

实验四：数据库与SQL语句

在实验三中，我已经爬取了ESI网站的学科数据并完成1-2题

在实验四中，我决定还是用学长提供的数据来完成3-7题

3. 将获取的数据导入到一个关系型数据库系统中（系统可以自选）

暑假我们刚学过SQL，那选择 SQL Server Management Studio 最顺手

先打开一个csv看一下，发现第1行是说明文字，第2行才是表头：

	A	B	C	D	E	F	G	H	I	J	K	L
1	Indicators	Results List	Institutions	Filter Results By:	ResearchFields	Filter Value(s):	AGRICULTURAL SCIENCES	Show: Top				
2		Institution	Countries/	Web of Sc	Cites	Cites/Paper	Top Papers					
3	1	CHINESE /	CHINA M/	15661	332254	21.22	251					
4	2	CHINESE /	CHINA M/	12222	223855	18.32	198					
5	3	UNITED S	USA	12564	220644	17.56	105					
6	4	CHINA AC	CHINA M/	10052	207779	20.67	166					
7	5	INRAE	FRANCE	9314	187838	20.17	118					
8	6	JIANGNAN	CHINA M/	8027	181325	22.59	166					
9	7	NORTHW	CHINA M/	9024	180634	20.02	162					
10	8	WAGENIN	NETHERLA	6819	172734	25.33	152					
11	9	CONSEJO	SPAIN	7656	167593	21.89	94					
12	10	NANJING	CHINA M/	6480	137921	21.28	99					
13	11	UNIVERSI	USA	6175	129479	20.97	110					
14	12	UNIVERSI	CHINA M/	5950	127682	21.46	95					
15	13	ZHEJIANG	CHINA M/	5313	125715	23.66	148					
16	14	EGYPTIAN	EGYPT	7861	117405	14.94	103					
17	15	INDIAN C	INDIA	12753	111262	8.72	66					
18	16	SOUTH C	CHINA M/	3524	108282	30.73	117					
19	17	CGIAR	FRANCE	5262	107193	20.37	69					
20	18	HUAZHON	CHINA M/	4855	97482	20.08	68					
21	19	UNIVERSI	BRAZIL	6169	95242	15.44	35					
22	20	JIANGSU	CHINA M/	4084	89040	21.8	176					
23	21	STATE UN	USA	5172	86527	16.73	62					
24	22	CENTRE N	FRANCE	4113	81348	19.78	60					
25	23	CORNELL	USA	3551	78744	22.18	59					

最后一行是 Copyright ?2025 Clarivate，也应该去掉：

	A	B	C	D	E	F	G	H	I	J	K	L
1355	1355	VELLOR	INDIA	135	3627	23.71	2					
1359	1356	MAE FAH	THAILAND	236	3627	15.37	2					
1360	1358	NAMIK K	TURKIYE	352	3625	10.3	1					
1361	1359	GANGNEU	SOUTH KO	268	3621	13.51	1					
1362	1360	TAMPERE	FINLAND	234	3619	15.47	0					
1363	1361	HOSPITAL	CANADA	202	3618	17.91	2					
1364	1362	BRAUNSC	GERMANY	188	3614	19.22	1					
1365	1362	ICAR - IN	INDIA	393	3614	9.2	6					
1366	1364	HUNGARI	HUNGARY	180	3605	20.03	2					
1367	1365	LIAONING	CHINA M/	333	3604	10.82	0					
1368	1366	DAIRYNZ	NEW ZEAL	239	3599	15.06	0					
1369	1367	SMITHSO	USA	209	3598	17.22	0					
1370	1367	BAYER CR	GERMANY	250	3598	14.39	2					
1371	1369	GENERAL	USA	107	3595	33.6	4					
1372	1370	STEFAN C	ROMANIA	161	3591	22.3	5					
1373	1371	KAGOSHII	JAPAN	306	3588	11.73	1					
1374	1372	UNIVERSI	USA	127	3580	28.19	4					
1375	1373	HOPITAL I	FRANCE	123	3573	29.05	3					
1376	1373	HEINRICH	GERMANY	136	3573	26.27	4					
1377	1375	UNIVERSI	SOUTH AF	184	3572	19.41	6					
1378	1376	BANGLAD	BANGLAD	170	3567	20.98	2					
1379	1376	NEAR EAST	TURKIYE	204	3567	17.49	2					
1380	1378	GRIFFITH I	AUSTRALI	205	3562	17.38	2					
1381	1378	UMM AL-	SAUDI AR	308	3562	11.56	3					
1382	1380	SAINT LO	USA	108	3561	32.97	4					
1383	1381	VIT VELL	INDIA	144	3558	24.71	2					
1384	Copyright ?2025 Clarivate											
1385												

SSMS 不能直接跳过第一行和最后一行，所以我们先把每个 CSV 的**第一行和最后一行删掉**，再导入

在 Powershell 执行下面命令：

```
$src="D:\ESI\download"; $dst="D:\ESI\clean"; New-Item -ItemType Directory -Force -Path $dst | Out-Null
Get-ChildItem $src -Filter *.csv | ForEach-Object {
    $lines = Get-Content $_.FullName
    Get-ChildItem $src -Filter *.csv
    $lines[1..($lines.Count-2)] | Set-Content -Encoding UTF8 (Join-Path $dst $_.Name)
}
```

