SQL Project – Google Store Visitor Data

BUAN 6320.006

Data Model

Assumptions/Notes About Data Entities and Relationships

We tried to make the tables in 3NF by having every non-key value fully functionally dependent on the primary key and all columns are determined by the primary key and not any non-primary key.

We deleted all values that do not have any meaning such as: '(not set)'; 'not available in demo dataset'

Columns we deleted and reason:

Columns that only have one value or data not available in demo dataset: cityId

latitude
longitude networkLocation
browserVersion browsersize
operatingsystemVersion
mobileDeviceBranding
mobileDeviceModel
mobileInputSelector
mobileDeviceInfo
mobileDeviceMarketingName flashVersion
language
screenColors
screenResolution visit

socialEngagementType

Redundant Columns that can be directly inferred from another non-key column without any external knowledge: Date: Can be taken directly from visitStartTime column in DateTime format

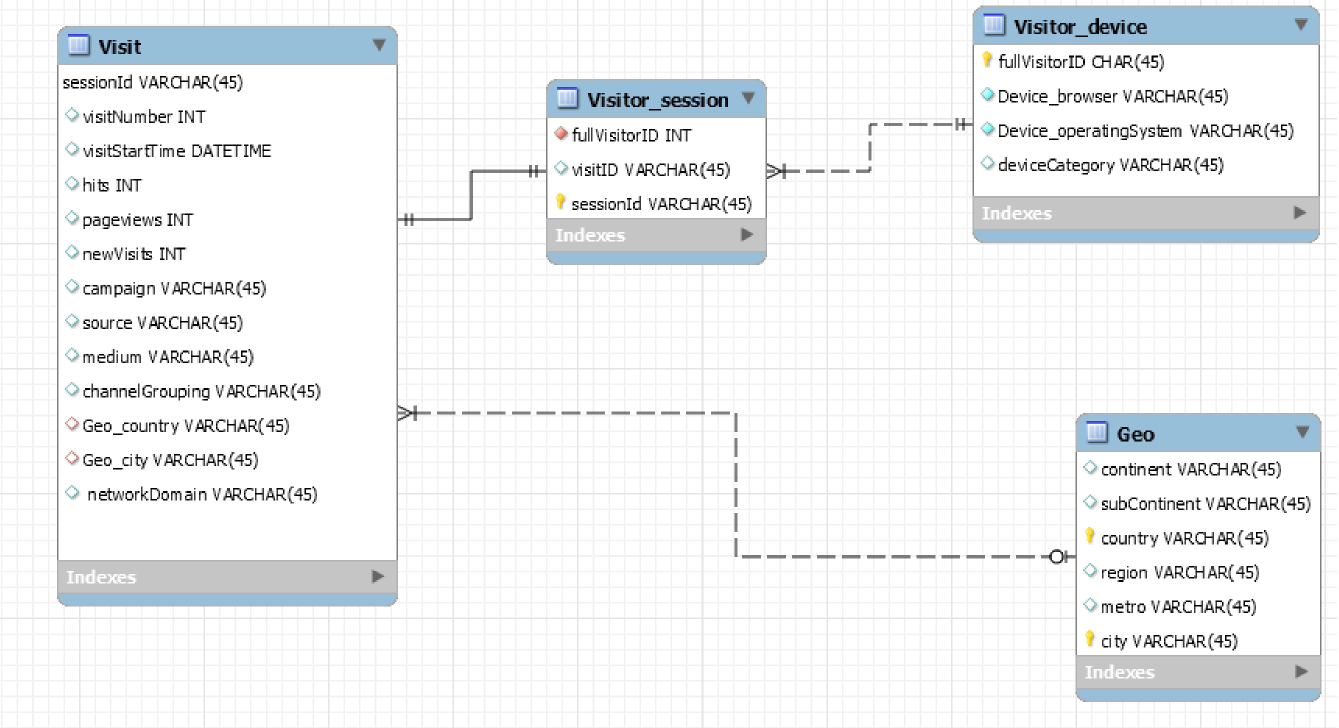
IsMobile: Can be inferred directly from the deviceCategory column

Assumption about relationships:

- VisitId and fullVisitorId together is unique for each visit session

- Sessionid is a combination of visitId and fullVisitorId so it is unique for each visit session
- We assume that google analytics recognize separate devices as different visitor
- Some sessions only have city recorded, some only have country recorded. We assume that for each combination of city and country, there's only one combination of continent, subcontinent, region,metro and we use city and country as composite primary key for GeoNetwork Table

Entity-Relationship Diagram

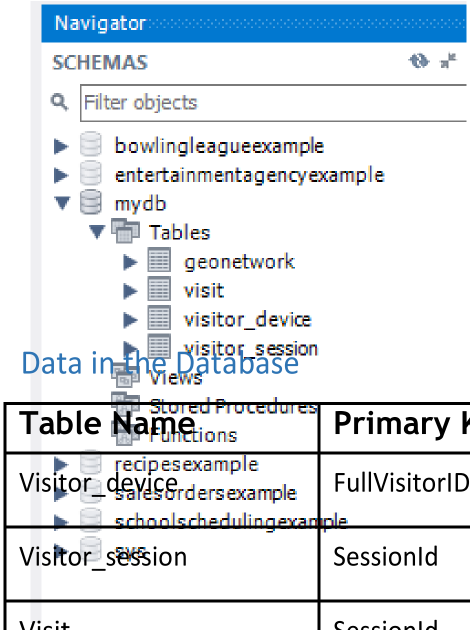


Physical Database

Assumptions/Notes About Data Set

- We changed the format of visitStartTime to DateTime
- fullvisitorID was changed to TEXT format

Screenshot of Physical Database objects



Navigator

SCHEMAS

Filter objects

- bowlingleagueexample
- entertainmentagencyexample
- mydb
  - Tables
    - geonetwork
    - visit
    - visitor_device
    - visitor_session
  - Views
  - Stored Procedures
  - Functions
- recipesexample
- salesordersexample
- schoolschedulingexample
- sys

Data in the Database

Table Name	Primary Key	Foreign Key	# of Rows in Table
Visitor_device	FullVisitorID		617570
Visitor_session	SessionId	FullVisitorID	804864
Visit	SessionId	SessionId City country	804684
GeoNetwork	city country		1724

SQL Queries

Query 1 Question

Were mobile devices users more socially engaged than non-mobile device users?

Notes/Comments About SQL Query and Results (Include # of Rows in Result)

All of the records in this database are Not Socially Engaged. Therefore we can see that there is no difference between mobile users and non-mobile device users. The reason might be that of more than 800,000 sessions recorded, no one is socially engaged in the store (which is hard to believe) or there might be mistake in the function that determine social engagement. This column is not helpful when we want to understand the customers's patterns and why some decide to visit again or buy product while some do not. Therefore, we deleted the column 'socialEngagementType' from this dataset and will not run this query.

Query 2

Question

Which user had the maximum number of visits and when?

Notes/Comments About SQL Query and Results (Include # of Rows in Result)

