Drinking from the Fire Hose: Tools for Teaching Big Data Concepts in the Introductory IT Classroom

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 is designated for advanced students in database or programming courses. This workshop will share several activities appropriate to incorporate Big Data concepts in the introductory IT classroom. Using free online resources from Google, the US Government, Wikipedia, IMDb, and other providers, participants will learn to query and analyze real world data, create infographics and visualizations, and understand the impact of Data as a Service as a model for cloud computing.

TARGETED ATTENDEES

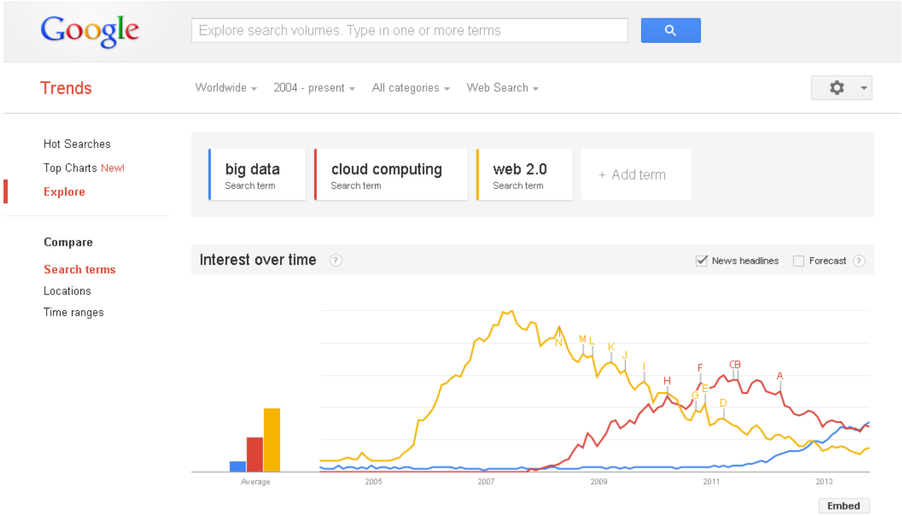
This workshop is suitable for instructors of introductory IT or database courses who wish to incorporate Big Data concepts. It is based on activities presented to 7 sections of IT 101 at the instructor's university during the spring 2013 semester.

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.



big data
Search term

cloud computing
Search term

web 2.0
Search term

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?

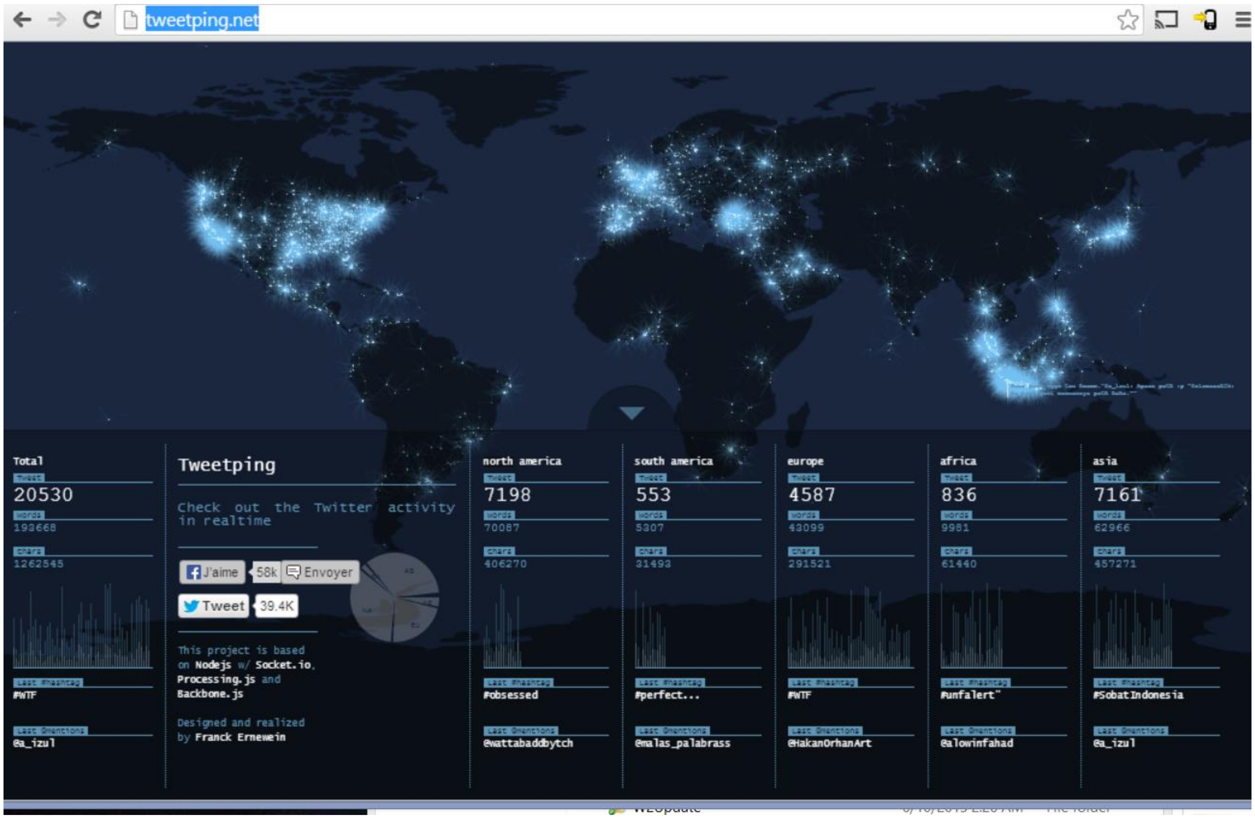
Collaborative filtering is the process of using data from many people to make decisions or offer advice. Can you think of other examples of collaborative filtering?

