Fundamentals of Big Data

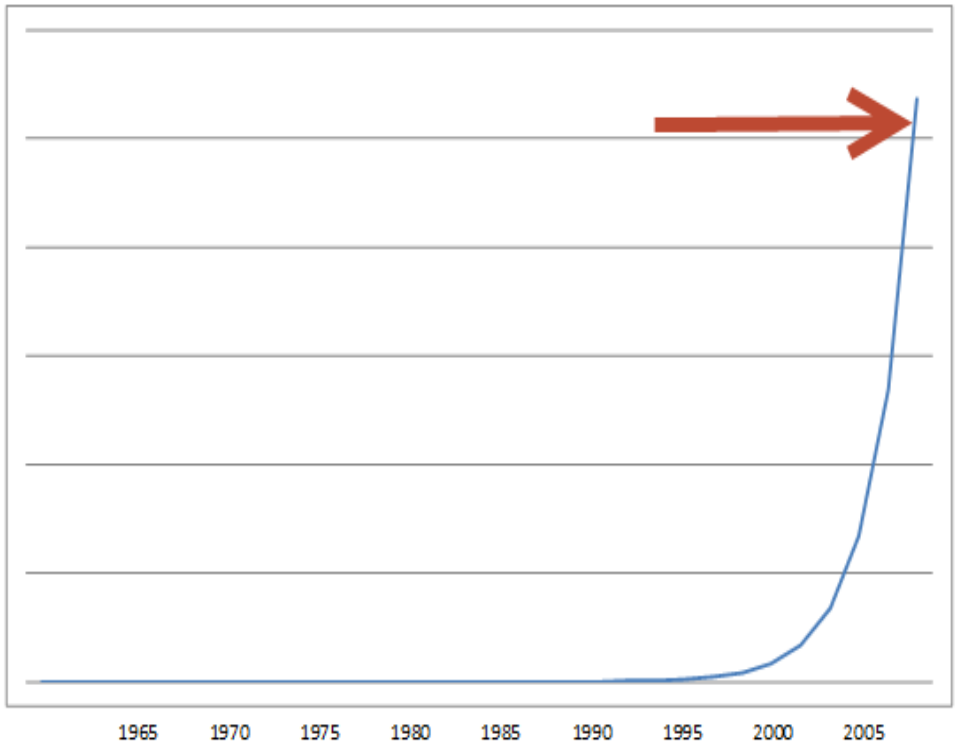
Mark DuBois

How did we get here?

- Moore's Law has had a major impact
  - But we may not really understand
- Consider a month of doubling your income (start with 1 penny)
  - Day 30 = $5 million, day 31…

Accelerator

- Moore's Law - Computer **processing** power doubles every 18 months



1.5 years
30 cycles
1967

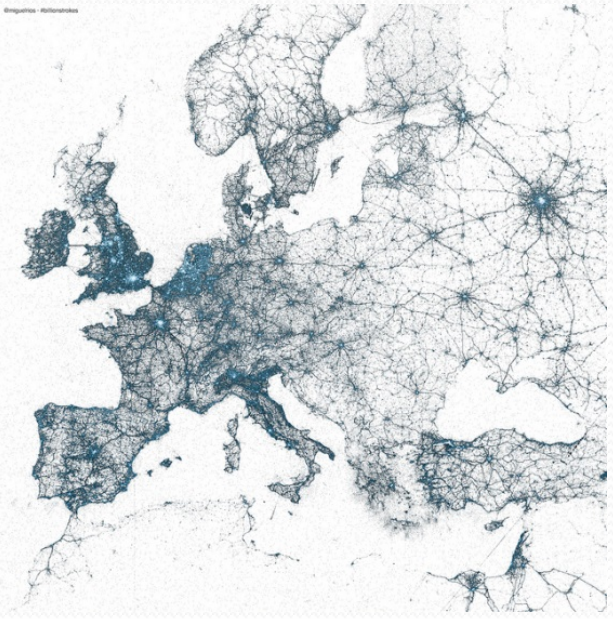
Accelerators

- **Bandwidth** - increasing even faster than Moore's Law
- **Storage** - doubling every 12 months (faster than processor and bandwidth)

- “In the years ahead, there will be only two kinds of people: those who see the waters receding as the giant prepares to blow - and those who don’t. Those who don’t will experience massive chaos and dislocation. Those who do will find **unprecedented opportunity**.” - Daniel Burrus [Flash Foresight book]

What is…

- Big data
  - Often involves the data we generate ourselves
  - Think digital "breadcrumbs" we drop
  - Consider Twitter - http://markdubois.me/TwitterGeog
  - Geotagged tweets
    - Since 2009