解释一下：

1. src 是原始文件路径，dst 是目标文件路径
2. New-Item 用于创建一个新目录，如果 D:\ESI\clean 不存在，会自动创建
3. Get-ChildItem \$src -Filter *.csv 会列出 src 中所有扩展名为 .csv 的文件
4. ForEach-Object 遍历每个文件
5. \$lines = Get-Content \$_.FullName 会把整个文件的内容读成一个字符串数组，每一行是一个元素
6. \$lines[1..(\$lines.Count-2)] 是数组切片语法，表示从第二行到倒数第二行，删除文件的第一行和最后一行
7. Set-Content 把上一步得到的“去掉第一行的内容”写入到新的文件中

可以看到，第一行已经成功去掉了：

	A	B	C	D	E	F	G	H	I	J	K	L
1		Institution	Countries	Web of Sc	Cites	Cites/Paper	Top Papers					
2	1	CHINESE	CHINA M	15661	332254	21.22	251					
3	2	CHINESE	CHINA M	12222	223855	18.32	198					
4	3	UNITED S	USA	12564	220644	17.56	105					
5	4	CHINA AC	CHINA M	10052	207779	20.67	166					
6	5	INRAE	FRANCE	9314	187838	20.17	118					
7	6	JIANGNA	CHINA M	8027	181325	22.59	166					
8	7	NORTHW	CHINA M	9024	180634	20.02	162					
9	8	WAGENIN	NETHERL	6819	172734	25.33	152					
10	9	CONSEJO	SPAIN	7656	167593	21.89	94					
11	10	NANJING	CHINA M	6480	137921	21.28	99					
12	11	UNIVERSI	USA	6175	129479	20.97	110					
13	12	UNIVERSI	CHINA M	5950	127682	21.46	95					
14	13	ZHEJIANG	CHINA M	5313	125715	23.66	148					
15	14	EGYPTIAN	EGYPT	7861	117405	14.94	103					
16	15	INDIAN C	INDIA	12753	111262	8.72	66					
17	16	SOUTH C	CHINA M	3524	108282	30.73	117					
18	17	CGIAR	FRANCE	5262	107193	20.37	69					
19	18	HUAZHON	CHINA M	4855	97482	20.08	68					
20	19	UNIVERSI	BRAZIL	6169	95242	15.44	35					
21	20	JIANGSU	CHINA M	4084	89040	21.8	176					
22	21	STATE UN	USA	5172	86527	16.73	62					
23	22	CENTRE N	FRANCE	4113	81348	19.78	60					
24	23	CORNELL	USA	3551	78744	22.18	59					
25	24	UNIVERSI	USA	3908	78346	20.05	64					

最后一行也成功去掉了：

	A	B	C	D	E	F	G	H	I	J	K	L
1365	1364	HONGKONG HONGKONG		100	3600	20.00	2					
1366	1365	LIAONING CHINA MA		333	3604	10.82	0					
1367	1366	DAIRYNZ NEW ZEAL		239	3599	15.06	0					
1368	1367	SMITHSON USA		209	3598	17.22	0					
1369	1367	BAYER CR GERMANY		250	3598	14.39	2					
1370	1369	GENERAL USA		107	3595	33.6	4					
1371	1370	STEFAN C ROMANIA		161	3591	22.3	5					
1372	1371	KAGOSHII JAPAN		306	3588	11.73	1					
1373	1372	UNIVERSI USA		127	3580	28.19	4					
1374	1373	HOPITAL I FRANCE		123	3573	29.05	3					
1375	1373	HEINRICH GERMANY		136	3573	26.27	4					
1376	1375	UNIVERSI SOUTH AF		184	3572	19.41	6					
1377	1376	BANGLAD BANGLAD		170	3567	20.98	2					
1378	1376	NEAR EAS TURKIYE		204	3567	17.49	2					
1379	1378	GRIFFITH I AUSTRALI		205	3562	17.38	2					
1380	1378	UMM AL- SAUDI AR		308	3562	11.56	3					
1381	1380	SAINT LOI USA		108	3561	32.97	4					
1382	1381	VIT VELLC INDIA		144	3558	24.71	2					
1383												
1384												
1385												
1386												
1387												
1388												
1389												
1390												
1391												

批量导入所有CSV文件，每个表格的名字来源于CSV文件名，也就是**学科名称**：

```
# === 参数设置 ===
$folder = "D:\ESI\clean"      # 存放 CSV 文件的文件夹
$server = "localhost"        # SQL Server 实例名
$database = "ESI"            # 目标数据库名

# === 遍历文件夹中所有 CSV 文件并逐个导入 ===
Get-ChildItem $folder -Filter *.csv | ForEach-Object {

    $file = $_.FullName          # 当前 CSV 文件的完整路径
    $table = "staging_" + ($_.BaseName -replace '[^A-Za-z0-9_]', '_') # 根据文件名生成表名
    $tmp = [IO.Path]::GetTempFileName() # 创建一个临时 SQL 文件

@"
-- 如果表不存在就创建
IF OBJECT_ID('dbo.$table') IS NULL
BEGIN
    CREATE TABLE dbo.$table(
        [Rank] NVARCHAR(64),
        [Institutions] NVARCHAR(512),
        [Countries/Regions] NVARCHAR(128),
        [Web of Science Documents] NVARCHAR(64),
        [Cites] NVARCHAR(64),
        [Cites/Paper] NVARCHAR(64),
        [Top Papers] NVARCHAR(64)
    );
END;

-- 清空旧数据
TRUNCATE TABLE dbo.$table;

-- 从 CSV 导入数据
BULK INSERT dbo.$table
FROM '$file'
WITH (
    FORMAT='CSV',          -- 使用原生 CSV 解析
    FIRSTROW=2,            -- 第 1 行是表头，从第 2 行开始导
    FIELDQUOTE='"',        -- 支持带引号的字段
    ROWTERMINATOR='0x0d0a', -- windows 换行符
    CODEPAGE='65001',      -- UTF-8 编码
    TABLOCK                -- 提升导入性能
);
```

```
-- 显示导入结果
SELECT '$table' AS TableName, COUNT(*) AS RowsLoaded FROM dbo.$table;
GO
"@ | Set-Content -Encoding UTF8 $tmp # 将 SQL 写入临时文件

# 执行临时 SQL 文件
sqlcmd -S $server -d $database -E -i $tmp

# 删除临时文件，保持干净
Remove-Item $tmp
}
```