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?



Tweetping

Check out the Twitter activity in realtime

This project is based on Nodejs w/ Socket.io, Processing.js and Backbone.js

Designed and realized by Franck Ernewein

	Total	north america	south america	europe	africa	asia
Tweet	20530	7198	553	4587	836	7161
Words	193668	70087	5307	43099	9981	62966
Chars	1262545	406270	31493	291521	61440	457271
Last #hashtag	#WTF	#obsessed	#perfect...	#WTF	#unfalert"	#SobatIndonesia
Last @mentions	@a_izul	@wattabaddbytch	@malas_palabrass	@HakanOrhanArt	@alowinfahad	@a_izul

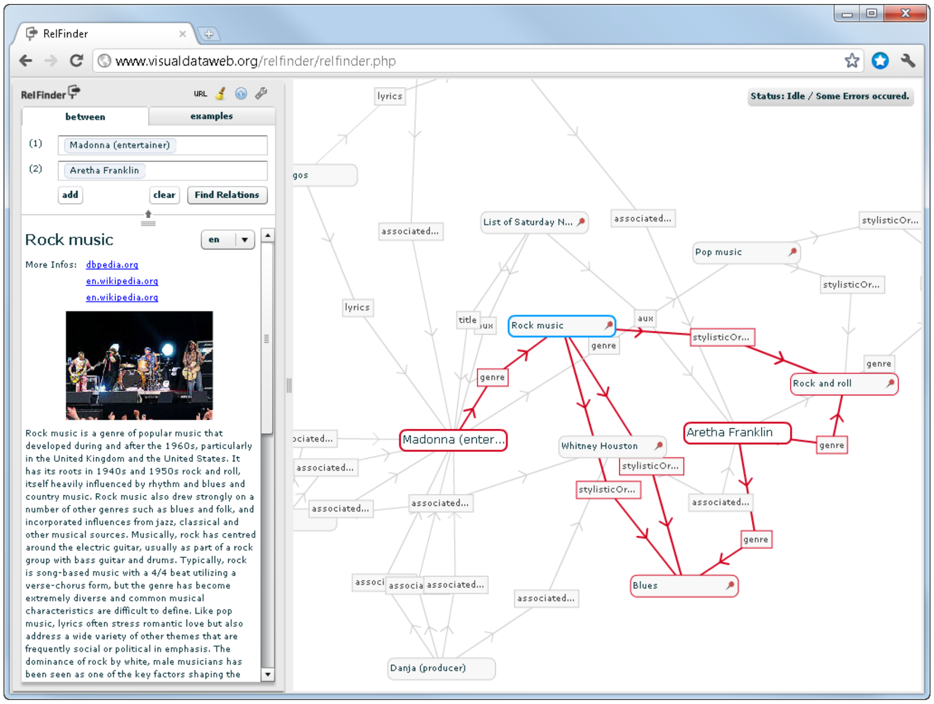
Other Twitter Visualizations:

- 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.



RelFinder

www.visualdataweb.org/relfinder/relfinder.php

Status: Idle / Some Errors occured.