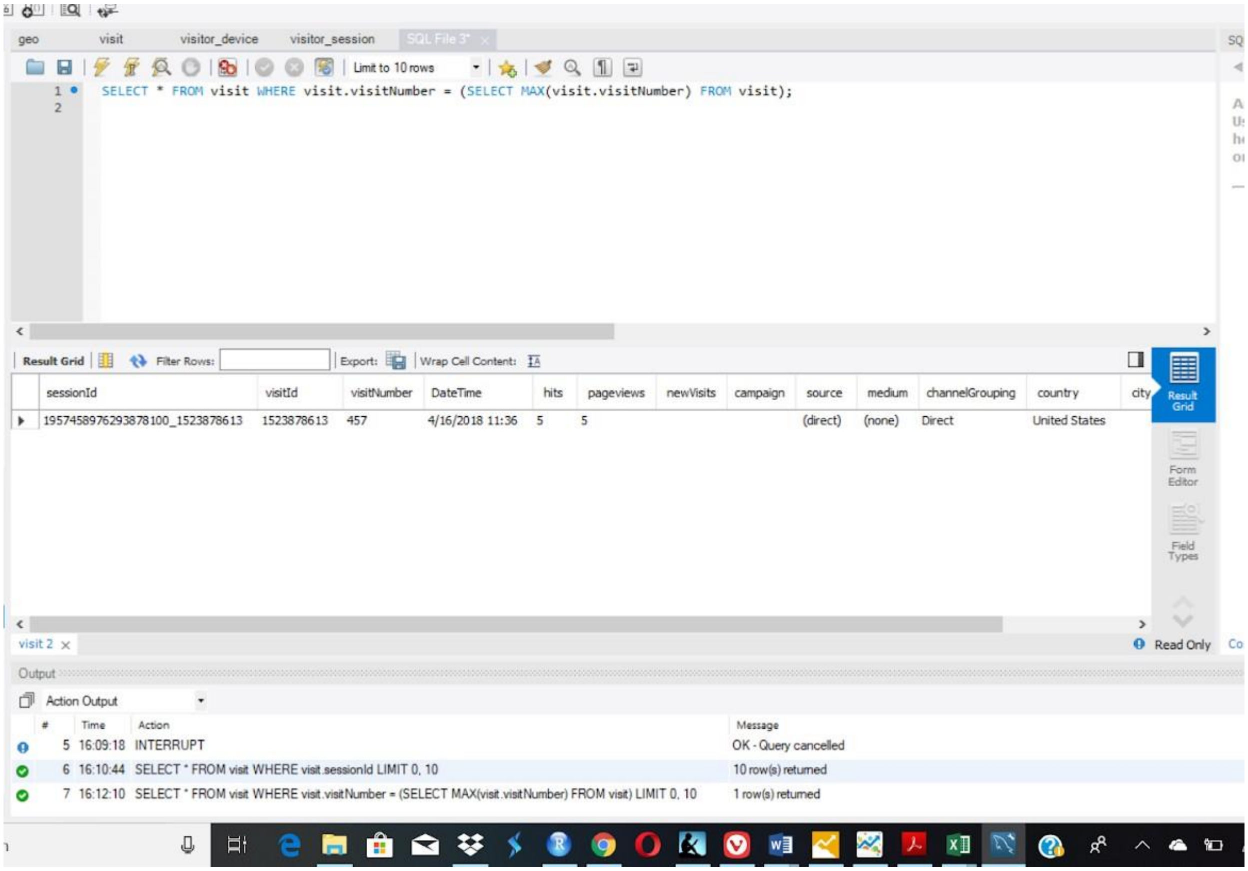
We saw a pattern that the sessionId is the combination of VisitorID and VisitID. From the first part of SessionId we can infer the VisitorID. Also, the visitnumber increment every time this user come back to the store.

Therefore, we try to find the session with the visitnumber that is highest in the visit table. From the sessionId and DateTime of this session we can see the user that had the maximum number of visit and the date and time this person reached this number of visit.

Translation

Select information of the session that has the visitnumber equal to the maximum visit number from the visit table

Screen Shot of SQL Query and Results



```sql
SELECT * FROM visit WHERE visit.visitNumber = (SELECT MAX(visit.visitNumber) FROM visit);
```

sessionId	visitId	visitNumber	DateTime	hits	pageviews	newVisits	campaign	source	medium	channelGrouping	country	city
1957458976293878100_1523878613	1523878613	457	4/16/2018 11:36	5	5			(direct)	(none)	Direct	United States	

```
#  Time      Action                                                                                      Message
5  16:09:18  INTERRUPT                                                                                   OK - Query cancelled
6  16:10:44  SELECT * FROM visit WHERE visit.sessionId LIMIT 0, 10                                       10 row(s) returned
7  16:12:10  SELECT * FROM visit WHERE visit.visitNumber = (SELECT MAX(visit.visitNumber) FROM visit) LIMIT 0, 10   1 row(s) returned
```

Query 3

Question

1. Is a Windows user more likely to visit the store than Macintosh user?

Notes/Comments About SQL Query and Results (Include # of Rows in Result)

