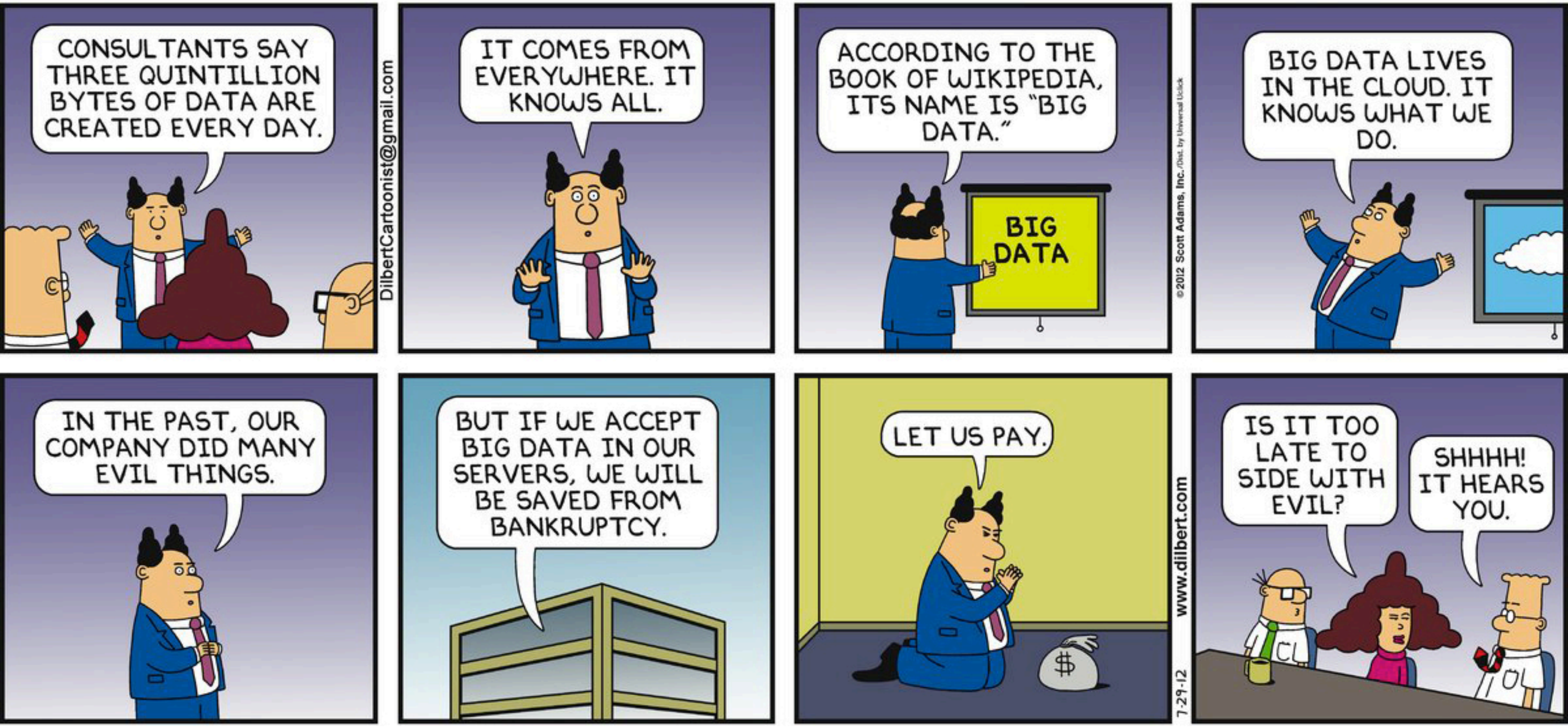
INTRODUCTION TO GOOGLE BIG QUERY

PHILADELPHIA MEDIA NETWORK

The Inquirer DAILY NEWS philly.com

Ramya Roopini Ravi
Data Engineer
Analytics @ PMN


CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
© 2012 Scott Adams, Inc. /Dist. by Universal Uclick
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
$
www.dilbert.com
7-29-12
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.