(1) Madonna (entertainer)

(2) Aretha Franklin

Rock music

More Infos: dbpedia.org en.wikipedia.org en.wikipedia.org

Rock music is a genre of popular music that developed during and after the 1960s, particularly in the United Kingdom and the United States. It has its roots in 1940s and 1950s rock and roll, itself heavily influenced by rhythm and blues and country music. Rock music also drew strongly on a number of other genres such as blues and folk, and incorporated influences from jazz, classical and other musical sources. Musically, rock has centred around the electric guitar, usually as part of a rock group with bass guitar and drums. Typically, rock is song-based music with a 4/4 beat utilizing a verse-chorus form, but the genre has become extremely diverse and common musical characteristics are difficult to define. Like pop music, lyrics often stress romantic love but also address a wide variety of other themes that are frequently social or political in emphasis. The dominance of rock by white, male musicians has been seen as one of the key factors shaping the

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

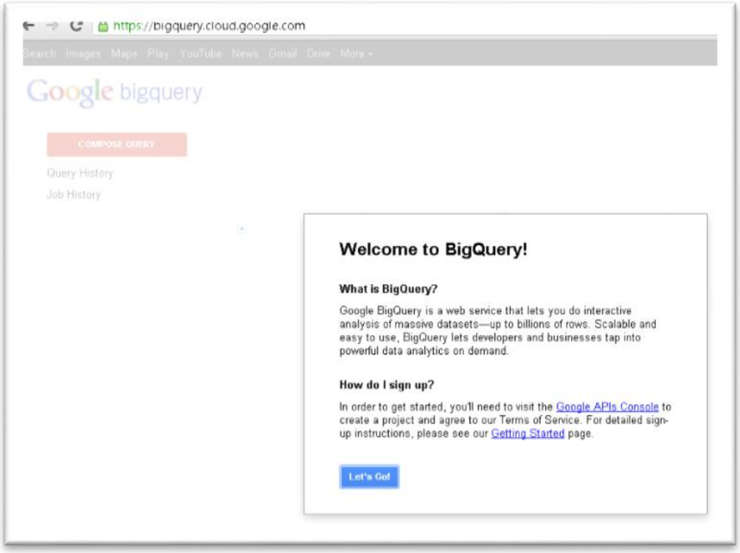
In this activity you will learn to create simple queries using Google BigQuery.

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.)



https://bigquery.cloud.google.com
Search Images Maps Play YouTube News Gmail Drive More
Google bigquery
COMPOSE QUERY
Query History
Job History

Welcome to BigQuery!

What is BigQuery?

Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.

How do I sign up?

In order to get started, you'll need to visit the Google APIs Console to create a project and agree to our Terms of Service. For detailed sign-up instructions, please see our Getting Started page.

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

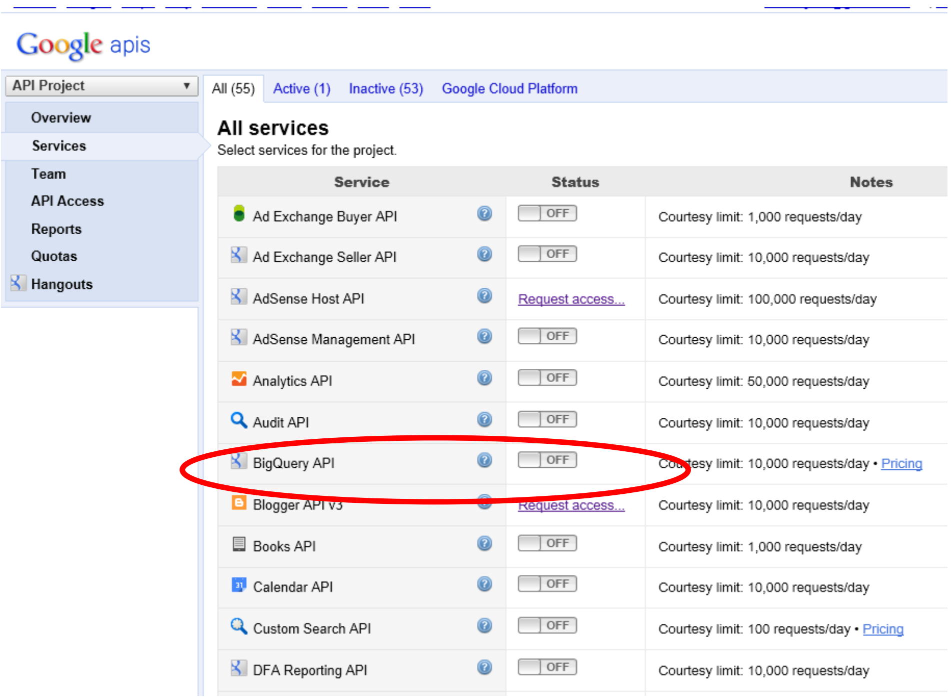
This is a good time to review (or introduce) cloud computing services (Software as a Service, Platform as a Service, and Infrastructure as a Service) with which students may be familiar, and introduce the concept of Data as a Service.

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



All services

Select services for the project.