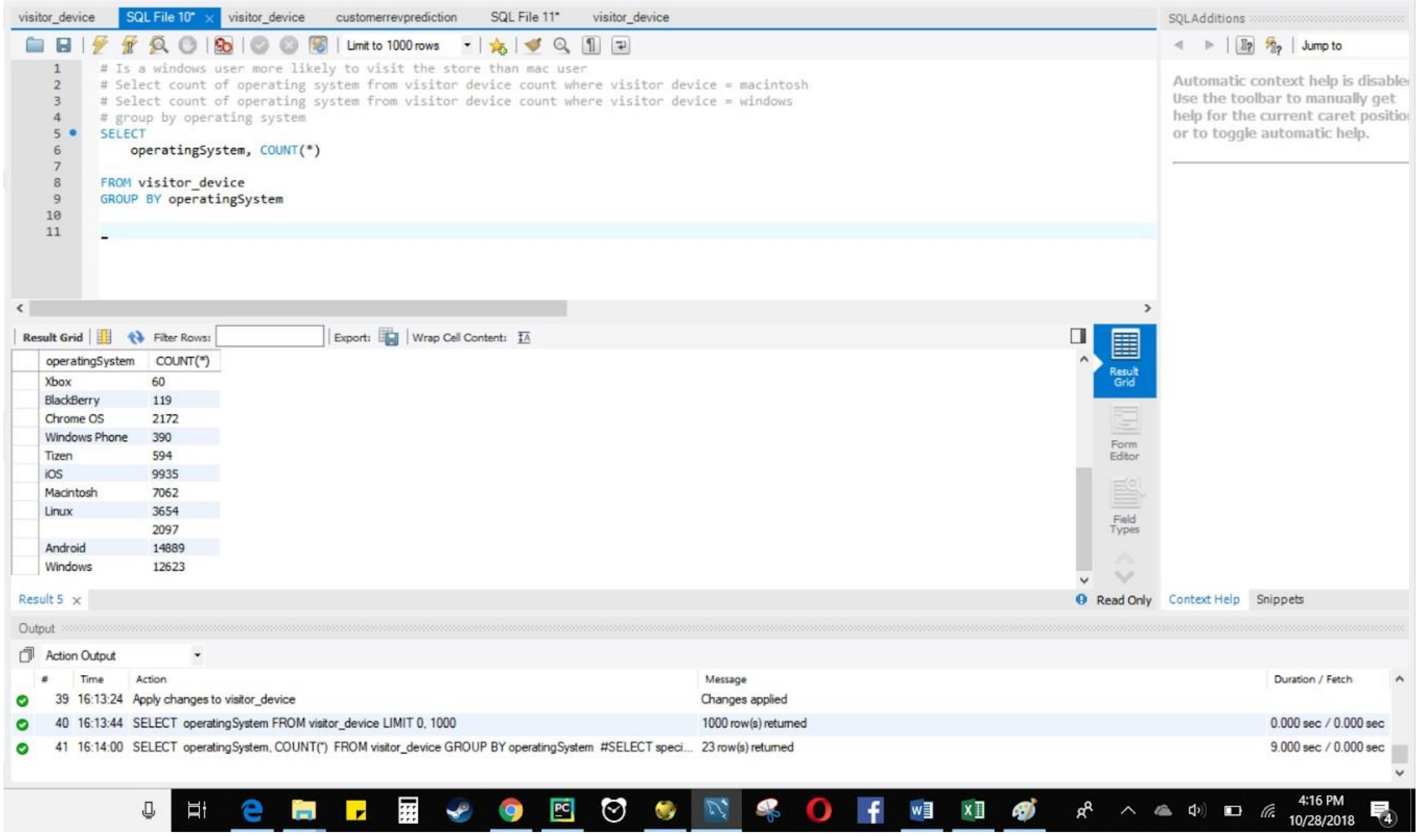
We will find out the number of times each operating system is used and compare windows user with Macintosh to see which have higher number of visits

# of Rows in Result: 23

Translation

select count of times each operating system was used

Screen Shot of SQL Query and Results



```
# Is a windows user more likely to visit the store than mac user
# Select count of operating system from visitor device count where visitor device = macintosh
# Select count of operating system from visitor device count where visitor device = windows
# group by operating system
SELECT
    operatingSystem, COUNT(*)

FROM visitor_device
GROUP BY operatingSystem

_
```

operatingSystem	COUNT(*)
Xbox	60
BlackBerry	119
Chrome OS	2172
Windows Phone	390
Tizen	594
iOS	9935
Macintosh	7062
Linux	3654
	2097
Android	14889
Windows	12623

#	Time	Action	Message	Duration / Fetch
39	16:13:24	Apply changes to visitor_device	Changes applied	
40	16:13:44	SELECT operatingSystem FROM visitor_device LIMIT 0, 1000	1000 row(s) returned	0.000 sec / 0.000 sec
41	16:14:00	SELECT operatingSystem, COUNT(*) FROM visitor_device GROUP BY operatingSystem #SELECT speci...	23 row(s) returned	9.000 sec / 0.000 sec

Query 4

Question

1. What was the average number of hits per unique visitor?

Notes/Comments About SQL Query and Results (Include # of Rows in Result)

The query was running for over an hour and still couldn't finish. The reason might be that the database is large and we're performing complex query with join and group by.

