COEN 380
GROUP PROJECT

Arthi Sivakumar
Nicholas Fong
Saurabh Somani
Shraddhaben Padariya
Vikas Shetty

Dataset

- MovieLens (https://grouplens.org/datasets/movielens/100k/)
- Data set consists of:
  - 100,000 ratings (1-5) from 943 users on 1682 movies.
  - Each user has rated at least 20 movies.
  - Simple demographic info for the users (age, gender, occupation, zip)
- Tables:
  - User (userid, age, gender, occupationid, zipcode)
  - Movie (movieid, title, release_date, genre)
  - Occupation (occupationid, occupation)
  - Zipcode (zipcode, city, state)
  - Data (userid, itemid, rating, timestamp)

Table	Column	Data Type	Length	Precision	Scale	Primary Key	Nullable	Default	Comment
USERS	USERID	NUMBER	22	-	-	1	-	-	-
	AGE	NUMBER	22	-	-	-	✓	-	-
	GENDER	VARCHAR2	1	-	-	-	✓	-	-
	OCCUPATIONID	NUMBER	22	-	-	-	✓	-	-
	ZIPCODE	VARCHAR2	30	-	-	-	✓	-	-
									1 - 5

COUNT(*)
943

Table	Column	Data Type	Length	Precision	Scale	Primary Key	Nullable	Default	Comment
OCCUPATION	OCCUPATIONID	NUMBER	22	-	-	1	-	-	-
	OCCUPATION	VARCHAR2	30	-	-	-	✓	-	-
									1 - 2

COUNT(*)
21

Table	Column	Data Type	Length	Precision	Scale	Primary Key	Nullable	Default	Comment
ZIPCODE	ZIPCODE	VARCHAR2	30	-	-	1	-	-	-
	CITY	VARCHAR2	30	-	-	-	✓	-	-
	STATE	VARCHAR2	30	-	-	-	✓	-	-
									1 - 3

COUNT(*)
42522

Table	Column	Data Type	Length	Precision	Scale	Primary Key	Nullable	Default	Comment
MOVIE	MOVIEID	NUMBER	22	-	-	1	-	-	-
	TITLE	VARCHAR2	255	-	-	-	✓	-	-
	RELEASE_DATE	VARCHAR2	30	-	-	-	✓	-	-
	UNKNOWN	NUMBER	22	-	-	-	✓	-	-
	ACTION	NUMBER	22	-	-	-	✓	-	-
	ADVENTURE	NUMBER	22	-	-	-	✓	-	-
	ANIMATION	NUMBER	22	-	-	-	✓	-	-
	CHILDRENS	NUMBER	22	-	-	-	✓	-	-
	COMEDY	NUMBER	22	-	-	-	✓	-	-
	CRIME	NUMBER	22	-	-	-	✓	-	-
	DOCUMENTARY	NUMBER	22	-	-	-	✓	-	-
	DRAMA	NUMBER	22	-	-	-	✓	-	-
	FANTASY	NUMBER	22	-	-	-	✓	-	-
	FILM_NOIR	NUMBER	22	-	-	-	✓	-	-
	HORROR	NUMBER	22	-	-	-	✓	-	-
	MUSICAL	NUMBER	22	-	-	-	✓	-	-
	MYSTERY	NUMBER	22	-	-	-	✓	-	-
	ROMANCE	NUMBER	22	-	-	-	✓	-	-
	SCI_FI	NUMBER	22	-	-	-	✓	-	-
	THRILLER	NUMBER	22	-	-	-	✓	-	-
	WAR	NUMBER	22	-	-	-	✓	-	-
	WESTERN	NUMBER	22	-	-	-	✓	-	-

1 - 22

COUNT(*)
1682

Table	Column	Data Type	Length	Precision	Scale	Primary Key	Nullable	Default	Comment
DATA	ID	NUMBER	22	-	-	1	-	-	-
	USERID	NUMBER	22	-	-	-	✓	-	-
	ITEMID	NUMBER	22	-	-	-	✓	-	-
	RATING	NUMBER	22	-	-	-	✓	-	-
	TIMESTAMP	NUMBER	22	-	-	-	✓	-	-

1 - 5

COUNT(*)
100000

Data Cleaning

- Removing unnecessary columns
- Removing duplicate rows
- Removing rows that don’t match foreign keys
- Removing commas in string values

Query 1

Select all the movies under Comedy and Crime genre

```
SELECT title, release_date
FROM movie
WHERE Comedy = 1 AND Crime = 1
```

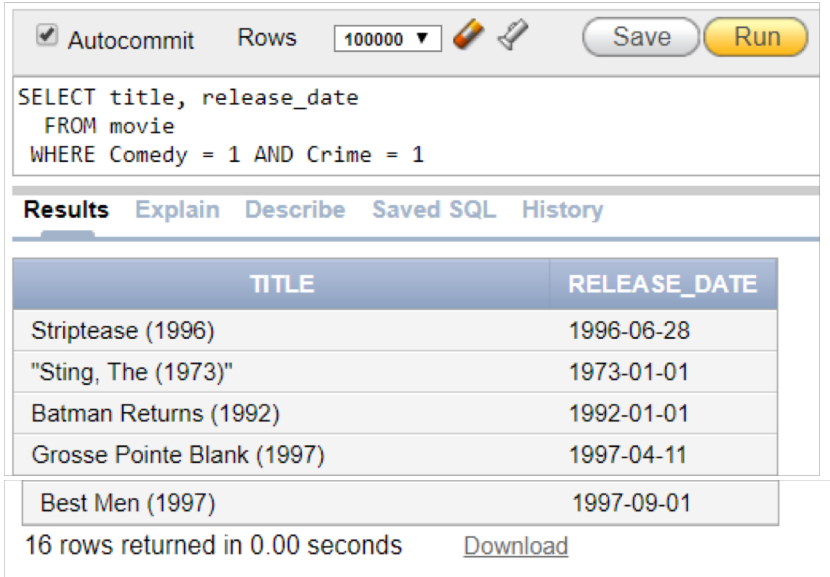
Results:

Rows fetched : 16

Oracle : 0.00 seconds

Hive: 11.368 seconds

Oracle and Hive query results



Autocommit Rows 100000 Save Run

```
SELECT title, release_date
  FROM movie
 WHERE Comedy = 1 AND Crime = 1
```

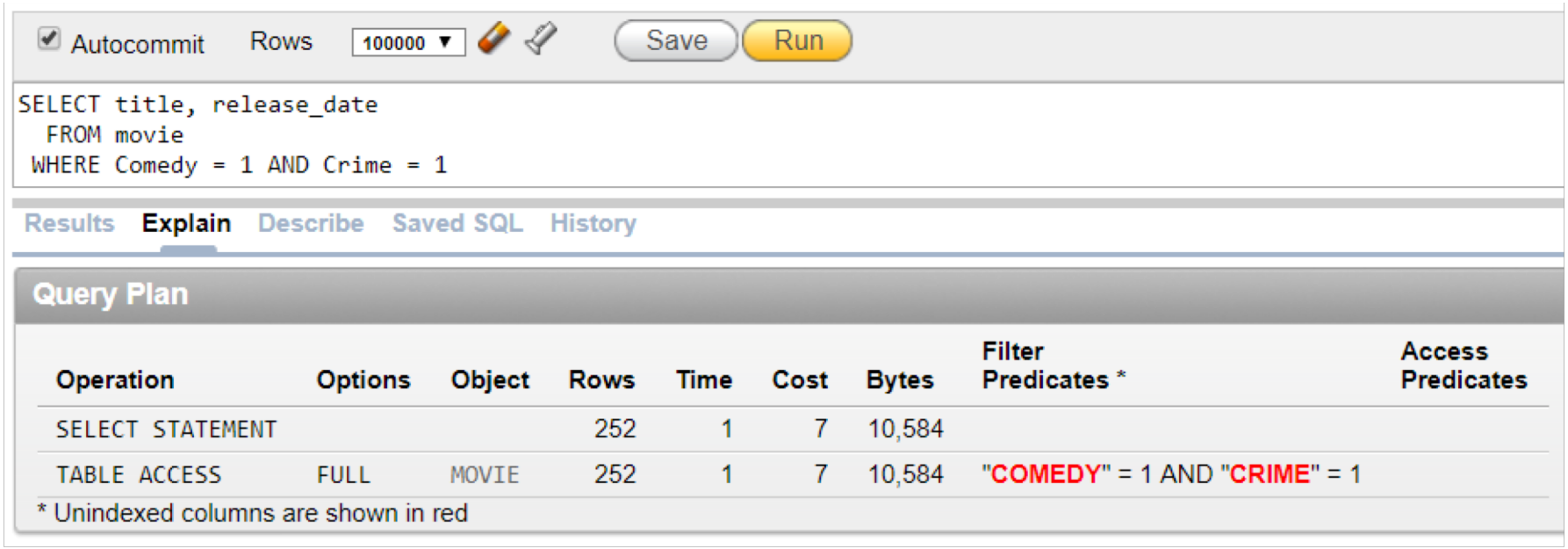
Results Explain Describe Saved SQL History

TITLE	RELEASE_DATE
Striptease (1996)	1996-06-28
"Sting, The (1973)"	1973-01-01
Batman Returns (1992)	1992-01-01
Grosse Pointe Blank (1997)	1997-04-11
Best Men (1997)	1997-09-01