Service	Status	Notes
Ad Exchange Buyer API	OFF	Courtesy limit: 1,000 requests/day
Ad Exchange Seller API	OFF	Courtesy limit: 10,000 requests/day
AdSense Host API	Request access...	Courtesy limit: 100,000 requests/day
AdSense Management API	OFF	Courtesy limit: 10,000 requests/day
Analytics API	OFF	Courtesy limit: 50,000 requests/day
Audit API	OFF	Courtesy limit: 10,000 requests/day
BigQuery API	OFF	Courtesy limit: 10,000 requests/day • Pricing
Blogger API v3	Request access...	Courtesy limit: 10,000 requests/day
Books API	OFF	Courtesy limit: 1,000 requests/day
Calendar API	OFF	Courtesy limit: 10,000 requests/day
Custom Search API	OFF	Courtesy limit: 100 requests/day • Pricing
DFA Reporting API	OFF	Courtesy limit: 10,000 requests/day

.

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.)

This is a good time to introduce the concept of a database, and querying a database using a query language such as SQL. For those familiar with SQL, mention the capabilities of BigQuery's SELECT statement. Students who do not know SQL should be able to follow simple queries and write their own by modifying the examples given.

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.

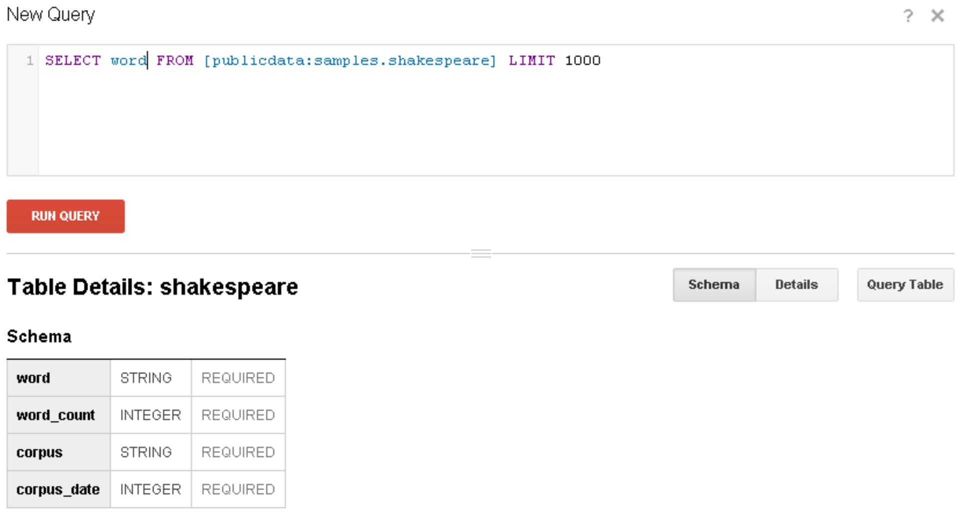
- *What makes this a big database?*
- *How many rows are in the Shakespeare database?*