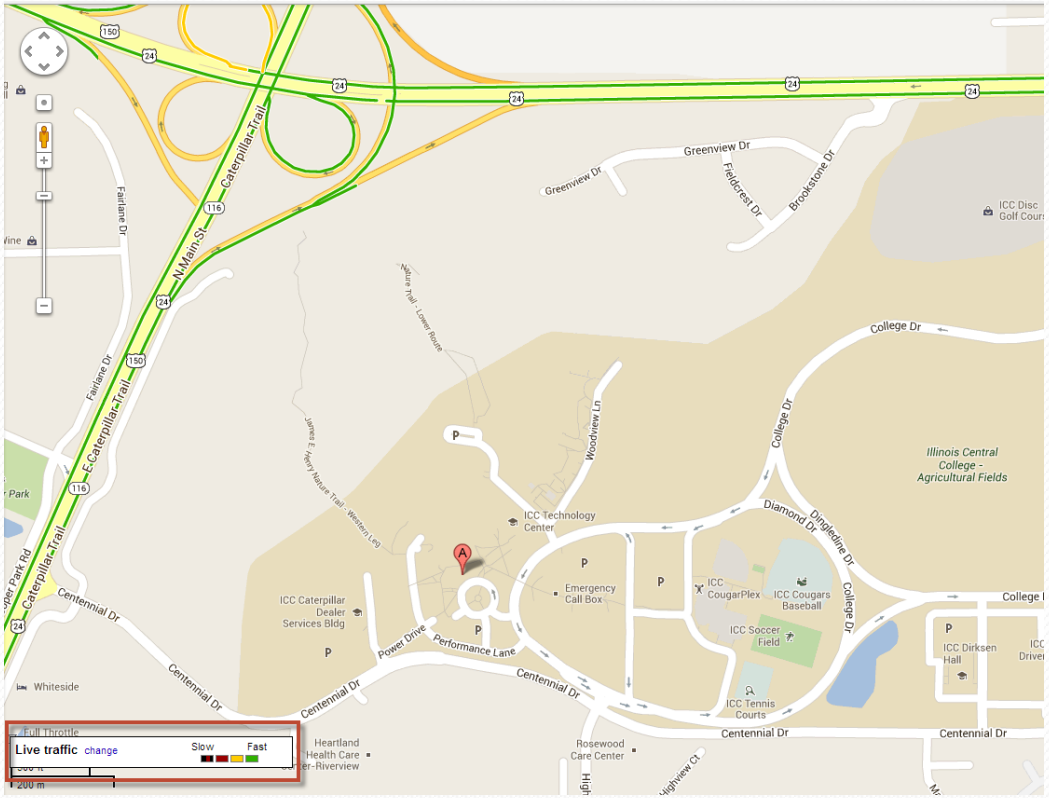
Anecdote

- Think of 1 – 5 drawer filing cabinet
- Now think of a room filled with 60 million of these
- That is how much data comes from all Wal-Mart stores every hour
  - That is “big data”

- Source: http://markdubois.me/BigData101

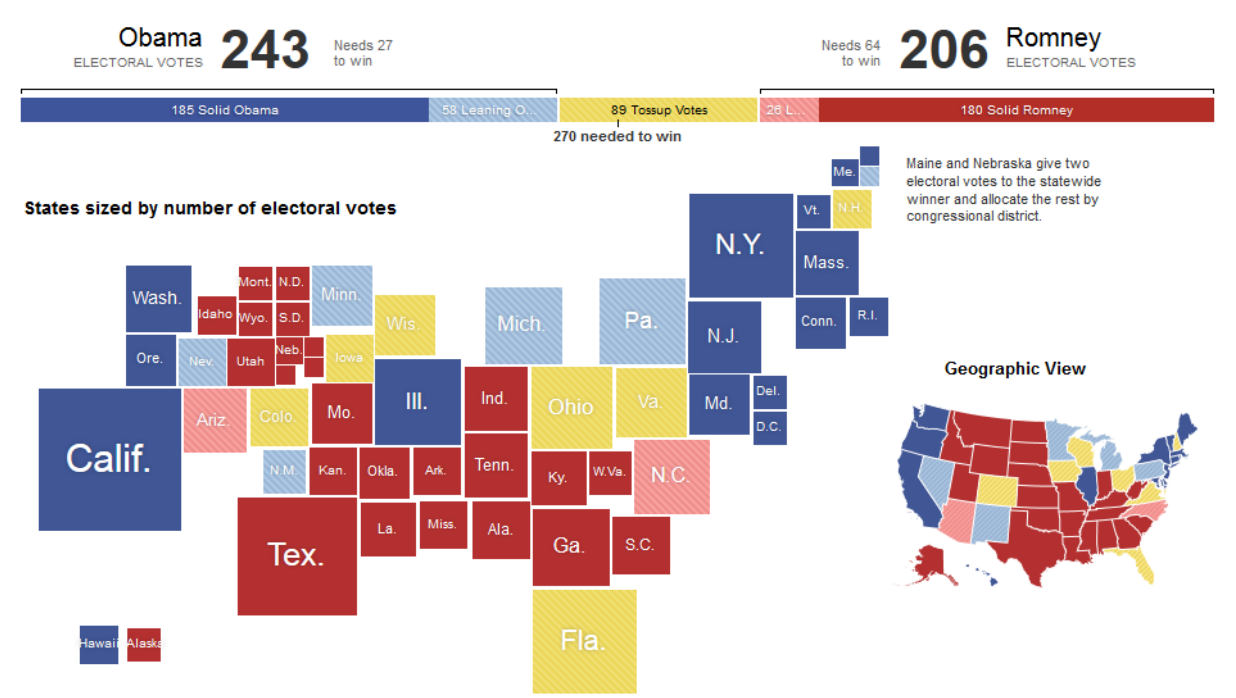
Another example

- Google maps with “traffic” generated from Droid phones – http://maps.google.com



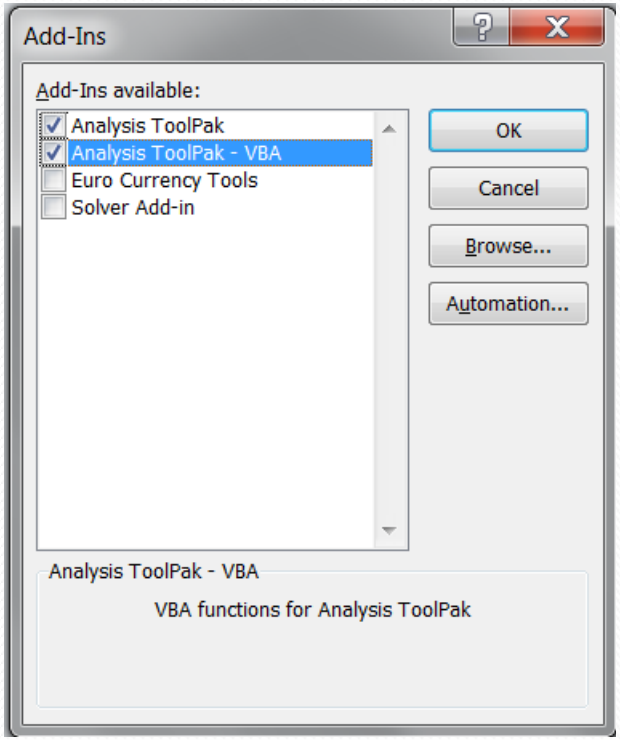
Predictive Analytics

- Nate Silver and 2012 US Presidential election
- http://fivethirtyeight.blogs.nytimes.com/

