

Fundamentals of Big Data

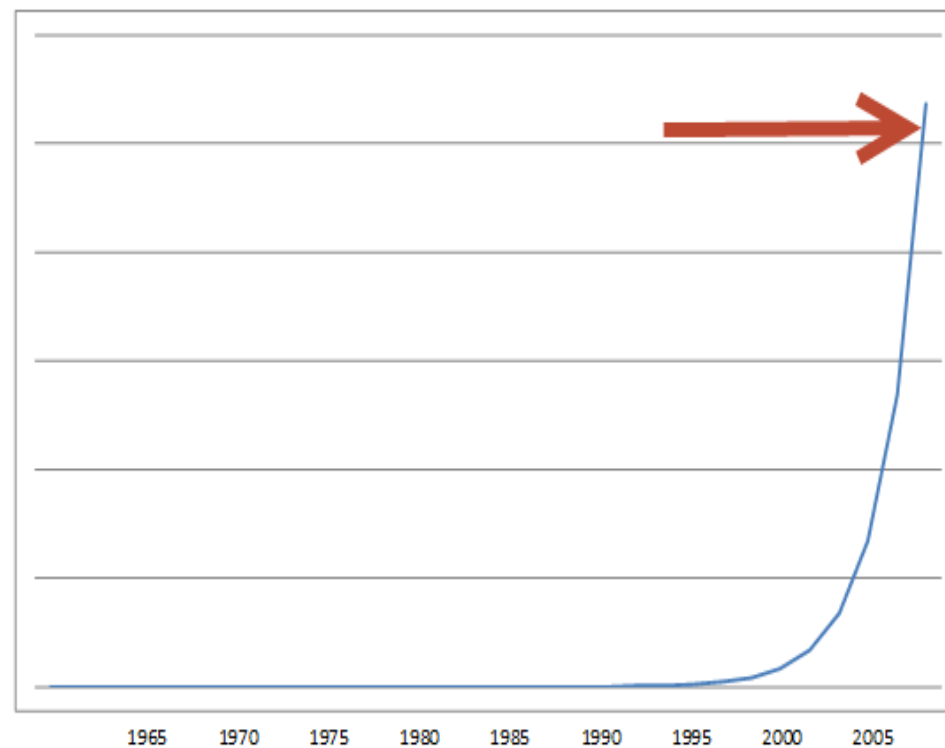
Mark DuBois

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

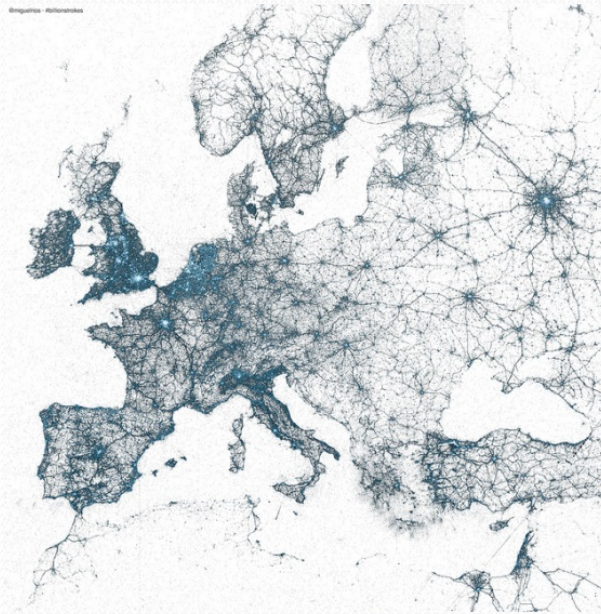
Accelerators

- **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



