữ liệu nằm trong miền giá trị nhỏ, giúp mô hình hội tụ nhanh hơn.

```
# Tạo tập dữ liệu chuỗi thời gian
def create_dataset(data, look_back=12):
    X, y = [], []
    for i in range(len(data) - look_back):
        X.append(data[i:i + look_back, 0])
        y.append(data[i + look_back, 0])
    return np.array(X), np.array(y)

look_back = 12
X, y = create_dataset(scaled_data, look_back)
X = X.reshape((X.shape[0], X.shape[1], 1))
```

- Sử dụng một khoảng thời gian 12 tháng trước đó (look_back = 12) làm đầu vào để dự đoán giá trị tháng tiếp theo.
- Mỗi tập mẫu (X) chứa 12 giá trị AQI gần nhất, và đầu ra (y) là giá trị AQI của tháng tiếp theo.

```

# Xây dựng mô hình LSTM
model = Sequential()
model.add(LSTM(50, return_sequences=True, input_shape=(look_back, 1)))
model.add(LSTM(50, return_sequences=False))
model.add(Dense(1))

# Biên dịch mô hình
model.compile(optimizer='adam', loss='mean_squared_error')

# Huấn luyện mô hình
model.fit(X, y, epochs=50, batch_size=32, validation_split=0.2, verbose=1)

# Dự báo cho tháng tiếp theo
last_data = scaled_data[-look_back:]
last_data = last_data.reshape(1, look_back, 1)
predicted_next_month = model.predict(last_data)

# Chuyển đổi giá trị dự báo về thang đo ban đầu
predicted_next_month = scaler.inverse_transform(predicted_next_month)

# Kết quả dự báo
print(f"Dự báo AQI cho tháng 1/2024: {predicted_next_month[0, 0]:.2f}")

```

Xây dựng và huấn luyện mô hình:

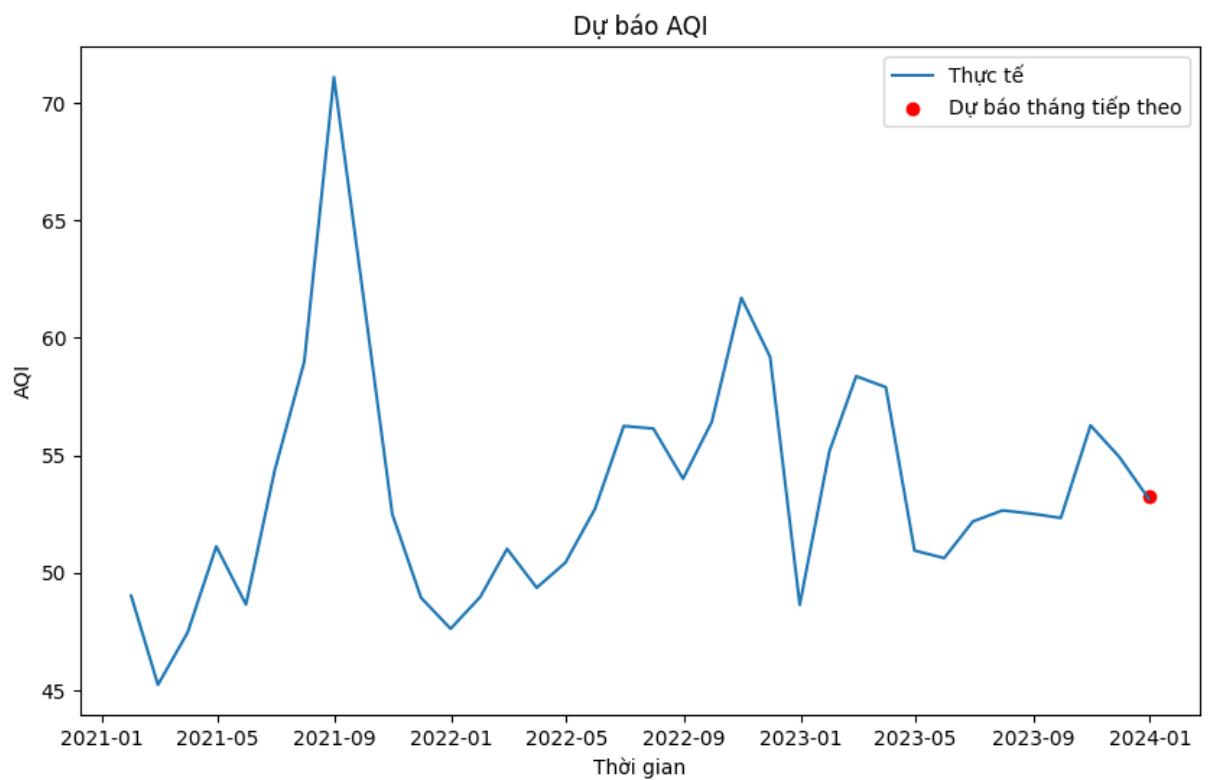
- Mô hình gồm hai lớp LSTM:
 - Lớp đầu tiên trả về chuỗi (return_sequences=True).
 - Lớp thứ hai xử lý đầu ra chuỗi thành một giá trị duy nhất.
 - Lớp Dense cuối cùng kết nối đầu ra với giá trị AQI cần dự đoán.
- Hàm mất mát (mean_squared_error) và tối ưu hóa (Adam) được sử dụng để giảm thiểu sai số.

