

Visualizing Big Data

Mark Frydenberg
Computer Information Systems Department
Bentley University

mfrydenberg@bentley.edu
@checkmark

OBJECTIVES

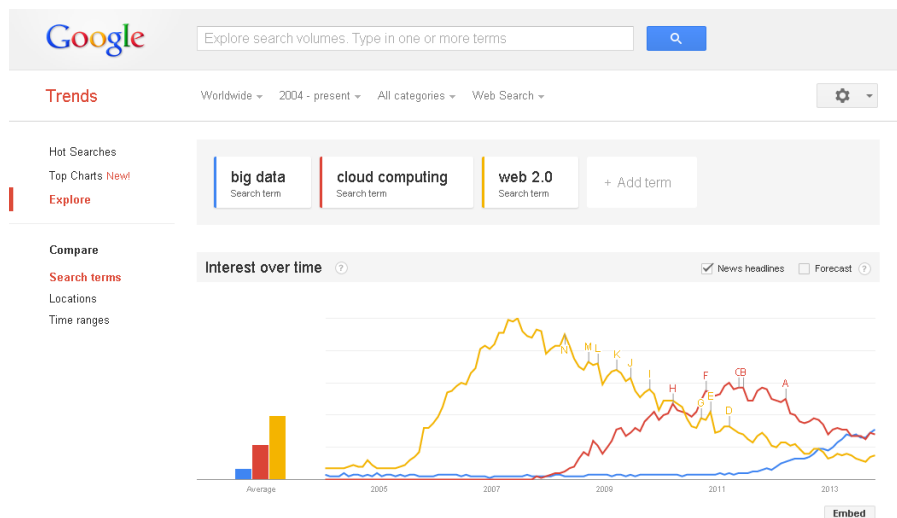
A flood of information online from tweets, news feeds, status updates, photos, government databases, private, and other sources contribute to the volume, velocity, and variety of Big Data artifacts available today. Yet introducing Big Data concepts and technologies in the classroom often waits for advanced students in database or programming courses. This workshop will share several activities appropriate to incorporate Big Data concepts as part of introductory technology concepts, web/Internet, or database courses. Using free online resources participants will learn to query and analyze real world data, create visualizations, and understand the impact of Data as a Service as a model for cloud computing.

Activity 1: Volume, Variety, Velocity

This activity introduces the concepts of volume, variety, and velocity when referring to big data through the use of Twitter, Google, and other big data sources.

Volume

Google Trends summarizes activity for specified search terms over time to give an indication of when they are popular. Visit Google Trends at <http://trends.google.com>. Search for big data, cloud computing, or other popular topics.



How does Google get all this information? Google archives every search term that users enter on Google. How else does Google use information based on everyone's search queries?