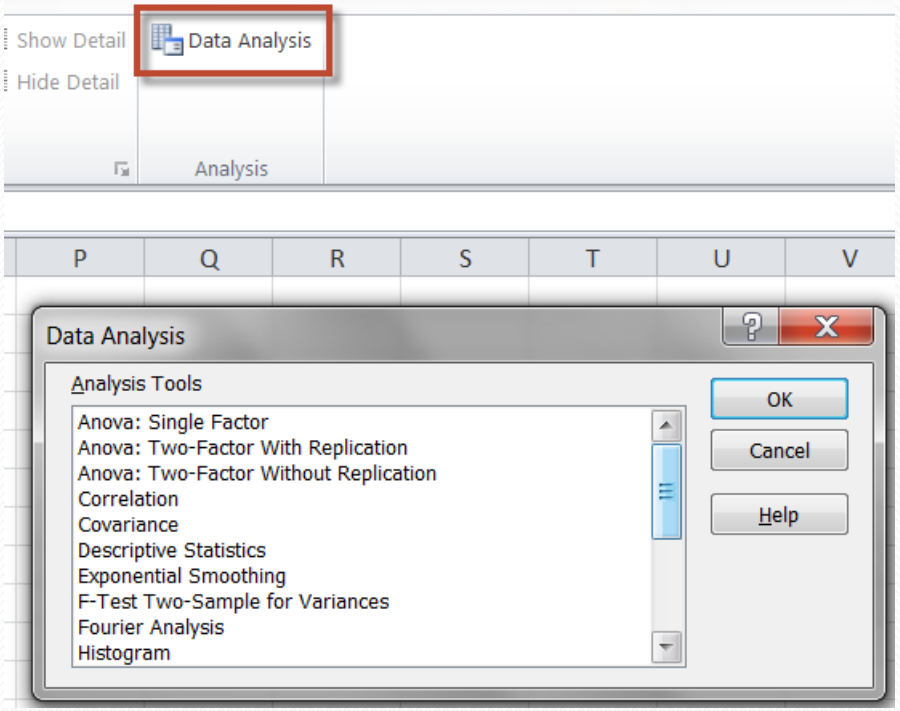


More tools than we think

- Excel Analysis toolpak

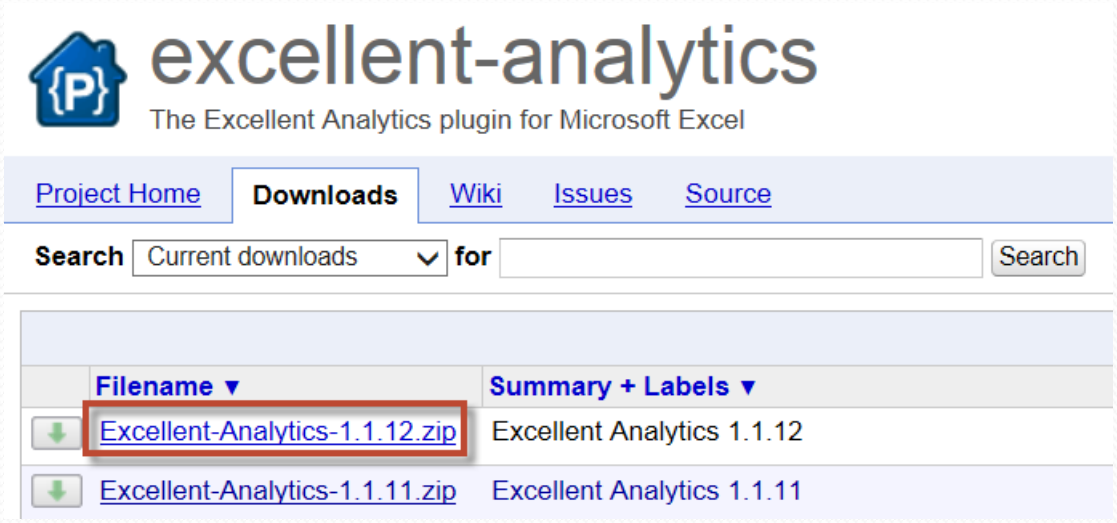


Results (data tab)