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



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

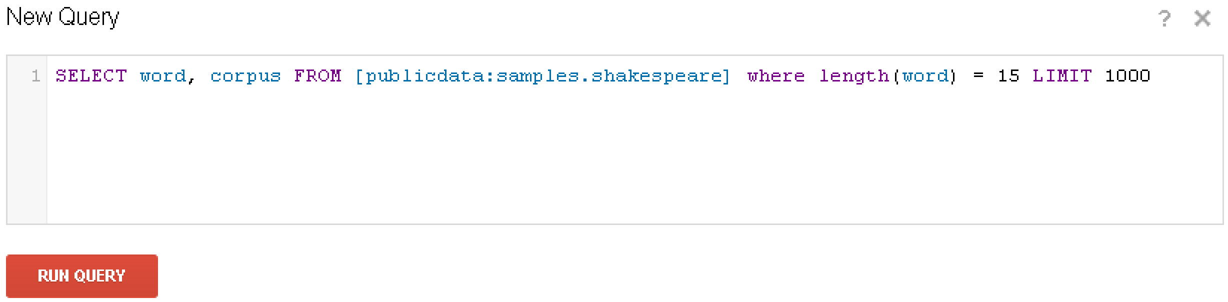
Table Details: shakespeare

Schema

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

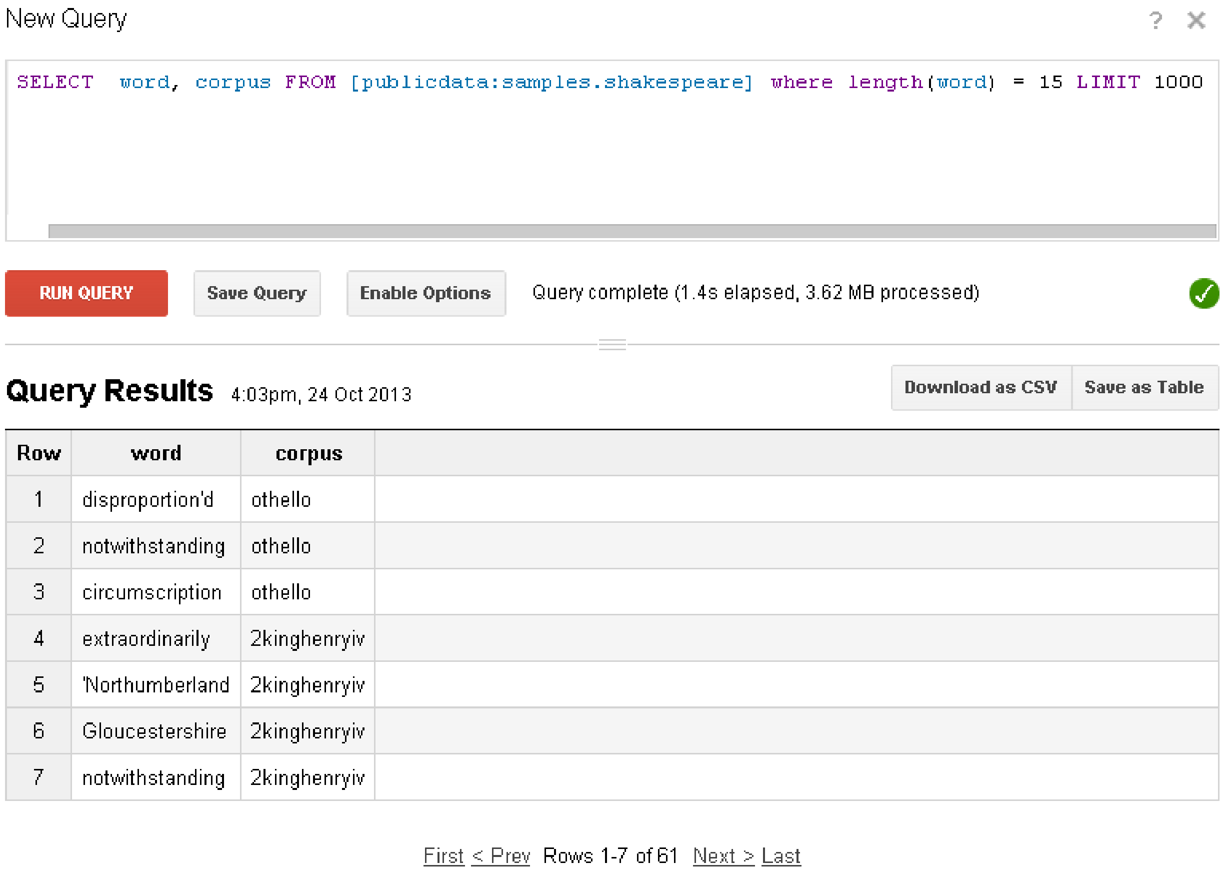
Click on the word corpus in the schema to add it to the query as well

Finish typing the rest of the command as shown:



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

Click the Run Query Button to run the query.



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

Query complete (1.4s elapsed, 3.62 MB processed) ✓

Query Results 4:03pm, 24 Oct 2013

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 15-letter word appears in Shakespeare’s King Henry VIII?

What 17-letter word appears in A Midsummer Night’s Dream?