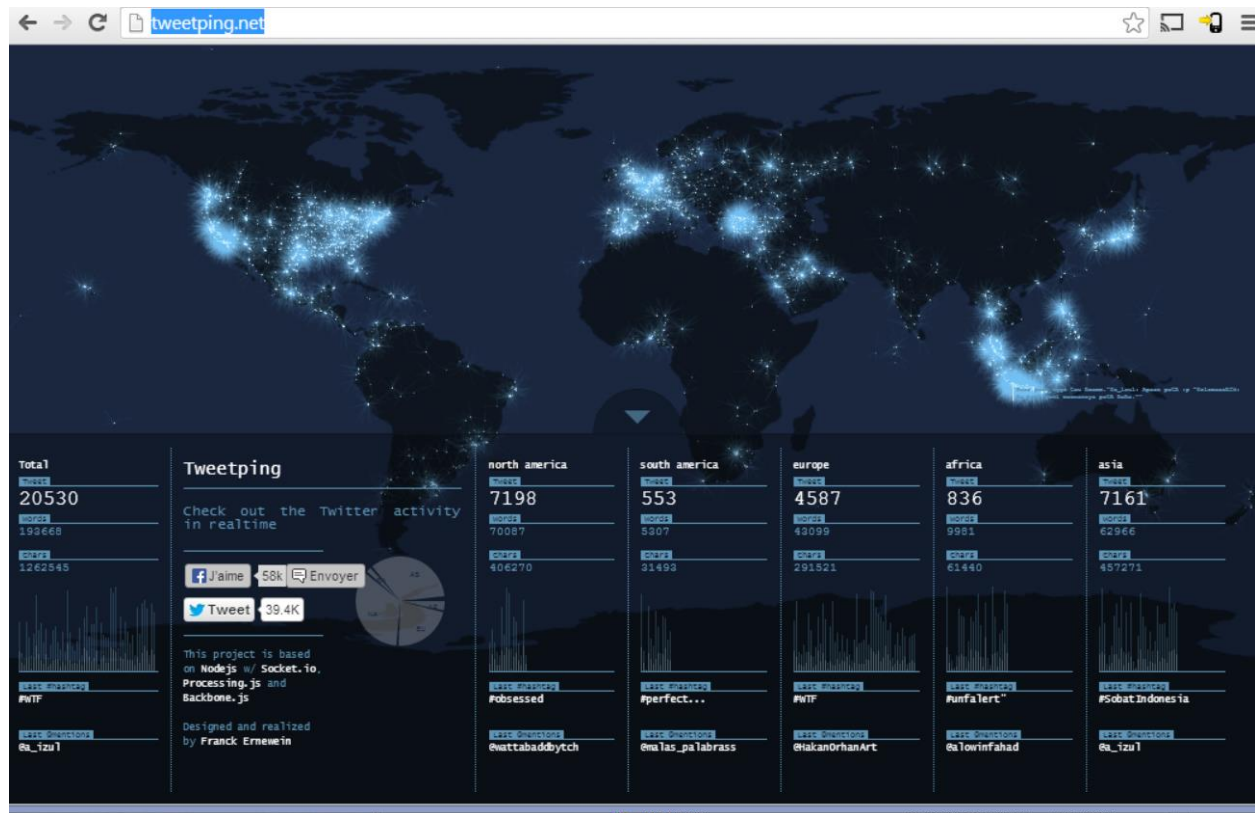
Collaborative filtering is the process of using data from many people to make decisions or offer advice. Can you think of other examples of collaborative filtering?

Velocity

Visit <http://tweetping.net/> for an example of a Twitter visualization in real time. (Or search online for other real time twitter visualizations).

What does this visualization tell you about Twitter's data?

How does twitter data meet the criteria of Volume, Variety, and Variety?



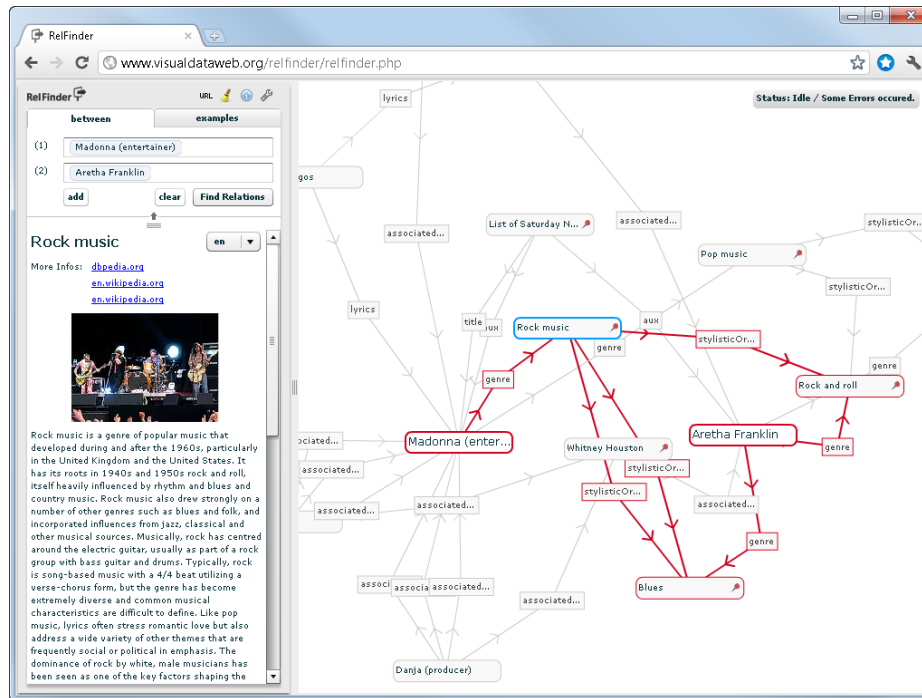
Other Twitter Visualizations:

- <https://www.mapbox.com/blog/visualizing-3-billion-tweets/>

Variety

Make use of the Relfinder app (located at <http://www.visualdataweb.org/relfinder.php>) to find relationships between seemingly unrelated topics.

This big data set is from dbPedia, which has indexed Wikipedia's content and stores it in a format known as RDF (representational data format). RDF uses triples (object 1, relationship, object 2) to store relationships between objects.



Activity 2: Analyze Data using Queries in Google BigQuery

Querying massive datasets can be time consuming and expensive without the right hardware and infrastructure. Google BigQuery solves this problem by enabling super-fast, SQL-like queries against append-only tables, using the processing power of Google's infrastructure. Google provides several sample databases for querying. Advanced users can upload their own databases. For a technical discussion of BigQuery, see <https://cloud.google.com/files/BigQueryTechnicalWP.pdf>.

WHAT YOU NEED:

- A Google ID
- Google Chrome browser
- An Internet connection

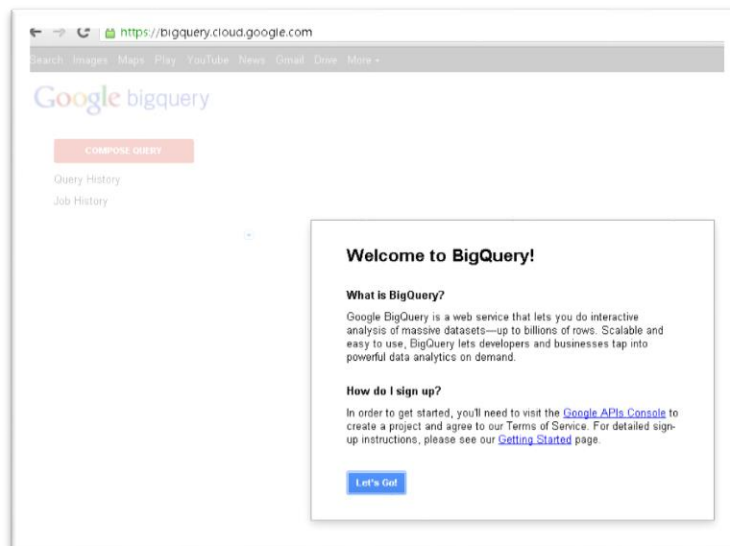
In this activity you will learn to create simple queries using Google BigQuery.

Setup Instructions

Sign in to Google using the Chrome browser.

Visit <http://bigquery.cloud.google.com>

If see the Welcome to BigQuery message below, it means you need to need to subscribe to the BigQuery service using your Google account. (If you don't see this message continue at Using BigQuery below.)



To subscribe to the service, click the Let's Go button, check the box to agree to terms of service; click accept.