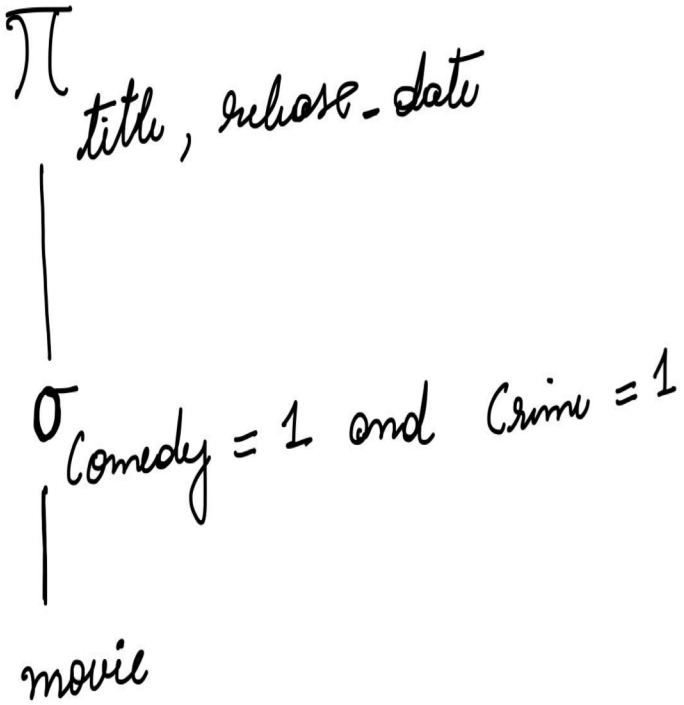
16 rows returned in 0.00 seconds Download

```
hive> SELECT title, release_date
    > FROM Movie_T1
    > WHERE Comedy = 1 and Crime=1;
Query ID = bigdata01_20181130200202_5e0b5761-7dfa-48ed-916e-f1beb435
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1542041472047_1995, Tracking URL = http://name1.h
Kill Command = /DCNFS/applications/cdh/5.12/app/hadoop-2.6.0-cdh5.12
Hadoop job information for Stage-1: number of mappers: 1; number of
2018-11-30 20:02:25,793 Stage-1 map = 0%,  reduce = 0%
2018-11-30 20:02:31,098 Stage-1 map = 100%,  reduce = 0%, Cumulative
MapReduce Total cumulative CPU time: 1 seconds 200 msec
Ended Job = job_1542041472047_1995
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 1.2 sec   HDFS Read: 136061
Total MapReduce CPU Time Spent: 1 seconds 200 msec
OK
From Dusk Till Dawn (1996)      1996-02-05
Batman Forever (1995)   1995-01-01
Mask The (1994) 1994-01-01
Striptease (1996)       1996-06-28
Sting The (1973)        1973-01-01
Batman Returns (1992)   1992-01-01
Grosse Pointe Blank (1997)      1997-04-11
Midnight in the Garden of Good and Evil (1997)  1997-01-01
Serial Mom (1994)       1994-01-01
Some Like It Hot (1959) 1959-01-01
Big Lebowski The (1998) 1997-12-26
C'est arriv0 pr0s de chez vous (1992)   1992-01-01
Best Men (1997) 1997-09-01
Carpool (1996)  1996-08-23
Twin Town (1997)        1997-05-30
Hana-bi (1997)  1998-03-20
Time taken: 11.368 seconds, Fetched: 16 row(s)
hive>
```

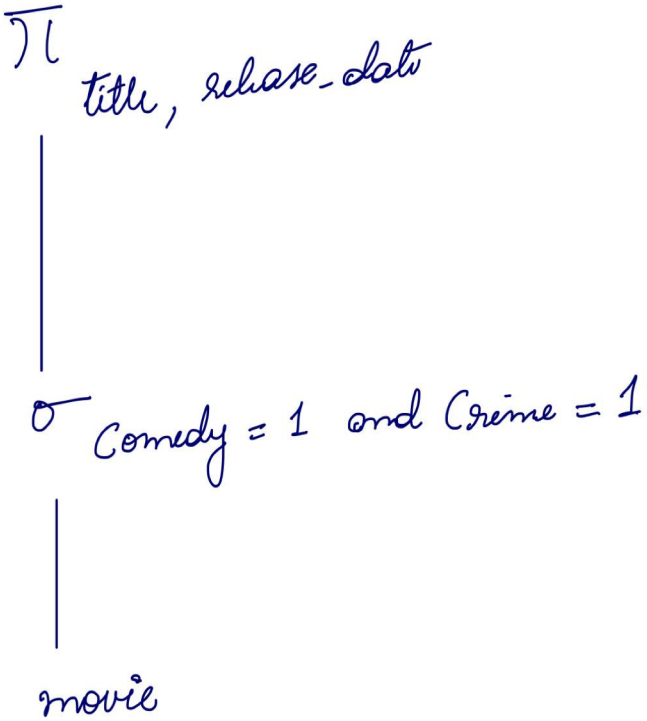
Oracle Explain Plan for Query 1



Our Explain Plan



Oracle's Explain Plan



Query 2

Most popular movies among women

```
SELECT m.movieid, m.title
FROM movie m
INNER JOIN (
    SELECT d.userid, d.itemid, d.rating
    FROM data d
    INNER JOIN (
        SELECT userid, gender
        FROM users
        WHERE gender = 'F') f
    ON d.userid = f.userid
    WHERE d.rating = 5) d
ON m.movieid = d.itemid
```

Result:

Rows fetched: 5975
Oracle : 0.09 seconds
Hive : 28.04 seconds

Oracle and Hive query results

Autocommit Rows 100000 Save Run

```
SELECT m.movieid,
       m.title
  FROM movie m
 INNER JOIN (
       SELECT d.userid,
              d.itemid,
              d.rating
         FROM data d
        INNER JOIN (
              SELECT userid,
                     gender
                FROM users
               WHERE gender = 'F'
            ) f
           ON d.userid = f.userid
        WHERE d.rating = 5
    ) d ON m.movieid = d.itemid
```

Results Explain Describe Saved SQL History

MOVIEID	TITLE
153	"Fish Called Wanda, A (1988)"
11	Seven (Se7en) (1995)
705	Singin' in the Rain (1952)
508	"People vs. Larry Flynt, The (1996)"
204	Back to the Future (1985)

5975 rows returned in 0.09 seconds Download

```
762     Beautiful Girls (1996)
7       Twelve Monkeys (1995)
224     Ridicule (1996)
278     Bed of Roses (1996)
265     Hunt for Red October The (1990)
128     Supercop (1992)
237     Jerry Maguire (1996)
591     Primal Fear (1996)
405     Mission: Impossible (1996)
303     Ulee's Gold (1997)
194     Sting The (1973)
181     Return of the Jedi (1983)
471     Courage Under Fire (1996)
815     One Fine Day (1996)
625     Sword in the Stone The (1963)
133     Gone with the Wind (1939)
993     Hercules (1997)
174     Raiders of the Lost Ark (1981)
143     Sound of Music The (1965)
109     Mystery Science Theater 3000: The Movie (1996)
174     Raiders of the Lost Ark (1981)
223     Sling Blade (1996)
176     Aliens (1986)
204     Back to the Future (1985)
Time taken: 28.04 seconds, Fetched: 5975 row(s)
hive>
```

Oracle Explain Plan for Query 2

Autocommit Rows 100000 Save Run

```
SELECT m.movieid,
       m.title
  FROM movie m
 INNER JOIN (
       SELECT d.userid,
              d.itemid,
              d.rating
         FROM data d
        INNER JOIN (
              SELECT userid,
                     gender
                FROM users
               WHERE gender = 'F'
            ) f
           ON d.userid = f.userid
        WHERE d.rating = 5
    ) d ON m.movieid = d.itemid
```

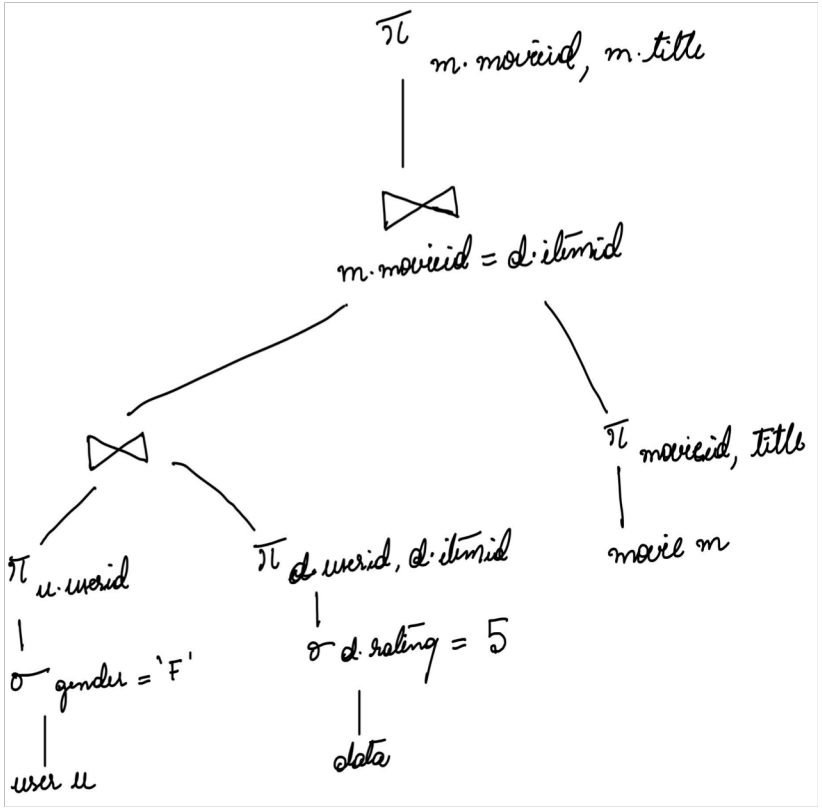
Results Explain Describe Saved SQL History

Query Plan

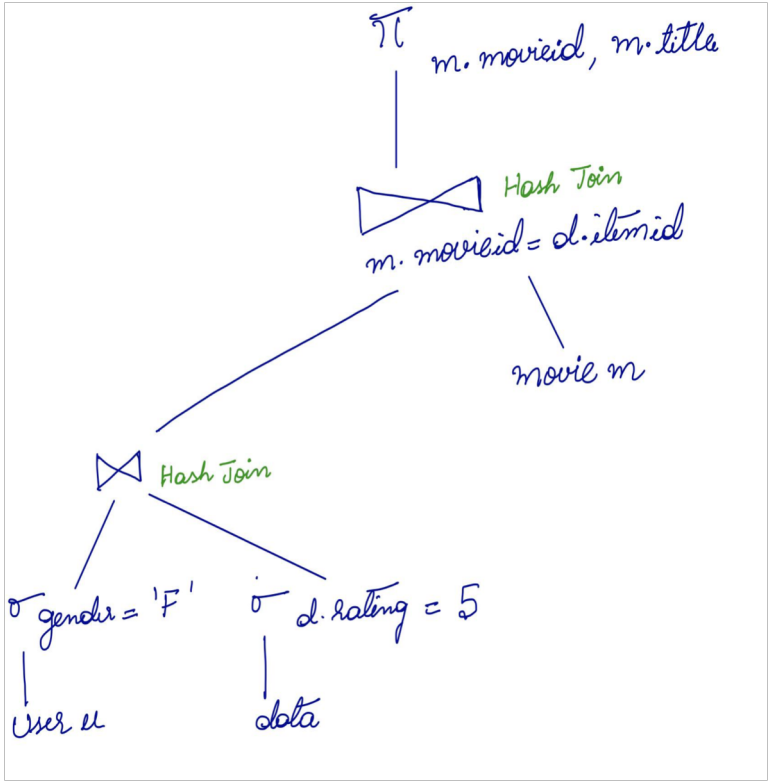
Operation	Options	Object	Rows	Time	Cost	Bytes	Filter Predicates *	Access Predicates
SELECT STATEMENT			6,438	2	148	296,148		
HASH JOIN			6,438	2	148	296,148		"M"."MOVIEID" = "D"."ITEMID"
TABLE ACCESS	FULL	MOVIE	1,682	1	7	48,778		
HASH JOIN			6,438	2	140	109,446		"D"."USERID" = "USERID"
TABLE ACCESS	FULL	USERS	273	1	3	1,638	"GENDER" = 'F'	
TABLE ACCESS	FULL	DATA	22,237	2	137	244,607	"D""RATING" = 5	