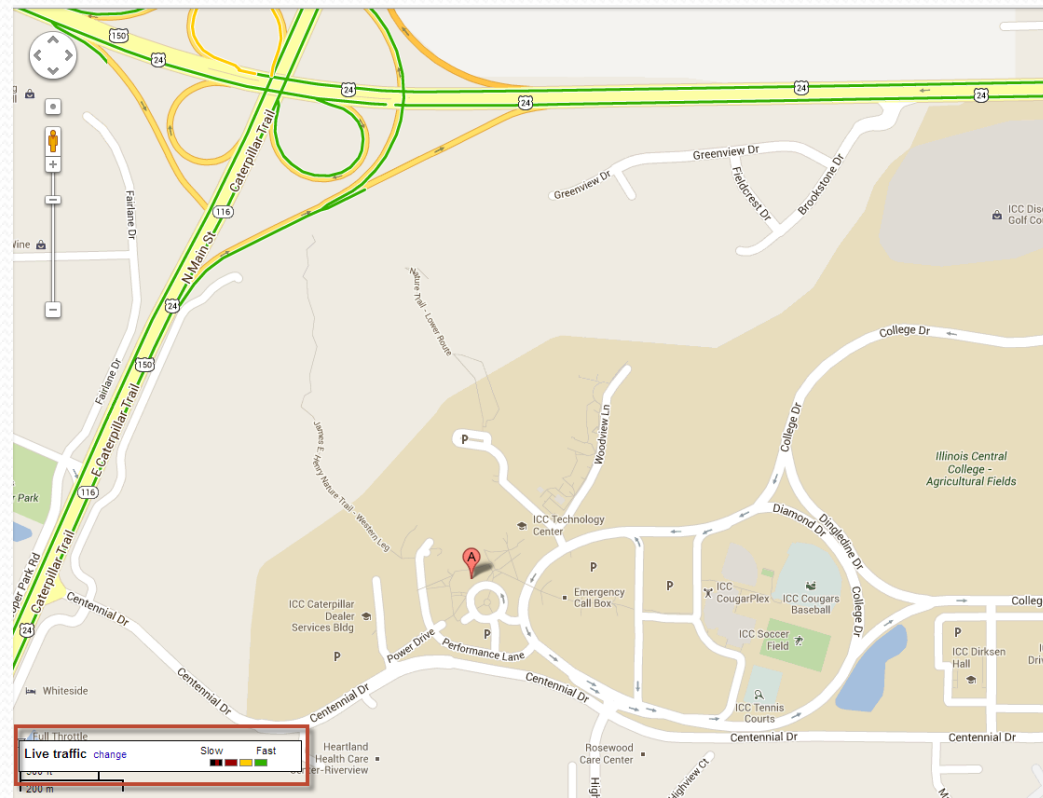
Anecdote

- 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/BigData01>

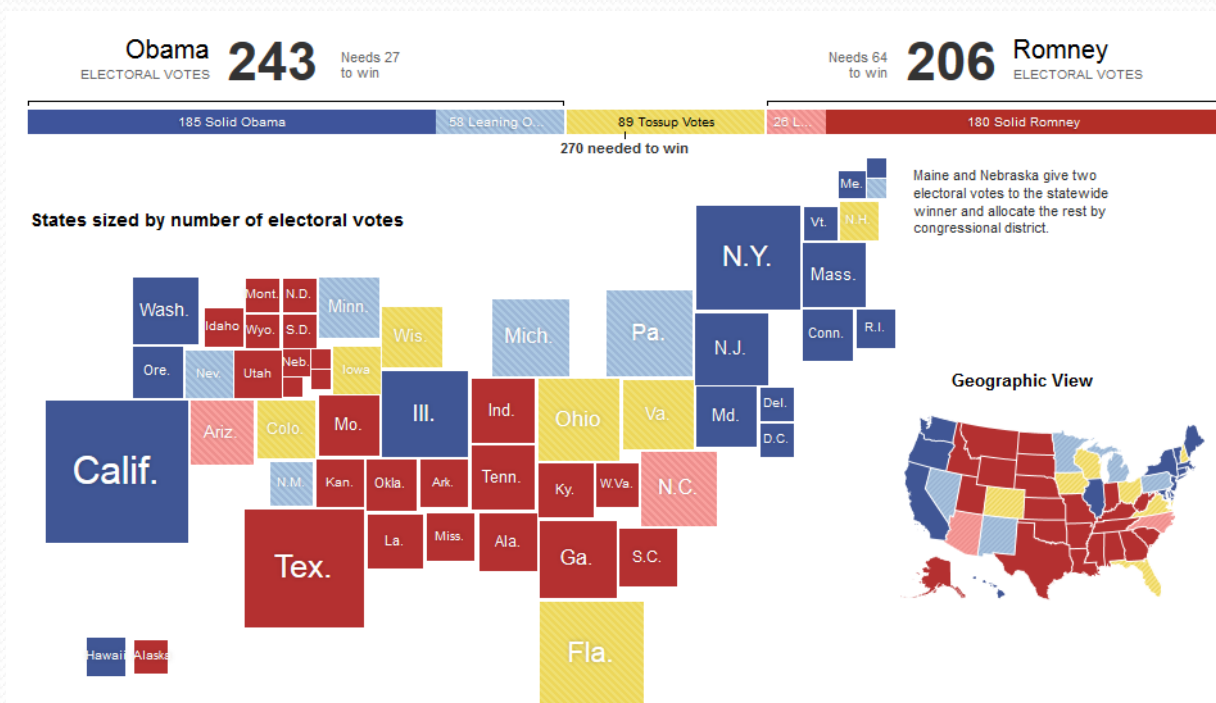
Another example

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



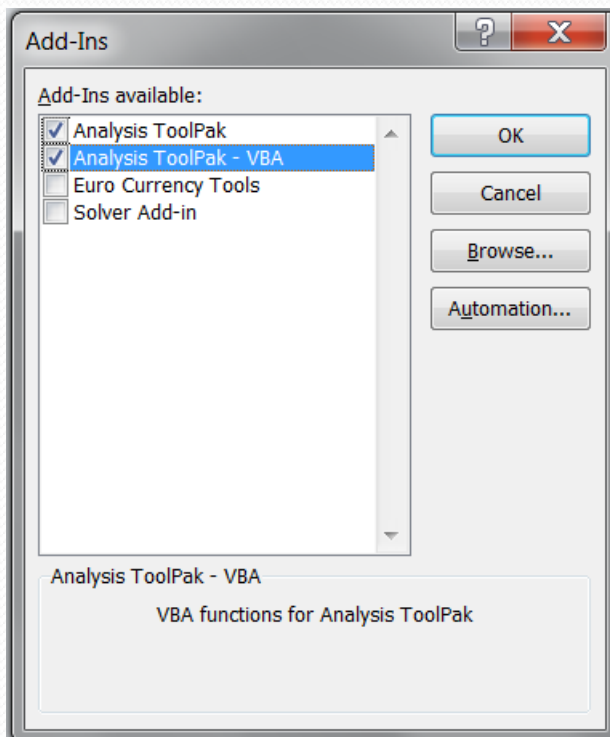
