Data Acquisition

Data Acquisition

- **What types of data are used in analysis**
- **Where you can find these data sets**
- **What are some common methods of accessing this data**

Data Acquisition

- **Where to source data**
- Acquisition techniques

Internal Data Sources

- **Most valuable information comes from your own organization**
- **There are many sources of data available internally**
  - Application databases (OLTP)
  - Data warehouses (OLAP)
  - Log files (Web, e-mail, and other applications)
  - Documents (file servers, intranet, Web site)
  - Sensors and network events

Freely Available Data Sources

- **External data is often used to augment a solution**
  - Geolocation for IP addresses in Web server logs
  - Demographic information about those locations
- **There are many sources of data available at no cost**
  - Some are public domain and some are copyrighted
  - Be sure to check the license to verify that your use is allowed

Freely Available Data Sources (cont'd)

U.S. Census Bureau	http://factfinder2.census.gov/
U.S. Executive Branch	http://www.data.gov/
U.K. Government	http://data.gov.uk/
E.U. Government	http://publicdata.eu/
The World Bank	http://data.worldbank.org/
Freebase	http://www.freebase.com/
Wikidata	http://meta.wikimedia.org/wiki/Wikidata
Amazon Web Services	http://aws.amazon.com/datasets
InfoChimps *	http://www.infochimps.com/marketplace

* Most data sets are available at no cost, but some have a fee

Commercial Data Sources

- **Many companies also offer data**
  - Usually for a fee, but sometimes available at no cost
  - Always be sure to check the license terms

Gnip	Social Media	http://gnip.com/
AC Nielsen	Media Usage	http://www.nielsen.com/
Rapleaf	Demographic	http://www.rapleaf.com/
ESRI	Geographic (GIS)	http://www.esri.com/
eBay	Auction	https://developer.ebay.com/
D&B	Business Entity	http://www.dnb.com/
Trulia	Real Estate	http://www.trulia.com/
Standard & Poor's	Financial	http://standardandpoors.com/

Data Acquisition

- Where to source data
- **Acquisition techniques**

Database Integration

- **Data internal to an organization is often kept in a database**
- **To access small samples, just export a subset to a local file**
  - Can do this programmatically or manually via query tool
  - Can also do this on command line, as shown below

```
$ cat 10k_customers.sql
select id, firstname, lastname, email, zipcode
into outfile '/user/jsmith/cust10k.csv'
fields terminated by ','
optionally enclosed by '"'
escaped by '\\'
lines terminated by '\n'
from customers limit 10000"

$ mysql -u jsmith -p mysecret < 10k_customers.sql
```

  - Invocation details will vary depending on database used

Other Internal Sources

- **Systems that produce data in the form of files are easily handled**
  - For a few small files, just copy them to a local filesystem
  - Larger file sets should be copied to HDFS instead

```
$ hadoop fs -put myfile.txt /bigdata/project/myfile.txt
```

  - This can be done manually or as part of a script
  - HDFS also supports a REST API through WebHDFS
- **Alternative: Use Flume to add data to HDFS as it's generated**
  - Can "tail" log files to capture lines as soon as they're written
  - Other sources: program execution, network port, and syslog
  - Write custom sources to integrate with legacy systems

Data Archive Downloads

- **External data sources are sometimes in the form of archives**
  - Delimited and fixed-width textfiles are most common type
  - Usually compressed to save storage space and bandwidth
- **These are usually hosted on Web or FTP sites**
  - Downloading is easy with your browser for small number of files
- **How do you automate download of many files?**
  - Use the `curl` or `wget` command line utilities

```
$ curl -i list_of_urls.txt

$ curl -O http://www.example.com/xyz[001-999].zip

$ curl -u jsmith:mysecret ftp://ftp.example.com/archive/bigfile.gz

$ wget --mirror http://www.example.com/data/ -o /home/jsmith
```

Data APIs

- **Many organizations offer data as services rather than downloads**
  - Some APIs are read-only, while others support data modification
  - Authentication is often required (register for an ID to use in calls)
- **Data APIs have several advantages over archive downloads**
  - The service maintains the data and can keep it updated
  - Usually cross-platform and cross-language (REST or SOAP)
  - Price may be based on only what you use
- **Access to data through APIs also has disadvantages**
  - Price or terms of service may change
  - Your application's availability depends on service availability
- **Data returned by an API is typically in XML or JSON format**

Screen Scraping

- **Sometimes the data is only available within a Web site**
  - You don't have access to the database powering the site
  - You only have access to the rendered pages themselves
- **You can acquire the data by "screen scraping"**
  - Programmatic access and parsing of page content
  - Fragile: your script may break when page changes
  - Should be viewed as a last resort

```
$ cat scraper.py
import urllib
from BeautifulSoup import BeautifulSoup

txt = urllib.urlopen("http://www.example.com/")
soup = BeautifulSoup(txt)

headings = soup.findAll("h2")
for heading in headings:
    print heading.string
```

Essential Points

- **The most valuable information is found within your organization**
- **There's a variety of data available from external sources that can help augment your solution**
- **External data is usually accessed as an archive or via an API**
  - Screen scraping is another option, but best avoided