DO WE HAVE ANY ACTIONABLE ANALYTICS FROM OUR BIG DATA IN THE CLOUD?
Dilbert.com DilbertCartoonist@gmail.com
YES, THE DATA SHOWS THAT MY PRODUCTIVITY PLUNGES WHENEVER YOU LEARN NEW JARGON.
1-9-13 © 2013 Scott Adams, Inc./Dist. by Universal Uclick
MAYBE IN–MEMORY COMPUTING WILL ACCEL– ERATE YOUR APPLICA– TIONS.
PLUNGE, PLUNGE, PLUNGE.

SQL

- SQL – Relational data
- Control fundamental business tasks
- INSERT, DELETE & UPDATE Transactions
- Have schema on write
- Operational tasks – Online Transaction Processing
- Normalized with many tables which means Data is not duplicated, and use Joins to get data from multiple tables

- Databases: SQL Server, MYSQL (RDBMS)
- https://www3.nd.edu/~oss/Data/ER-DiagramSourceForge_1page.pdf

NoSQL

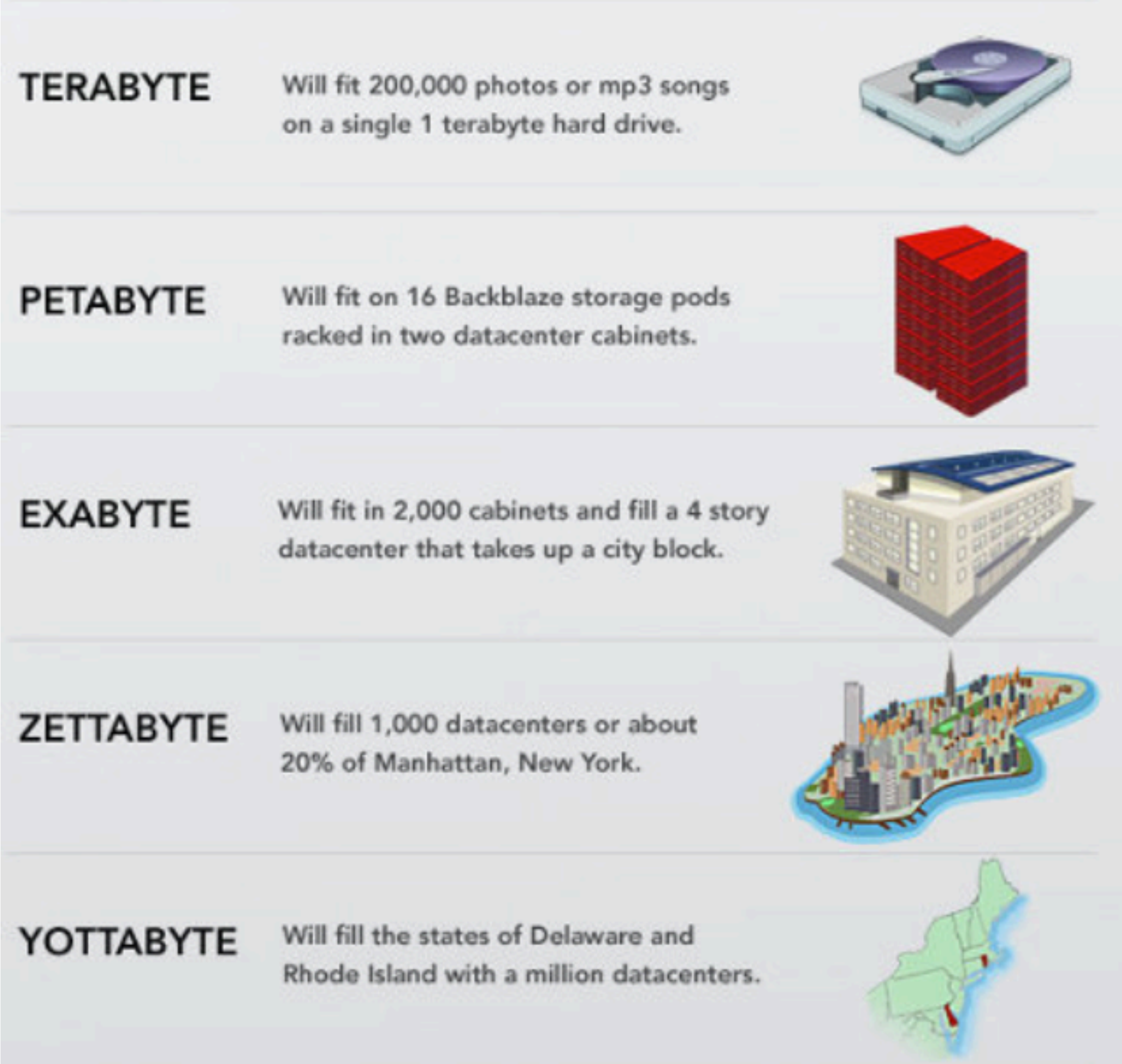
- History – Big players (AWS / Google)
- NoSQL word originated from a hashtag after a meetup in SFO
- It means not relational database
- Help with planning, problem solving and decision support
- Perform Data Mining techniques
- Online analytical processing system
- De-normalized and few tables
- Data comes from several sources as well as OLTP data