Predictive Analytics

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

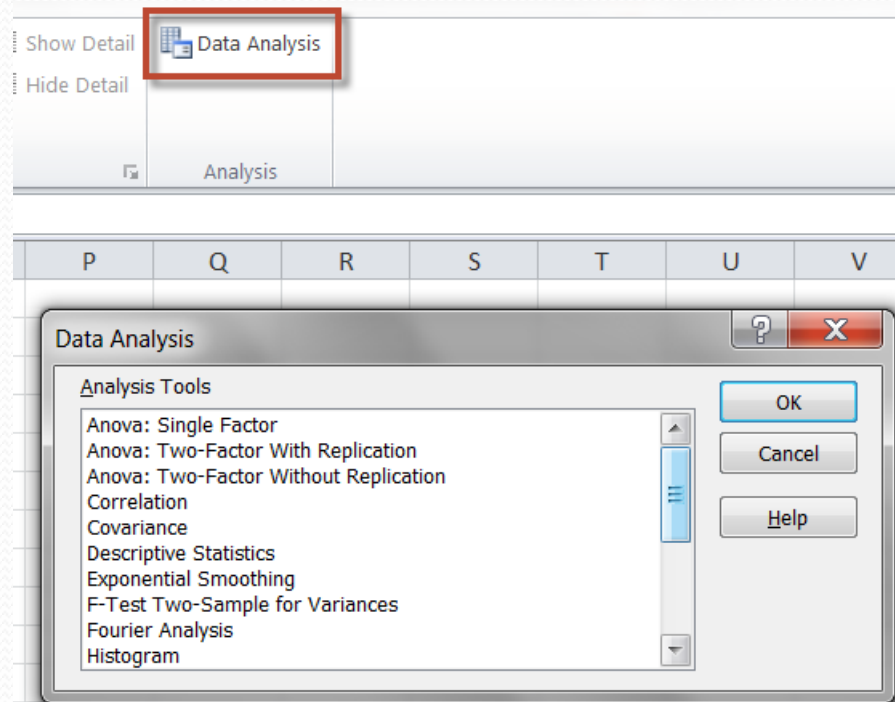


More tools than we think

- Excel Analysis toolpak

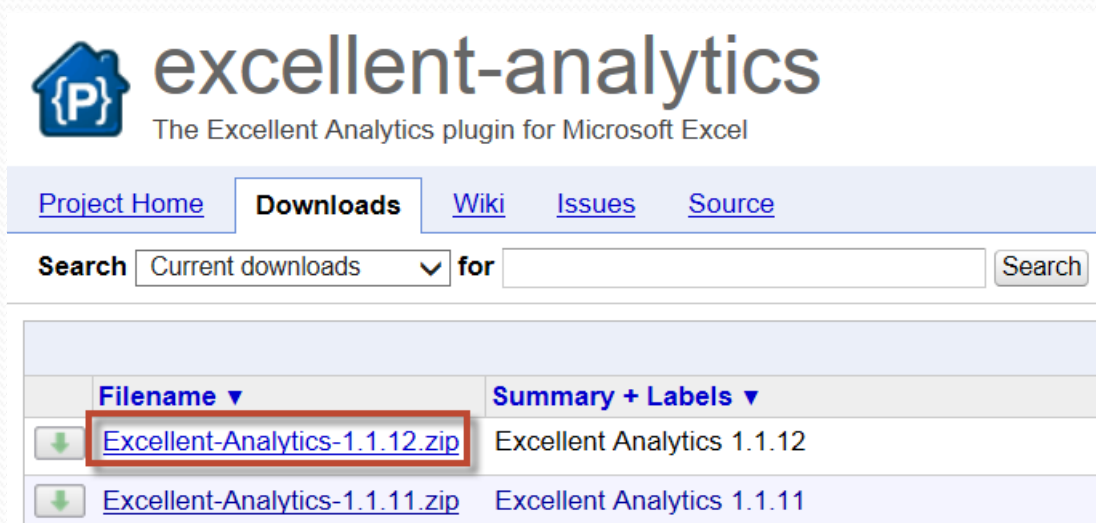


Results (data tab)