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
```

Natality (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/stnlst.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?

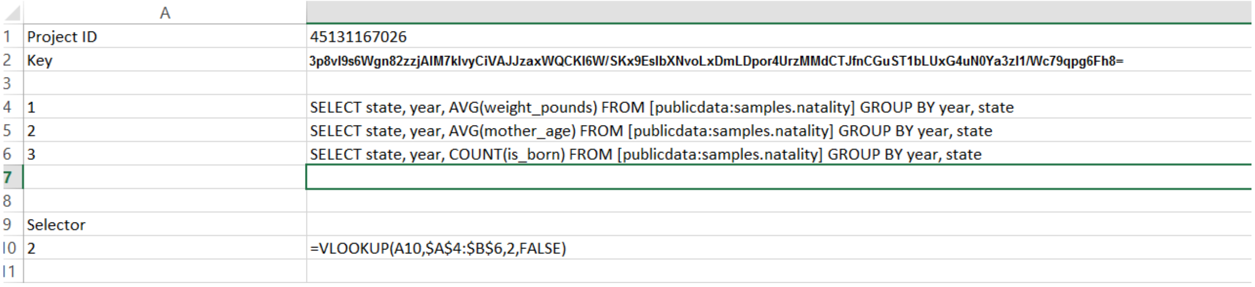
Activity 3: Creating Visualizations of BigQuery Data using Microsoft Excel

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, sparklines).

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

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.



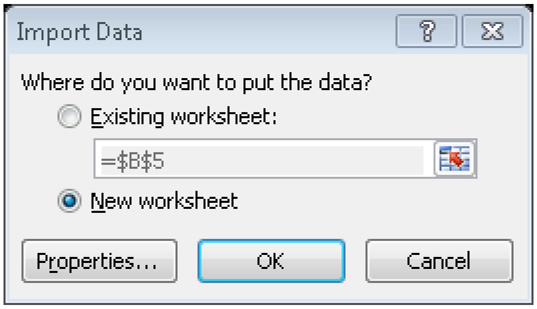
	A	
1	Project ID	45131167026
2	Key	3p8vl9s6Wgn82zzjAIM7klvyCiVAJJzaxWQCKI6W/SKx9EslbXNvoLxDmLDpor4UrzMMdCTJfnCGuST1bLUxG4uN0Ya3zI1/Wc79qpg6Fh8=
3		
4	1	SELECT state, year, AVG(weight_pounds) FROM [publicdata:samples.natality] GROUP BY year, state
5	2	SELECT state, year, AVG(mother_age) FROM [publicdata:samples.natality] 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.



Import Data

Where do you want to put the data?

○ Existing worksheet:

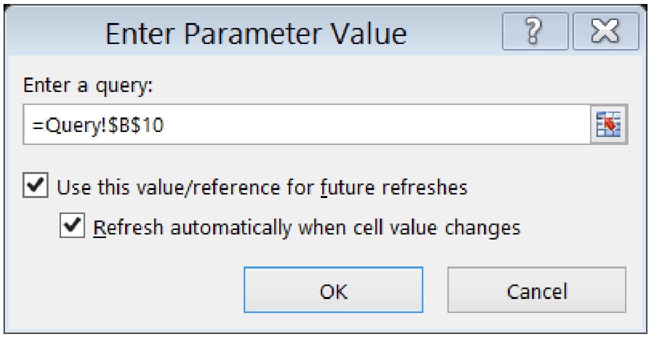
=$B$5

● New worksheet

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 a query:

=Query!$B$10

☑ Use this value/reference for future refreshes

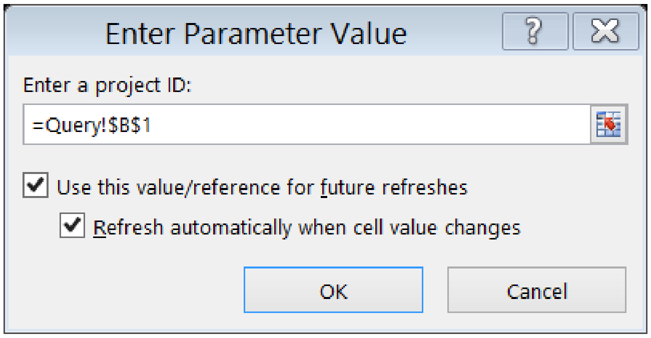
☑ Refresh automatically when cell value changes

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.

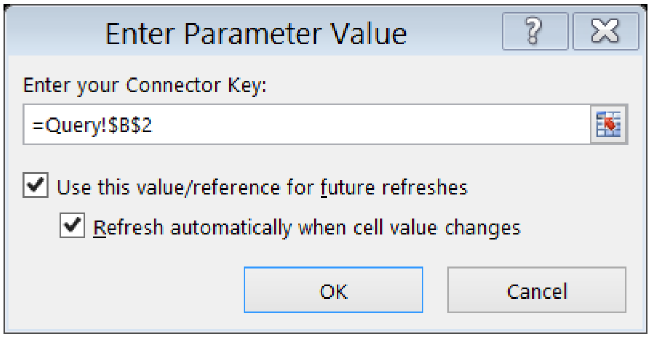


Enter a project ID:

=Query!$B$1

☑ Use this value/reference for future refreshes

☑ Refresh automatically when cell value changes



Enter your Connector Key:

=Query!$B$2

☑ Use this value/reference for future refreshes

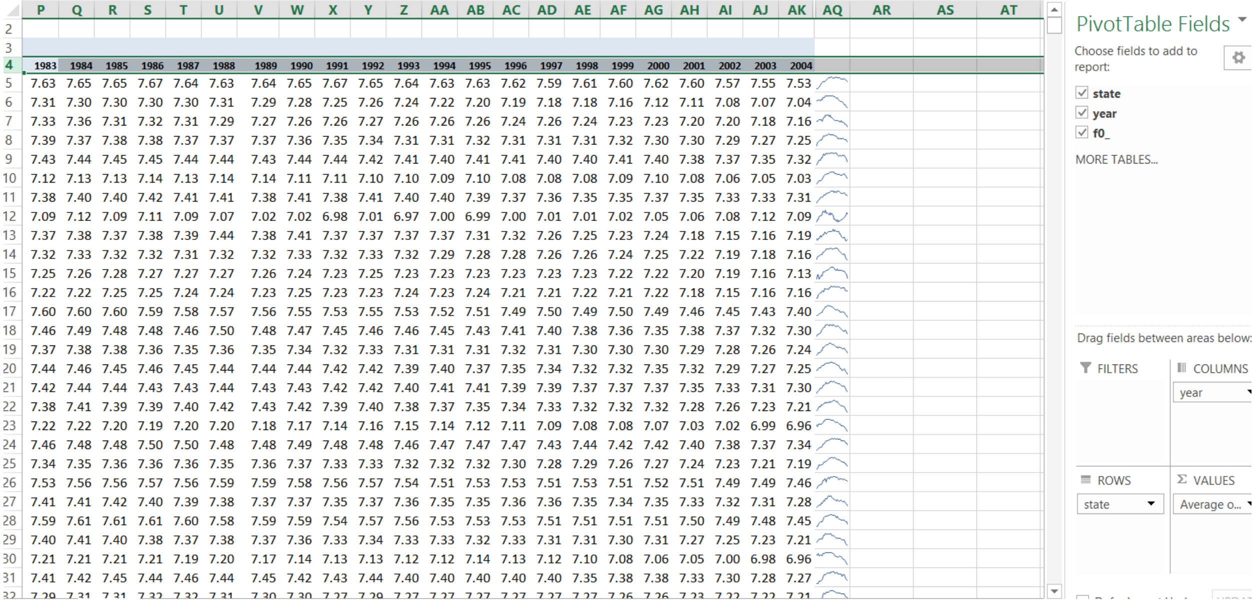
☑ Refresh automatically when cell value changes

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.



	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AQ	AR	AS	AT
2																										
3																										
4	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004				
5	7.63	7.65	7.65	7.67	7.64	7.63	7.64	7.65	7.67	7.65	7.64	7.63	7.63	7.62	7.59	7.61	7.60	7.62	7.60	7.57	7.55	7.53				