* Unindexed columns are shown in red

Our Explain Plan



Oracle's Explain Plan



Query 3

Average movie rating for each gender

```
SELECT u.gender, avg(d.rating) AS avg_rating
FROM users u
INNER JOIN data d
ON u.userid = d.userid
GROUP BY u.gender;
```

Result:

Rows fetched : 2
Oracle : 0.04 seconds
Hive : 39.7 seconds

Oracle and Hive query results

```
SELECT u.gender,
       avg(d.rating) AS avg_rating
  FROM users u
 INNER JOIN data d
    ON u.userid = d.userid
 GROUP BY u.gender
```

GENDER	AVG_RATING
M	3.52928898464853218421761378938863452734
F	3.53150738150738150738150738150738150738

2 rows selected. 0.04 seconds

```
hive> SELECT u.gender,
    >        avg(d.rating) AS avg_rating
    >   FROM Users_T1 u
    >  INNER JOIN Data_T1 d
    >     ON u.userid = d.userid
    >  GROUP BY u.gender
    > ;
Query ID = bigdata01_20181117102929_41c8d7fe-79b8-4687-b9fd-17cfdfb74f49
Total jobs = 1
Execution log at: /tmp/bigdata01/bigdata01_20181117102929_41c8d7fe-79b8-4687-b9fd-17cfdfb74f4
2018-11-17 10:29:45     Starting to launch local task to process map join;      maximum memor
2018-11-17 10:29:46     Dump the side-table for tag: 0 with group count: 943 into file: file:
4837-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile00--.hashtable
2018-11-17 10:29:46     Uploaded 1 File to: file:/tmp/bigdata01/5faf8012-2872-4f9f-a4a5-8fdee
n-mapfile00--.hashtable (21670 bytes)
2018-11-17 10:29:46     End of local task; Time Taken: 1.049 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1542041472047_0211, Tracking URL = http://name1.hadoop.dc.engr.scu.edu:808
Kill Command = /DCNFS/applications/cdh/5.12/app/hadoop-2.6.0-cdh5.12.1/bin/hadoop job  -kill
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-11-17 10:30:01,635 Stage-2 map = 0%,  reduce = 0%
2018-11-17 10:30:07,233 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.77 sec
2018-11-17 10:30:20,163 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 6.39 sec
MapReduce Total cumulative CPU time: 6 seconds 390 msec
Ended Job = job_1542041472047_0211
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 6.39 sec   HDFS Read: 2091299 HDFS Write:
Total MapReduce CPU Time Spent: 6 seconds 390 msec
OK
F       3.5315073815073816
M       3.5292889846485322
Time taken: 39.7 seconds, Fetched: 2 row(s)
```

Oracle Explain Plan for Query 3

Autocommit Rows 200 Save Run

```
SELECT u.gender,
       avg(d.rating) AS avg_rating
  FROM users u
 INNER JOIN data d
    ON u.userid = d.userid
 GROUP BY u.gender
```

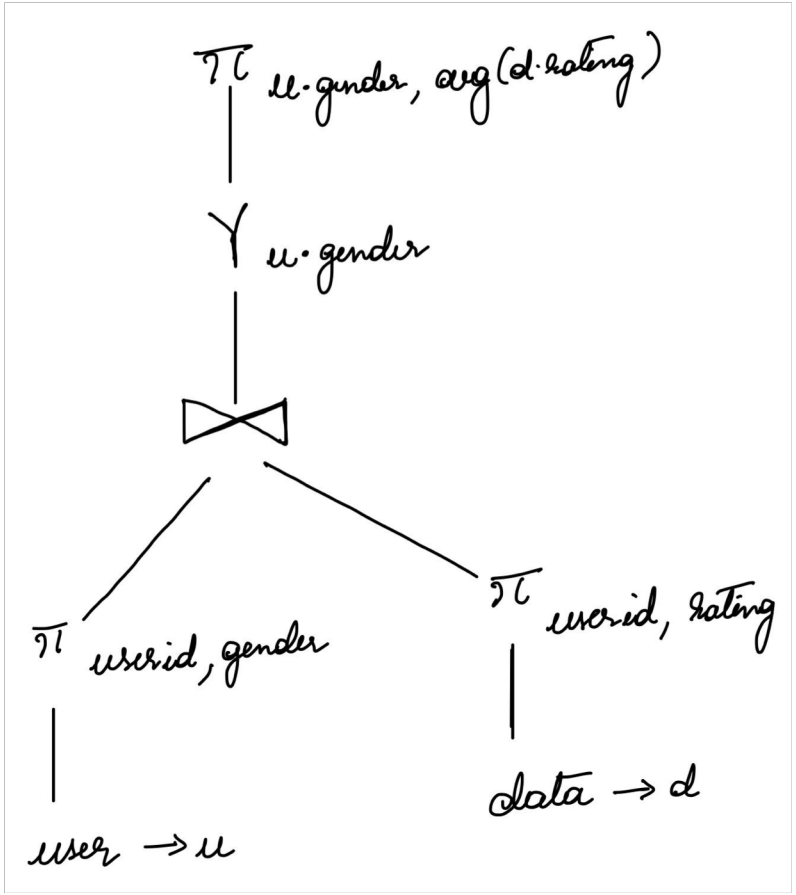
Results **Explain** Describe Saved SQL History

Query Plan

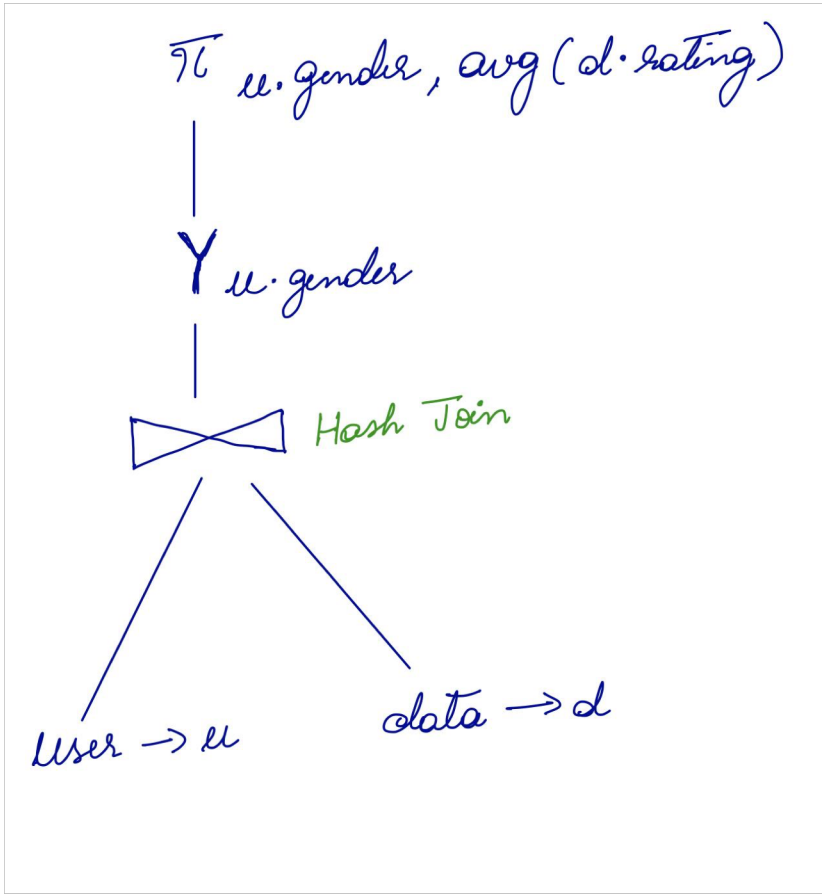
Operation	Options	Object	Rows	Time	Cost	Bytes	Filter Predicates *	Access Predicates
SELECT STATEMENT			2	2	144	26		
HASH	GROUP BY		2	2	144	26		
HASH JOIN			100,000	2	140	1,300,000		"U"."USERID" = "D"."USERID"
TABLE ACCESS	FULL	USERS	943	1	3	5,658		
TABLE ACCESS	FULL	DATA	100,000	2	137	700,000		

* Unindexed columns are shown in red

Our Explain Plan



Oracle's Explain Plan



Query 4

Average rating of movies

```
SELECT m.title, avg(d.rating) AS avg_rating
FROM data d
        INNER JOIN movie m
        ON movieid = itemid
GROUP BY m.title;
```

Result:

Rows fetched: 1664 rows
Oracle : 0.06 seconds
Hive : 33.27 seconds

Oracle and Hive query results

Autocommit Rows 100000 Save Run

```
SELECT m.title,
       avg(d.rating) AS avg_rating
  FROM data d
 INNER JOIN movie m
    ON movieid = itemid
 GROUP BY m.title
```