More Excel Tools

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

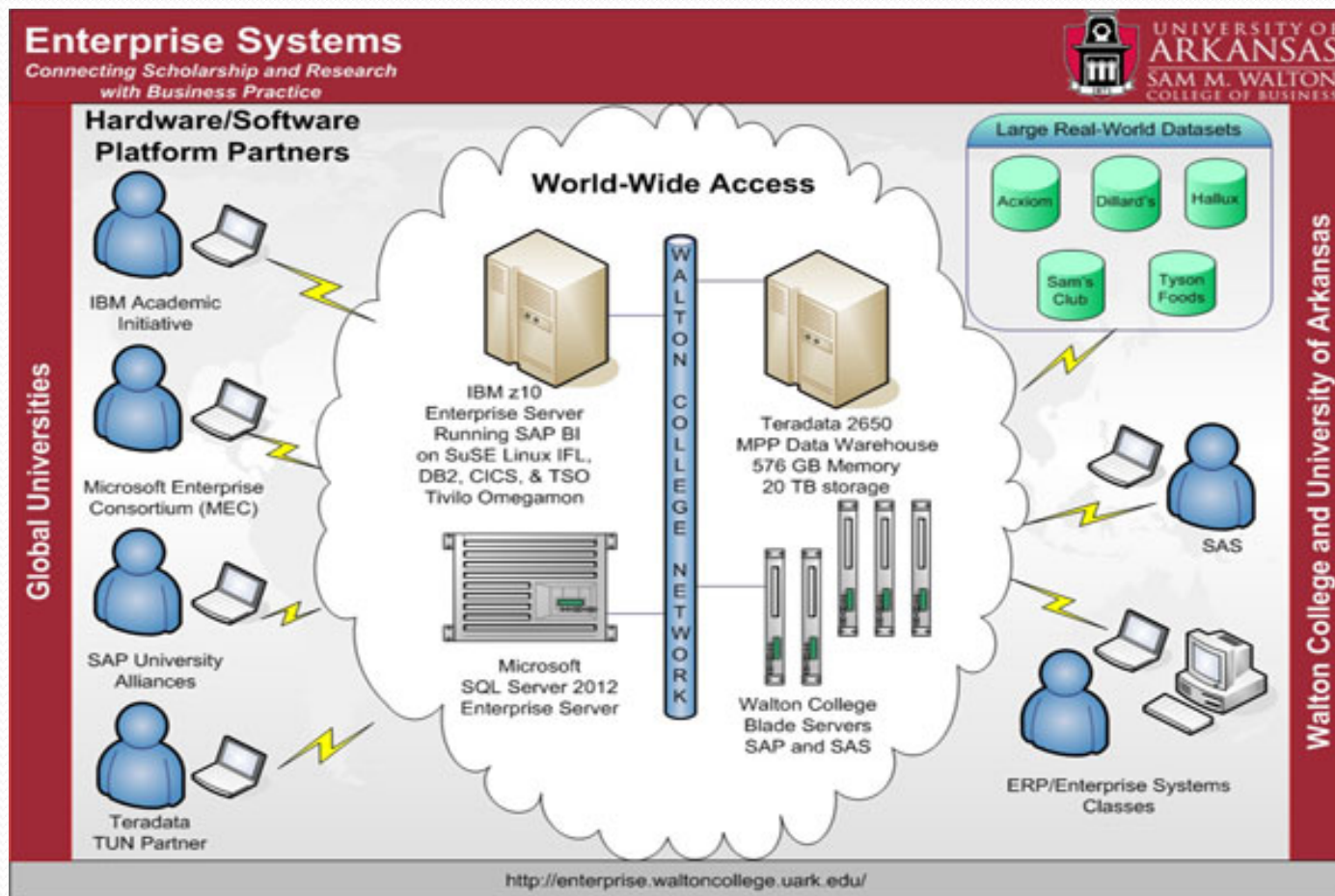


The screenshot shows the GitHub repository page for 'excellent-analytics'. The page title is 'excellent-analytics' with the subtitle 'The Excellent Analytics plugin for Microsoft Excel'. The navigation menu includes 'Project Home', 'Downloads', 'Wiki', 'Issues', and 'Source'. A search bar is present with the text 'Search Current downloads for' and a 'Search' button. Below the search bar is a table of download links:

Filename ▼	Summary + Labels ▼
Excellent-Analytics-1.1.12.zip	Excellent Analytics 1.1.12
Excellent-Analytics-1.1.11.zip	Excellent Analytics 1.1.11