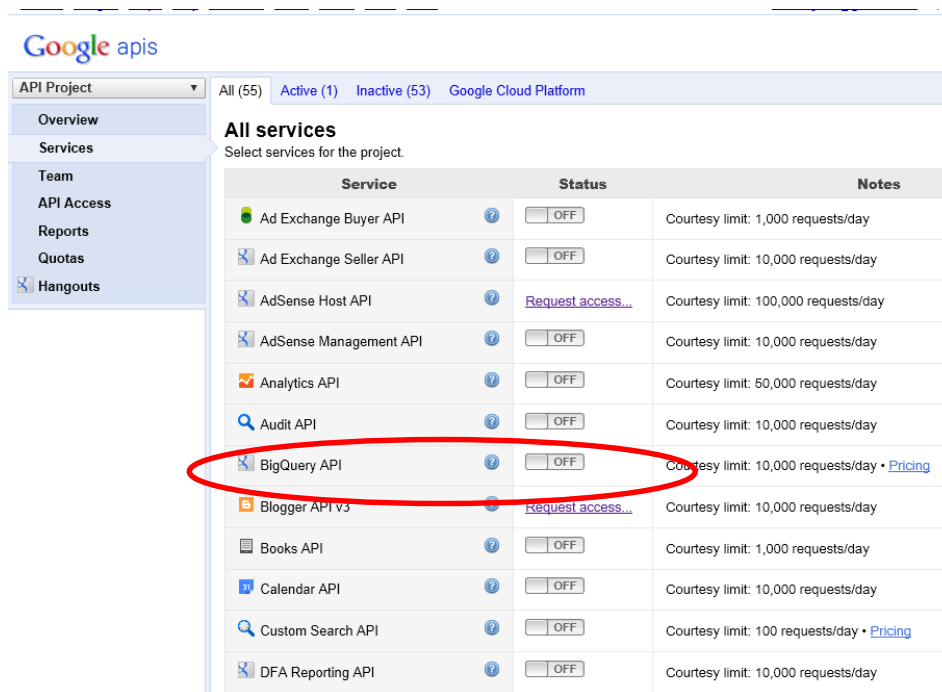
How does Data as a Service fit into the cloud computing model?

More generally, what is the difference between an application and a service?

How do service providers charge user for consuming cloud-based services?

On the Google api's page, click the Services Link.

In the All services panel, slide the BigQuery API switch to ON



Check the box to agree to any terms of service.

Return to bigquery.cloud.google.com (or click the ? near BigQuery api, which should also take you there.)

What is SQL?

How do you query a database?

BigQuery uses a variation of SQL's SELECT statement, including these keywords:

- SELECT
- WITHIN
- FROM
- JOIN
- WHERE
- GROUP BY
- HAVING
- ORDER BY
- LIMIT

See <https://developers.google.com/bigquery/query-reference> for the Query Reference.

Your First BigQuery Query

In this query we will examine the Shakespeare database.

Click the Public Data Samples tab at the left to list tables.

Click Shakespeare under the publicdata:samples list.

View the schema which shows the names of the fields that are available. Click Details.

- *What makes this a big database?*
- *How many rows are in the Shakespeare database?*

Let's find the number of 15 letter words in all of Shakespeare by querying the database.

Click the Query Table button to set up a query.

Click the Schema button to view the schema (database description) again.

Click on the word "word" in the Schema to add it to the query after the SELECT command

New Query ? ×

```
1 SELECT word FROM [publicdata:samples.shakespeare] LIMIT 1000
```

RUN QUERY

Table Details: shakespeare Schema Details Query Table

Schema

word	STRING	REQUIRED
word_count	INTEGER	REQUIRED
corpus	STRING	REQUIRED
corpus_date	INTEGER	REQUIRED

Click on the word corpus in the schema to add it to the query as well

Finish typing the rest of the command as shown:

New Query

? X

```
1 SELECT word, corpus FROM [publicdata:samples.shakespeare] where length(word) = 15 LIMIT 1000
```

RUN QUERY

Click the Run Query Button to run the query.

New Query

? X

```
SELECT word, corpus FROM [publicdata:samples.shakespeare] where length(word) = 15 LIMIT 1000
```

RUN QUERY

Save Query

Enable Options

Query complete (1.4s elapsed, 3.62 MB processed)



Query Results 4:03pm, 24 Oct 2013

Download as CSV

Save as Table

Row	word	corpus	
1	disproportion'd	othello	
2	notwithstanding	othello	
3	circumscription	othello	
4	extraordinarily	2kinghenryiv	
5	'Northumberland	2kinghenryiv	
6	Gloucestershire	2kinghenryiv	
7	notwithstanding	2kinghenryiv	

[First](#) [< Prev](#) Rows 1-7 of 61 [Next >](#) [Last](#)

- *Where is this database stored?*
- *How big is it?*
- *How long did it take to perform the query?*
- *What might happen if you tried to use Microsoft Access to query this database?*

More Queries to Try

Shakespeare Data Queries

What are the names of all of Shakespeare's writings?

```
SELECT corpus FROM [publicdata:samples.shakespeare] group by corpus;
```

What 15-letter word appears in Shakespeare's King Henry VIII? (kinghenryviii)

What 17-letter word appears in A Midsummer Night's Dream? (midsummersnightsdream)

Wikipedia Data Queries

How many rows appear in the Wikipedia database?

