

MAT259

Assignment 1:Data Querying

Submitted by Mohit Hingorani

Harry Potter is one of the best selling fantasy literature novels of all time. It is estimated that over 400 million books have been sold till date!

Question is how many kids make it through the series? How many kids actually like Harry Potter? How many kids lose interest? Which ones are the most popular?

The last book was published in July 2007. Hence I can safely assume that all the readers are reading the books for the first time. I am interested in finding out how many kids started reading Harry Potter books in the last year, and how many finished the entire series.

The query is straightforward: I am querying for only Harry Potter books in the adult or juvenile section which were checkout on or after 2013-01-21.

The 7 SQL commands with the number of rows (i.e number of checkouts) and the query time is as follows:

```
***** 1 *****
select * from inraw where title = "Harry Potter and the Sorcerers stone" and
(itemtype= 'acbk' or itemtype ='jcbk') and cout >='2013-01-21'
```

361 rows 151ms

```
***** 2 *****
select * from inraw where title = "Harry Potter and the chamber of secrets" and
(itemtype= 'acbk' or itemtype ='jcbk') and cout >='2013-01-21'
```

275 rows 124ms

***** 3 *****

select * from inraw where title = "Harry Potter and the prisoner of azkaban" and
(itemtype= 'acbk' or itemtype ='jcbk') and cout >='2013-01-21'

240 rows 109 ms

***** 4 *****

select * from inraw where title = "Harry Potter and the goblet of fire" and
(itemtype= 'acbk' or itemtype ='jcbk') and cout >='2013-01-21'

262 rows 117ms

***** 5 *****

select * from inraw where title = "Harry Potter and the order of the phoenix" and
(itemtype= 'acbk' or itemtype ='jcbk') and cout >='2013-01-21'

229 rows 152 ms

***** 6 *****

select * from inraw where title = "Harry Potter and the Half Blood Prince" and
(itemtype= 'acbk' or itemtype ='jcbk') and cout >='2013-01-21'

234 rows 124ms

***** 7 *****

select * from inraw where title = "Harry Potter and the deathly hallows" and
(itemtype= 'acbk' or itemtype ='jcbk') and cout >='2013-01-21'

255 rows 125 ms

Results:

The results were quite interesting: There is a sharp drop in the reader's interest, but the ones who read the first 3 books usually read all of them.

The first book in the series was checked out 361 times in the last year and the last book in the series was checked out 255 times: A pretty impressive retention rate.

Book name	Checkouts	Time
1. <i>Harry Potter and the Philosopher's Stone</i>	361 rows	151ms
2. <i>Harry Potter and the Chamber of Secrets</i>	275 rows	124ms
3. <i>Harry Potter and the Prisoner of Azkaban</i>	240 rows	109 ms
4. <i>Harry Potter and the Goblet of Fire</i>	262 rows	117ms
5. <i>Harry Potter and the Order of the Phoenix</i>	229 rows	152 ms
6. <i>Harry Potter and the Half-Blood Prince</i>	234 rows	124ms
7. <i>Harry Potter and the Deathly Hallows</i>	255 rows	125 ms

It is interesting to note that there is a small spike at #4 Harry Potter and the Goblet of Fire & the final book of the series: Harry Potter and the Deathly Hallows. It would be interesting to further research the data using check-in, check-out times, number of pages in the book, average time a reader takes to complete the book etc and find out why certain books were checked out more than the others something I intend to do for my future projects.

The link to the CSV files is here:

<https://www.dropbox.com/sh/ggqqf8qlg3jvjb0/c3fKxmjlY2/csv>