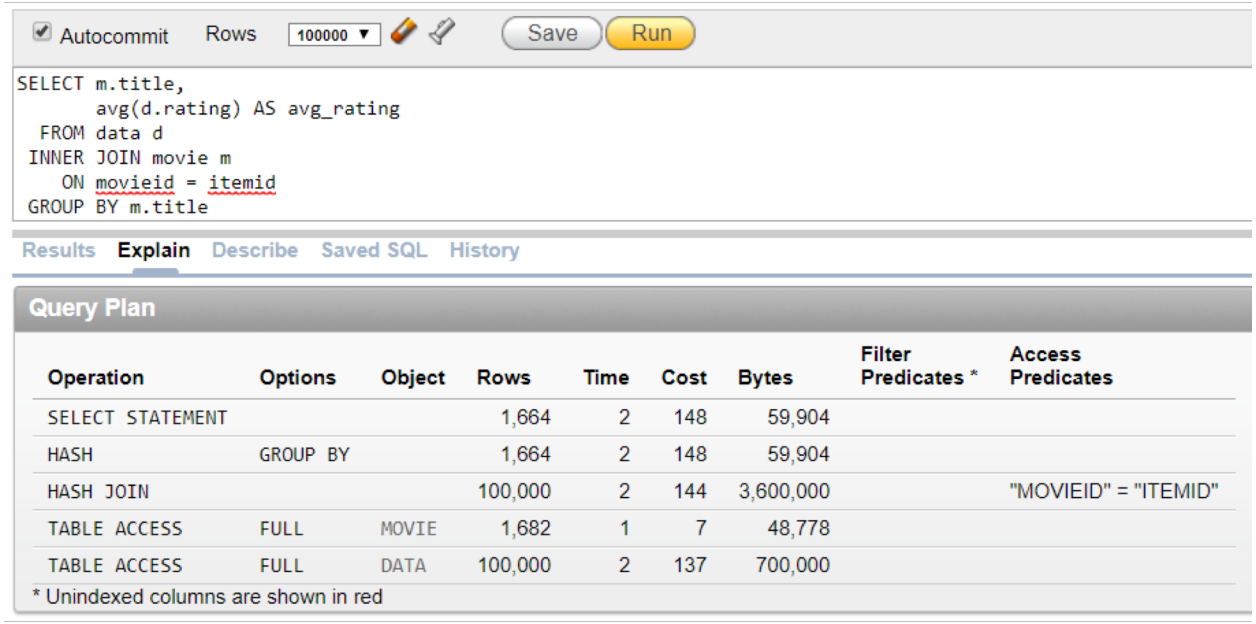
Results Explain Describe Saved SQL History

TITLE	AVG_RATING
Twister (1996)	3.2150170648464163822525597269624573378 8
Singin' in the Rain (1952)	3.9927007299270072992700729927007299270 072992701
"People vs. Larry Flynt, The (1996)"	3.5953488372093023255813953488372093023 3
Fled (1996)	2.6470588235294117647058823529411764705 9
Boys in Venice (1996)	1

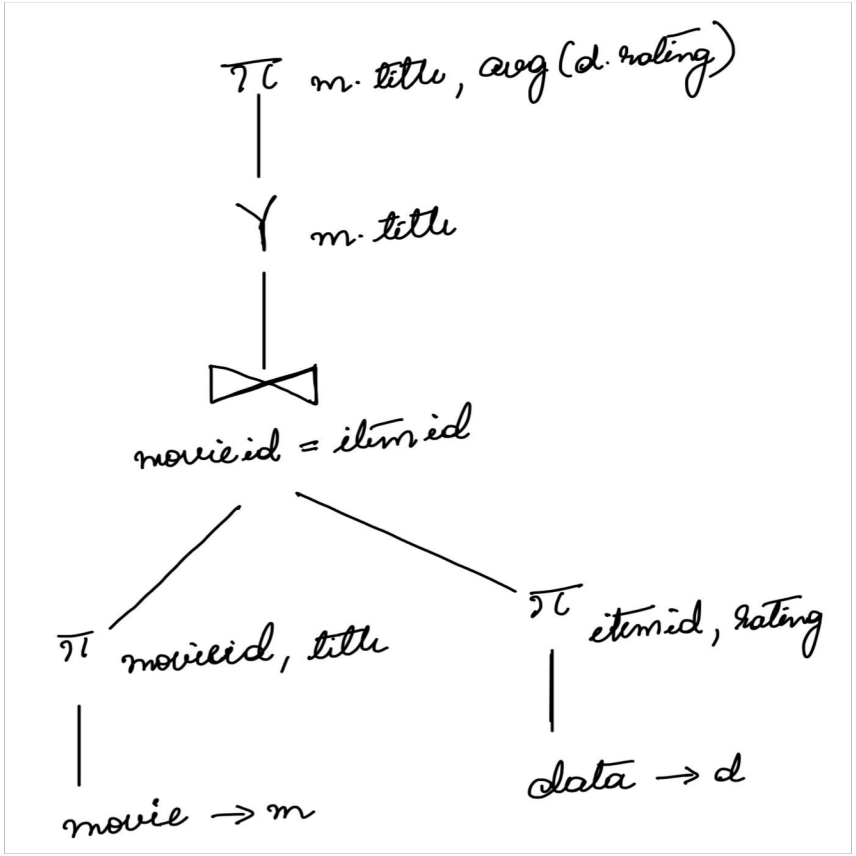
1664 rows returned in 0.06 seconds Download

```
Witness (1985)  4.0
Wizard of Oz The (1939) 4.0772357723577235
Wolf (1994)     2.701492537313433
Woman in Question The (1950)    1.0
Women The (1939)        3.6666666666666665
Wonderful Horrible Life of Leni Riefenstahl The (1993)  4.0
Wonderland (1997)       3.2
Wooden Man's Bride The (Wu Kui) (1994)  2.6666666666666665
World of Apu The (Apur Sansar) (1959)   4.0
Wrong Trousers The (1993)       4.466101694915254
Wyatt Earp (1994)       3.1
Yankee Zulu (1994)      1.0
Year of the Horse (1997)        3.2857142857142856
You So Crazy (1994)     3.0
Young Frankenstein (1974)       3.945
Young Guns (1988)       3.207920792079208
Young Guns II (1990)    2.772727272727273
Young Poisoner's Handbook The (1995)    3.341463414634146
Zeus and Roxanne (1997) 2.1666666666666665
unknown 3.4444444444444446
⊥ k÷ldum klaka (Cold Fever) (1994)      3.0
Time taken: 33.27 seconds, Fetched: 1664 row(s)
hive>
```

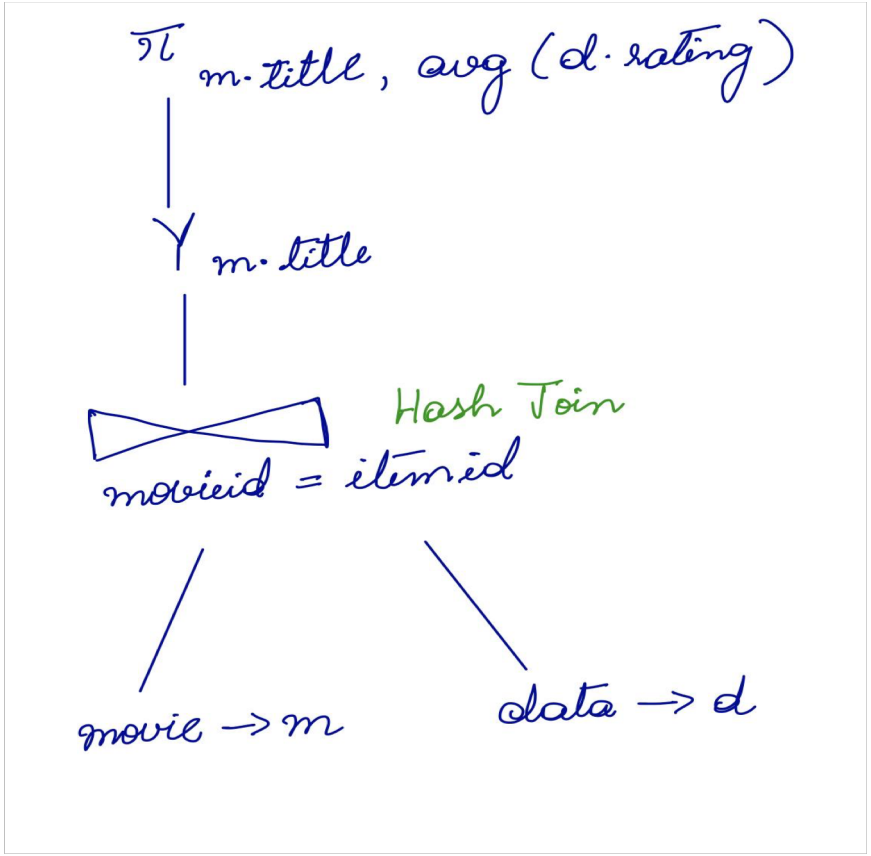
Oracle Explain Plan for Query 4



Our Explain Plan



Oracle's Explain Plan



Query 5

Full information about our users

```
SELECT u.*, z.city, z.state
FROM users u
LEFT JOIN zipcode z
ON z.zipcode = u.zipcode
```

Result:

Rows fetched : 943
Oracle : 0.14 seconds
Hive : 36.393 seconds

Oracle and Hive query result

Autocommit Rows 100000 Save Run

```
SELECT u.*,
       z.city,
       z.state
  FROM users u
  LEFT JOIN zipcode z
    ON z.zipcode = u.zipcode
```

Results Explain Describe Saved SQL History

USERID	AGE	GENDER	OCCUPATIONID	ZIPCODE	CITY	STATE
79	39	F	1	3755	HANOVER	NH
495	29	M	5	3052	LITCHFIELD	NH
210	39	M	5	3060	NASHUA	NH
891	51	F	1	3062	NASHUA	NH
782	21	F	2	33205	-	-

