基于Hadoop的豆瓣电电影数据分析

1. 数据爬虫

从豆瓣电影利用webdriver动态爬取数据，

数据项介绍：

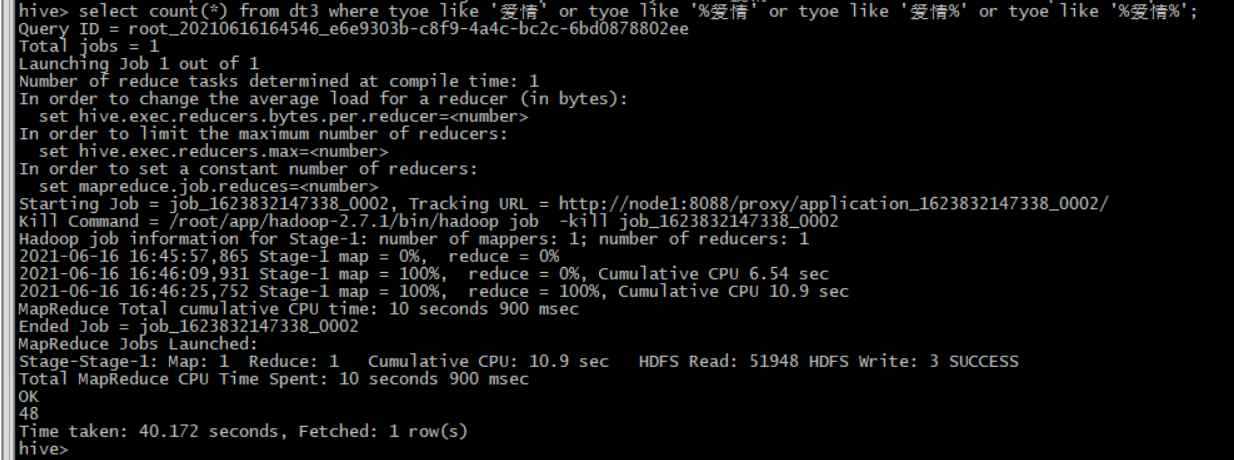
Img:爬取电影图片链接

Title:爬取数据的电影名

Score:对电影的评分

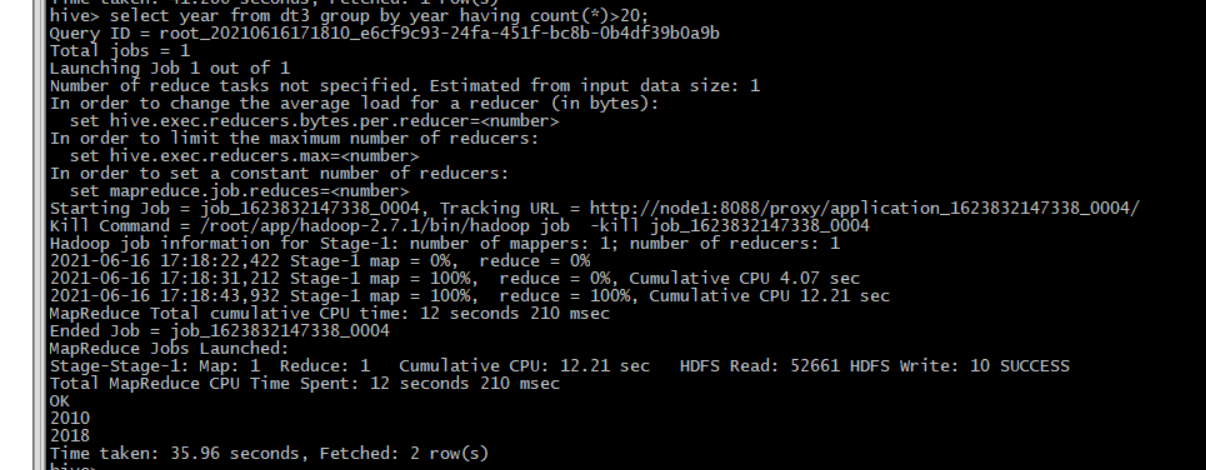
1. 求爱情类型电影的数量

select count(\*) from dt3 where type like '爱情' or type like '%爱情' or type like '爱情%' or type like '%爱情%';