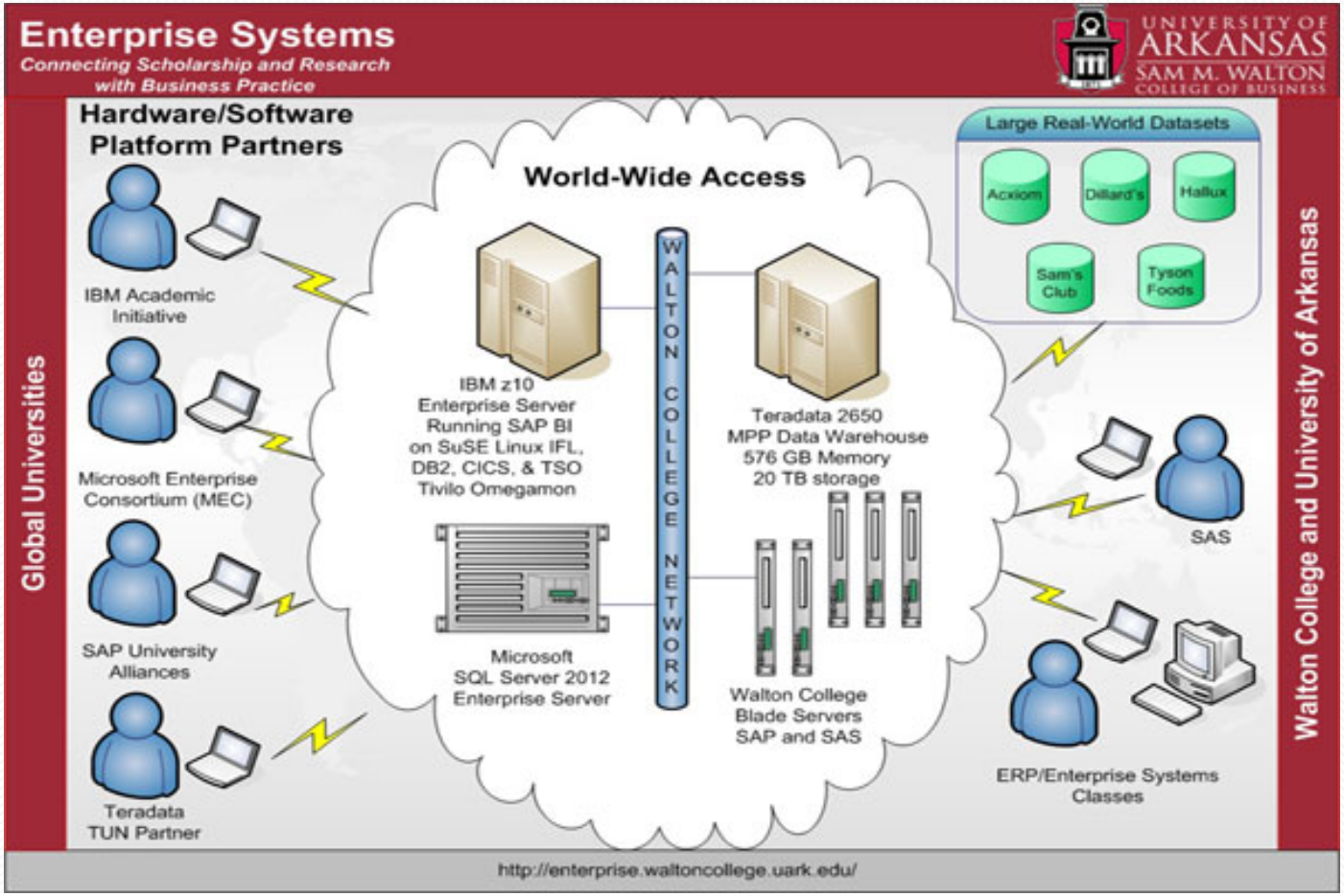
More Excel Tools

- Microsoft
- Plugin for Analytics - http://markdubois.me/GAExcel
- (pull Google Analytics into Excel)
- .msi file



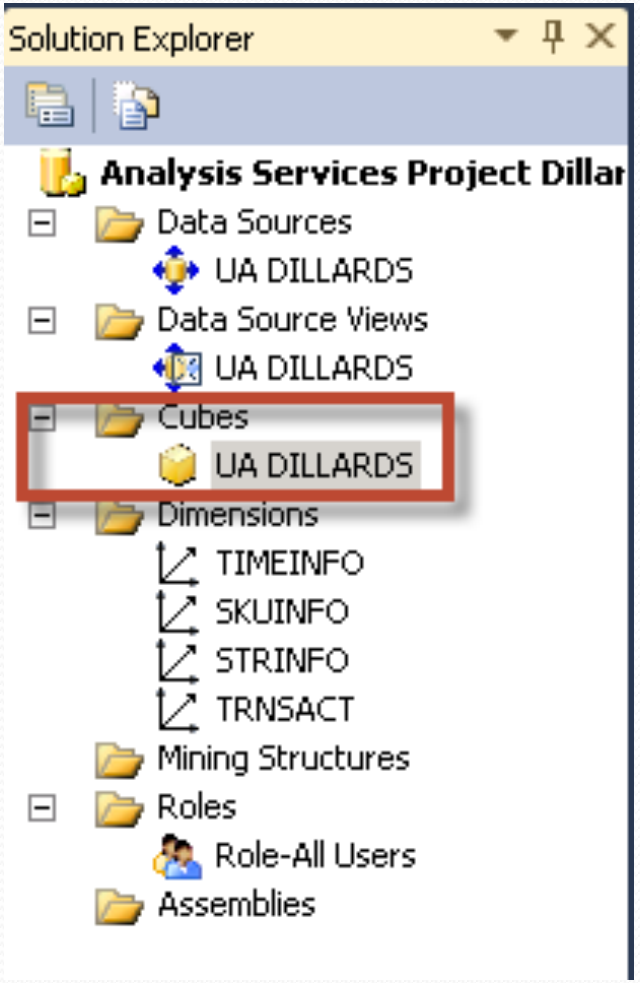
Large amounts of data

- University of Arkansas – Walton School of Business



Visual Studio Interface

- Running SQL 2010 tools
- Windows 2008 R2 Server
  - Define Data Source
  - Define Source View
  - Configure SQL Cube