943 rows returned in 0.14 seconds Download

```
hive> SELECT u.*, z.city, z.state
    > FROM Users_T1 u LEFT JOIN Zipcode_T1 z ON z.zipcode = u.zipcode;
Query ID = bigdata01_20181119184444_95d0a15f-cef9-4c63-a1e8-e495bf6c26b9
Total jobs = 1
Execution log at: /tmp/bigdata01/bigdata01_20181119184444_95d0a15f-cef9-4c
2018-11-19 06:44:29     Starting to launch local task to process map join;
2018-11-19 06:44:30     Dump the side-table for tag: 1 with group count: 4
2479300490271552590-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile51--.h
2018-11-19 06:44:30     Uploaded 1 File to: file:/tmp/bigdata01/d122aa22-8
tage-3/MapJoin-mapfile51--.hashtable (1647495 bytes)
2018-11-19 06:44:30     End of local task; Time Taken: 1.453 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1542041472047_1300, Tracking URL = http://name1.hadoop.
Kill Command = /DCNFS/applications/cdh/5.12/app/hadoop-2.6.0-cdh5.12.1/bin
Hadoop job information for Stage-3: number of mappers: 1; number of reduce
2018-11-19 18:44:46,506 Stage-3 map = 0%,  reduce = 0%
2018-11-19 18:45:00,614 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 6
MapReduce Total cumulative CPU time: 6 seconds 530 msec
Ended Job = job_1542041472047_1300
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 6.53 sec   HDFS Read: 24910 HDFS W
Total MapReduce CPU Time Spent: 6 seconds 530 msec
OK
1       24      M       20      85711   TUCSON  AZ
4       24      M       20      43537   MAUMEE  OH
44      26      M       20      46260   INDIANAPOLIS    IN
77      30      M       20      29379   UNION   SC
143     42      M       20      8832    KEASBEY NJ
197     55      M       20      75094   PLANO   TX
244     28      M       20      80525   FORT COLLINS    CO
294     34      M       20      92110   SAN DIEGO       CA
311     32      M       20      73071   NORMAN  OK
325     48      M       20      2139    CAMBRIDGE       MA
441     50      M       20      55013   CHISAGO CITY    MN
456     24      M       20      31820   MIDLAND GA
427     51      M       3       85258   SCOTTSDALE      AZ
841     45      M       3       47401   BLOOMINGTON     IN
845     64      M       3       97405   EUGENE  OR
935     42      M       3       66221   OVERLAND PARK   KS
Time taken: 36.393 seconds, Fetched: 943 row(s)
```

Oracle Explain plan for query 5

Autocommit Rows 100000 Save Run

```
SELECT u.*,
       z.city,
       z.state
  FROM users u
  LEFT JOIN zipcode z
    ON z.zipcode = u.zipcode
```

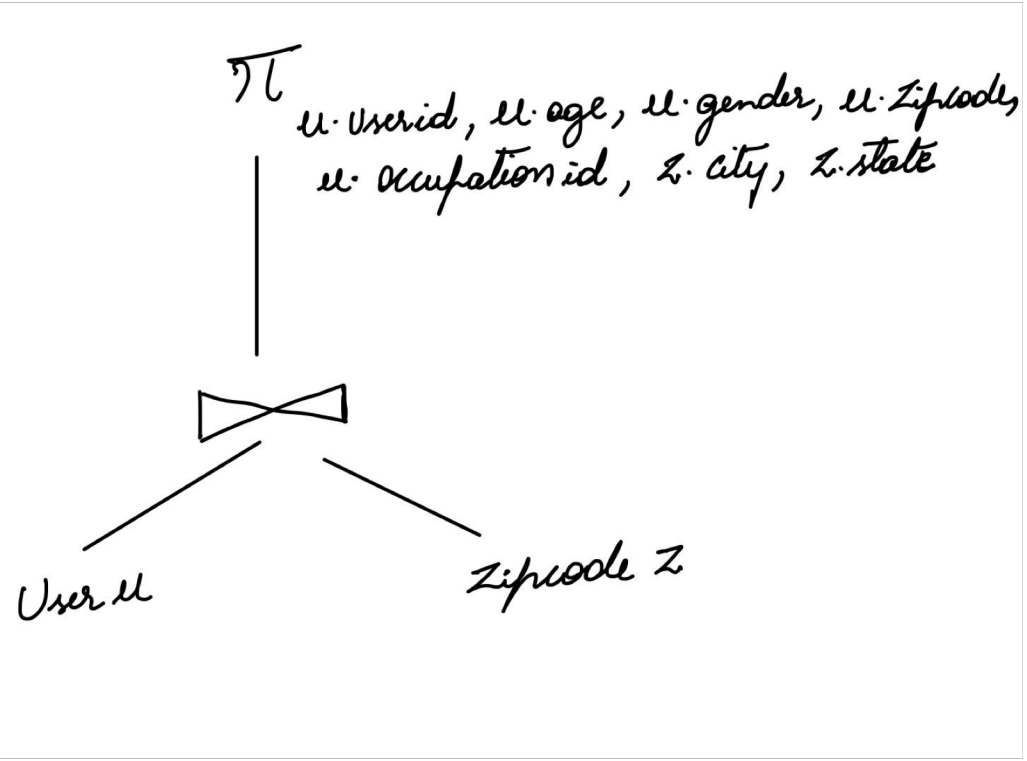
Results **Explain** Describe Saved SQL History

Query Plan

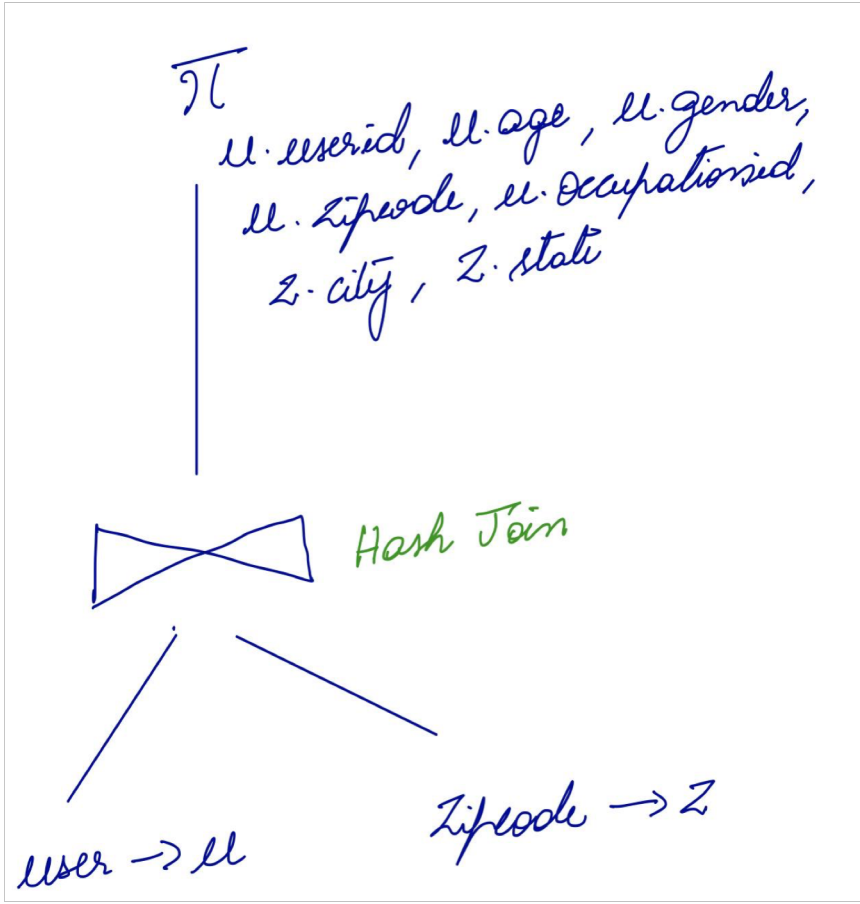
Operation	Options	Object	Rows	Time	Cost	Bytes	Filter Predicates *	Access Predicates
SELECT STATEMENT			943	1	72	34,891		
HASH JOIN	OUTER		943	1	72	34,891		"Z"."ZIPCODE"(+) = "U"."ZIPCODE"
TABLE ACCESS	FULL	USERS	943	1	3	16,974		
TABLE ACCESS	FULL	ZIPCODE	42,522	1	68	807,918		

* Unindexed columns are shown in red

Our Explain Plan



Oracle's Explain Plan



Query 6

Average rating by occupation

```
SELECT o.occupation, avg(d.rating)
FROM users u
        INNER JOIN data d ON u.userid = d.userid
        INNER JOIN occupation o ON u.occupationid = o.occupationid
GROUP BY o.occupation
```

Result:

Rows fetched: 21
Oracle: 0.15 seconds
Hive: 41.735 seconds

Oracle and Hive query results

Autocommit Rows 100000 Save Run

```
SELECT o.occupation,
       avg(d.rating)
  FROM users u
 INNER JOIN data d
    ON u.userid = d.userid
 INNER JOIN occupation o
    ON u.occupationid = o.occupationid
 GROUP BY o.occupation
```

Results Explain Describe Saved SQL History

OCCUPATION	AVG(D.RATING)
other	3.552377379724280221326080840288849291 94
student	3.515143234503802887461857266475383704 51
technician	3.532230462065031374786081003993154592 13
administrator	3.635646476801711458751169942505682577 88
entertainment	3.441050119331742243436754176610978520 29

21 rows returned in 0.15 seconds Download

```
hive> SELECT o.occupation, avg(d.rating)
    > FROM Users_T1 u
    > INNER JOIN Data_T1 d ON u.userid = d.userid
    > INNER JOIN Occupation_T1 o ON u.occupationid = o.occupationid
    > GROUP BY o.occupation;
Query ID = bigdata01_20181119182222_6baf1d78-a233-45a0-9518-4407bea3d1cd
Total jobs = 1
Execution log at: /tmp/bigdata01/bigdata01_20181119182222_6baf1d78-a233-45a0-9518-4
2018-11-19 06:22:23     Starting to launch local task to process map join;      max
2018-11-19 06:22:24     Dump the side-table for tag: 1 with group count: 21 into fi
266901795769836-1/-local-10005/HashTable-Stage-3/MapJoin-mapfile21--.hashtable
2018-11-19 06:22:24     Uploaded 1 File to: file:/tmp/bigdata01/d122aa22-8e4b-4023-
age-3/MapJoin-mapfile21--.hashtable (855 bytes)
2018-11-19 06:22:24     Dump the side-table for tag: 0 with group count: 943 into f
7266901795769836-1/-local-10005/HashTable-Stage-3/MapJoin-mapfile30--.hashtable
2018-11-19 06:22:24     Uploaded 1 File to: file:/tmp/bigdata01/d122aa22-8e4b-4023-
age-3/MapJoin-mapfile30--.hashtable (20722 bytes)
2018-11-19 06:22:24     End of local task; Time Taken: 1.034 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1542041472047_1298, Tracking URL = http://name1.hadoop.dc.engr.s
Kill Command = /DCNFS/applications/cdh/5.12/app/hadoop-2.6.0-cdh5.12.1/bin/hadoop j
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2018-11-19 18:22:31,464 Stage-3 map = 0%,  reduce = 0%
2018-11-19 18:22:46,396 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 8.06 sec
2018-11-19 18:23:00,244 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 12.62 se
MapReduce Total cumulative CPU time: 12 seconds 620 msec
Ended Job = job_1542041472047_1298
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 12.62 sec   HDFS Read: 2093507 H
Total MapReduce CPU Time Spent: 12 seconds 620 msec
OK
administrator   3.6356464768017114
artist  3.653379549393414
doctor  3.688888888888889
student 3.5151432345038027
technician      3.5322304620650313
writer  3.3757225433526012
Time taken: 41.735 seconds, Fetched: 21 row(s)
```