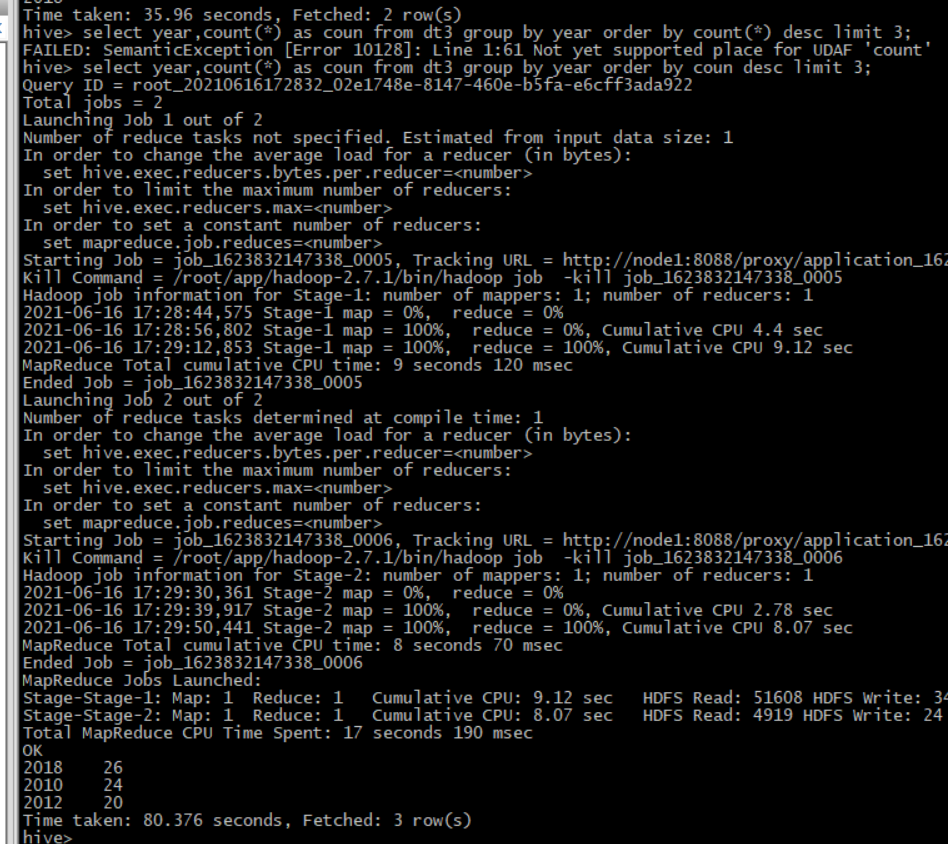
1. 找出总电影数目超过20的年份

select year from dt3 group by year having count(\*)>20;



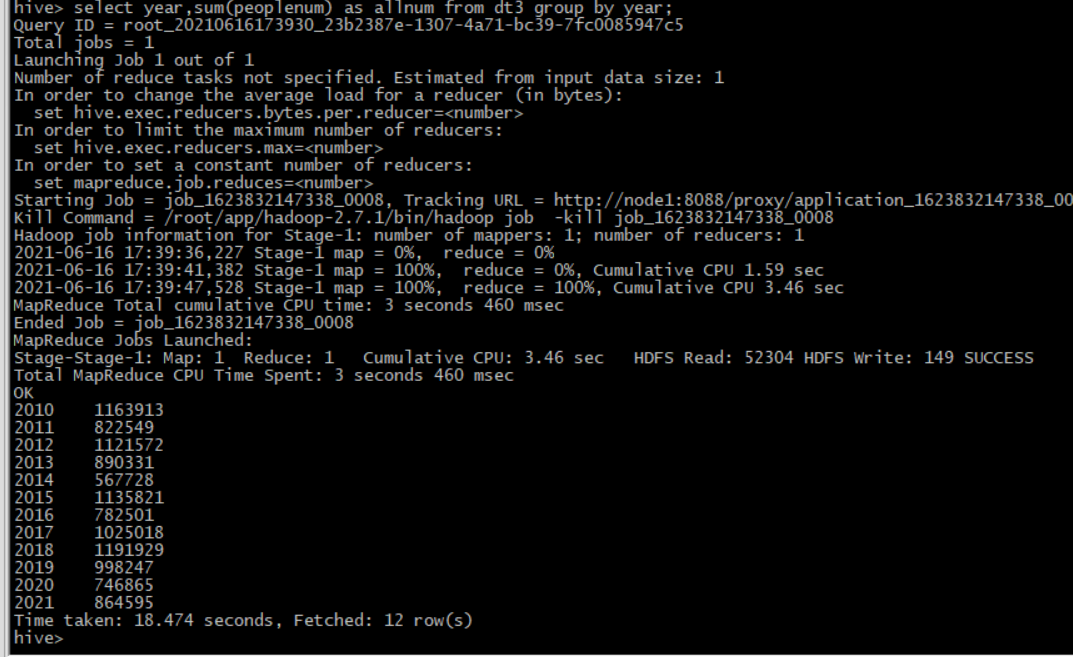
1. 统计上映电影数目最多的三年

select year,count(\*) as coun from dt3 group by year order by count(\*) desc limit 3;