Hint: Click on the Wikipedia database in the list under publicdata:samples.

How many titles in Wikipedia contain the word "Massachusetts"?

Use this query to find out:

```
SELECT count(title) FROM [publicdata:samples.wikipedia] WHERE title CONTAINS "Massachusetts"
```

Perform the following analysis of your search results:

- How long does this query take to run?
- How many titles are found?
- What percentage of articles in the Wikipedia database match this criteria?

How many titles in Wikipedia contain numeric characters?

```
SELECT count(*) from [publicdata:samples.wikipedia]  
WHERE REGEXP_MATCH (title, '[0-9]*') AND wp_namespace = 0;
```

Perform the following analysis of your search results:

- How long does this query take to run?
- How many titles are found?
- What percentage of articles in the Wikipedia database match this criteria?

Write a query to list titles of articles in Wikipedia contain the letters in GOOGLE in order?

```
SELECT title, sum (num_characters) as num_characters
FROM [publicdata:samples.wikipedia]
WHERE
  regexp_match(title, r'^G.*o.*o.*g.*l.*e$')
GROUP BY title
ORDER BY num_characters DESC
LIMIT 1000
```

Nativity (Births) Data Queries

The Natality database contains historical information about birthrates in the United States. Determine how many baby girls were born in 1999.

- Can you write the query?

What does this (more complicated) query of the Natality database tell you?

```
SELECT state, year,
       AVERAGE(mother_age) as avg_age,
       AVERAGE(weight_pounds) as avg_weight
FROM [publicdata:samples.natality]
WHERE state IS NOT NULL
      AND ever_born = 1
      AND mother_age IS NOT NULL
      AND weight_pounds IS NOT NULL
GROUP BY year, state
```

Weather Data

Now let's evaluate weather data. According to <http://www.wetterzentrale.de/klima/stn1st.html>, station number 725090 is Boston's Logan Airport. Write a query to tell you the temperature on January 1 in Boston for each year on record?

Can you write a query to determine the years for which temperatures in Boston were over 50 degrees on January 1?

Activity 3: Creating Visualizations of BigQuery Data using Microsoft Excel

In this activity you will use BigQuery to query a large database, and then import the resulting data to Excel to make use of its data visualization capabilities (charts, graphs, spark lines).

Download the BigQuery internet query file `connector.iqy` from <https://bigquery-connector.appspot.com/>. Follow the instructions to create a key.

Option 1 (easier Excel skills – charts and graphs)

- Using Google BigQuery, write a query to see how many times the word “love” appears in all of Shakespeare’s works.

```
SELECT corpus, word_count, word FROM [publicdata:samples.shakespeare] WHERE word="love" ORDER BY word_count DESC
```

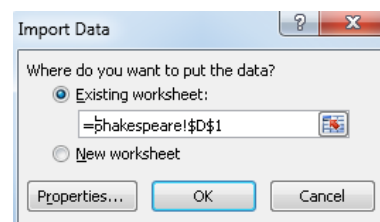
We will use this query to create a more general query to find how many times any word a user enters appears in all of Shakespeare's works.