Oracle Explain plan for Query 6

Autocommit Rows 100000 Save Run

```
SELECT o.occupation,
       avg(d.rating)
  FROM users u
 INNER JOIN data d
    ON u.userid = d.userid
 INNER JOIN occupation o
    ON u.occupationid = o.occupationid
 GROUP BY o.occupation
```

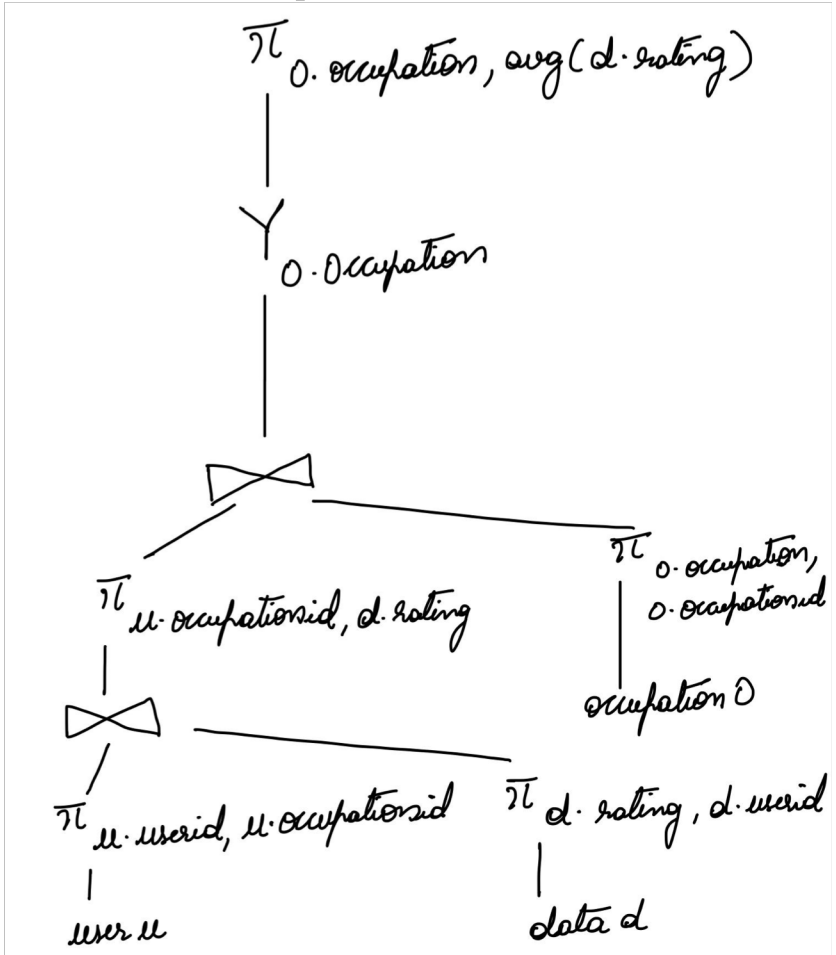
Results **Explain** Describe Saved SQL History

Query Plan

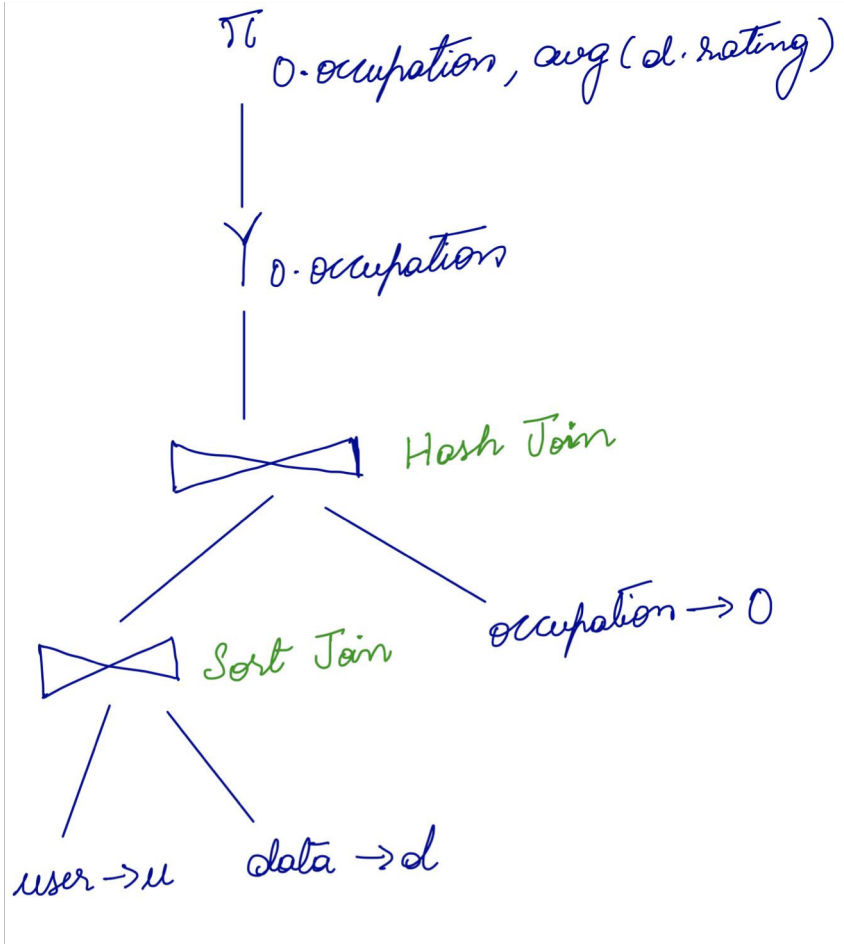
Operation	Options	Object	Rows	Time	Cost	Bytes	Filter Predicates *	Access Predicates
SELECT STATEMENT			21	2	147	546		
HASH	GROUP BY		21	2	147	546		
HASH JOIN			100,000	2	143	2,600,000		"U"."USERID" = "D"."USERID"
MERGE JOIN			943	1	6	17,917		
TABLE ACCESS	BY INDEX ROWID	OCCUPATION	21	1	2	252		
INDEX	FULL SCAN	OCCUPATION_PK	21	1	1			
SORT	JOIN		943	1	4	6,601	"U"."OCCUPATIONID" = "O"."OCCUPATIONID"	"U"."OCCUPATIONID" = "O"."OCCUPATIONID"
TABLE ACCESS	FULL	USERS	943	1	3	6,601		
TABLE ACCESS	FULL	DATA	100,000	2	137	700,000		

* Unindexed columns are shown in red

Our Explain Plan



Oracle's Explain Plan



Query 7

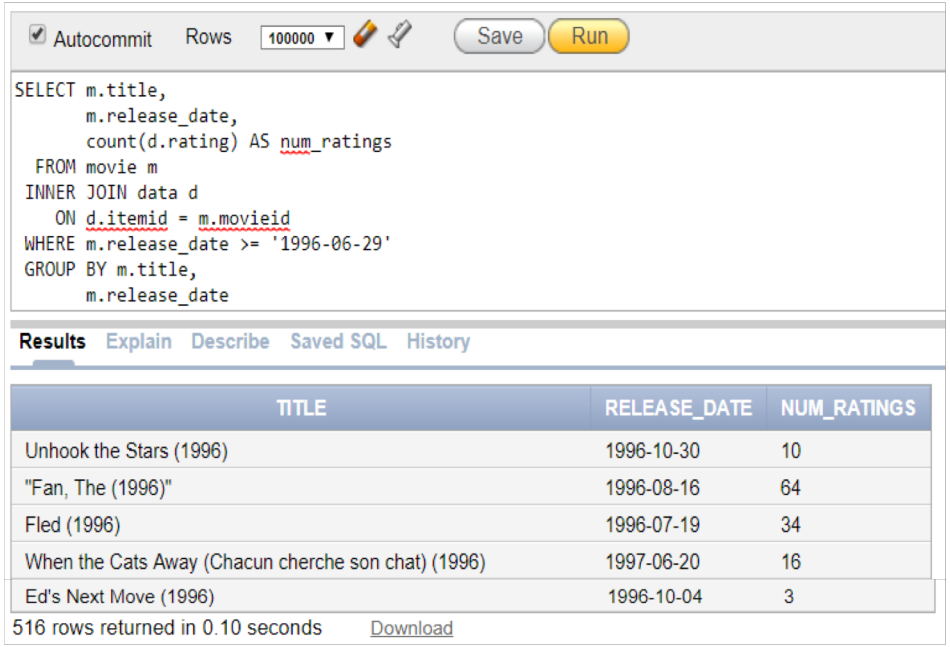
Number of ratings for movies released since Nicholas was born

```
SELECT m.title, m.release_date,
count(d.rating) AS num_ratings
FROM movie m
        INNER JOIN data d
        ON d.itemid = m.movieid
WHERE m.release_date >= '1996-06-29'
GROUP BY m.title, m.release_date
```

Result:

Rows fetched : 516
Oracle : 0.10 seconds
Hive : 43.264 seconds

Oracle and Hive query results



Autocommit Rows 100000 Save Run

```
SELECT m.title,
       m.release_date,
       count(d.rating) AS num_ratings
  FROM movie m
 INNER JOIN data d
    ON d.itemid = m.movieid
 WHERE m.release_date >= '1996-06-29'
 GROUP BY m.title,
       m.release_date
```