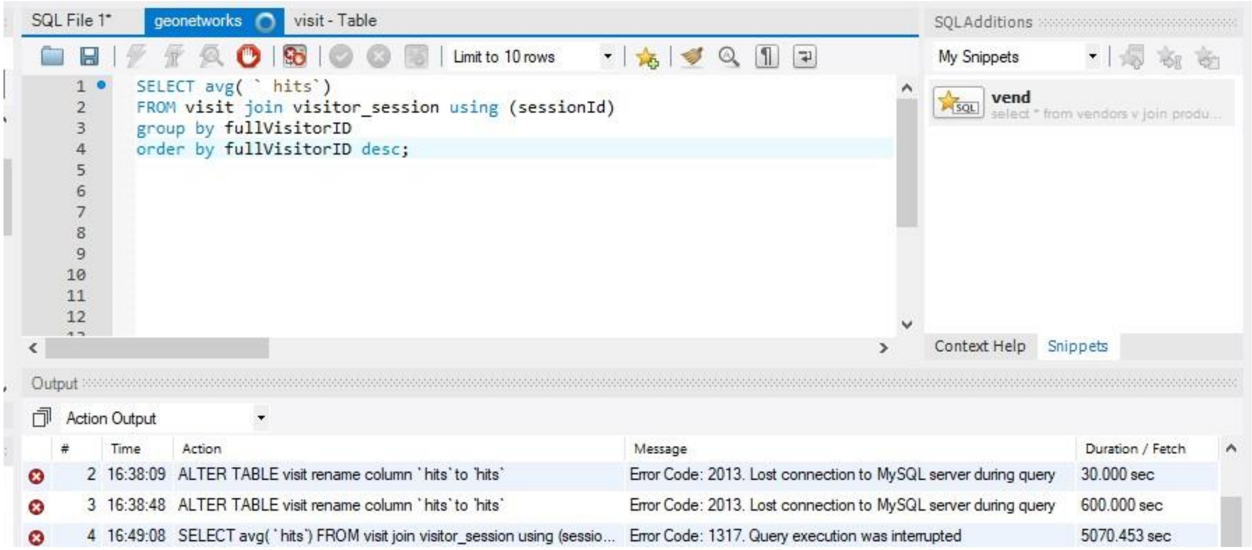
Another reason might be that our assumption is wrong and the sessionId is not unique, thus making it hard to have a join table.

Given enough time and assume that the data model is correct, this query should work and give the average number of hits per visitor. The number of rows will be the number of visitor Id.

Translation

select the average number of hits by each visitor from table visit inner join table visitor_session, using sessionId.

Screen Shot of SQL Query and Results



```sql
SELECT avg( ` hits`)
FROM visit join visitor_session using (sessionId)
group by fullVisitorID
order by fullVisitorID desc;
```

#	Time	Action	Message	Duration / Fetch
2	16:38:09	ALTER TABLE visit rename column ` hits` to `hits`	Error Code: 2013. Lost connection to MySQL server during query	30.000 sec
3	16:38:48	ALTER TABLE visit rename column ` hits` to `hits`	Error Code: 2013. Lost connection to MySQL server during query	600.000 sec
4	16:49:08	SELECT avg(` hits`) FROM visit join visitor_session using (sessio...	Error Code: 1317. Query execution was interrupted	5070.453 sec

Query 5

Question

1. Provide a breakdown of store hits by region

Notes/Comments About SQL Query and Results (Include # of Rows in Result)

Similar to the last one, the query was running for over an hour and still couldn’t finish. The reason might be that the database is large and we’re performing complex query with join and group by.