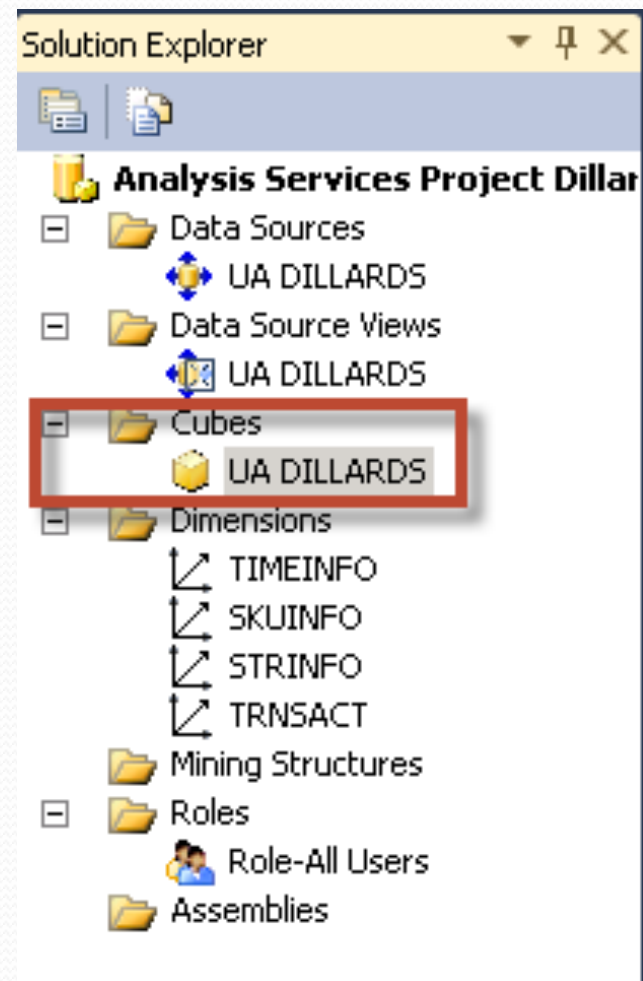
Large amounts of data

- 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

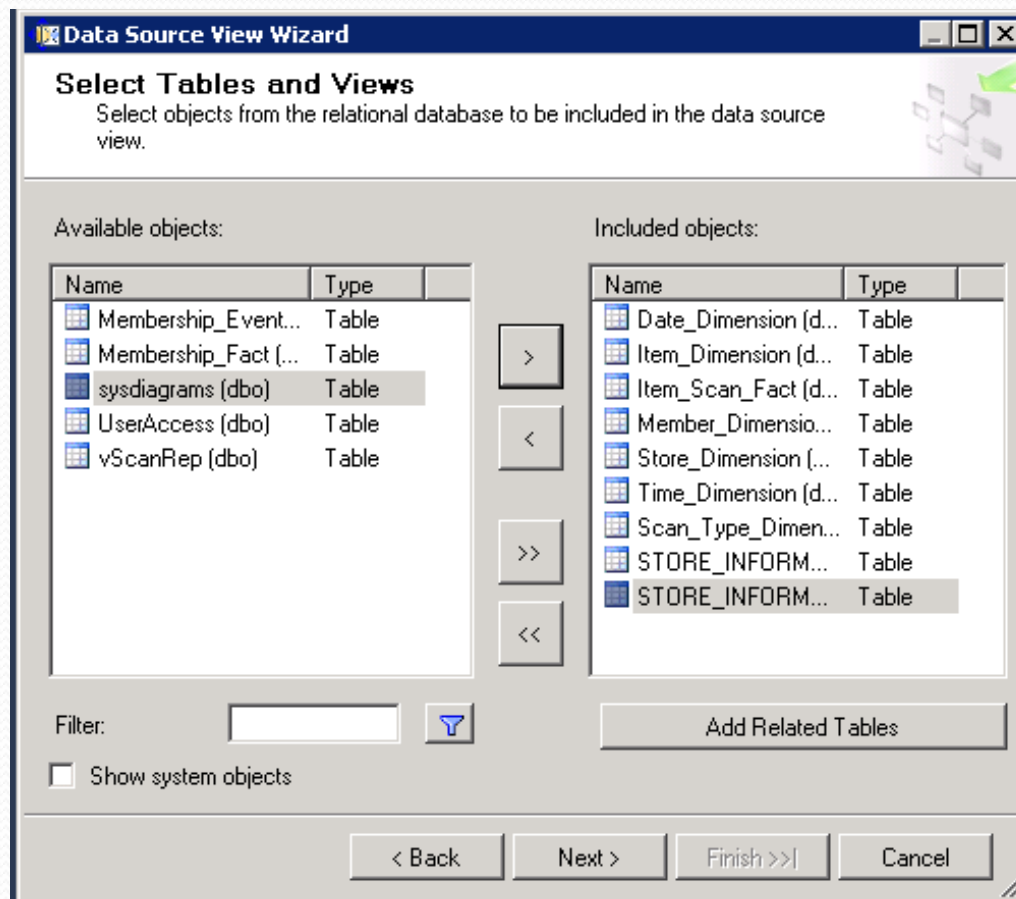


SQL What?