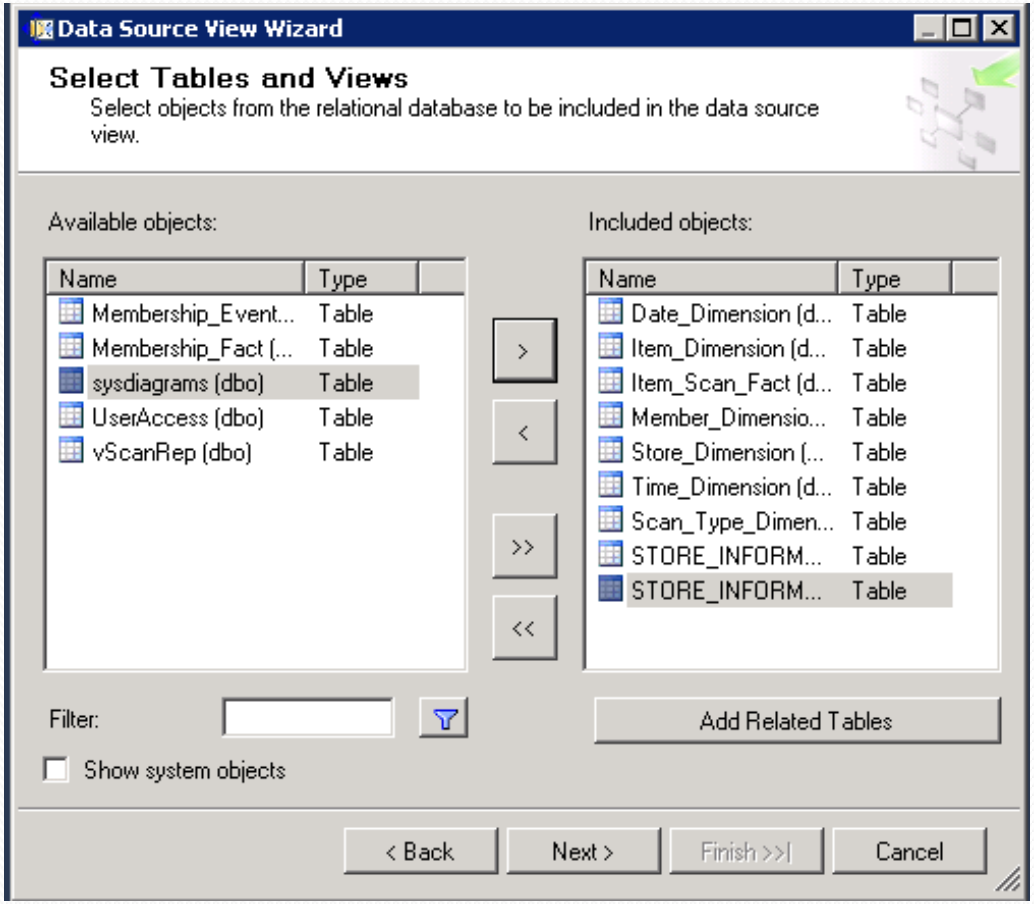
SQL What?

- Cube – multidimensional structure
  - Dimensions – define structure of the cube
  - Measures – numerical values of interest