6	7.31	7.30	7.30	7.30	7.30	7.31	7.29	7.28	7.25	7.26	7.24	7.22	7.20	7.19	7.18	7.18	7.16	7.12	7.11	7.08	7.07	7.04				
7	7.33	7.36	7.31	7.32	7.31	7.29	7.27	7.26	7.26	7.27	7.26	7.26	7.26	7.24	7.26	7.24	7.23	7.23	7.20	7.20	7.18	7.16				
8	7.39	7.37	7.38	7.38	7.37	7.37	7.37	7.36	7.35	7.34	7.31	7.31	7.32	7.31	7.31	7.31	7.32	7.30	7.30	7.29	7.27	7.25				
9	7.43	7.44	7.45	7.45	7.44	7.44	7.43	7.44	7.44	7.42	7.41	7.40	7.41	7.41	7.40	7.40	7.41	7.40	7.38	7.37	7.35	7.32				
10	7.12	7.13	7.13	7.14	7.13	7.14	7.14	7.11	7.11	7.10	7.10	7.09	7.10	7.08	7.08	7.08	7.09	7.10	7.08	7.06	7.05	7.03				
11	7.38	7.40	7.40	7.42	7.41	7.41	7.38	7.41	7.38	7.41	7.40	7.40	7.39	7.37	7.36	7.35	7.35	7.37	7.35	7.33	7.33	7.31				
12	7.09	7.12	7.09	7.11	7.09	7.07	7.02	7.02	6.98	7.01	6.97	7.00	6.99	7.00	7.01	7.01	7.02	7.05	7.06	7.08	7.12	7.09				
13	7.37	7.38	7.37	7.38	7.39	7.44	7.38	7.41	7.37	7.37	7.37	7.37	7.31	7.32	7.26	7.25	7.23	7.24	7.18	7.15	7.16	7.19				
14	7.32	7.33	7.32	7.32	7.31	7.32	7.32	7.33	7.32	7.33	7.32	7.29	7.28	7.28	7.26	7.26	7.24	7.25	7.22	7.19	7.18	7.16				
15	7.25	7.26	7.28	7.27	7.27	7.27	7.26	7.24	7.23	7.25	7.23	7.23	7.23	7.23	7.23	7.23	7.22	7.22	7.20	7.19	7.16	7.13				
16	7.22	7.22	7.25	7.25	7.24	7.24	7.23	7.25	7.23	7.23	7.24	7.23	7.24	7.21	7.21	7.22	7.21	7.22	7.18	7.15	7.16	7.16				
17	7.60	7.60	7.60	7.59	7.58	7.57	7.56	7.55	7.53	7.55	7.53	7.52	7.51	7.49	7.50	7.49	7.50	7.49	7.46	7.45	7.43	7.40				
18	7.46	7.49	7.48	7.48	7.46	7.50	7.48	7.47	7.45	7.46	7.46	7.45	7.43	7.41	7.40	7.38	7.36	7.35	7.38	7.37	7.32	7.30				
19	7.37	7.38	7.38	7.36	7.35	7.36	7.35	7.34	7.32	7.33	7.31	7.31	7.31	7.32	7.31	7.30	7.30	7.30	7.29	7.28	7.26	7.24				
20	7.44	7.46	7.45	7.46	7.45	7.44	7.44	7.44	7.42	7.42	7.39	7.40	7.37	7.35	7.34	7.32	7.32	7.35	7.32	7.29	7.27	7.25				
21	7.42	7.44	7.44	7.43	7.43	7.44	7.43	7.43	7.42	7.42	7.40	7.41	7.41	7.39	7.39	7.37	7.37	7.37	7.35	7.33	7.31	7.30				
22	7.38	7.41	7.39	7.39	7.40	7.42	7.43	7.42	7.39	7.40	7.38	7.37	7.35	7.34	7.33	7.32	7.32	7.32	7.28	7.26	7.23	7.21				