所有表格已成功导入 SSMS:

Rank	Institutions	Countries/Regions	Web of Science Documents	Cites	Cites/Paper	Top Papers
1	CHINESE ACADEMY OF SCIENCES	CHINA MAINLAND	15661	332254	21.22	251
2	CHINESE ACADEMY OF AGRICULTURAL SCIENCES	CHINA MAINLAND	12222	223855	18.32	198
3	UNITED STATES DEPARTMENT OF AGRICULTURE (USDA)	USA	12564	220644	17.56	105
4	CHINA AGRICULTURAL UNIVERSITY	CHINA MAINLAND	10052	207779	20.67	166
5	INRAE	FRANCE	9314	187838	20.17	118
6	JIANGNAN UNIVERSITY	CHINA MAINLAND	8027	181325	22.59	166
7	NORTHWEST A&F UNIVERSITY - CHINA	CHINA MAINLAND	9024	180634	20.02	162
8	WAGENINGEN UNIVERSITY & RESEARCH	NETHERLANDS	6819	172734	25.33	152
9	CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS...	SPAIN	7656	167593	21.89	94
10	NANJING AGRICULTURAL UNIVERSITY	CHINA MAINLAND	6480	137921	21.28	99
11	UNIVERSITY OF CALIFORNIA SYSTEM	USA	6175	129479	20.97	110
12	UNIVERSITY OF CHINESE ACADEMY OF SCIENCES, CAS	CHINA MAINLAND	5950	127682	21.46	95
13	ZHEJIANG UNIVERSITY	CHINA MAINLAND	5313	125715	23.66	148
14	EGYPTIAN KNOWLEDGE BANK (EKB)	EGYPT	7861	117405	14.94	103
15	INDIAN COUNCIL OF AGRICULTURAL RESEARCH (ICAR)	INDIA	12753	111262	8.72	66
16	SOUTH CHINA UNIVERSITY OF TECHNOLOGY	CHINA MAINLAND	3524	108282	30.73	117
17	OGIAR	FRANCE	5262	107193	20.37	69
18	HUAZHONG AGRICULTURAL UNIVERSITY	CHINA MAINLAND	4855	97482	20.08	68
19	UNIVERSIDADE DE SAO PAULO	BRAZIL	6169	95242	15.44	35
20	JIANGSU UNIVERSITY	CHINA MAINLAND	4084	89040	21.80	176
21	STATE UNIVERSITY SYSTEM OF FLORIDA	USA	5172	86527	16.73	62
22	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (C...	FRANCE	4113	81348	19.78	60
23	CORNELL UNIVERSITY	USA	3551	78744	22.18	59

4. 优化关系型数据，并整理一个合理的schema

为什么要优化？

为了方便查询，优化做的好，5-7题会非常简单

怎么优化？

目前是一个学科一个表格，我们要创建一个**总表**，整合所有数据

这个总表要增加一列，记录这行数据来源于哪个学科的排名表格：

```
USE ESI;
GO
-- 创建一个统一的总表，用来整合所有 staging_* 数据
IF OBJECT_ID('dbo.esi_rankings') IS NOT NULL DROP TABLE dbo.esi_rankings;
CREATE TABLE dbo.esi_rankings(
    id INT IDENTITY(1,1) PRIMARY KEY,          -- 自增主键
    discipline NVARCHAR(128) NOT NULL,         -- 学科名（来自表名）
    [rank] INT NULL,                           -- 排名
    institution NVARCHAR(255) NULL,            -- 机构名
    country_region NVARCHAR(128) NULL,         -- 国家或地区
    docs INT NULL,                             -- 论文数
    cites INT NULL,                           -- 引用数
    cites_per_paper DECIMAL(10,2) NULL,       -- 每篇引用数
    top_papers INT NULL                       -- 高被引论文数
);
GO
```