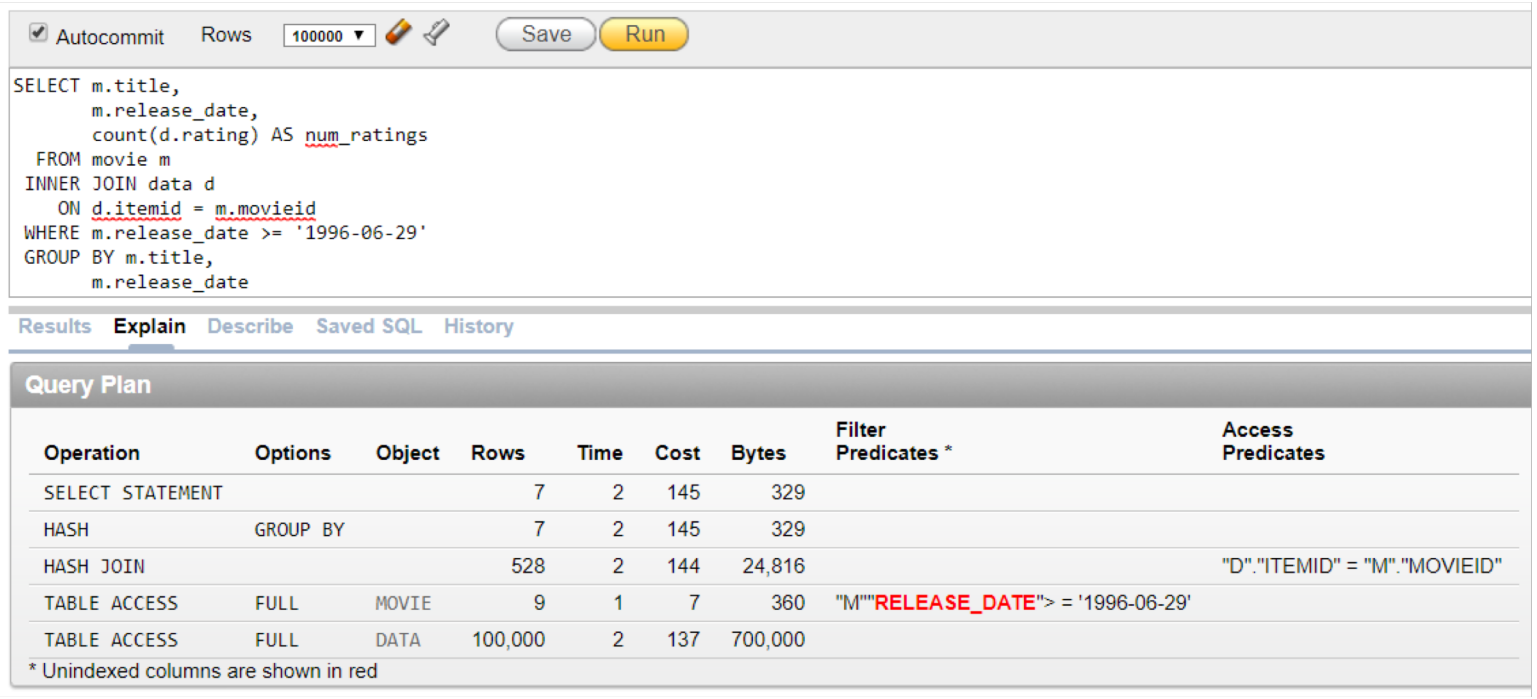
Results Explain Describe Saved SQL History

TITLE	RELEASE_DATE	NUM_RATINGS
Unhook the Stars (1996)	1996-10-30	10
"Fan, The (1996)"	1996-08-16	64
Fled (1996)	1996-07-19	34
When the Cats Away (Chacun cherche son chat) (1996)	1997-06-20	16
Ed's Next Move (1996)	1996-10-04	3

516 rows returned in 0.10 seconds Download

```
Visitors The (Visiteurs Les) (1993)     1996-07-19      2
Volcano (1997)  1997-04-25      219
Wag the Dog (1997)      1998-01-09      137
Waiting for Guffman (1996)      1997-01-31      47
Walking and Talking (1996)      1996-07-12      8
Warriors of Virtue (1997)       1997-05-02      10
Washington Square (1997)        1997-01-01      34
Wedding Bell Blues (1996)       1997-06-13      1
Wedding Singer The (1998)       1998-02-13      72
Welcome To Sarajevo (1997)      1997-01-01      22
When We Were Kings (1996)       1997-02-14      44
When the Cats Away (Chacun cherche son chat) (1996)     1997-06-20      16
Whole Wide World The (1996)     1996-12-25      6
Wife The (1995) 1996-07-26      1
Wild America (1997)     1997-07-04      9
Wild Things (1998)      1998-03-14      11
William Shakespeare's Romeo and Juliet (1996)   1996-10-25      106
Wings of the Dove The (1997)    1997-01-01      75
Winter Guest The (1997) 1997-01-01      9
Wishmaster (1997)       1997-01-01      27
Wonderland (1997)       1997-01-01      10
Year of the Horse (1997)        1997-01-01      7
Zeus and Roxanne (1997) 1997-01-10      6
Time taken: 43.264 seconds, Fetched: 516 row(s)
hive>
```

Oracle Explain plan



Autocommit Rows 100000 Save Run

```
SELECT m.title,
       m.release_date,
       count(d.rating) AS num_ratings
  FROM movie m
 INNER JOIN data d
    ON d.itemid = m.movieid
 WHERE m.release_date >= '1996-06-29'
 GROUP BY m.title,
       m.release_date
```

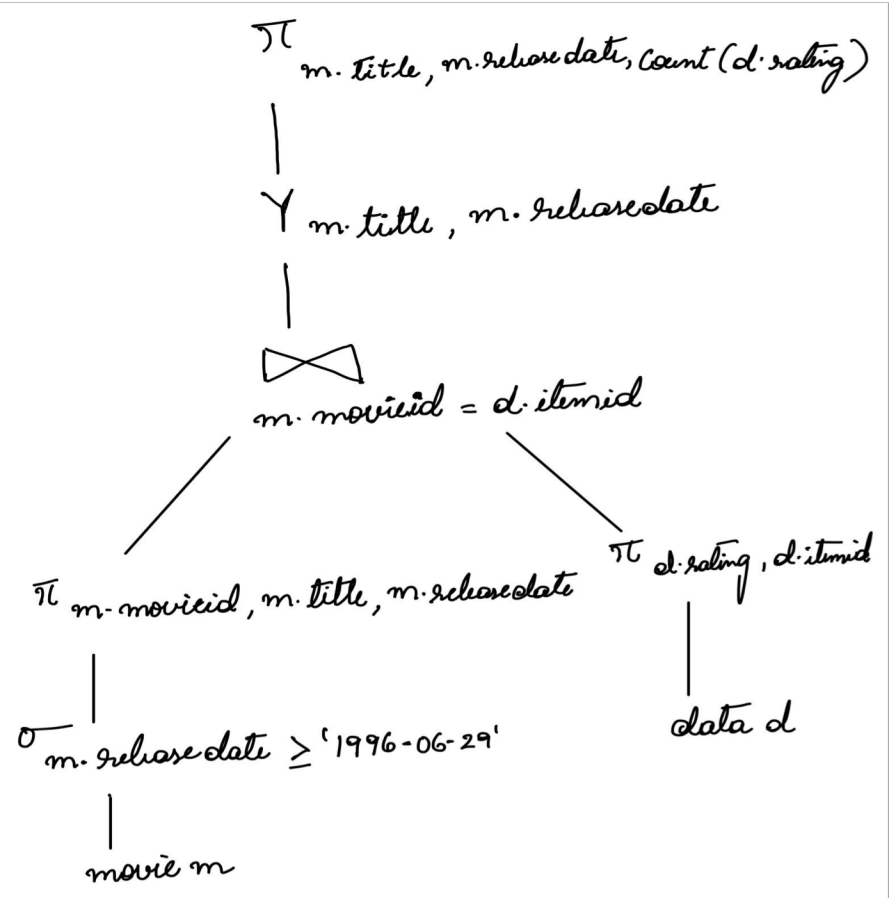
Results Explain Describe Saved SQL History

Query Plan

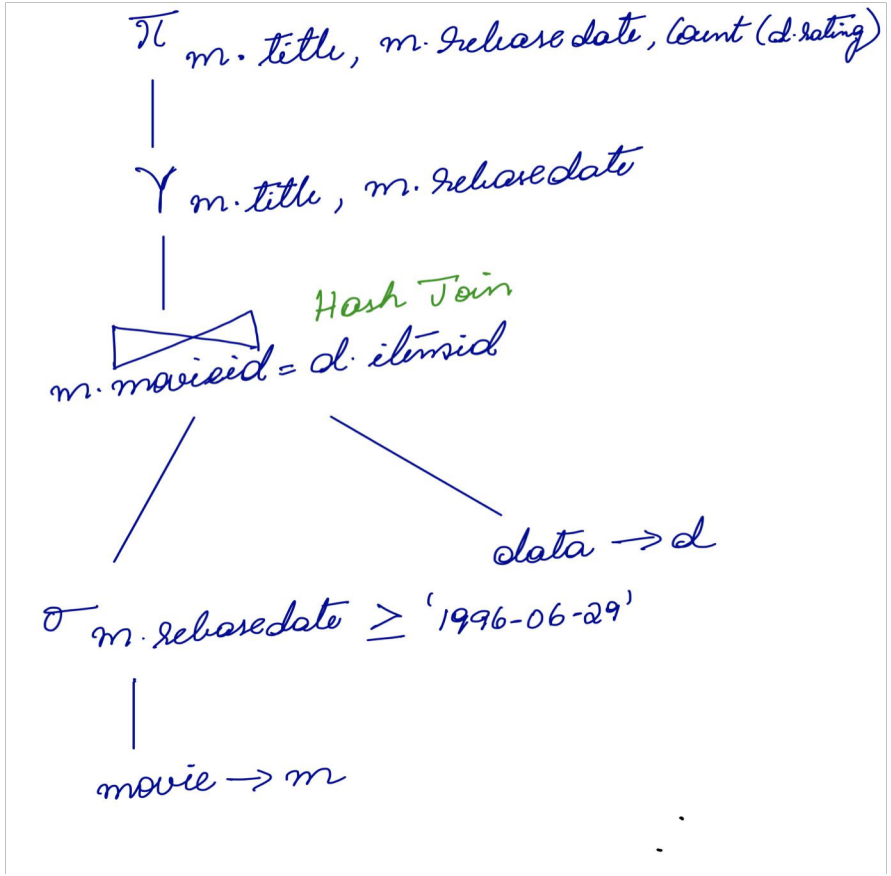
Operation	Options	Object	Rows	Time	Cost	Bytes	Filter Predicates *	Access Predicates
SELECT STATEMENT			7	2	145	329		
HASH	GROUP BY		7	2	145	329		
HASH JOIN			528	2	144	24,816		"D"."ITEMID" = "M"."MOVIEID"
TABLE ACCESS	FULL	MOVIE	9	1	7	360	"M""RELEASE_DATE"> = '1996-06-29'	
TABLE ACCESS	FULL	DATA	100,000	2	137	700,000		