- 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



Establish relationships

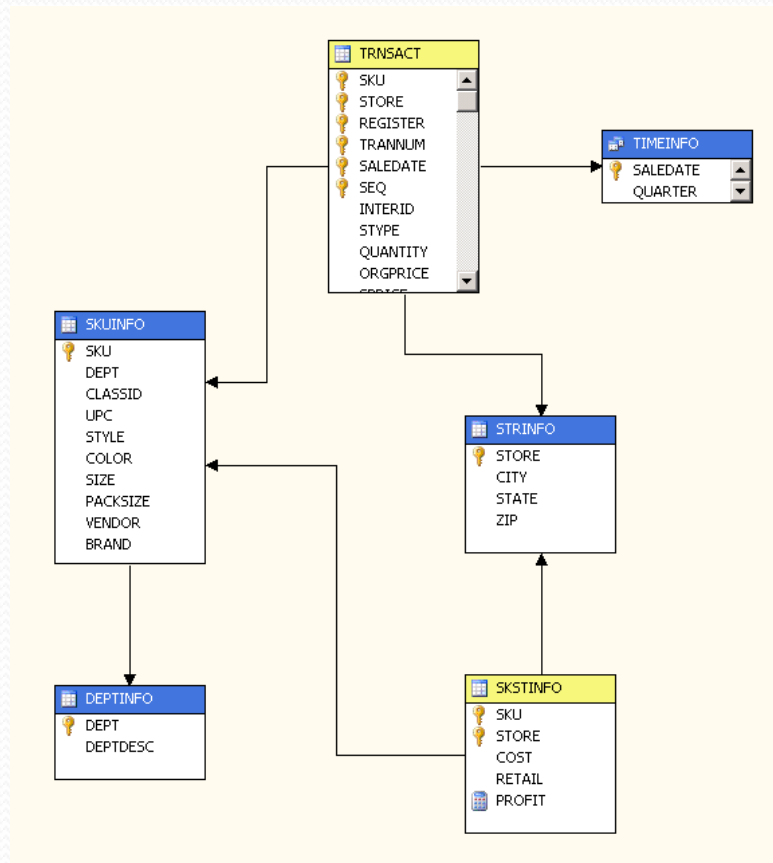
The screenshot displays a database schema with several tables and their relationships. The tables shown are:

- STORE_INFORMATION**: StoreID, Store_Nbr, Store_Name, Align_Sub_Division_..., Region_Nbr, District_Nbr, Open_Date, Store_Type, Street_Addr, City.
- Time_Dimension**: Time_Key, Time_Description, AM_PM.
- Member_Dimension**: Member_Key, Member_Number, Zip_Code, Member_Type, Member_Status_Code, Elite_Status_Code, Bus_CR_Type_Code, Secondary_Card_Count, Qualify_Org_Code, Complimentary_Card_Count.
- Item_Dimension**: Item_Key, Item_Number, Category_Number, Sub_Category_Number, Primary_Description, Secondary_Description, Color_Description, Size_Description, Item_Status_Code, Finline.
- Item_Scan_Fact**: Visit_Number, Store_Key, Item_Key, Member_Key, Transaction_Type_Key, Transaction_Date_Key, Transaction_Store_Time_Key, Transaction_Base_Time_Key, Item_Quantity, Total_Scan_Amount.
- Store_Dimension**: Store_Key, Store_Number, Store_Name, Subdivision_Nu..., Region_Number, District_Number, Open_Date_Key, Store_Type, Street_Address, City.

Relationships are indicated by arrows: Item_Scan_Fact is linked to Item_Dimension, Store_Dimension, and Time_Dimension. The 'Specify Relationship' dialog box is open, showing the configuration for a relationship between Item_Scan_Fact (Source) and Time_Dimension (Destination). The Source Column is Transaction_Store_Time_Key and the Destination Column is Time_Key. The dialog also includes a 'Reverse' button and a 'Description' field.