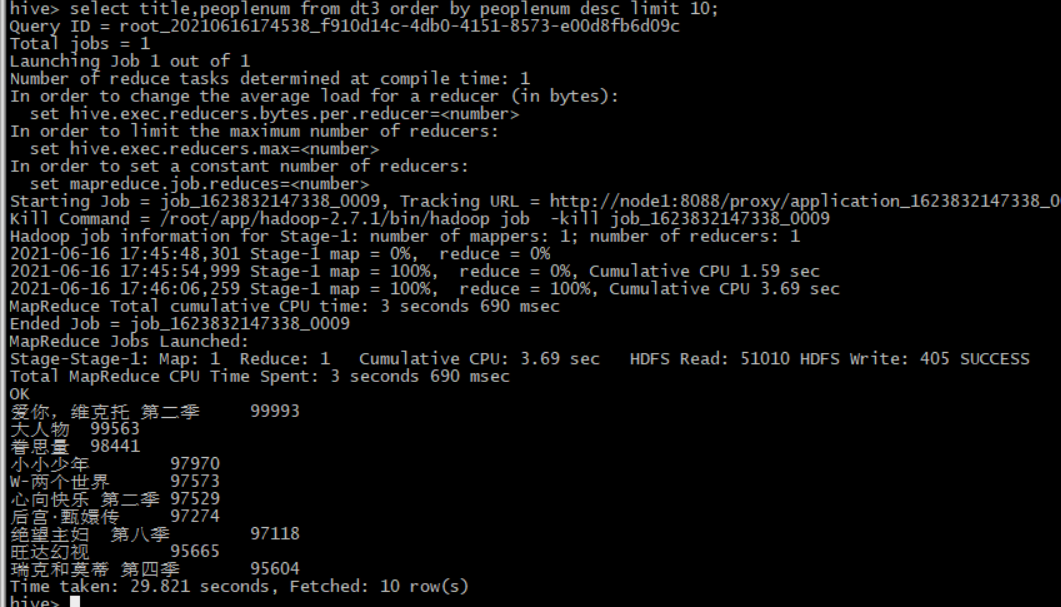
1. 统计每年上映电影的总观看量

select year,sum(peoplenum) as allnum from dt3 group by year;



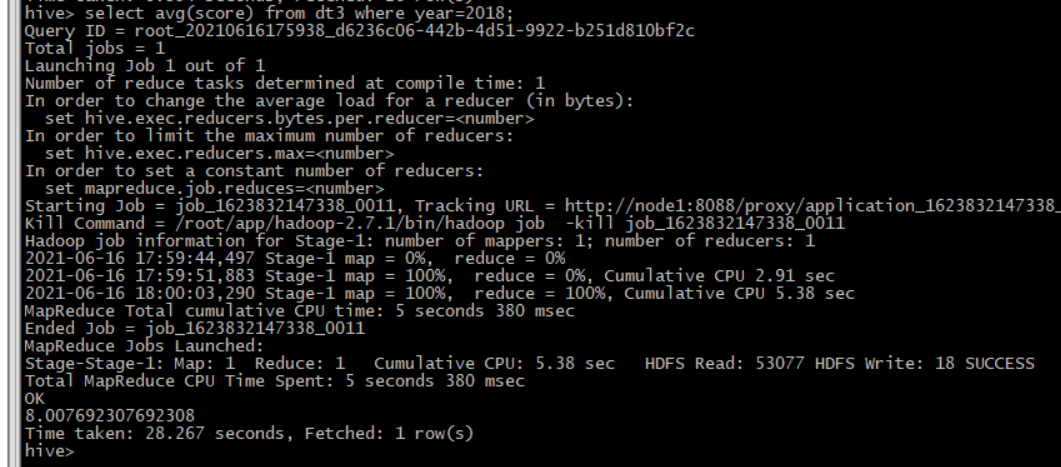
1. 查找dt3表的电影中票房前10的电影

select title,peoplenum from dt3 order by peoplenum desc limit 10;



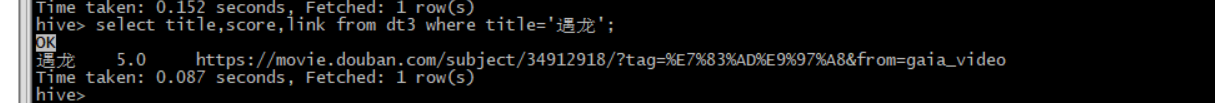
1. 求2018年上映电影的平均评分

select avg(score) from dt3 where year=2018;