然后我们要遍历每个表格，从表名提取学科名，再把学科名结合其他信息，插入总表中：

```
-- 遍历所有 staging_* 表, 将数据导入 esi_rankings
DECLARE @t SYSNAME, @sql NVARCHAR(MAX), @disc NVARCHAR(128);

-- 获取所有 staging_* 表名
DECLARE cur CURSOR LOCAL FAST_FORWARD FOR
SELECT name FROM sys.tables WHERE name LIKE 'staging\_%' ESCAPE '\';

OPEN cur; FETCH NEXT FROM cur INTO @t;
WHILE @@FETCH_STATUS = 0
BEGIN
    -- 从表名提取学科名, 例如 staging_PHYSICS -> "PHYSICS"
    SET @disc = REPLACE(SUBSTRING(@t, LEN('staging\_')+1, 200), '_', ' ');

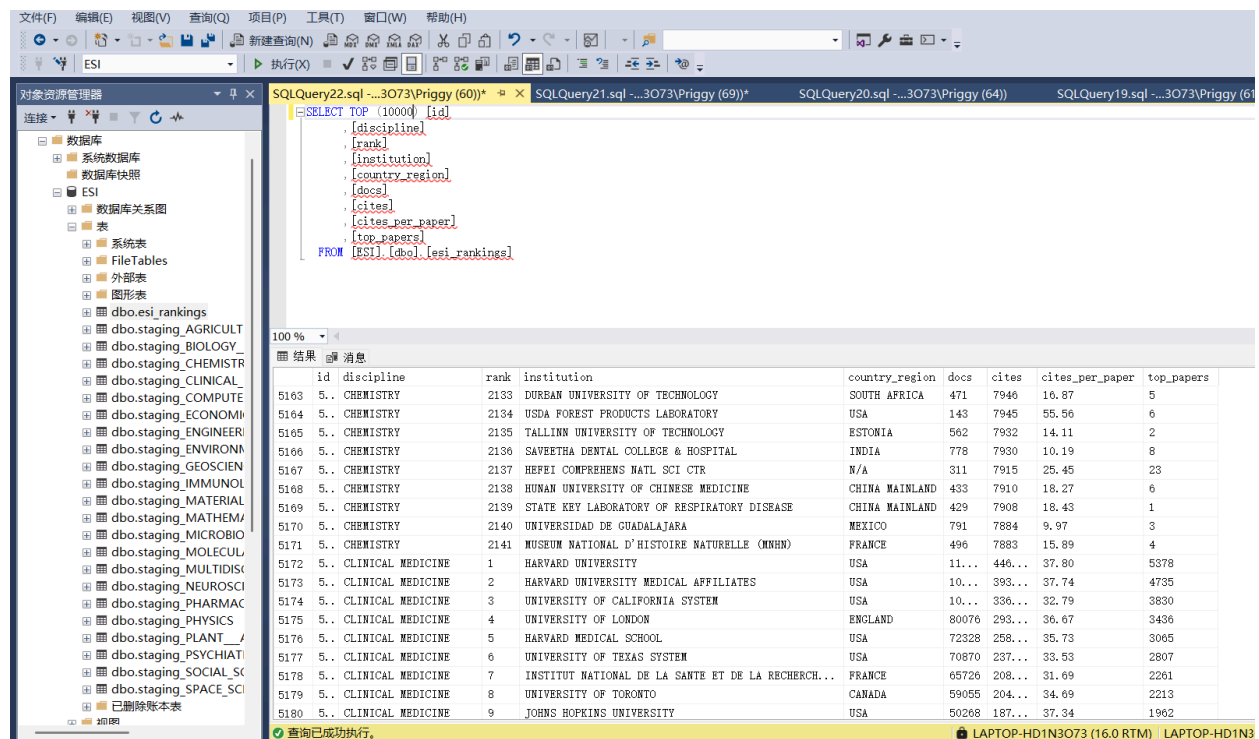
    -- 构造动态 SQL 插入语句
    SET @sql = '
        INSERT INTO dbo.esi_rankings(discipline,
[rank],institution,country_region,docs,cites,cites_per_paper,top_papers)
        SELECT ''' + @disc + ''',
            TRY_CONVERT(INT,[Rank]),
            [Institutions],
            [Countries/Regions],
            TRY_CONVERT(INT,[web of Science Documents]),
            TRY_CONVERT(INT,[Cites]),
            TRY_CONVERT(DECIMAL(10,2),[Cites/Paper]),
            TRY_CONVERT(INT,[Top Papers])
        FROM dbo.' + QUOTENAME(@t) + '
        WHERE [Institutions] IS NOT NULL AND [Institutions] NOT LIKE ''Copyright%'';
    ';
    EXEC(@sql);

    FETCH NEXT FROM cur INTO @t;
END
CLOSE cur; DEALLOCATE cur;
GO
```

为了方便后续查询, 添加学科和机构索引:

```
-- 为后续查询 (第5-7题) 建立索引
CREATE INDEX IX_esi_rank_disc ON dbo.esi_rankings(discipline); -- 按学科快速筛选
CREATE INDEX IX_esi_rank_inst ON dbo.esi_rankings(institution); -- 按机构快速查询
GO
```

最终的总表如下:



id	discipline	rank	institution	country_region	docs	cites	cites_per_paper	top_papers
5163	CHEMISTRY	2133	DURBAN UNIVERSITY OF TECHNOLOGY	SOUTH AFRICA	471	7946	16.87	5
5164	CHEMISTRY	2134	USDA FOREST PRODUCTS LABORATORY	USA	143	7945	55.56	6
5165	CHEMISTRY	2135	TALLINN UNIVERSITY OF TECHNOLOGY	ESTONIA	562	7932	14.11	2
5166	CHEMISTRY	2136	SAVEETHA DENTAL COLLEGE & HOSPITAL	INDIA	778	7930	10.19	8
5167	CHEMISTRY	2137	HEFEI COMPREHENS NATL SCI CTR	N/A	311	7915	25.45	23
5168	CHEMISTRY	2138	HUNAN UNIVERSITY OF CHINESE MEDICINE	CHINA MAINLAND	433	7910	18.27	6
5169	CHEMISTRY	2139	STATE KEY LABORATORY OF RESPIRATORY DISEASE	CHINA MAINLAND	429	7908	18.43	1
5170	CHEMISTRY	2140	UNIVERSIDAD DE GUADALAJARA	MEXICO	791	7884	9.97	3
5171	CHEMISTRY	2141	MUSEUM NATIONAL D'HISTOIRE NATURELLE (MNHN)	FRANCE	496	7883	15.89	4
5172	CLINICAL MEDICINE	1	HARVARD UNIVERSITY	USA	11...	446...	37.80	5378
5173	CLINICAL MEDICINE	2	HARVARD UNIVERSITY MEDICAL AFFILIATES	USA	10...	393...	37.74	4735
5174	CLINICAL MEDICINE	3	UNIVERSITY OF CALIFORNIA SYSTEM	USA	10...	336...	32.79	3830
5175	CLINICAL MEDICINE	4	UNIVERSITY OF LONDON	ENGLAND	80076	293...	36.67	3436
5176	CLINICAL MEDICINE	5	HARVARD MEDICAL SCHOOL	USA	72328	258...	35.73	3065
5177	CLINICAL MEDICINE	6	UNIVERSITY OF TEXAS SYSTEM	USA	70870	237...	33.53	2807
5178	CLINICAL MEDICINE	7	INSTITUT NATIONAL DE LA SANTE ET DE LA RECHERCH...	FRANCE	65726	208...	31.69	2261
5179	CLINICAL MEDICINE	8	UNIVERSITY OF TORONTO	CANADA	59055	204...	34.69	2213
5180	CLINICAL MEDICINE	9	JOHNS HOPKINS UNIVERSITY	USA	50268	187...	37.34	1902