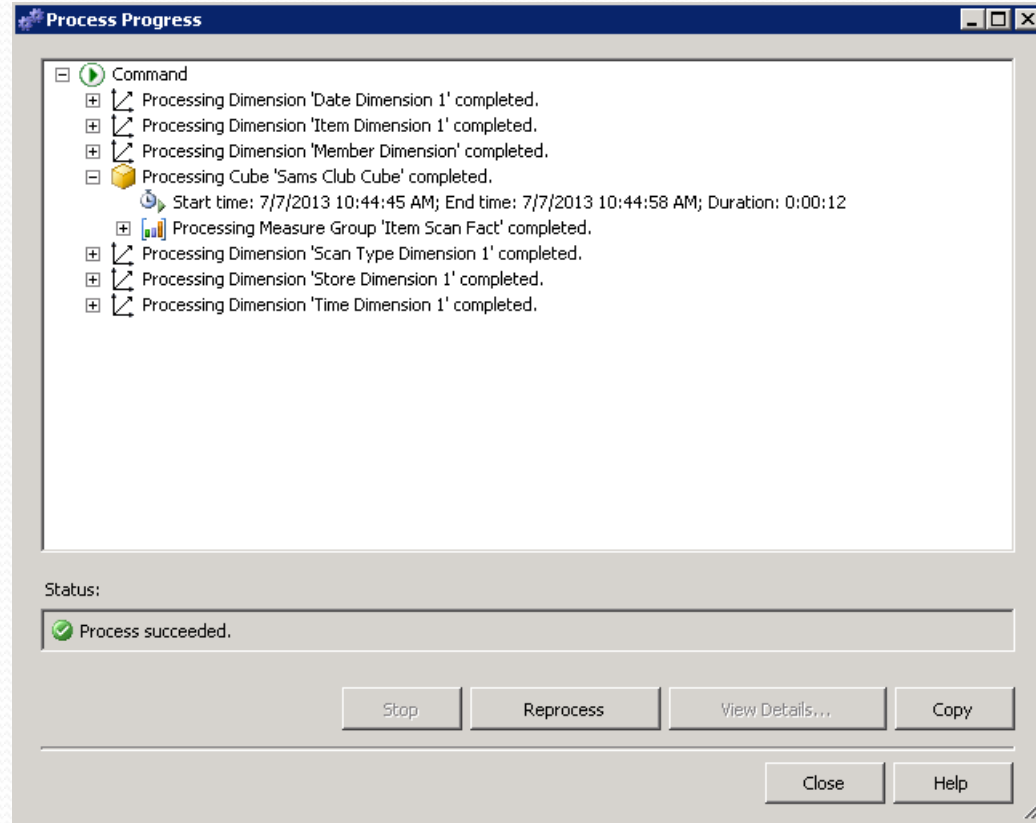
Dillard's Store Data

- Example of SQL Cube



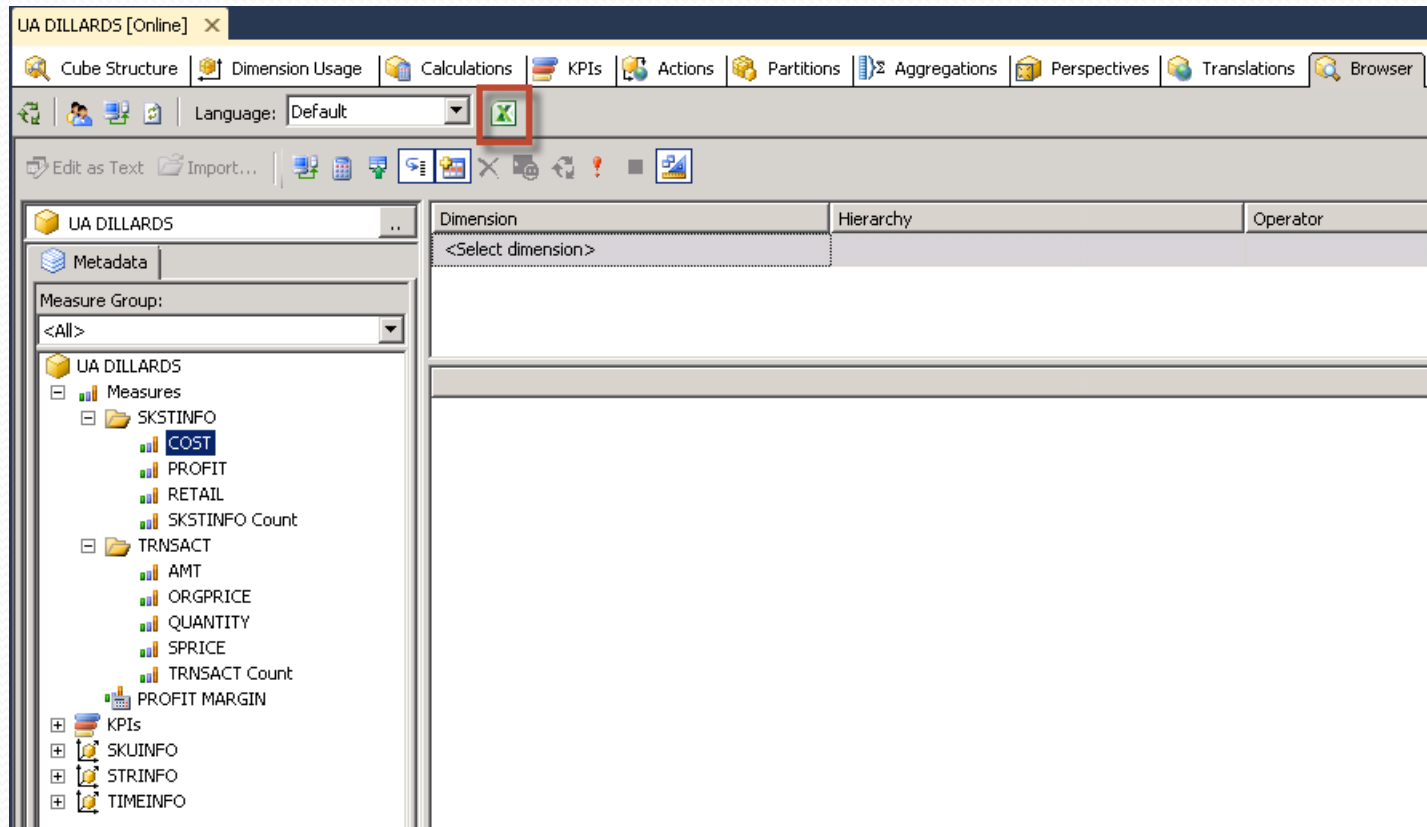
Process

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



Using Dillard's Cube

- Open data in Excel



Excel Pivot Table

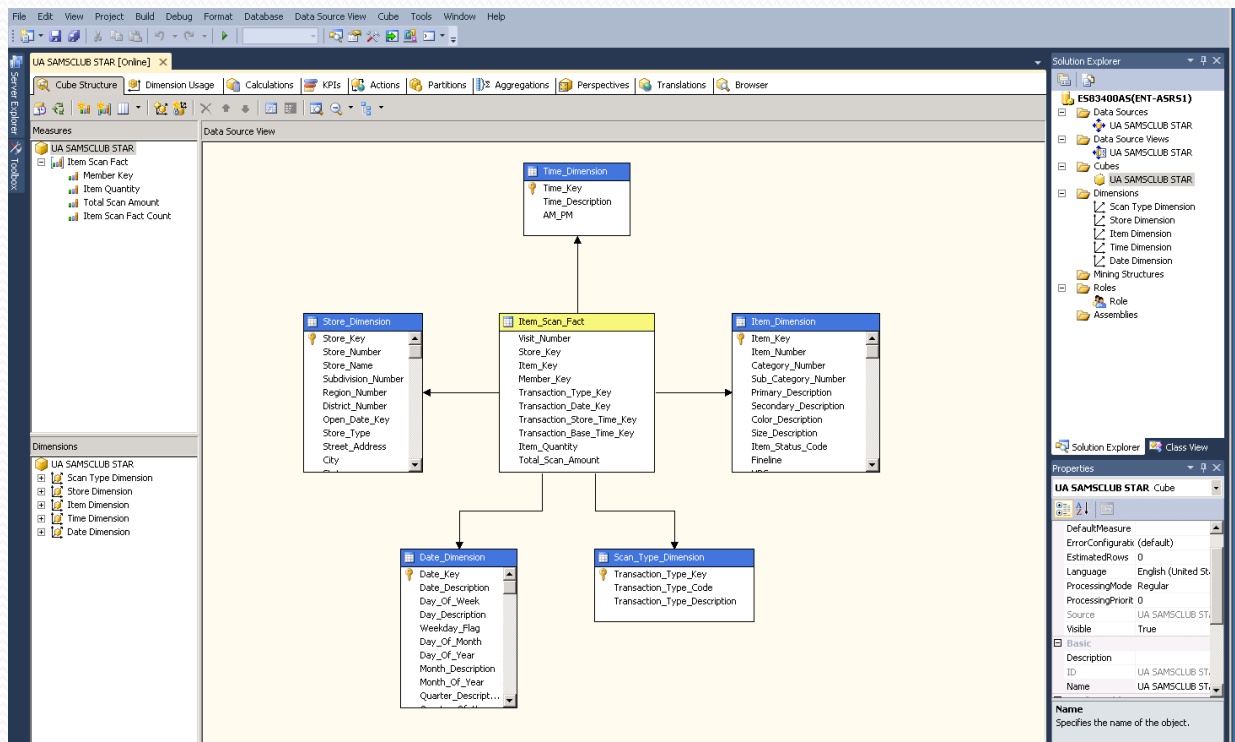
- For further analysis
- Table is tied to data stores

The screenshot displays the Microsoft Excel interface with a PivotTable and the PivotTable Field List task pane. The PivotTable is located in the range A77:F77 and shows data for various cities, with 'FAYETTEVILLE' selected. The PivotTable Field List task pane is open on the right, showing the 'Show fields related to:' dropdown set to '(All)'. The task pane lists several data sources: SKSTINFO (with 'COST' checked), TRNSACT, KPIs, SKUINFO, and STRINFO (with 'Geography' checked). The 'Report Filter' area is empty, and the 'Column Labels' area is empty. The 'Row Labels' area contains 'CITY', and the 'Values' area contains 'COST'. A tooltip for 'FAYETTEVILLE (CITY)' is visible over the selected cell in the PivotTable, showing 'STATE: AR' and 'Row: FAYETTEVILLE'.

	A	B	C	D	E	F
75	FAIRVIEW PARK	\$1,780,703.08				