- Open Excel. Create a worksheet named **Shakespeare**, with your project ID and key.
 - Allow the user to enter a word in cell B4.
 - In cell B6, form a query to find out how many times it appears in each of Shakespeare's works. Use the string concatenation operator & to form a query that includes the word entered in cell B4. Note the use of double quotation marks to represent a single quotation mark:

```
= "SELECT corpus, word_count, word FROM [publicdata:samples.shakespeare] where word=" & B4 & " " order by word_count desc"
```

	A	B
1	Project ID	240656651949
2	Key	3p8vI9s6Wgn82zzjAlM7kF5gKg+Cj9SYduf/76QkzS/bSP9HQEMduoSCAwenVaqUFSBXRSSCunzwANg42wJ1+rh3VaK/WvTMWF5KuyZK403RjF5dgyOZeJtcelbFyZLk
3		
4	Word	love
5		
6	Query	SELECT corpus, word_count, word FROM [publicdata:samples.shakespeare] where word="love" order by word_count desc
7		
8		
9		

- Use BigQuery Connector to import this data into Excel, placing it beginning in cell D1.
 - Go to Data -> Get external data -> Existing Connections -> Connections on this Computer -> browse for more -> (navigate to downloads) -> select connector.iqy
 - On the Import Data dialog box, click the New Worksheet selector to put the data the existing worksheet in cell D1.



- c. Click the OK button to continue to the next step.
- d. On the Enter a Parameter value dialog box, locate the cell (=Shakespeare!B6) containing the query to run using BigQuery.
- e. Place a check in the box labeled Use this value/reference for future refreshes, so that when we change the query, Excel will automatically go out to BigQuery to import the latest data.
- f. Place a check in the Refresh automatically when cell value changes box.
- g. Click the OK button to continue.
- h. Follow similar steps to enter the project ID (cell B1) and the BigQuery key (cell B2). Be sure to check both boxes in the Enter Parameter Value dialog boxes.

Enter Parameter Value

Enter a query:

=Shakespeare!\$B\$6

☒ Use this value/reference for future refreshes

☒ Refresh automatically when cell value changes

OK Cancel

Enter Parameter Value

Enter a project ID:

=Shakespeare!\$B\$1

☒ Use this value/reference for future refreshes

☒ Refresh automatically when cell value changes

OK Cancel

Enter Parameter Value

Enter your Connector Key:

=Shakespeare!\$B\$2

☒ Use this value/reference for future refreshes

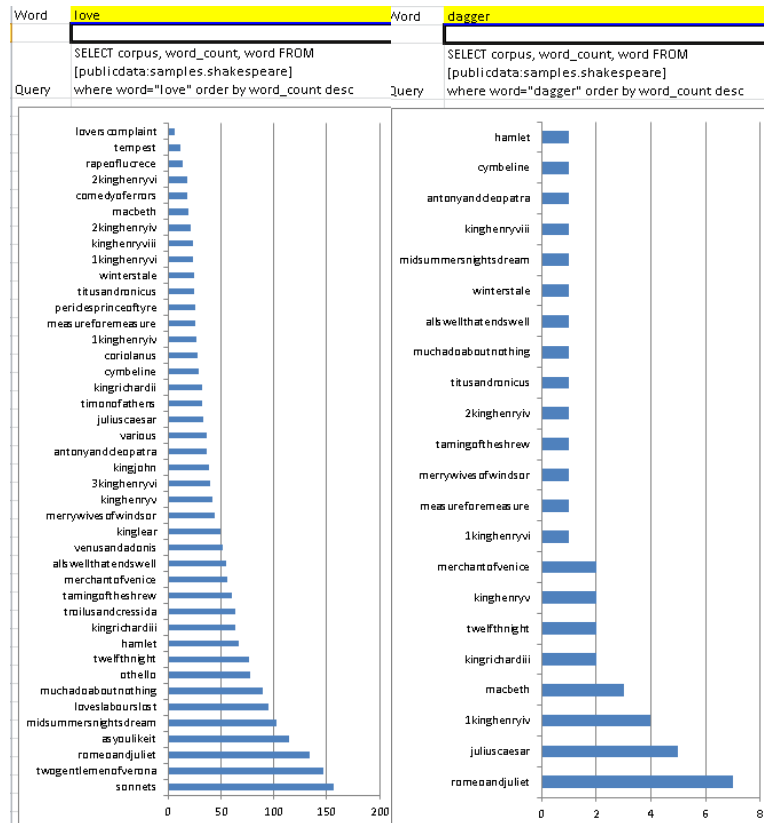
☒ Refresh automatically when cell value changes

OK Cancel

Excel invokes BigQuery to import the data from the query results into Excel.

