Introduction to Data Science

Week 1, Lecture 2

Jeff Hammerbacher
January 19, 2012
Lecture Outline

- 1. The Standard Architecture for Analytics
- 2. Some History
- 3. Instantiations
- 4. Examples
- 5. Exercise
1. The Standard Architecture
Application Database
Application Database → ETL → Data Warehouse → Business Intelligence
Data Warehouse

Application Database

ETL

Data Warehouse

Business Intelligence

Analytics
2. Some History
The Evolution of Data Warehousing

Teradata

- Founded in 1979
  - Collaboration between Citigroup and Caltech
- Shipped first product to Wells Fargo in 1983
- First terabyte-sized database at Wal-mart in 1992
- Announced 5 customers with over 1 petabyte, 35 over 100 terabytes in 2008

Thursday, January 19, 12
The Evolution of Data Warehousing

Best Practices

  - Barry Devlin and Paul Murphy, IBM
  - Coins the term “business data warehouse”
- 1991: “Building the Data Warehouse”
  - Bill Inmon
  - “Top-down” design
- 1996: “The Data Warehouse Toolkit”
  - Ralph Kimball
  - “Bottom-up” design
3. Instantiations
The Businessperson

- Data Sources
  - Web pages
  - Excel
- ETL
  - Copy and paste
- Data Warehouse
  - Excel
- Business Intelligence and Analytics
  - Excel functions
  - Excel charts
  - Visual Basic!!
The Programmer

- Data Sources
  - Web scraping, web services API
  - Excel spreadsheet exported as CSV
  - Database queries
- ETL
  - wget, curl, Beautiful Soup, lxml
- Data Warehouse
  - Flat files
- Business Intelligence and Analytics
  - Numpy, Matplotlib, R
The Enterprise

- Data Sources
  - Application databases
  - Intranet files
  - Application server log files
- ETL
  - Informatica, IBM DataStage, Ab Initio, Talend
- Data Warehouse
  - Teradata, Oracle, IBM DB2, Microsoft SQL Server
- Business Intelligence and Analytics
  - Business Objects, Cognos, Microstrategy
  - SAS, SPSS, R
The Web Company

- Data Sources
  - Application databases
  - Logs from the services tier
  - Web crawl data

- ETL
  - Flume, Sqoop, Pig, Crunch, Oozie

- Data Warehouse
  - Hadoop/Hive

- Business Intelligence and Analytics
  - Custom dashboards: Argus, BirdBrain

R
4. Examples
Data Preparation

Examples from Spring 2011

- HTML tables
  - Kiplinger college information
- File downloads
  - Stack Overflow
  - Basketball Reference
  - Pirate Bay, Last.fm
  - Iraq Body Count
- REST APIs
  - YouTube, Twitter, Yelp, GitHub
Data Presentation
Examples from Spring 2011

Stack Overflow Visualizations

Tag Popularity over the years

Normalized
Data Presentation
Examples from Spring 2011

Average Admit Rate by State

Percent Admitted
- < 75%
- < 65%
- < 55%
- < 45%
- < 35%
- No Data

Number of Schools
- 40
- 30
- 20
- 10
- 0

Thursday, January 19, 12
Data Presentation
Examples from Spring 2011
Data Products
Examples from Spring 2011

- Stack Overflow tag recommendation and response time prediction
- Locating ethnic food in ethnic neighborhoods
- Building optimal NBA teams
- Recommending new musical artists
- Prioritize emergency calls in Seattle
- Finding the right college for you
Observation

Examples from Spring 2011

- Factors that predict top college rankings
  - SAT scores
  - 4-year graduation rate
  - Freshmen retained
- Factors that influence popularity of a video on YouTube
- Impact of playing in the NBA summer league on next season’s performance
- Impact of distance from police station on crime levels
- Impact of alcohol consumption on unemployment
## Experimentation