5. 通过写SQL语句，获取华东师范大学在各个学科中的排名

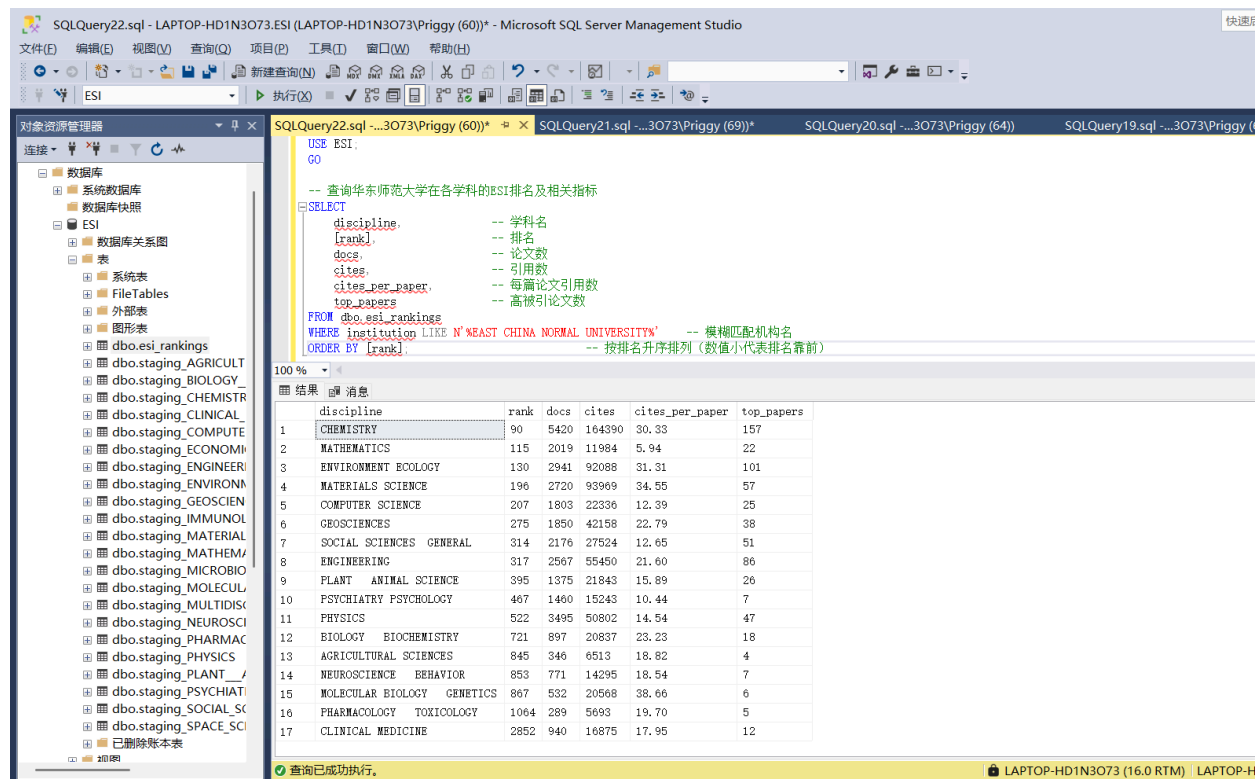
注意，华东师范大学在表格中是英文名称，即 EAST CHINA NORMAL UNIVERSITY

从 dbo.esi_rankings 表格中，选择机构名为 EAST CHINA NORMAL UNIVERSITY 的行，按 rank 从小到大展示

```
USE ESI;
GO

-- 查询华东师范大学在各学科的ESI排名及相关指标
SELECT
    discipline,          -- 学科名
    [rank],              -- 排名
    docs,                -- 论文数
    cites,               -- 引用数
    cites_per_paper,     -- 每篇论文引用数
    top_papers           -- 高被引论文数
FROM dbo.esi_rankings
WHERE institution LIKE N'%EAST CHINA NORMAL UNIVERSITY%' -- 模糊匹配机构名
ORDER BY [rank];      -- 按排名升序排列（数值小代表排名靠前）
```

运行结果如下：



The screenshot shows the Microsoft SQL Server Management Studio interface. The left pane displays the '对象资源管理器' (Object Explorer) with the 'ESI' database selected. The central pane shows the SQL query being executed, which is the same query as in the previous block. The bottom pane displays the results of the query, showing a table with 7 columns: discipline, rank, docs, cites, cites_per_paper, and top_papers. The results are sorted by rank in ascending order.