Chuyển đổi kết quả về thang đo ban đầu và in ra kết quả.

```

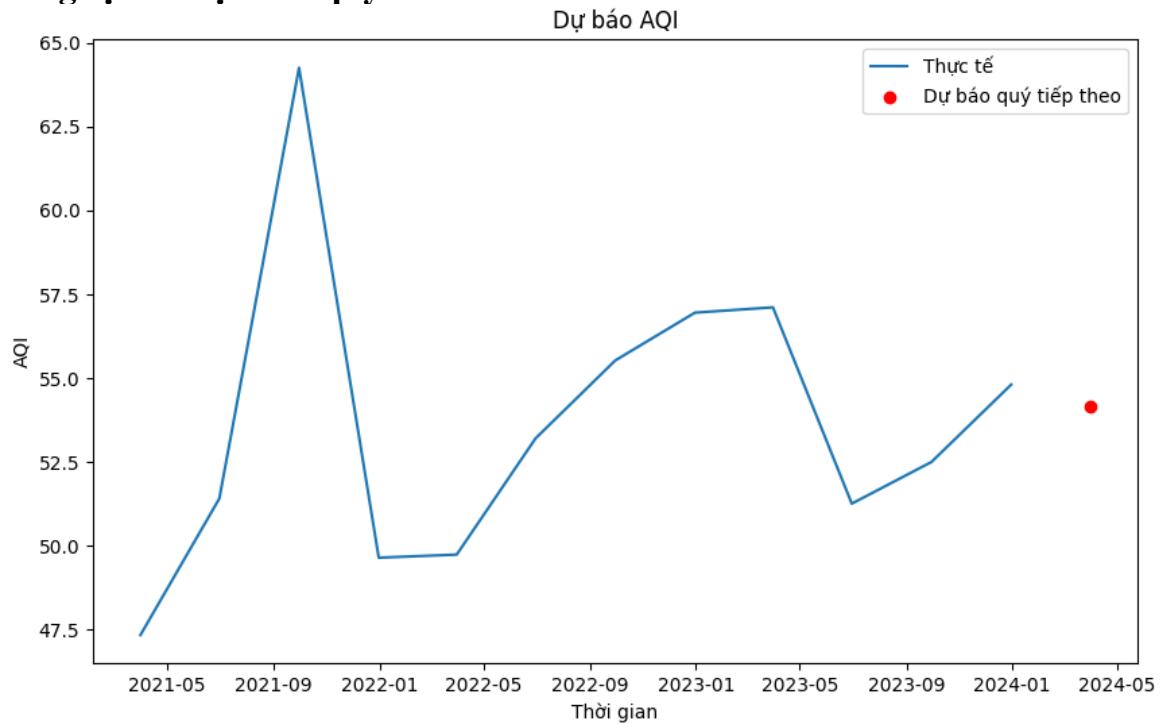
# Vẽ biểu đồ
plt.figure(figsize=(10, 6))
plt.plot(monthly_data.index, scaler.inverse_transform(scaled_data), label='Thực tế')
plt.scatter(monthly_data.index[-1], predicted_next_month[0, 0], color='r', marker='o', label='Dự báo tháng tiếp theo')
plt.title('Dự báo AQI')
plt.xlabel('Thời gian')
plt.ylabel('AQI')
plt.legend()
plt.show()

```



Dự báo AQI cho tháng 1/2024: 53.24

Tương tự với dự toán quý 1/2024



Dự báo AQI cho quý 1/2024: 54.2

10. Kết luận

- Tổng kết về phân tích:

+ Chất lượng không khí tại các quận của Hoa Kỳ trong năm 2023 có sự biến động đáng kể, phụ thuộc vào nhiều yếu tố như hoạt động công nghiệp, giao thông, điều kiện thời tiết và các biện pháp kiểm soát ô nhiễm. Đây là một vấn đề đáng quan ngại, với nhiều khu vực, đặc biệt là các quận ở bang California, New York... ghi nhận mức độ ô nhiễm cao.

+ Tuy nhiên, với các biện pháp kiểm soát và giám sát chặt chẽ, nhiều khu vực đã đạt được những cải thiện đáng kể trong việc giảm thiểu ô nhiễm không khí, bảo vệ sức khỏe cộng đồng và môi trường.

- Về đồ án thực hành:

+ Thông qua đồ án thực hành, nhóm đã thực hiện được:

- (1) Nạp dữ liệu từ file csv. vào SQL Server thông qua qui trình ETL.
- (2) Tạo lập được qui trình nạp dữ liệu Source → Stage → NDS → DDS → OLAP.
- (3) Thực hiện biến đổi, tạo các chiều dữ liệu để hình thành OLAP cube.
- (4) Truy vấn dữ liệu bằng MDX.
- (5) Xây dựng Dashboard trong PowerBI.
- (6) Thực hiện data mining bằng mô hình LSTM(Long Short-Term Memory)

+ Tuy vậy, bài làm của nhóm vẫn còn những hạn chế như sau:

- (1) Truy vấn dữ liệu MDX chưa thành công khi thống kê các giá trị trung bình, lớn nhất, nhỏ nhất bằng các hàm AVG(), MAX(), MIN() do khi thực hiện, kết quả nhận được là tổng các chỉ số AQI.
- (2) Dashboard chưa quá đa dạng các biểu đồ.

III. Tài Liệu tham khảo

- Slide bài giảng của gv: Hồ Thị Hoàng Vy, Tiết Gia Hồng, Nguyễn Ngọc Minh Châu
- NDS,DDS: [Link](#)
- Olap,MDX:
 - + https://youtu.be/vtEUxJzeF2A?si=JnGL_C9Z4LQ-Eng5
 - + <https://www.red-gate.com/simple-talk/databases/sql-server/bi-sql-server/creating-a-date-dimension-in-an-analysis-services-ssas-cube/>
- Data mining: <https://trituevietvn.com/chi-tiet/su-dung-mang-lstm-long-short-term-memory-de-du-doan-so-lieu-huong-thoi-gian-123>