

Hive作业

辅助资料

- 基本**SQL**教程:
- <https://www.w3school.com.cn/sql/index.asp>
- <https://www.liaoxuefeng.com/wiki/1177760294764384>
- Hive DDL:
- <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL>

数据解释-hive_sql_test1.t_user表

- **t_user**观众表共**6000+**条数据
- 字段为: **UserID, Sex, Age, Occupation, Zipcode**
- 字段中文解释: 用户**id**, 性别, 年龄, 职业, 邮编

	t_user.userid	t_user.sex	t_user.age	t_user.occupation	t_user.zipcode
1	1	F	1	10	48067
2	2	M	56	16	70072
3	3	M	25	15	55117
4	4	M	45	7	02460
5	5	M	25	20	55455
6	6	F	50	9	55117
7	7	M	35	1	06810
8	8	M	25	12	11413
9	9	M	25	17	61614
10	10	F	35	1	95370


数据解释-hive_sql_test1.t_movie表

- **t_movie**电影表共**3000+**条数据
- 字段为: **MovieID, MovieName, MovieType**
- 字段中文解释: 电影**ID**, 电影名, 电影类型

	t_movie.movieid	t_movie.moviename	t_movie.movietype
1	1	Toy Story (1995)	Animation Children's Comedy
2	2	Jumanji (1995)	Adventure Children's Fantasy
3	3	Grumpier Old Men (1995)	Comedy Romance
4	4	Waiting to Exhale (1995)	Comedy Drama
5	5	Father of the Bride Part II (1995)	Comedy
6	6	Heat (1995)	Action Crime Thriller
7	7	Sabrina (1995)	Comedy Romance
8	8	Tom and Huck (1995)	Adventure Children's
9	9	Sudden Death (1995)	Action
10	10	GoldenEye (1995)	Action Adventure Thriller

数据解释-hive_sql_test1.t_rating表

- **t_rating**影评表**100万+**条数据
- 字段为: **UserID, MovieID, Rate, Times**
- 字段中文解释: 用户**ID**, 电影**ID**, 评分, 评分时间



t_rating.userid		t_rating.movieid	t_rating.rate	t_rating.times
1	1	1193	5	978300760
2	1	661	3	978302109
3	1	914	3	978301968
4	1	3408	4	978300275
5	1	2355	5	978824291
6	1	1197	3	978302268
7	1	1287	5	978302039
8	1	2804	5	978300719
9	1	594	4	978302268
10	1	919	4	978301368

题目一

简单：展示电影ID为**2116**这部电影各年龄段的平均影评分

INFO : Completed executing command(queryId=hive_20210802153200_beb2a6f3-7f1f-49c1-9d86-584ff07b976b); Time taken: 39.145 seconds
INFO : OK

查询历史记录

保存的查询

结果 (7)



	age	avgrate
1	1	3.2941176470588234
2	18	3.3580246913580245
3	25	3.436548223350254
4	35	3.2278481012658227
5	45	2.8275862068965516
6	50	3.32
7	56	3.5

题目二

中等：找出男性评分最高且评分次数超过**50**次的**10**部电影，展示电影名，平均影评分和评分次数

	sex	name	avgrate	total
1	M	Sanjuro (1962)	4.639344262295082	61
2	M	Godfather, The (1972)	4.583333333333333	1740
3	M	Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	4.576628352490421	522
4	M	Shawshank Redemption, The (1994)	4.560625	1600
5	M	Raiders of the Lost Ark (1981)	4.520597322348094	1942
6	M	Usual Suspects, The (1995)	4.518248175182482	1370
7	M	Star Wars: Episode IV - A New Hope (1977)	4.495307167235495	2344
8	M	Schindler's List (1993)	4.49141503848431	1689
9	M	Paths of Glory (1957)	4.485148514851486	202
10	M	Wrong Trousers, The (1993)	4.478260869565218	644

题目三（选做）

困难：找出影评次数最多的女士所给出最高分的**10**部电影的**平均**影评评分，展示电影名和**平均**影评评分（可使用多行**SQL**）

t.moviename	t.avgrate
Close Shave, A (1995)	4.52054794520548
Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963)	4.4498902706656915
Rear Window (1954)	4.476190476190476
It Happened One Night (1934)	4.280748663101604
Crying Game, The (1992)	3.7314890154597236
Trust (1990)	4.188888888888889
Duck Soup (1933)	4.21043771043771
Night on Earth (1991)	3.747422680412371
Roger & Me (1989)	4.0739348370927315
Being John Malkovich (1999)	4.125390450691656

作业提交

在HUE上使用student账户跑SQL

截图上缴作业，要求SQL，如图queryId和结果在一张截图里
(PS: 下图的sql和结果没有截全)

```
6 group by c.moviename
7 having total >= 50
8 order by avgrate desc
```

```
limit 10
```

```
WARN :
```

```
INFO : Query ID = hive_20210802153445_b65aca39-a945-45b0-8274-26bc738c54af
```

```
INFO : Total jobs = 2
```

```
INFO : Starting task [Stage-9:MAPREDLOCAL] in serial mode
```

job_1626505659980_0396

查询历史记录

保存的查询

结果 (10)

	sex	name	avgrate	total
1	M	Sanjuro (1962)	4.639344262295082	61
2	M	Godfather, The (1972)	4.583333333333333	1740
3	M	Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	4.576628352490421	522
4	M	Shawshank Redemption, The (1994)	4.560625	1600
5	M	Raiders of the Lost Ark (1981)	4.520597322348094	1942

其他

在**HUE**上只有**student**一个账户，大家记得跑完清除历史记录！

```
INFO : 2021-08-02 15:35:39,468 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 12.73 sec
INFO : 2021-08-02 15:35:45,657 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 15.9 sec
```

查询历史记录



保存的查询

2 分钟前

清除查询历史记录

```
select "M" as sex, c.moviename as name, avg(a.rate) as c
```

提交链接及截止时间

- 作业提交链接: <https://jinshuju.net/f/DW0aL1>
- 作业截止日期: 8 月 8 日 23:59 前

附加作业：GeekFileFormat

请优先完成前面三个作业，**Hive**的练习更多的是对于**HQL**的使用，对于完成的同学，可以试着写一个**Hive**的**FileFormat**：**GeekFileFormat**

要求：

- 实现两个类：**GeekTextInputFormat**和**GeekTextOutputFormat**
- 建表时使用**create table ... stored as geek**来创建**GeekFormat**表
- 该表的文件类型为文本类型，非二进制类型
- 类似**Base64TextInputFormat**和**Base64TextOutputFormat**，**GeekFormat**也是用于加密
- 解密规则如下：文件中出现任何的**geek**，**geeeek**，**geeeeeeeeeeeek**等单词时，进行过滤，即删除该单词。**gek**需要保留。字母中连续的“e”最大长度为**256**个。
 - 例如： **This notebook can be geeeek used to geek install gek on all geeeek worker nodes, run data generation, and create the TPCDS geeeeeeeeeek database.**
 - 解密为： **This notebook can be used to install gek on all worker nodes, run data generation, and create the TPCDS database.**
- 【附加的附加】加密规则如下：文件输出时每随机**2**到**256**个单词，就插入一个**gee...k**，字母**e**的个数等于前面出现的非**gee...k**单词的个数。
 - 例如： **This notebook can be used to install gek on all worker nodes, run data generation, and create the TPCDS database.**
 - 加密为： **This notebook can be geeeek used to geek install gek on all geeeek worker nodes, run data generation, and create the TPCDS geeeeeeeeeek database.**