	discipline	rank	docs	cites	cites_per_paper	top_papers
1	CHEMISTRY	90	5420	164390	30.33	157
2	MATHEMATICS	115	2019	11984	5.94	22
3	ENVIRONMENT ECOLOGY	130	2941	92088	31.31	101
4	MATERIALS SCIENCE	196	2720	93969	34.55	57
5	COMPUTER SCIENCE	207	1803	22336	12.39	25
6	GEOSCIENCES	275	1850	42158	22.79	38
7	SOCIAL SCIENCES GENERAL	314	2176	27524	12.65	51
8	ENGINEERING	317	2567	55450	21.60	86
9	PLANT ANIMAL SCIENCE	395	1375	21843	15.89	26
10	PSYCHIATRY PSYCHOLOGY	467	1460	15243	10.44	7
11	PHYSICS	522	3495	50802	14.54	47
12	BIOLOGY BIOCHEMISTRY	721	897	20837	23.23	18
13	AGRICULTURAL SCIENCES	845	346	6513	18.82	4
14	NEUROSCIENCE BEHAVIOR	853	771	14295	18.54	7
15	MOLECULAR BIOLOGY GENETICS	867	532	20568	38.66	6
16	PHARMACOLOGY TOXICOLOGY	1064	289	5693	19.70	5
17	CLINICAL MEDICINE	2852	940	16875	17.95	12

为了便于老师检查，我想把查询结果输出到一个文件中

把SQL代码保存为 SQLQuery5.sql，然后在 PowerShell 中运行：

```
sqlcmd -S localhost -d ESI -E -i "D:\ESI\SQLQuery5.sql" -o "D:\ESI\result5.csv" -s "," -w -f 65001
```

文件在附件中供老师检查（5-7题都有，为了节省篇幅，后面不再说明）：

	A	B	C	D	E	F	G	H	I	J	K	L
1	已将数据库上下文更改为 "ESI"。											
2	discipline	rank	docs	cites	cites_per_paper	top_papers						
3	-----	----	----	-----	-----	-----						
4	CHEMISTRY	90	5420	164390	30.33	157						
5	MATHEMATICS	115	2019	11984	5.94	22						
6	ENVIRONMENTAL	130	2941	92088	31.31	101						
7	MATERIALS	196	2720	93969	34.55	57						
8	COMPUTER	207	1803	22336	12.39	25						
9	GEOSCIENCES	275	1850	42158	22.79	38						
10	SOCIAL SCIENCES	314	2176	27524	12.65	51						
11	ENGINEERING	317	2567	55450	21.6	86						
12	PLANT AND ANIMAL	395	1375	21843	15.89	26						
13	PSYCHIATRY	467	1460	15243	10.44	7						
14	PHYSICS	522	3495	50802	14.54	47						
15	BIOLOGY	721	897	20837	23.23	18						
16	AGRICULTURE	845	346	6513	18.82	4						
17	NEUROSCIENCE	853	771	14295	18.54	7						
18	MOLECULAR	867	532	20568	38.66	6						
19	PHARMACOLOGY	1064	289	5693	19.7	5						
20	CLINICAL	2852	940	16875	17.95	12						
21												
22	(17 行受影响)											
23												
24												

6. 通过写SQL语句，获取中国（大陆地区）大学在各个学科中的表现

如何限定范围（中国大陆）？

限定 country_region 为 CHINA MAINLAND 即可

如何衡量在各个学科中的表现？

看各学科上榜机构数量、平均排名、各学科被引论文总数、各学科论文总数、各学科引用总数、平均每篇引用

这些指标中，平均排名是最权威的，所有按平均排名升序展示

```
USE ESI;
GO

-- 第6题：分析中国大陆高校在各学科的表现
SELECT
    discipline, -- 学科
    COUNT(*) AS institution_count, -- 上榜机构数量
    AVG([rank]) AS avg_rank, -- 平均排名
    SUM(top_papers) AS total_top_papers, -- 各学科高被引论文总数
    SUM(docs) AS total_docs, -- 各学科论文总数
    SUM(cites) AS total_cites, -- 各学科引用总数
    ROUND(SUM(cites) * 1.0 / NULLIF(SUM(docs), 0), 2) AS avg_cites_per_paper -- 平均每篇引用
FROM dbo.esi_rankings
WHERE country_region LIKE N'%CHINA MAINLAND%' -- 限定中国大陆高校
GROUP BY discipline
ORDER BY avg_rank; -- 按平均排名升序排列
```

运行结果如下：

USE ESI;
GO

-- 第6题：分析中国大陆高校在各学科的表现

```
SELECT
    discipline, -- 学科
    COUNT(*) AS institution_count, -- 上榜机构数量
    AVG(rank) AS avg_rank, -- 平均排名
    SUM(top_papers) AS total_top_papers, -- 各学科高被引论文总数
    SUM(docs) AS total_docs, -- 各学科论文总数
    SUM(cites) AS total_cites, -- 各学科引用总数
    ROUND(SUM(cites) * 1.0 / NULLIF(SUM(docs), 0), 2) AS avg_cites_per_paper -- 平均每篇引用
FROM dbo.esi_rankings
WHERE country_region LIKE N'%CHINA MAINLAND%' -- 限定中国大陆高校
GROUP BY discipline
ORDER BY avg_rank -- 按平均排名升序排列
```