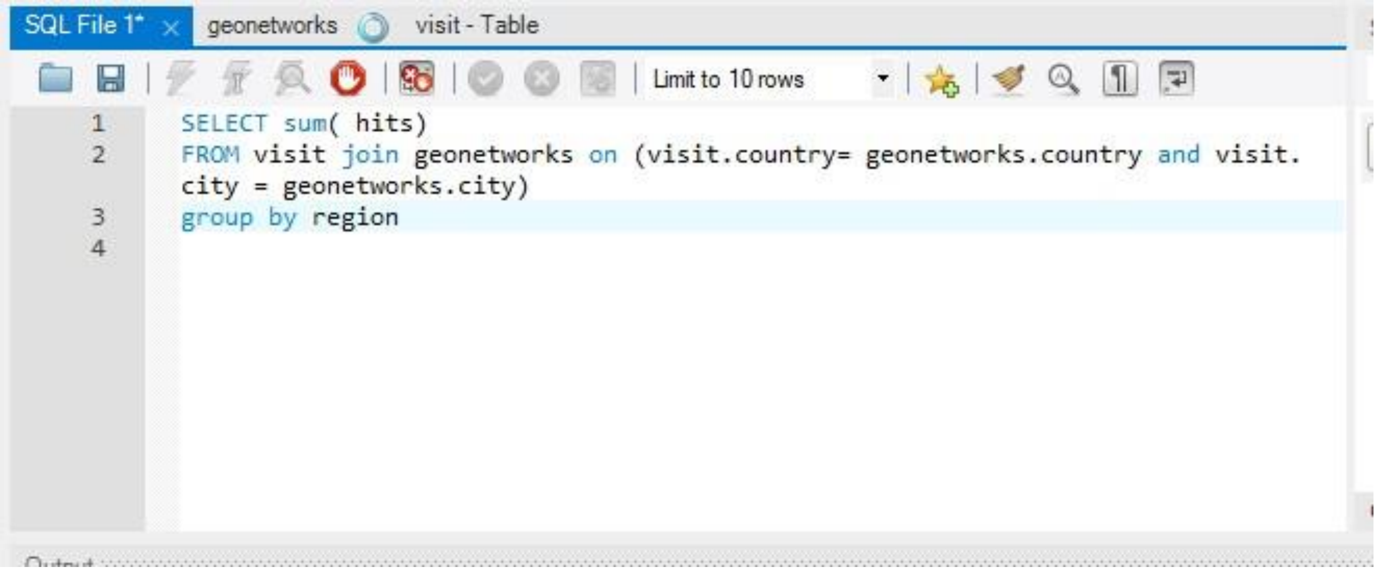
Another reason might be that our assumption is wrong and the combination of country and city is not unique, thus making it hard to have a join table.

Given enough time and assume that the data model is correct, this query should work and give the number of hits by region

Translation

select summation of hits grouped by region

Screen Shot of SQL Query and Results



```
SELECT sum( hits)
FROM visit join geonetworks on (visit.country= geonetworks.country and visit.city = geonetworks.city)
group by region
```