76	FARMINGTON	\$1,202,805.19				
77	FAYETTEVILLE	\$3,070,446.75				
78	FLAGSTAFF	FAYETTEVILLE (CITY) STATE: AR Row: FAYETTEVILLE 2.73				
79	FLORENCE	6.96				
80	FLORISSANT	24,033,377.67				
81	FORT SMITH	\$2,062,553.70				
82	FORT WORTH	\$2,275,196.35				
83	FRANKLIN	\$4,384,590.29				
84	FRIENDSWOOD	\$4,177,112.56				
85	FT MYERS	\$2,752,691.68				
86	FT. LAUDERDALE	\$2,885,629.29				
87	FT. WORTH	\$3,629,412.94				
88	GAINESVILLE	\$1,780,526.88				
89	GASTONIA	\$1,663,595.59				
90	GLEN ALLEN	\$1,324,142.03				
91	GLENDALE	\$3,340,509.33				
92	GOODLETTSVILLE	\$2,456,739.93				
93	GRAND ISLAND	\$1,124,410.01				
94	GREELEY	\$1,342,586.86				
95	GREENSBORO	\$2,567,663.39				
96	GREENVILLE	\$3,038,334.64				
97	GRETNA	\$2,955,394.03				
98	HAMMOND	\$1,442,556.48				
99	HARLINGEN	\$1,842,421.36				
100	HATTIESBURG	\$2,036,097.54				
101	HELENA	\$901,638.14				
102	HENDERSON	\$3,194,288.94				
103	HICKORY	\$1,573,559.53				
104	HIGH POINT	\$1,192,297.66				
105	HOT SPRINGS	\$1,540,986.08				
106	HOUMA	\$2,230,156.28				
107	HOUSTON	\$14,354,509.67				
108	HUMBLE	\$2,172,993.14				
109	HUNTSVILLE	\$4,146,076.66				

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