	A	B	C	D	E	F
1	Project ID	240856651949		sonnets	157	love
2	Key	3p8u9s6Wgn82zzjAlM7kF5gKg+Cj9SYduf/760kz5.bSP9H QEMduoSCAwenVaqUF5BXRSSCunzwAng42wJ1+rh3Va KWW1MWF5KuyZK403Rf5dgyOZeJtcclbFyZLk		twogentlemenofverona	147	love
3				romeoandjuliet	134	love
4	Word	love		asyoulikeit	114	love
5				midsummersnightsdrea	102	love
6	Query	SELECT corpus, word_count, word FROM [publicdata:samples.shakespeare] where word="love" order by word_count desc		loveslabourslost	95	love
7				muchadoaboutnothing	89	love
8				othello	78	love
9				twelfthnight	77	love
10				hamlet	67	love
11				kingrichardiii	63	love
12				trollusandressida	63	love
13				tamingoftheshrew	60	love
14				merchantofvenice	56	love
15				allswellthatendswell	55	love
16				venusandadonis	52	love
17				kinglear	49	love
18				merrywivesofwindsor	44	love
19				kinghenryv	42	love
20				3kinghenryvi	40	love
21				kingjohn	39	love
22				antonyanddeopatra	37	love
23				various	37	love
24				juliuscaesar	33	love
25				timonofathens	32	love
26				kingrichardii	32	love
27				cymbeline	29	love
28				coriolanus	28	love
29				1kinghenryiv	27	love
30				measureforemeasure	26	love
31				peridesprinceof tyre	26	love
32				titusandronicus	25	love
33				winterstale	25	love

4. Create a horizontal bar chart to graph data from columns D and E. Enter a new word into cell B3 and watch the bar chart update.



Option 2 (more advanced Excel skills – pivot tables and spark lines)

Launch Excel.

Create a worksheet named Query with your project ID, key, and three queries as shown below. The values in A4:B6 allow us to identify each query by number, and the formula in cell B10 allows us to specify the query to run in BigQuery by entering its number in cell A10.

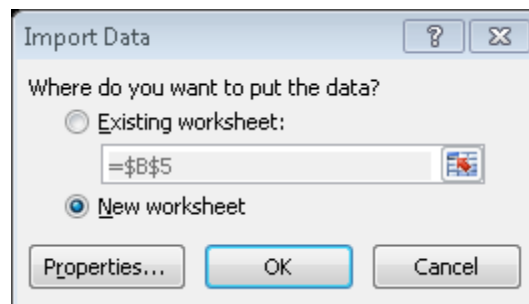
	A	
1	Project ID	45131167026
2	Key	3p8vI9s6Wgn82zzjAlM7klvyCiVAJJzaxWQCKI6W/SKx9EslbXNvoLxDmLDpor4UrzMMdCTJfnCGuST1bLUXG4uN0Ya3z11/Wc79pg6Fh8=
3		
4	1	SELECT state, year, AVG(weight_pounds) FROM [publicdata:samples.natality] GROUP BY year, state
5	2	SELECT state, year, AVG(mother_age) FROM [publicdata:samples.natality] GROUP BY year, state
6	3	SELECT state, year, COUNT(is_born) FROM [publicdata:samples.natality] GROUP BY year, state
7		
8		
9	Selector	
10	2	=VLOOKUP(A10,\$A\$4:\$B\$6,2,FALSE)
11		

Instruct Excel to use the connector.iqy file.

Go to Data -> get external data -> Existing Connections -> Connections on this Computer -> browse for more -> (navigate to downloads) -> select connector.iqy

On the Import Data dialog box, click the New Worksheet selector to put the data on a new worksheet.

Click the OK button to continue to the next step.



On the Enter a Parameter value dialog box, locate the cell (=Sheet1!B10) containing the query to run using BigQuery.

Place a check in the box labelled Use this value/reference for future refreshes, so that when we switch between queries, Excel will automatically go out to BigQuery to import the latest data.