Query 6

Question

1. Visitors from which country visited the store the most?

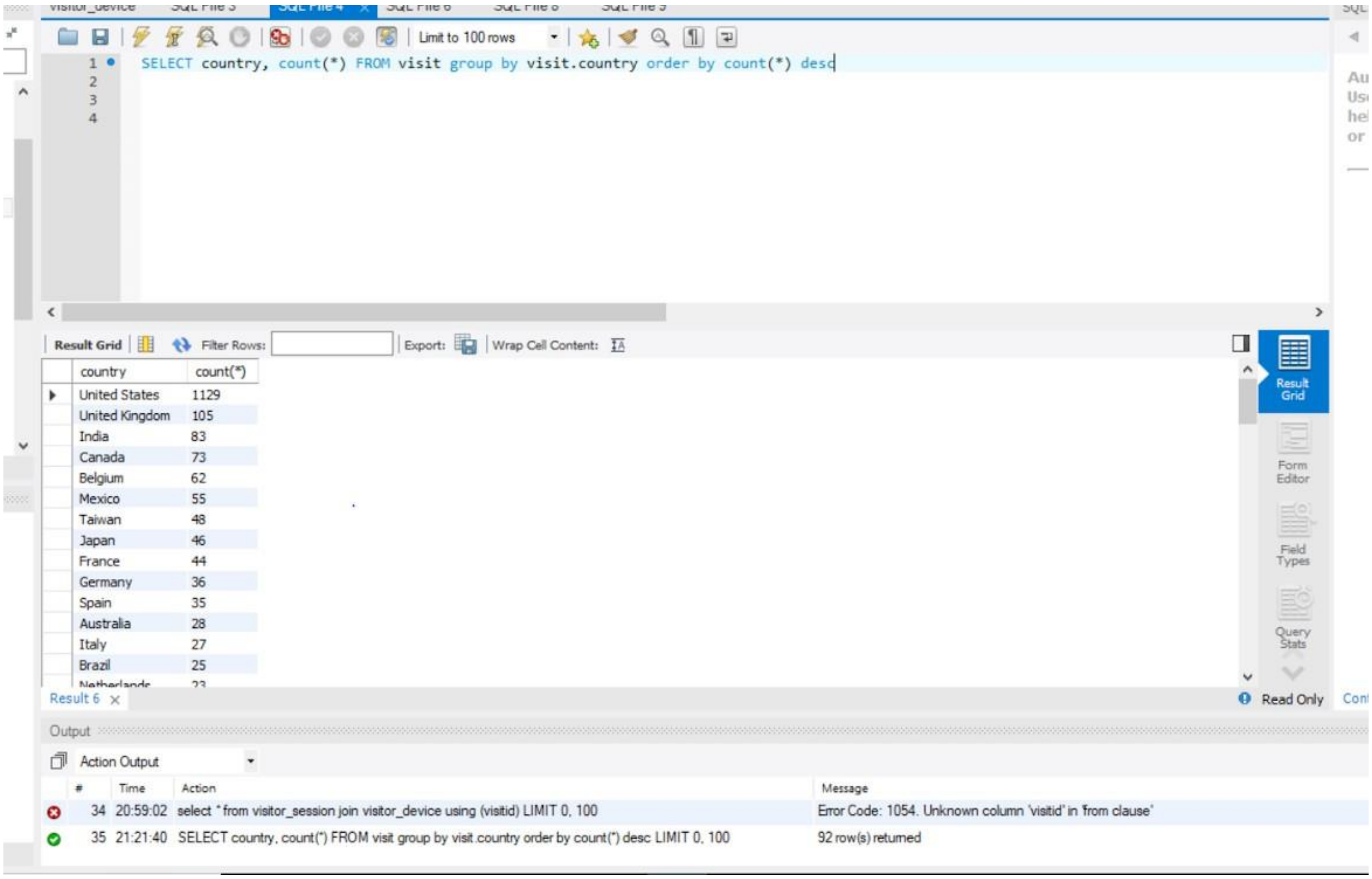
Notes/Comments About SQL Query and Results (Include # of Rows in Result)

We calculated the time people from each country visited the store and sorted by descending order. The result shows that people from the U.S. visited the store the most.

# of rows in Result: 92

Translation

Screen Shot of SQL Query and Results



```
SELECT country, count(*) FROM visit group by visit.country order by count(*) desc
```

country	count(*)
United States	1129
United Kingdom	105
India	83
Canada	73
Belgium	62
Mexico	55
Taiwan	48
Japan	46
France	44
Germany	36
Spain	35
Australia	28
Italy	27
Brazil	25
Netherlands	23

```
#   Time      Action                                                                                   Message
34  20:59:02  select * from visitor_session join visitor_device using (visitid) LIMIT 0, 100              Error Code: 1054. Unknown column 'visitid' in 'from clause'
35  21:21:40  SELECT country, count(*) FROM visit group by visit.country order by count(*) desc LIMIT 0, 100   92 row(s) returned
```

Query 7

Question

1. How many users used only Macintosh devices to visit the store?

Notes/Comments About SQL Query and Results (Include # of Rows in Result)