23	7.22	7.22	7.20	7.19	7.20	7.20	7.18	7.17	7.14	7.16	7.15	7.14	7.12	7.11	7.09	7.08	7.08	7.07	7.03	7.02	6.99	6.96				
24	7.46	7.48	7.48	7.50	7.50	7.48	7.48	7.49	7.48	7.48	7.46	7.47	7.47	7.47	7.43	7.44	7.42	7.42	7.40	7.38	7.37	7.34				
25	7.34	7.35	7.36	7.36	7.36	7.35	7.36	7.37	7.33	7.33	7.32	7.32	7.32	7.30	7.28	7.29	7.26	7.27	7.24	7.23	7.21	7.19				
26	7.53	7.56	7.56	7.57	7.56	7.59	7.59	7.58	7.56	7.57	7.54	7.51	7.53	7.53	7.51	7.53	7.51	7.52	7.51	7.49	7.49	7.46				
27	7.41	7.41	7.42	7.40	7.39	7.38	7.37	7.37	7.35	7.37	7.36	7.35	7.35	7.36	7.36	7.35	7.34	7.35	7.33	7.32	7.31	7.28				
28	7.59	7.61	7.61	7.61	7.60	7.58	7.59	7.59	7.54	7.57	7.56	7.53	7.53	7.53	7.51	7.51	7.51	7.51	7.50	7.49	7.48	7.45				
29	7.40	7.41	7.40	7.38	7.37	7.38	7.37	7.36	7.33	7.34	7.33	7.33	7.32	7.33	7.31	7.31	7.30	7.31	7.27	7.25	7.23	7.21				
30	7.21	7.21	7.21	7.21	7.19	7.20	7.17	7.14	7.13	7.13	7.12	7.12	7.14	7.13	7.12	7.10	7.08	7.06	7.05	7.00	6.98	6.96				
31	7.41	7.42	7.45	7.44	7.46	7.44	7.45	7.42	7.43	7.44	7.40	7.40	7.40	7.40	7.40	7.35	7.38	7.38	7.33	7.30	7.28	7.27				
32	7.29	7.31	7.31	7.32	7.32	7.31	7.30	7.30	7.27	7.29	7.27	7.27	7.27	7.27	7.27	7.27	7.26	7.26	7.23	7.22	7.22	7.21				

PivotTable Fields

Choose fields to add to report:

☑ state
☑ year
☑ f0_

Drag fields between areas below:

FILTERS

COLUMNS: year

ROWS: state

VALUES: Average o...

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?

Activity 4. Analyze Big Data using BigQuery with Google Spreadsheets

You also can use Google BigQuery to update Google Spreadsheets with data from Google BigQuery, and then make use of the visualization features of Google Spreadsheets to create visualizations.

Follow the steps in the tutorial from Michael Manoochehri from the Google BigQuery team. This tutorial takes about 20 minutes to complete.

https://developers.google.com/apps-script/articles/bigquery_tutorial

If you complete this tutorial, compare using BigQuery with the connector vs using BigQuery in Google Spreadsheets?