	discipline	institution_count	avg_rank	total_top_papers	total_docs	total_cites	avg_cites_per_paper
1	MULTIDISCIPLINARY	17	125	118	2665	134168	50.3400000000000
2	SPACE SCIENCE	10	151	778	45825	993908	21.6900000000000
3	MATHEMATICS	86	190	2367	124005	942827	7.60000000000000
4	ECONOMICS BUSINESS	55	287	1372	62617	988422	15.7900000000000
5	COMPUTER SCIENCE	190	378	5882	325041	5330017	16.4000000000000
6	MICROBIOLOGY	92	427	1347	81415	1672848	20.5500000000000
7	PHYSICS	112	554	9832	551518	9516603	17.2600000000000
8	GEOSCIENCES	177	571	7381	475790	8302420	17.4500000000000
9	MOLECULAR BIOLOGY GENETICS	115	600	3318	232132	6896738	29.7100000000000
10	NEUROSCIENCE BEHAVIOR	71	614	1591	115210	2180233	18.9200000000000
11	IMMUNOLOGY	84	624	1050	80843	1681473	20.8000000000000
12	PHARMACOLOGY TOXICOLOGY	182	625	2556	217471	3502876	16.1100000000000
13	AGRICULTURAL SCIENCES	251	639	4615	274214	5080975	18.5300000000000
14	MATERIALS SCIENCE	375	668	21769	1119272	32354021	28.9100000000000

查询已成功执行。

为了便于老师检查，我想把查询结果输出到一个文件中

把SQL代码保存为 SQLQuery6.sql，然后在 PowerShell 中运行：

```
sqlcmd -S localhost -d ESI -E -i "D:\ESI\SQLQuery6.sql" -o "D:\ESI\result6.csv" -s "," -w -f 65001
```

7. 通过写SQL语句，分析全球不同区域在各个学科中的表现

全球不同区域需要我们手动划分，按国家或地区填充所属大洲

```
-- 给表添加 region_group 字段
ALTER TABLE dbo.esi_rankings ADD region_group NVARCHAR(64);
GO

-- 按国家或地区填充所属大洲
UPDATE dbo.esi_rankings SET region_group = 'Asia'
WHERE country_region LIKE N'%CHINA%' OR country_region LIKE N'%JAPAN%' OR country_region LIKE N'%KOREA%' OR
country_region LIKE N'%INDIA%';

UPDATE dbo.esi_rankings SET region_group = 'Europe'
WHERE country_region IN (N'UNITED KINGDOM', N'FRANCE', N'GERMANY', N'ITALY', N'SPAIN', N'NETHERLANDS');

UPDATE dbo.esi_rankings SET region_group = 'North America'
WHERE country_region IN (N'USA', N'CANADA', N'MEXICO');

UPDATE dbo.esi_rankings SET region_group = 'South America'
WHERE country_region IN (N'BRAZIL', N'ARGENTINA', N'CHILE');

UPDATE dbo.esi_rankings SET region_group = 'Oceania'
WHERE country_region IN (N'AUSTRALIA', N'NEW ZEALAND');

UPDATE dbo.esi_rankings SET region_group = 'Africa'
WHERE country_region IN (N'EGYPT', N'SOUTH AFRICA');
GO
```

这样，表格右侧就会增加一列 region_group，便于区分全球不同区域：

对象资源管理器

SQLQuery24.sql ...3073\Priggy (55)

```

SELECT TOP (1000) [id]
, [discipline]
, [rank]
, [institution]
, [country_region]
, [docs]
, [cites]
, [cites_per_paper]
, [top_papers]
, [region_group]
FROM [ESI].[dbo].[esi_rankings]

```

100 %

结果 消息

	id	discipline	rank	institution	country_region	docs	cites	cites_per_paper	top_papers	region_group
1	1	AGRICULTURAL SCIENCES	1	CHINESE ACADEMY OF SCIENCES	CHINA MAIN	12222	223855	18.32	251	Asia
2	2	AGRICULTURAL SCIENCES	2	CHINESE ACADEMY OF AGRICULTURAL SCIENCES	CHINA MAINLAND	12564	220644	17.56	105	North America
3	3	AGRICULTURAL SCIENCES	3	UNITED STATES DEPARTMENT OF AGRICULTURE (USDA)	USA	10052	207779	20.67	166	Asia
4	4	AGRICULTURAL SCIENCES	4	CHINA AGRICULTURAL UNIVERSITY	CHINA MAINLAND	9314	187838	20.17	118	Europe
5	5	AGRICULTURAL SCIENCES	5	INRAE	FRANCE	8027	181325	22.59	166	Asia
6	6	AGRICULTURAL SCIENCES	6	JIANGNAN UNIVERSITY	CHINA MAINLAND	9024	180634	20.02	162	Asia
7	7	AGRICULTURAL SCIENCES	7	NORTHWEST A&F UNIVERSITY - CHINA	CHINA MAINLAND	6819	172734	25.33	152	Europe
8	8	AGRICULTURAL SCIENCES	8	WAGENINGEN UNIVERSITY & RESEARCH	NETHERLANDS	7656	167593	21.89	94	Europe
9	9	AGRICULTURAL SCIENCES	9	CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS...	SPAIN	6480	137921	21.28	99	Asia
10	10	AGRICULTURAL SCIENCES	10	WUZHONG AGRICULTURAL UNIVERSITY	CHINA MAINLAND	6175	129479	20.97	110	North America
11	11	AGRICULTURAL SCIENCES	11	UNIVERSITY OF CALIFORNIA SYSTEM	USA	5950	127682	21.46	95	Asia
12	12	AGRICULTURAL SCIENCES	12	UNIVERSITY OF CHINESE ACADEMY OF SCIENCES, CAS	CHINA MAINLAND	5313	125715	23.66	148	Asia
13	13	AGRICULTURAL SCIENCES	13	ZHEJIANG UNIVERSITY	CHINA MAINLAND	7861	117405	14.94	103	Africa
14	14	AGRICULTURAL SCIENCES	14	EGYPTIAN KNOWLEDGE BANK (EKB)	EGYPT	12753	111262	8.72	66	Asia
15	15	AGRICULTURAL SCIENCES	15	INDIAN COUNCIL OF AGRICULTURAL RESEARCH (ICAR)	INDIA	3524	108282	30.73	117	Asia
16	16	AGRICULTURAL SCIENCES	16	SOUTH CHINA UNIVERSITY OF TECHNOLOGY	CHINA MAINLAND	5262	107193	20.37	69	Europe
17	17	AGRICULTURAL SCIENCES	17	OGIAR	FRANCE	4855	97482	20.08	68	Asia
18	18	AGRICULTURAL SCIENCES	18	HUAZHONG AGRICULTURAL UNIVERSITY	CHINA MAINLAND	6169	95242	15.44	35	South America
19	19	AGRICULTURAL SCIENCES	19	UNIVERSIDADE DE SAO PAULO	BRAZIL					