  - Similar to a pivot table in Excel

Define data source

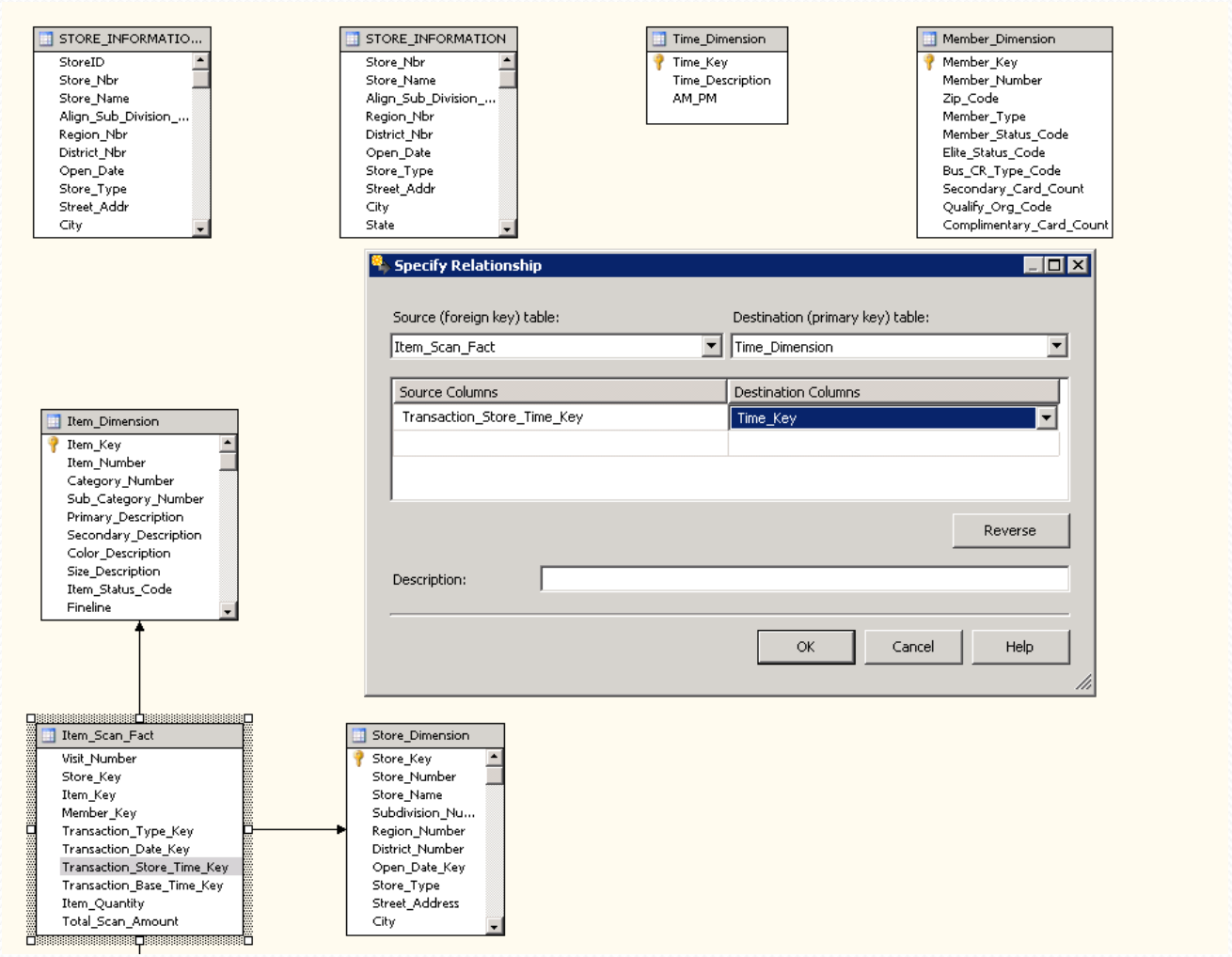
- Run wizard and authenticate



Cube Wizard

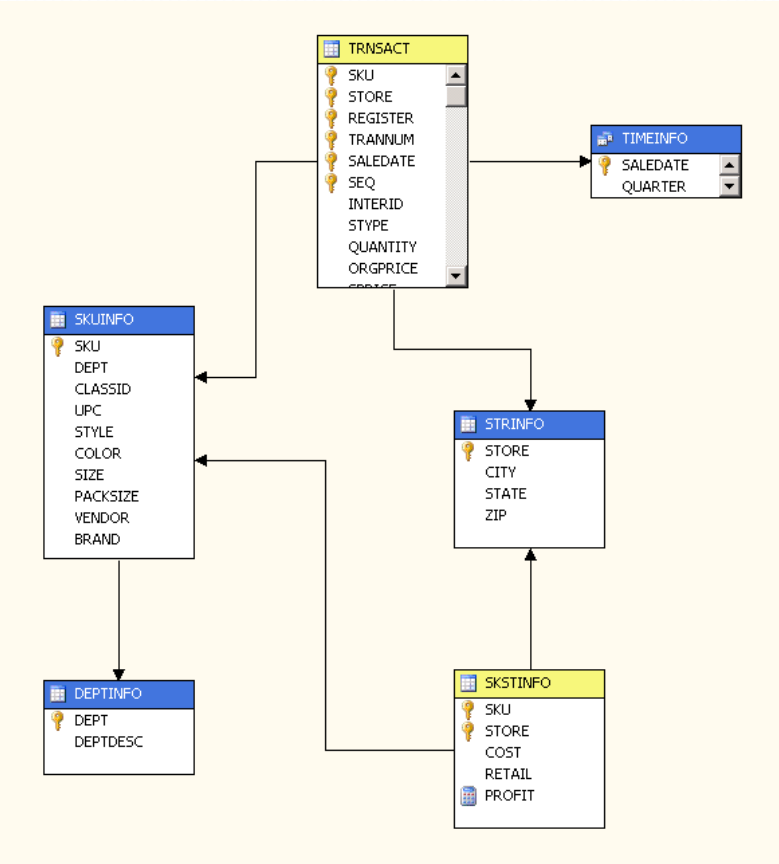

Cube Wizard
Completing the Wizard
Name the cube, review its structure, and then click Finish to save the cube.
Cube name:
Sams Club Cube
Preview:
Measure groups
Item Scan Fact
Item Quantity
Total Scan Amount
Item Scan Fact Count
Dimensions
Scan Type Dimension
Store Dimension
Item Dimension
Time Dimension
Date Dimension
Member Dimension
< Back
Next >
Finish
Cancel