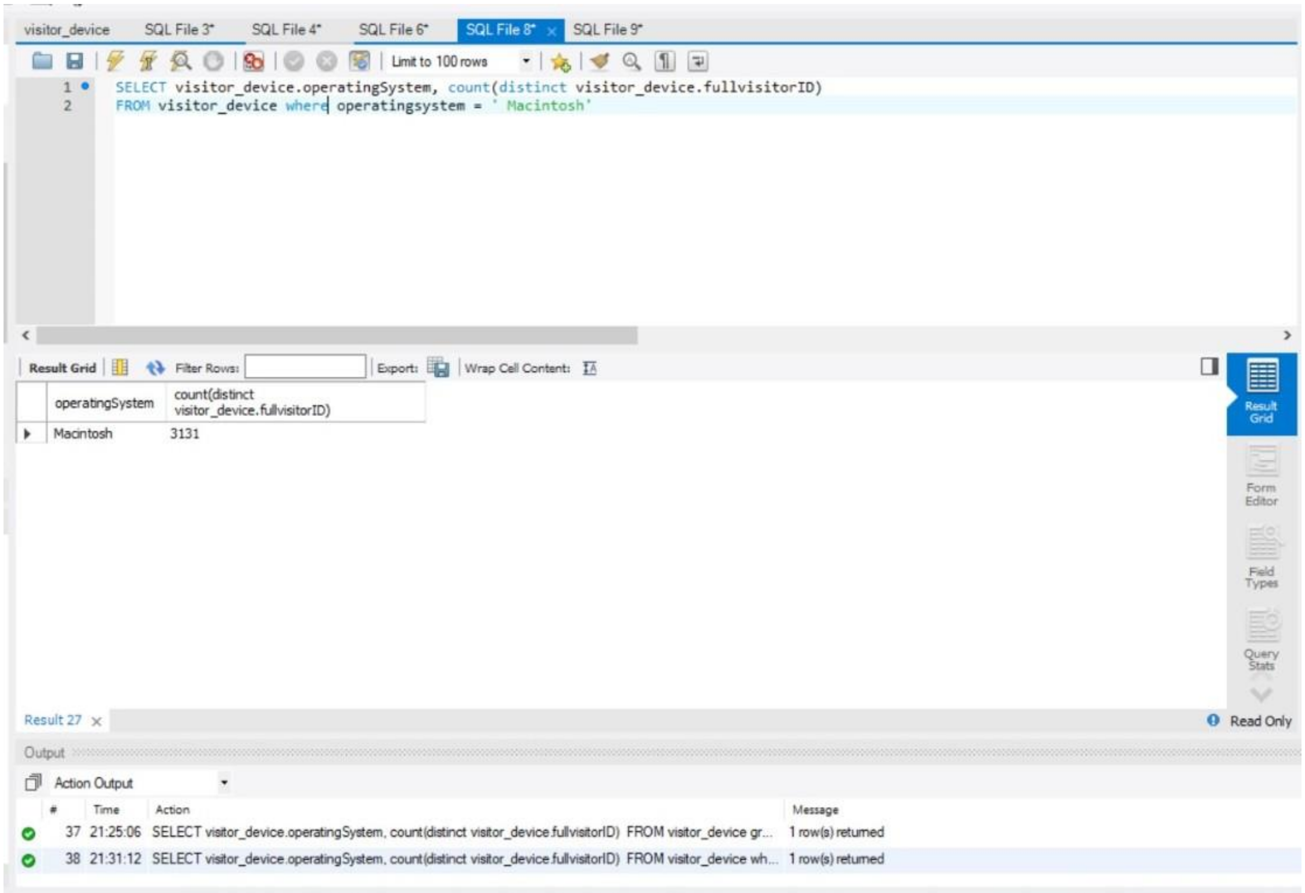
Our assumption is that google analytics doesn’t recognize the person that uses the device, it can only recognizes the device separately. Therefore we counted the number of macintosh devices that visited.

# of rows in result: 1

Translation

select the operation system and the count of visitors that visited the store when the operating system is Macintosh

Screen Shot of SQL Query and Results



```sql
SELECT visitor_device.operatingSystem, count(distinct visitor_device.fullvisitorID)
FROM visitor_device where operatingsystem = ' Macintosh'
```

operatingSystem	count(distinct visitor_device.fullvisitorID)
Macintosh	3131

```
   #   Time      Action                                                                                               Message
✓ 37  21:25:06  SELECT visitor_device.operatingSystem, count(distinct visitor_device.fullvisitorID)  FROM visitor_device gr...  1 row(s) returned
✓ 38  21:31:12  SELECT visitor_device.operatingSystem, count(distinct visitor_device.fullvisitorID)  FROM visitor_device wh...  1 row(s) returned
```