* Unindexed columns are shown in red

Our Explain Plan



Oracle's Explain Plan



Query 8

Select movies watched from users who live in Palo Alto and are not an engineer

```
SELECT DISTINCT m.title
FROM movie m
WHERE m.movieid IN (
        SELECT d.itemid
        FROM data d
        WHERE d.userid IN (
                SELECT u.userid
                FROM users u
                WHERE u.occupationid IN (
                        SELECT o.occupationid
                        FROM occupation o
                        WHERE o.occupation != 'engineer')
                        AND u.zipcode IN (
                                SELECT z.zipcode
                                FROM zipcode z
                                WHERE z.city = 'PALO ALTO')))
```

Result:

Rows fetched: 53

Oracle : 0.10 seconds

Hive : 44.28 seconds

Oracle and Hive query results

Autocommit Rows 100000 ▼ Save Run

```
SELECT DISTINCT m.title
  FROM movie m
 WHERE m.movieid IN (
       SELECT d.itemid
         FROM data d
        WHERE d.userid IN (
              SELECT u.userid
                FROM users u
               WHERE u.occupationid IN (
                     SELECT o.occupationid
                       FROM occupation o
                      WHERE o.occupation != 'engineer'
                   )
                 AND u.zipcode IN (
                     --zipcodes of Santa Clara
                     SELECT z.zipcode
                       FROM zipcode z
                      WHERE z.city = 'PALO ALTO'
              )
       )
    )
```

Results Explain Describe Saved SQL History

TITLE
Kiss the Girls (1997)
Bean (1997)
Career Girls (1997)
"Smile Like Yours, A (1997)"
Seven Years in Tibet (1997)

53 rows selected. 0.10 seconds

```
hive> SELECT DISTINCT m.title
    > FROM Users_T1 u, Occupation_T1 o, Zipcode_T1 z, Movi
    > WHERE m.movieid = d.itemid AND d.userid = u.userid A
Warning: Map Join MAPJOIN[44][bigTable=?] in task 'Stage-5
Query ID = bigdata01_20181119190808_481a8ce2-aa0f-4184-b05
Total jobs = 1
Execution log at: /tmp/bigdata01/bigdata01_20181119190808_
2018-11-19 07:08:51     Starting to launch local task to p
2018-11-19 07:08:52     Dump the side-table for tag: 1 wit
2018-11-19 07:08:52     Uploaded 1 File to: file:/tmp/bigd
2018-11-19 07:08:52     Dump the side-table for tag: 1 wit
2018-11-19 07:08:52     Uploaded 1 File to: file:/tmp/bigd
2018-11-19 07:08:52     Dump the side-table for tag: 1 wit
2018-11-19 07:08:52     Uploaded 1 File to: file:/tmp/bigd
2018-11-19 07:08:52     Dump the side-table for tag: 1 wit
2018-11-19 07:08:52     Uploaded 1 File to: file:/tmp/bigd
2018-11-19 07:08:52     End of local task; Time Taken: 1.5
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input
In order to change the average load for a reducer (in byte
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1542041472047_1327, Tracking URL = http
Kill Command = /DCNFS/applications/cdh/5.12/app/hadoop-2.6
Hadoop job information for Stage-5: number of mappers: 1;
2018-11-19 19:09:01,452 Stage-5 map = 0%,  reduce = 0%
2018-11-19 19:09:15,304 Stage-5 map = 100%,  reduce = 0%,
2018-11-19 19:09:30,203 Stage-5 map = 100%,  reduce = 100%
MapReduce Total cumulative CPU time: 11 seconds 590 msec
Ended Job = job_1542041472047_1327
MapReduce Jobs Launched:
Stage-Stage-5: Map: 1  Reduce: 1   Cumulative CPU: 11.59 s
Total MapReduce CPU Time Spent: 11 seconds 590 msec
OK
"Birdcage
"English Patient
"Game
Shadow Conspiracy (1997)
That Thing You Do! (1996)
Vegas Vacation (1997)
Volcano (1997)
Time taken: 44.28 seconds, Fetched: 53 row(s)
hive>
```

Oracle Explain Plan for Query 8

Autocommit Rows 100000 Save Run

```
SELECT DISTINCT m.title
  FROM movie m
 WHERE m.movieid IN (
       SELECT d.itemid
         FROM data d
        WHERE d.userid IN (
              SELECT u.userid
                FROM users u
               WHERE u.occupationid IN (
                     SELECT o.occupationid
                       FROM occupation o
                      WHERE o.occupation != 'engineer'
                   )
                 AND u.zipcode IN (
                     SELECT z.zipcode
                       FROM zipcode z
                      WHERE z.city = 'PALO ALTO'
              )
        )
    )
```

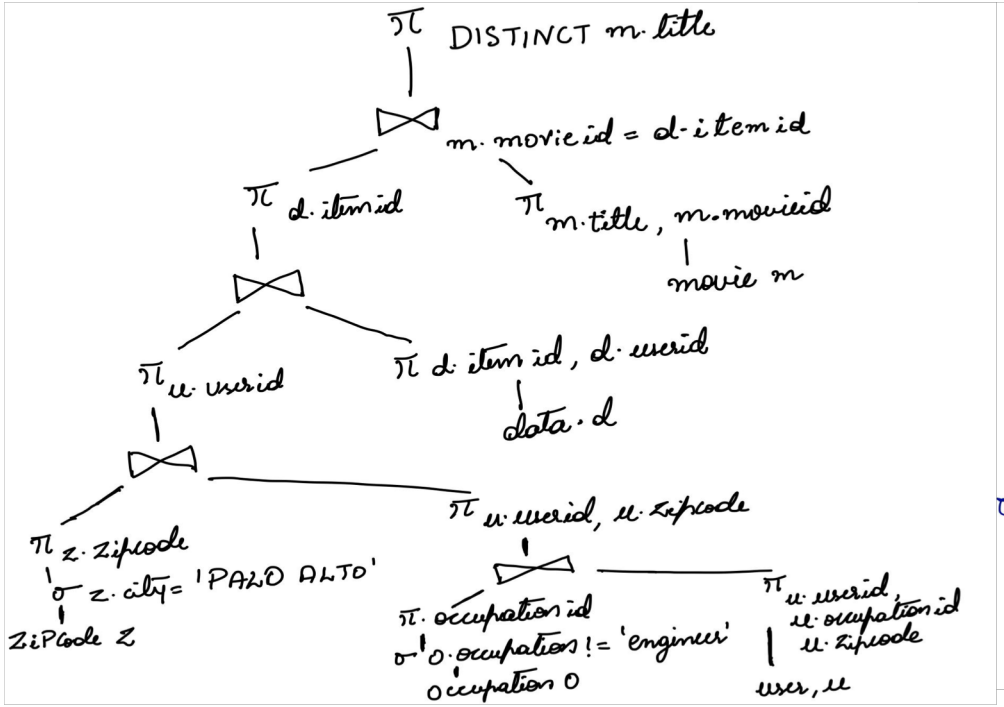
Results Explain Describe Saved SQL History

Query Plan

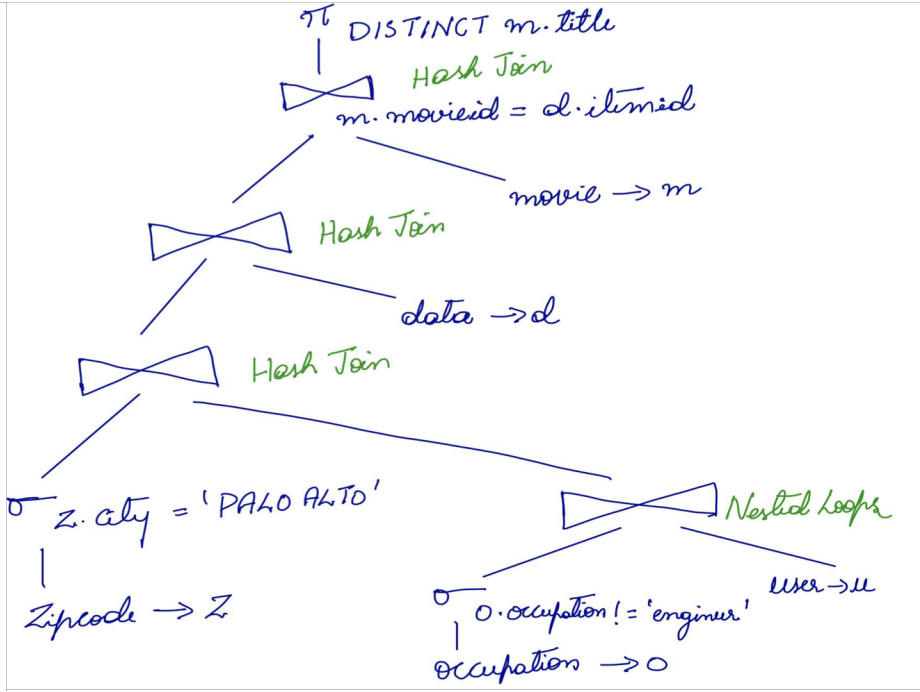
Operation	Options	Object	Rows	Time	Cost	Bytes	Filter Predicates *	Access Predicates
SELECT STATEMENT			263	3	221	20,514		
HASH	UNIQUE		263	3	221	20,514		
HASH JOIN			263	3	220	20,514		"M"."MOVIEID" = "D"."ITEMID"
HASH JOIN			263	3	212	12,887		"D"."USERID" = "U"."USERID"
NESTED LOOPS								
NESTED LOOPS			2	1	75	82		
HASH JOIN			3	1	72	87		"U"."ZIPCODE" = "Z"."ZIPCODE"
TABLE ACCESS	FULL	ZIPCODE	2	1	68	32	"Z""CITY" = 'PALO ALTO'	
TABLE ACCESS	FULL	USERS	943	1	3	12,259		
INDEX	UNIQUE SCAN	OCCUPATION_PK	1	1	0			"U"."OCCUPATIONID" = "O"."OCCUPATIONID"
TABLE ACCESS	BY INDEX ROWID	OCCUPATION	1	1	1	12	"O""<>'engineer'	
TABLE ACCESS	FULL	DATA	100,000	2	137	800,000		
TABLE ACCESS	FULL	MOVIE	1,682	1	7	48,778		

* Unindexed columns are shown in red

Our Explain Plan



Oracle's Explain Plan



Conclusion: Oracle is faster than Hive

- Oracle uses index
- Data is small enough for Oracle to run efficiently
- Hive follows schema on read, better suited for large analytic processing
- Hive has to do shuffling of the data
- Hive has lots of overhead

Q&A

Thank You!!!