Establish relationships



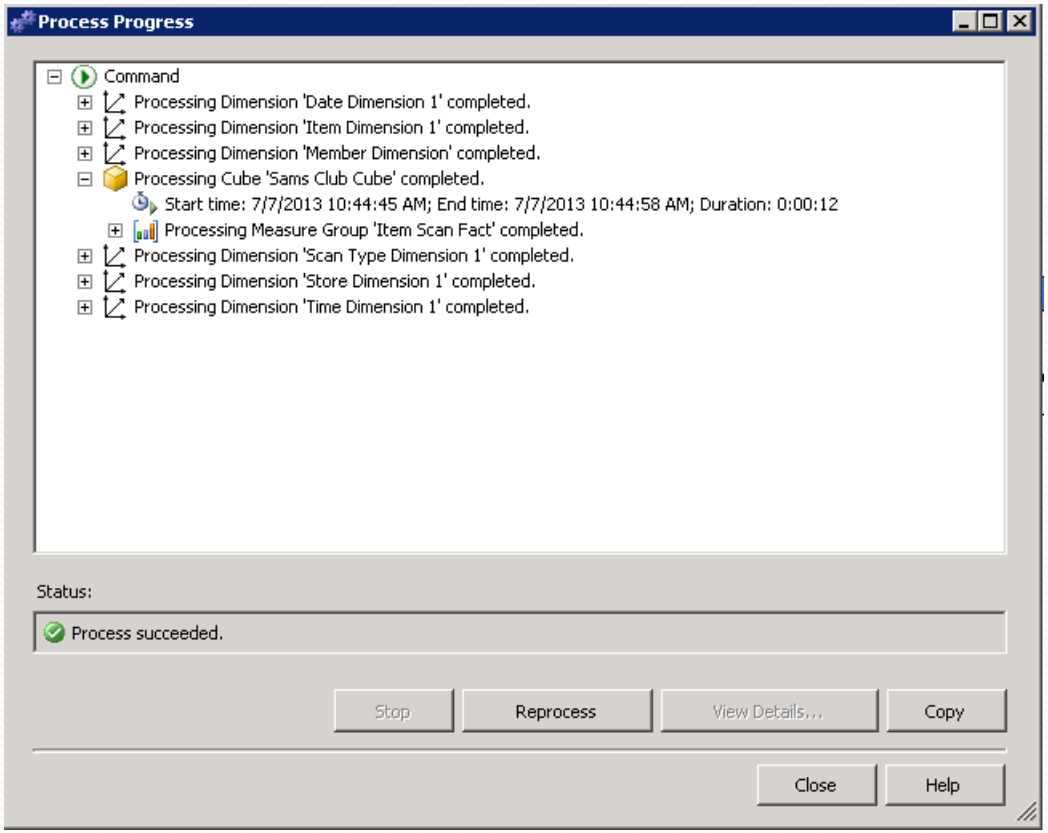
Dillard's Store Data

- Example of SQL Cube



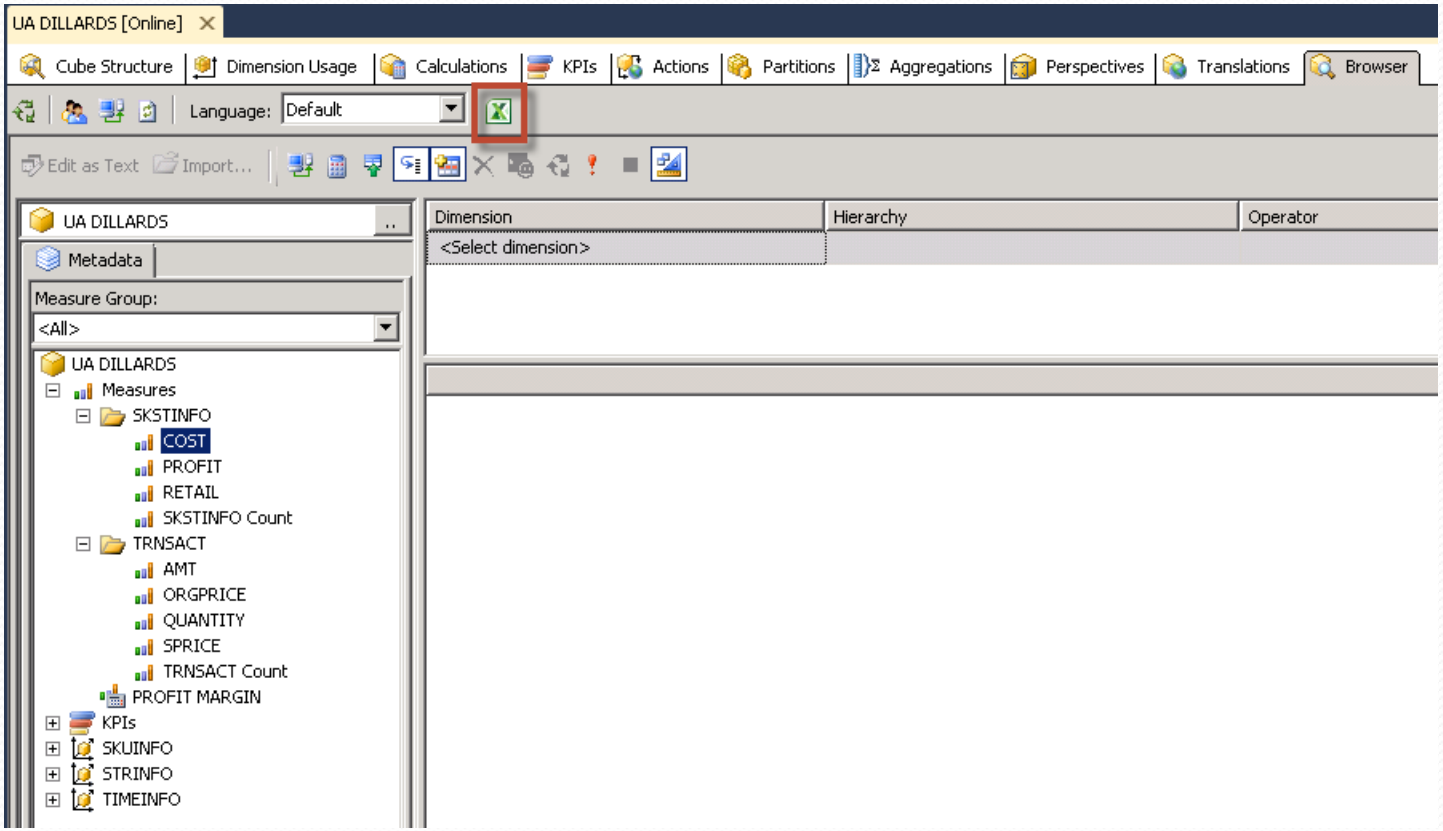
Process

- Once cube is created, must process (time consuming)