Click the OK button to continue.

Enter Parameter Value

Enter a query:

=Query!\$B\$10

☒ Use this value/reference for future refreshes

☒ Refresh automatically when cell value changes

OK Cancel

Place a check in the Use this value/reference for future refreshes box.

Place a check in the Refresh automatically when cell value changes box.

Click the OK button to continue.

Follow similar steps to enter the project ID and the BigQuery key.

Enter Parameter Value

Enter a project ID:

=Query!\$B\$1

☒ Use this value/reference for future refreshes

☒ Refresh automatically when cell value changes

OK Cancel

Enter Parameter Value

Enter your Connector Key:

=Query!\$B\$2

☒ Use this value/reference for future refreshes

☒ Refresh automatically when cell value changes

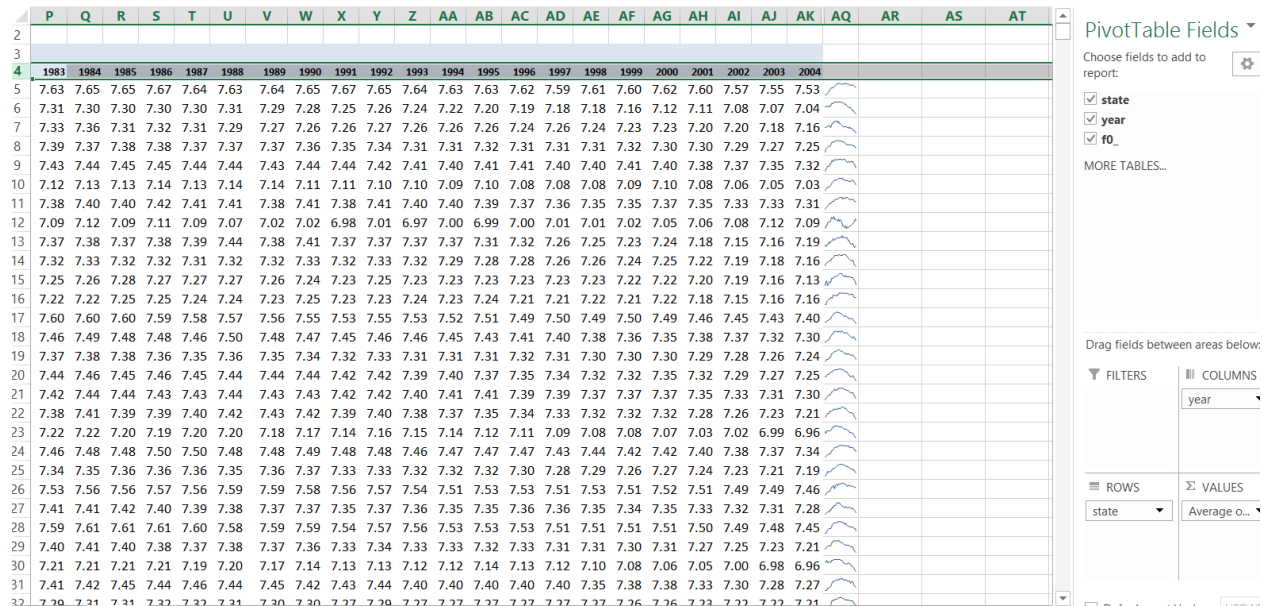
OK Cancel

Excel invokes BigQuery to import the data from the query results into Excel in the new sheet.

Return to the Query sheet, and in the selector box, enter a number (1, 2, or 3) for another query. View the data that results. Convince yourself that the data changes.

ANALYSIS AND VISUALIZATION

Once you have the data loaded in excel, use a pivot table to analyze the results. Your pivot table might look something like this, which includes a spark line visualization.



Try to create a similar visualization of big data yourself. Copy the Query sheet and replace the queries from the Natality database with queries you write (or have already written) from one of the other databases, and import the results into Excel using the BigQuery connector. Then create a chart, graph, spark lines, or other visualization of the data.

Why is it preferable to do the data mining on the Google server, and the visualization tasks on your computer?