- Ex: Dynamo DB, Big Table, Graph DB, Mongo DB

Size of the Digital Universe

- Bit
- Byte
- Kilobyte
- Megabyte
- Gigabyte
- Terabyte
- Petabyte
- Etc.

- Do you know the largest data size and who uses it?



TERABYTE	Will fit 200,000 photos or mp3 songs on a single 1 terabyte hard drive.
PETABYTE	Will fit on 16 Backblaze storage pods racked in two datacenter cabinets.
EXABYTE	Will fit in 2,000 cabinets and fill a 4 story datacenter that takes up a city block.
ZETTABYTE	Will fill 1,000 datacenters or about 20% of Manhattan, New York.
YOTTABYTE	Will fill the states of Delaware and Rhode Island with a million datacenters.

So what is Big Data?

Big Data characteristics

- Volume
- Veracity
- Velocity
- Variety
- Cloud

Cloud Storage

- Low maintenance
- Distributed
- Power outage doesn’t matter
- Always On
- Fast
- Nothing to configure
- Replication

Features of Cloud by some providers:

- Edge Caching
- Robust

Google Cloud Storage

Components:

- Project
- Buckets
- Objects (Object itself and Metadata)

Access to Google Cloud Storage SDK:

- Terminal/ Command line

Steps:

- Go to cloud.google.com on Python 2.7x or higher (Better to have 2.7 version and have an environment set for Python 3.5
- Install the SDK: https://cloud.google.com/sdk/
- ON Mac - Use Terminal and paste: curl https://sdk.cloud.google.com | bash

- Go to https://console.cloud.google.com
- Check for the free bucket under Resources
- Start to upload files into the bucket by clicking on create transfer
  - Select source as TSV files (Public accessible object URL's)
  - Select destination bucket

Big Query

- Google Big Query : Easily store and analyze big data in Google's infrastructure
- Query to mine massive datasets in seconds
- Simple loading and exporting
- Big Query is not a database
- Does not support joins like SQL. It supports Joins if one table is very small compared to the other