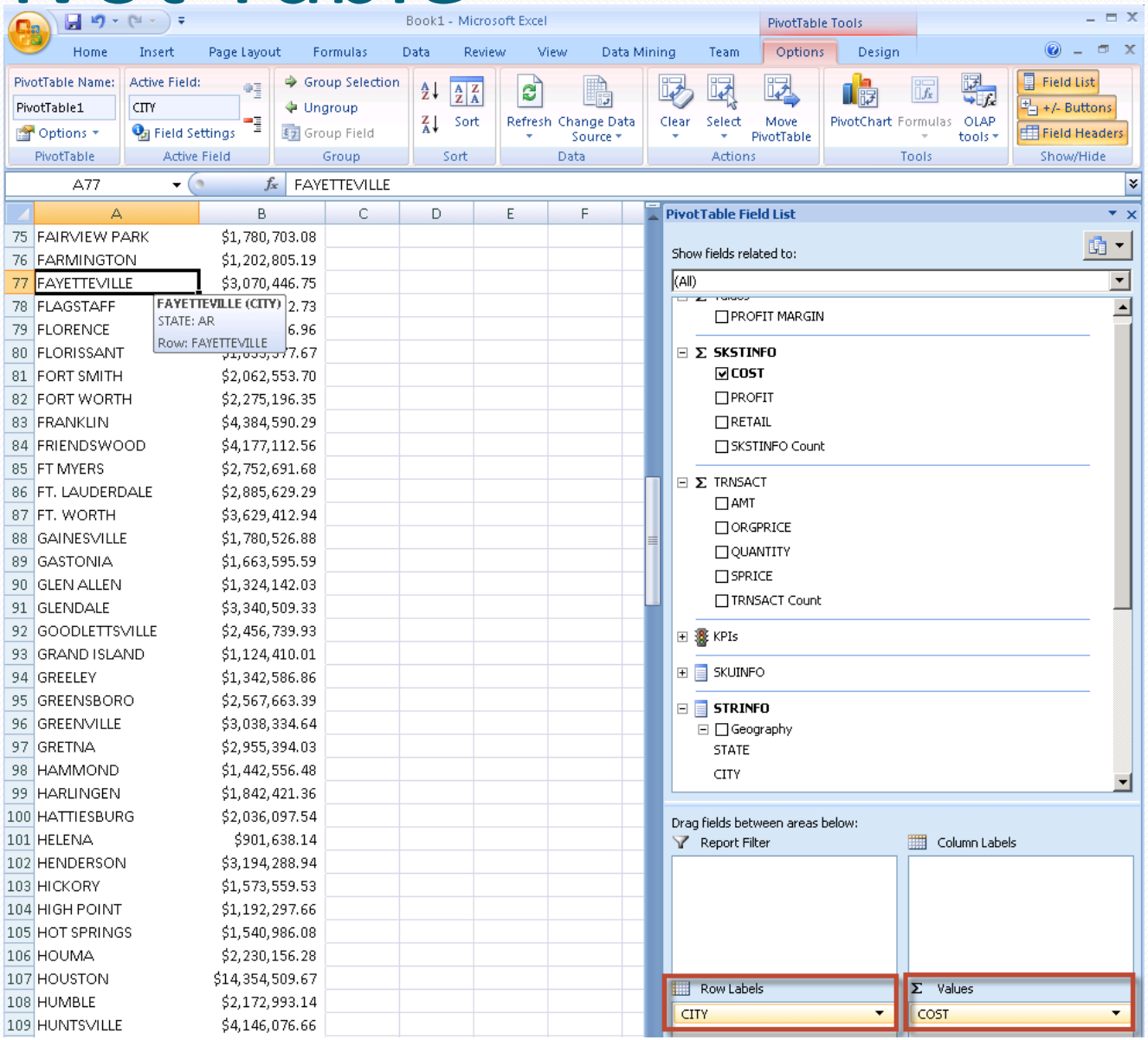
Using Dillard's Cube

- Open data in Excel



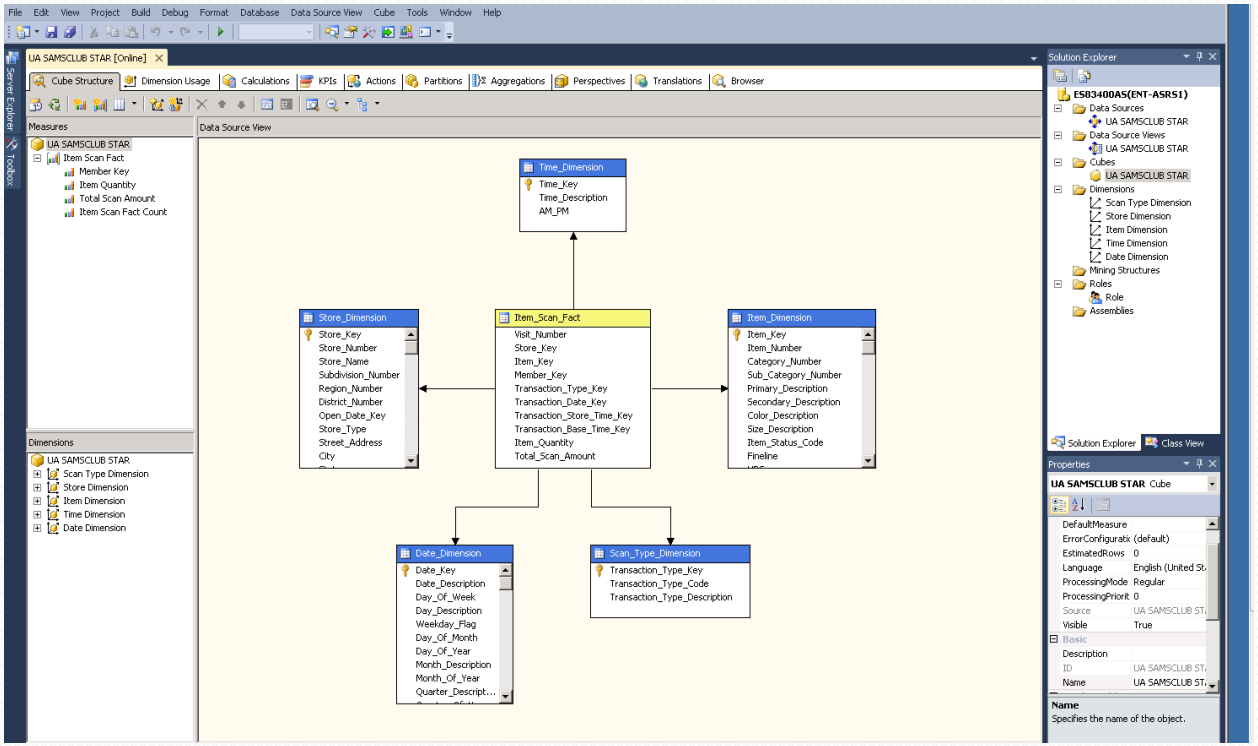
Excel Pivot Table

- For further analysis
- Table is tied to data stores



SQL Cubes

- Sam's Club data (one table 120 million+ rows)



Brief Introduction to Big Data

- Examined how big data is used
  - Twitter geolocation
  - Android phone data for Google maps
- Examined initial analysis for business intelligence
  - Excel tools (and Google Analytics)
  - Sam's Club and Dillard's data
  - SQL Cubes
  - Excel pivot tables
  - Use these tools to ask questions

Questions

- We likely don't have ability to remote desktop into UA data center
- Just trying to show you some of the things one can do
- Takes a fair amount of math and training to actually use data for business intelligence