BINF 8211/6211
Design and Implementation of Bioinformatics Databases
Lecture #17

Dr. D. Andrew Carr
Dept. Bioinformatics and Genomics UNCC
Spring 2016

Review Relational Model

Today

- SQL Review
- New SQL
  - Data loading
  - Collation
- XML/RDF

- What does the CAST function do?
- What does the CASE function do?
- What MySQL command will let you import data from a CSV file?
- How do you change the type of separator to be recognized in the file?
- How does this command interpret a carriage return or new line?
- What is a common transformation of sequence databases that requires piping data to an external application and then bringing back the results?
- In the OBO what type of syntax is used? What type of semantics?
- What is a stanza?
- What 3 keywords indicate the type of stanza?
- How do you indicate an Authority?

Review

CAST is used to convert the results of an evaluation/expression to a different storage class than the original data had. It cannot irreversibly change the way you can interpret the data, so some transformations are not allowed (a real or integer value cannot be changed to alphanumeric).

CASE gives and IF-THEN-ELSE loop type of fucntionality – it is placedi n the SELECT statements target list. You know you are done when you see END. The ELSE is optional and gives a default value when every WHEN has evaluated as FALSE.

.import fileA.csv TABLEname (include PATH to file name)
.separator ";" or whatever you want, a CR or NL signals the end of the row (tuple).

We have talked in some detail about sequence trimming and alignment, also sequence comparisons and sequence feature annotation.
In the OBO a name-value pair is used (also called a tag format), the syntax uses controlled vocabularies described in namespaces. The semantics in the namespaces are a subset of the Web Ontology Language.
In the OBO language, a stanza is the chunk of information that comes after the header and is lead with one of the keywords which are Term, TypeDef and Instance.
An Authority is enclosed in square brackets after the value in the name-vale pair.

MySQL

- Import Data
  - Internal
    - load data local infile 'file.csv' into table table
    - fields terminated by ','
    - enclosed by '"'
    - lines terminated by '\n'
    - (column1, column2, column3,...)
  - External
    - mysqlimport --columns='head -n 1 $yourfile' --ignore-lines=1 dbname $yourfile

MySQL

- Show all tables in the database
- SHOW [FULL] TABLES [{FROM | IN} *db_name*] [LIKE '*pattern*' | WHERE *expr*]

MySQL Export to file

- Export data
  - Easy
- SELECT * FROM Cars INTO OUTFILE '/tmp/cars.csv' FIELDS TERMINATED BY ',';

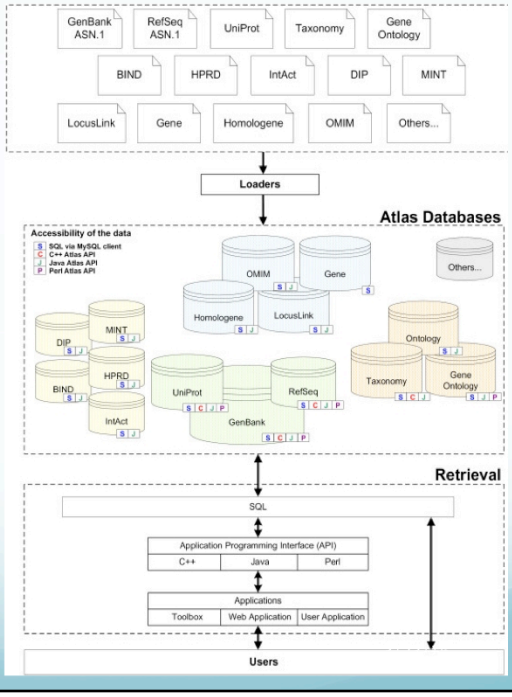
Data: ETL
Extract, Transform, Load

- Extract:
  - Acquiring the data from the raw source.
  - Different from transformation – This is pulling the data from source.
- Transform:
  - From defined order
    - FASTA →
    - FASTQ →
    - CEL or CDF
- Load:
  - Moving the data into the DB

E = extract
T = transform

L = Load

ETL is going to take up most of your time



UNCC BINF 8211 Weller

Create a model.
Find the data.
Map to your data model.
Iterate.

Your data may come from a large number of resources (literature, Web sites, other databases, lab device output, analysis programs, etc)

Data validation: you may have to change units, change precision, compute a standard value, remove columns, rearrange columns, recombine columns.

Then you can actually load the columns.

Loading – if you can program you can parse a file to retrieve values from a particular field in each row and insert that into the correct database table, looping to insert serially. If you do not know how to do this then you can use the .import command from a CSV file, but then the file must map to the table exactly.