查询已成功执行。

LAPTOP-HD1N3073 (16.0 RTM) LAPTOP-HD1N3073\Priggy...

之后的思路和第6题很相似，只不过 GROUP BY discipline 变为 GROUP BY region_group, discipline，从而将每个区域都展示一遍：

```

USE ESI;
GO

```

-- 第7题：分析全球不同区域在各学科的表现

```

SELECT

```

```

    region_group,                -- 区域（亚洲、欧洲、北美等）
    discipline,                  -- 学科
    COUNT(*) AS institution_count, -- 上榜机构数
    AVG([rank]) AS avg_rank,      -- 平均排名
    SUM(top_papers) AS total_top_papers, -- 高被引论文总数
    SUM(docs) AS total_docs,      -- 论文总数
    SUM(cites) AS total_cites,    -- 引用总数
    ROUND(SUM(cites)*1.0/NULLIF(SUM(docs),0),2) AS avg_cites_per_paper -- 平均每篇引用

```

```

FROM dbo.esi_rankings
WHERE region_group IS NOT NULL
GROUP BY region_group, discipline
ORDER BY region_group, avg_rank;

```

运行结果如下：

对象资源管理器

SQLQuery24.sql ...3073\Priggy (55)*

```

SELECT
    region_group,                -- 区域（亚洲、欧洲、北美等）
    discipline,                  -- 学科
    COUNT(*) AS institution_count, -- 上榜机构数
    AVG([rank]) AS avg_rank,      -- 平均排名
    SUM(top_papers) AS total_top_papers, -- 高被引论文总数
    SUM(docs) AS total_docs,      -- 论文总数
    SUM(cites) AS total_cites,    -- 引用总数
    ROUND(SUM(cites)*1.0/NULLIF(SUM(docs),0),2) AS avg_cites_per_paper -- 平均每篇引用
FROM dbo.esi_rankings
WHERE region_group IS NOT NULL
GROUP BY region_group, discipline
ORDER BY region_group, avg_rank;

```

100 %

结果 消息

	region_group	discipline	institution_count	avg_rank	total_top_papers	total_docs	total_cites	avg_cites_per_paper
13	Africa	AGRICULTURAL SCIENCES	26	729	284	23036	343630	14.920000000000000
14	Africa	MOLECULAR BIOLOGY GENETICS	6	794	88	5455	164566	30.170000000000000
15	Africa	PLANT ANIMAL SCIENCE	41	850	1096	53105	720980	13.580000000000000
16	Africa	BIOLOGY BIOCHEMISTRY	23	941	309	28307	564315	19.940000000000000
17	Africa	MATERIALS SCIENCE	27	1068	213	36737	732479	19.940000000000000
18	Africa	ENVIRONMENT ECOLOGY	36	1092	583	38955	812124	20.850000000000000
19	Africa	SOCIAL SCIENCES GENERAL	25	1094	416	32769	354639	10.820000000000000
20	Africa	CHEMISTRY	44	1251	656	85181	1535456	18.030000000000000
21	Africa	ENGINEERING	53	1322	1120	79909	1421619	17.790000000000000
22	Africa	CLINICAL MEDICINE	53	2796	2441	99658	2829332	28.390000000000000
23	Asia	MULTIDISCIPLINARY	23	133	140	3472	165784	47.750000000000000
24	Asia	SPACE SCIENCE	18	144	1541	74117	1840336	24.830000000000000
25	Asia	MATHEMATICS	95	193	2500	141531	1037522	7.330000000000000
26	Asia	ECONOMICS BUSINESS	65	294	1488	73012	1139002	15.600000000000000
27	Asia	COMPUTER SCIENCE	262	418	6696	400648	6365399	15.890000000000000
28	Asia	MICROBIOLOGY	133	453	1584	108320	2145050	19.800000000000000
29	Asia	PHYSICS	207	554	14401	823560	15662866	19.020000000000000
30	Asia	GEOSCIENCES	250	608	8416	566526	9858409	17.400000000000000
31	Asia	MOLECULAR BIOLOGY GENETICS	184	653	4339	297123	9286170	31.250000000000000

查询已成功执行。

LAPTOP-HD1N3073 (16.0 RTM) LAPTOP-HD1N3073\Priggy...

为了便于老师检查，我想把查询结果输出到一个文件中

把SQL代码保存为 SQLQuery7.sql，然后在 PowerShell 中运行：

```
sqlcmd -S localhost -d ESI -E -i "D:\ESI\SQLQuery7.sql" -o "D:\ESI\result7.csv" -s "," -w -f 65001
```