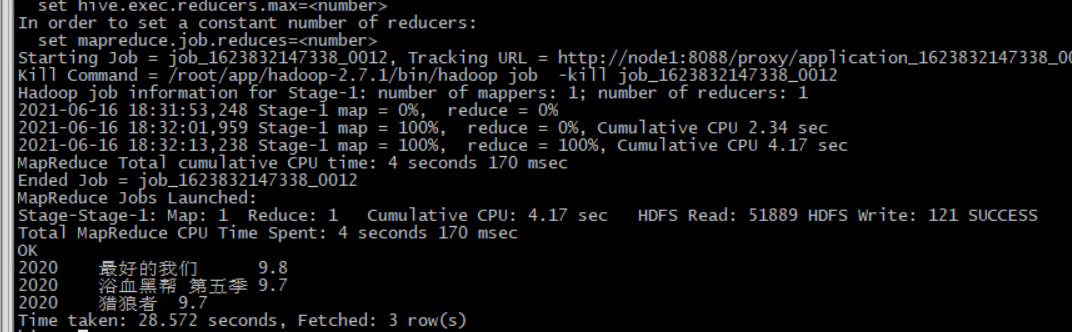
1. 查找电视剧遇龙的评分和观看网址

select title,score,link from dt3 where title=’遇龙’;



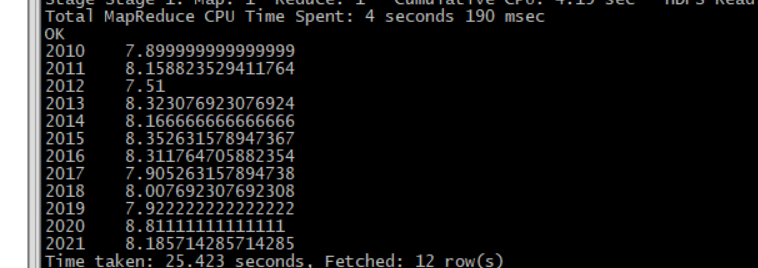
1. 统计2020年上映电影中评分最高的前三部电影

select year,title,score from dt3 where year=2020 order by score desc limit 3;



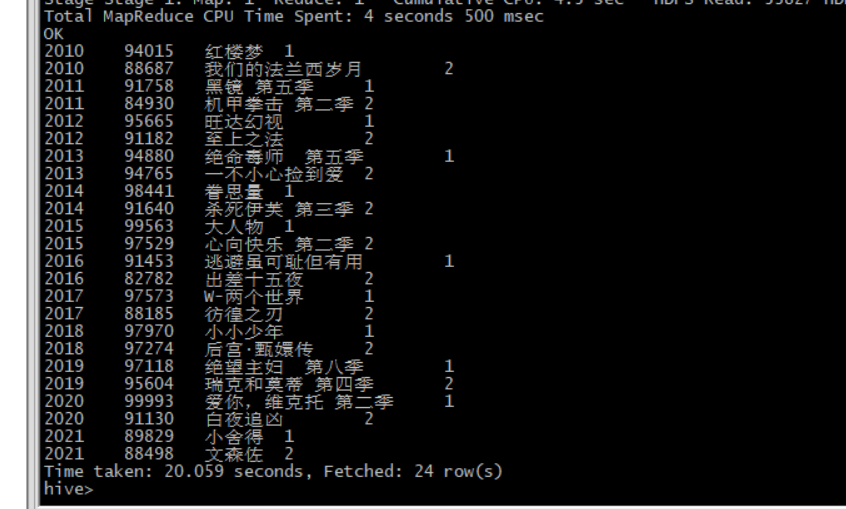
1. 统计每年上映电影的平均评分

select year,avg(score) as allavg from dt3 group by year;



10.统计每年观看人数前二的电影

select \* from (select year,peoplenum,title,row\_number() over(partition by year order by peoplenum desc) as num from dt3) k where num<=2;



Top-N

1.统计每年观看人数前二的电影

2.统计每年电影评分最高的电影

3.统计观看人数最多的三年

4.统计每年出版的电影的数目最多的三个

5.统计不同类型的电影的数目最多的两个

可视化

1. 统计每年评分5分以上和5分及以下所占的数量(柱状图)
2. 统计每年上映电影的平均评分(折线图)
3. 每年上映电影的总观看量(象形柱状图)
4. 统计每年上映的电影数量和平均评分(表格)