Data Extraction

- Formalization: Pulling data from its raw source
  - Time consuming
  - No Trivial – Depending on the data format
    - Binary – compression
      - Format within the compression.
      - How the data is packed.
  - Tools
    - BioPython
    - R
    - Bio Perl

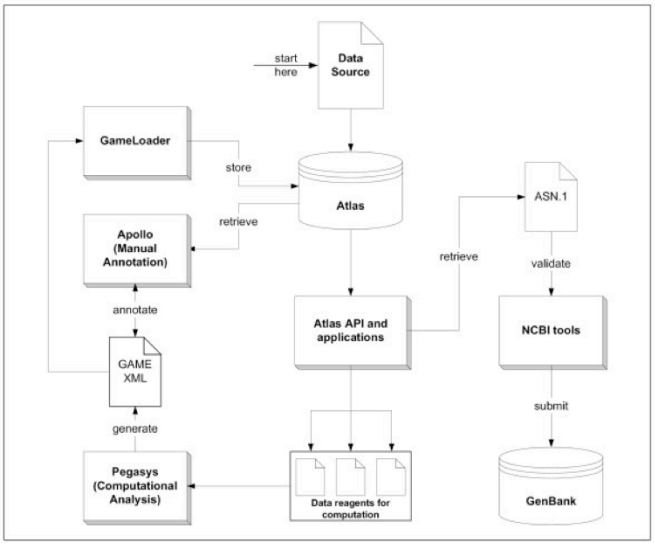
Lots for tools are available for this, but it may be necessary to reference the APIU.

Data Extraction

- The Extract function is fairly obvious if not always easy.
- Data transformation task types– is the goal of the database to update base data frequently, to update derived data frequently, or to allow user-derivations at unspecified intervals?
- "In business applications the volumes of base data are greater" –MOST SCIENCE
  - NGS – HUGE DATA PROBLEM
  - A hybrid system (pipe data to external analysis packages) is common in Bioinformatics (example on next page).

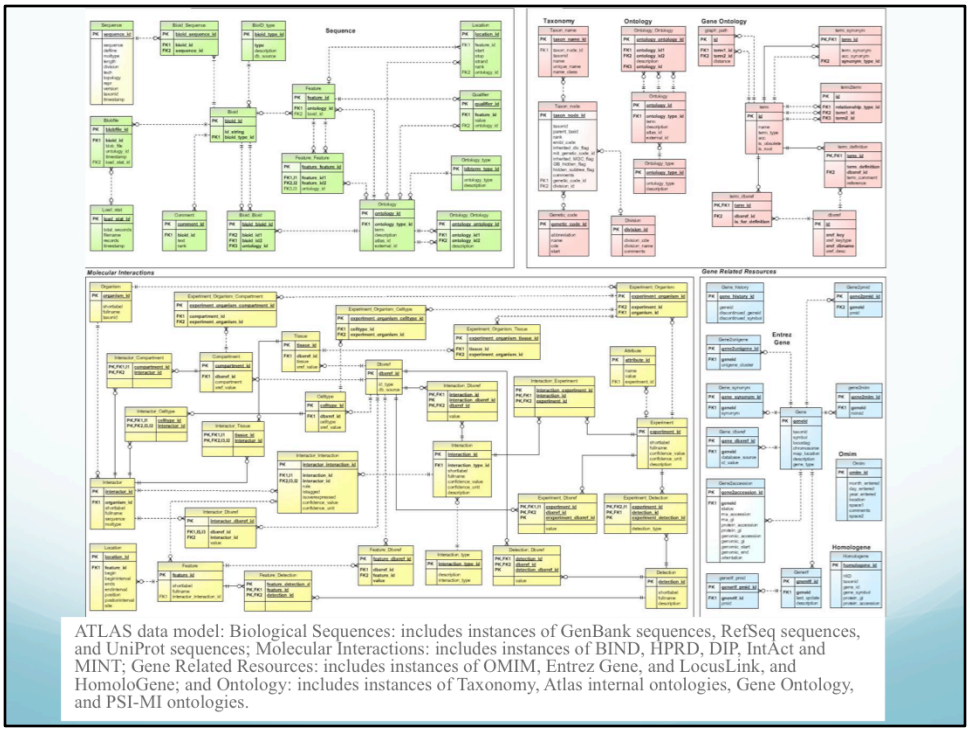
In practical experience this can also be pulling data directly from a measurement device. – NMR – PCR – Sequencer

You should find data as early in the process as possible and download the files and open them – what do you have to work with, do you understand the meaning of all the elements.
Data transformation is often a combination of transforming files and the order of data as well as the data itself.



Genome annotation tasks: storage of annotations, and data transformation for submission. This shows how Pegasys, Apollo, NCBI tools and Atlas can be used for data processing. Pegasys system outputs annotations in GAME XML for import into Apollo. The biological features are stored in the Atlas **Feature** tables can be and exported in GFF3 for import into the NCBI submission tools for validation, and submission to GenBank.

One of the most common tasks in Bioinformatics databases is to create the desired types of sequence feature annotation using a set of software applications that pipe data from one algorithm to another, with output at specified points. The output is captured and stored. The output may be tagged or it may be implicitly ordered, we are going to talk about several of the tagging methods.