**Examples from Spring 2011**

<table>
<thead>
<tr>
<th>Variations</th>
<th>Never Signed Up</th>
<th>Signed Up, Never Donated</th>
<th>Previously Donated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DONATE NOW</strong></td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td><strong>PLEASE DONATE</strong></td>
<td>+2.3%</td>
<td>+27.8%</td>
<td>+16.3%</td>
</tr>
<tr>
<td><strong>WHY DONATE?</strong></td>
<td>-27.8%</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>DONATE AND GET A GIFT</strong></td>
<td>+15.2%</td>
<td>-24.6%</td>
<td>+11.9%</td>
</tr>
<tr>
<td><strong>CONTRIBUTE</strong></td>
<td>+8.5%</td>
<td>+2.9%</td>
<td>+18.4%</td>
</tr>
</tbody>
</table>
Data Preparation

Examples from Facebook

- Application databases
- Web server logs
- Event logs
- API server logs
- Ad server logs
- Search server logs
- Advertisement landing page content
- Wikipedia
- Images and video
Data Presentation
Examples from Facebook

- Argus
- Insights
- Magellan
- Growth Report
- Group Intelligence
- Pre-sales and post-sales reports
- Visualizations
Data Products
Examples from Facebook

- Platform application reputation scoring system
- Language classifier
- Classifier for bad users and bad content
- Lexicon
- People you may know
- Applications you might like
- Insights
- Inventory Estimation service
Observation

Examples from Facebook

- “Tipping point” for college networks
- Impact of new feature launch on user growth and activity
- Impact of PR on user growth and activity
- User navigation
- User activation and deactivation
- Motivations for content production
- Content diffusion
- User happiness
- Search quality
Experimentation
Examples from Facebook

- Index.php
- Invitation emails
  - Message content
  - HTML versus plain text
- Deactivation
- Number of ads per page
- Different versions of the PYMK algorithm
- Content of an empty News Feed
- Page load time
Data Preparation

Examples from outside of Facebook

- Collection
  - Seismic surveys
  - Ion Torrent: low-cost genome sequencing
  - Biopsy: removal of tissue for the purposes of measurement
  - LSST: 3,200 megapixel telescope
  - CPI: basket of goods and services in 87 urban areas

- Integration
  - Google Refine
  - Needlebase
Data Presentation

Examples from outside of Facebook

- USPTO Patent Dashboard
- OPOWER energy usage report
- ESPN GameCast drive visualizations for football games
- The wifi and power meters on my laptop
- Every New York Times infographic
- Sense.us
- Stock charts
- Radar
Data Products

Examples from outside of Facebook

- Acxiom: enrichment of consumer data
- Fair Isaac: FICO credit score
- Bloomberg: market data, news, simple analysis
- Factual: consolidated data sets with standard API
- DNAnexus: genome analysis-as-a-service
- WeatherBill: weather insurance
- Google: web search
- Carfax
- Player draft ratings
Observation
Examples from outside of Facebook

- OK Cupid
  - Message frequency versus attractiveness
  - Needed to look at distribution of attractiveness, not just average
- British Doctors Study
  - Smoking causes lung cancer
- Nanex Flash Crash Analysis
  - Found evidence of mini-crashes
  - Determined “quote-stuffing” might have a negative impact on the markets
Experimentation

Examples from outside of Facebook

- Clinical trials
  - Toxicity
  - Effectiveness
- Netflix: “simple trumps complete”
  - Tools for input of taste preferences
  - UI for iPhone
  - Streaming encoding quality
- Retail product pricing
- Manufacturing quality
5. Exercise
Your Turn!

CDO for a Day

- Your team has been tasked with enabling your organization to “compete on analytics”
  - 1. Define the top three priorities of the organization
  - 2. Determine the data sources you’d like to collect
  - 3. Highlight the largest data integration challenges you’ll face
  - 4. Determine the most important data to present to your organization
  - 5. What data products could you build?
  - 6. What studies could you run to answer the most pressing questions for the organization?
  - 7. Suggest some experiments to run to help guide the organization towards their goals
Your Turn!
CDO for a Day

- Organizations
  - 1. Hospital
  - 2. University
  - 3. Equities trading desk
  - 4. Retail chain
  - 5. Foursquare
  - 6. Police department
  - 7. CIA
  - 8. Yourself (QS)