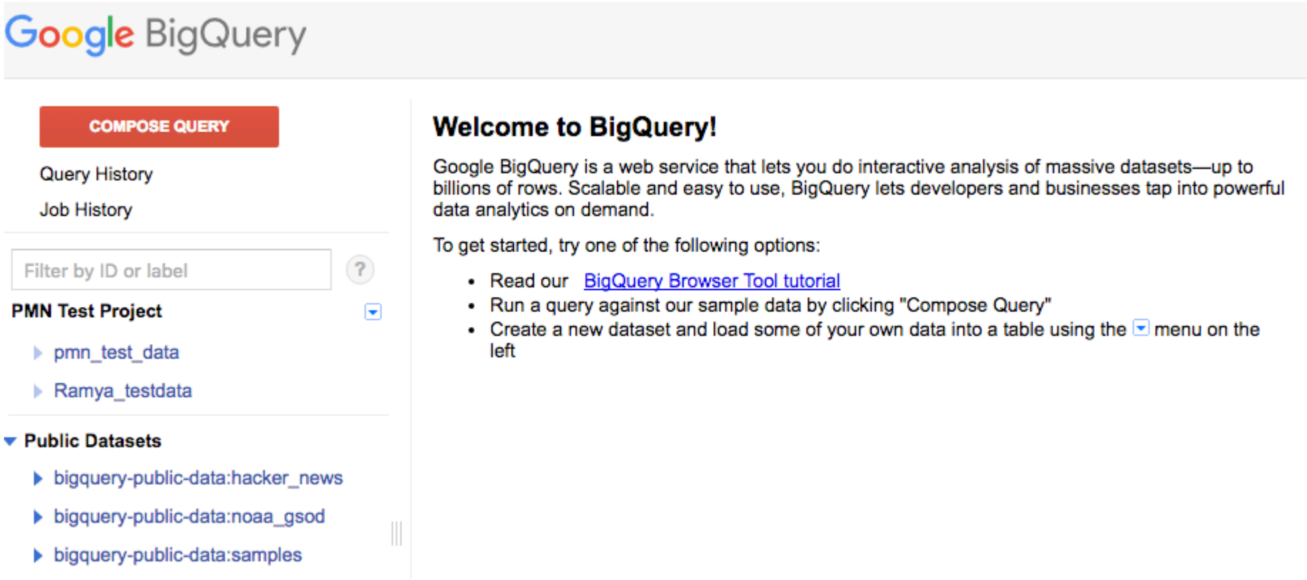
Querying tools

- Browser
- Command line tool - bq
- REST api - can be tapped from virtually anywhere.
- Excel connector

Write Big Queries using the console

Create an account on Google and sign up for Google cloud

Go to the Big Query Cloud url



Welcome to 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.

To get started, try one of the following options:

- Read our BigQuery Browser Tool tutorial
- Run a query against our sample data by clicking "Compose Query"
- Create a new dataset and load some of your own data into a table using the ▾ menu on the left

Writing Queries

- Query NOAA public dataset
- http://www1.ncdc.noaa.gov/pub/data/gsod/readme.txt
- https://bigquery.cloud.google.com/dataset/bigquery-public-data:noaa_gsod

- On the Query explorer, we will start writing queries to perform some operations on the data.

Demo

Find the Station, Temperature and dew point for the year 2000 when there was hail.

```
SELECT station_number,mean_temp,mean_dew_point,hail
        FROM [bigquery-public-data:samples.gsod]
        WHERE year = 2000 AND
        hail = true
```

Return the average temperatures by month and year for the NOAA dataset and Sort by year

```
SELECT year, month, AVG(mean_temp) AS avgtemp
        FROM [bigquery-public-data:samples.gsod]
        GROUP BY year, month, ORDER BY YEAR DESC
```

```
SELECT year, month, AVG(mean_temp) AS avgtemp
        FROM [bigquery-public-data:samples.gsod]
        GROUP BY year, month, ORDER BY YEAR
```

Writing Queries with Python using Pandas and bq

- Install python 2.7
- From terminal, type
  - bash Anaconda2-4.1.1-MacOSX-x86_64.sh   - Install Anaconda for scientific computing
  - curl https://sdk.cloud.google.com | bash
  - pip install bq - Install bq connector for Python
  - pip install google-python-api-client
  - Ipython notebook

Demo