ATLAS data model: Biological Sequences: includes instances of GenBank sequences, RefSeq sequences, and UniProt sequences; Molecular Interactions: includes instances of BIND, HPRD, DIP, IntAct and MINT; Gene Related Resources: includes instances of OMIM, Entrez Gene, and LocusLink, and HomoloGene; and Ontology: includes instances of Taxonomy, Atlas internal ontologies, Gene Ontology, and PSI-MI ontologies.

The ATLAS pipeline of annotation and analysis programs has its own data model, shown here – you can see the types of information that can be produced from sequence data as the initial input.

If you map from your own data model to what a particular application and pipeline of applications produces you will find loading data from a particular step is easier.

Data Transformation

- QC of the data.
  - Cleaning
- Incomplete
  - Missing records – you expect a record and it is not present
  - Missing fields – a value is needed
  - Skipped records – a system deliberately does not capture some records
- Incorrect
  - Something causes an incorrect code/value to be assigned to a record
  - Calculations/aggregations programmed improperly
  - Duplicates (records or attributes)
  - Incorrect pairing of codes

Data Transformation

- Incomprehensible
  - Multiple value attributes when mapping from source to warehouse
  - Odd field formatting or record layouts that vary
  - Undefined codes
  - Sources that are spreadsheets and processing files
  - Hierarchical and M:N sources
- Inconsistent
  - Units, codes:
    - time variance of coded entities, overlapping codes, different codes with the same meaning,
    - different business rules, aggregating rules, granularity.,
    - use of attribute, interval boundaries, formatting

QC: Validating data put into the database

- Once the data has been inserted, part of the process is to validate that what you expected is what happened.
- COUNT rows
- Make sure constraints were followed (data types, range limits, default to, etc.)
- Make sure FK works (JOINS, in-line views and subqueries).
- Carry out *queries that verify business rules*.

Transform means at least data cleansing. What type of errors should you look for (know these)?

UNCC BINF 8211 Weller 3/23/16

When we think of data transformation most scientists are considering derived data – often involving a standardization or normalization step, and perhaps applying some trimming to remove outliers, perhaps assignment to classes or similar simplification or pattern discovery. In the database world transformation is even more basic: you are nearly always changing the data at least a little: rounding floating point numbers, for example, perhaps changing the scale from the original measurement, and making sure the information is complete, accurate, interpretable and consistent. This is often called data cleansing – it can alter the shape of the data (if you graph it, for example).

Loading data into the database

- Generally, you add data to your database by creating a tab-delimited file and then doing a bulk insert.
- Using line-by-line INSERT statements is tedious and inefficient.
- You can write code to loop through the INSERT statements.
- Most database MS have commands to allow bulk loading.
- For large data sets you often need to write code to transform the original data to a specific delimited format that corresponds to the table structure –
  - This is often the MOST time-consuming part of making a database.

Part of the lecture last Thursday was to demonstrate and give you starter code to see how the looping code can work by several mechanisms.

- Once the data has been inserted, part of the process is to validate that what you expected is what happened.
- COUNT rows
- Make sure constraints were followed (data types, range limits, default to, etc.)
- Make sure FK works (JOINS, in-line views and subqueries).
- Carry out *queries that verify business rules*.

ETL-V: Validating data put into the database

Symantics, Syntactic Communication -- Ontology

- Syntactic Web –
  - Web as it exists
- Symantic Web
  - Consortium W3C of scientists pushing to codify language across the W3.
  - Developed RDF
    - Resource Description Framework
    - "RDF is an attempt to support effective creation, exchange and use of annotations on the Web."
    - Annotations don't supply semantics
    - Thus – Ontologies

RDF

- RDF was a first attempt at SW.
  - Started in 1999 updated 2014 to RDF 1.1
  - Incomplete
  - these issues include 'too few entailments', 'contradiction classes' and 'size of the universe'
  - Previous attempts to extend lead to indecipherable differences and "undecibaliltiy."
- RDFS (FA) more recent attempt to fix the issues while being true to OWL DL.

RDF

- RDF is a XML based language
  - Explicit markup vocabulary
  - Searchable with SPARQL
    - SQL like language

RDF Example

- concepts in RDF are taken from logic and linguistics
  - subject-predicate and subject-predicate-object structures have meanings similar to, yet distinct from, the uses of those terms in RDF.
- This example demonstrates:

  In the English language statement 'New York has the postal abbreviation NY' , 'New York' would be the subject, 'has the postal abbreviation' the predicate and 'NY' the object.

  Encoded as an RDF triple, the subject and predicate would have to be resources named by URIs. The object could be a resource or literal element. For example, in the N-Triples form of RDF, the statement might look like:

- <urn:x-states:New%20York> <http://purl.org/dc/terms/alternative> "NY" .

Example taken from Wikipedia 2016

OWL DL

- Web and Science overlap but domain knowledge and terminology complicates the communication space.
- OWL was created to specifically handle ontologies