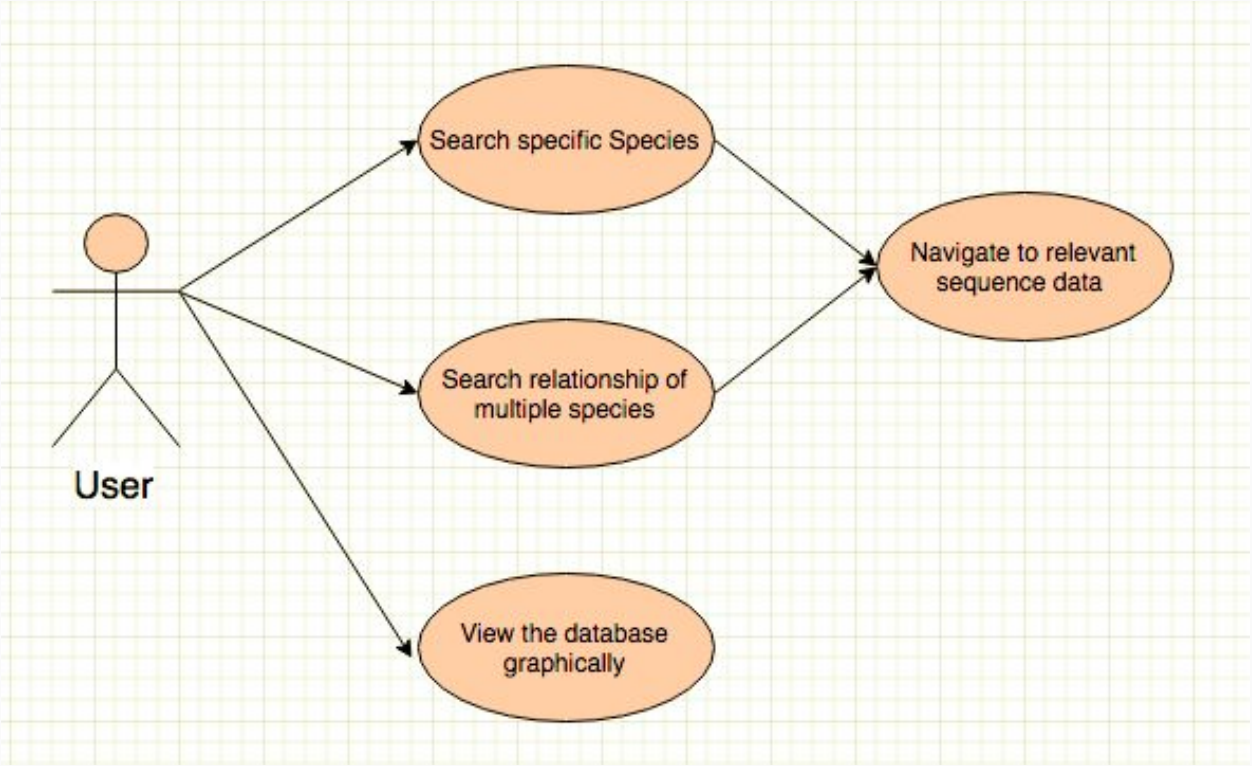
CSYE 7200 Project -
NCBI Taxonomy Search Engine

Team Member: Zhilong Hou, Li Ma
Professor: Robin Hillyard

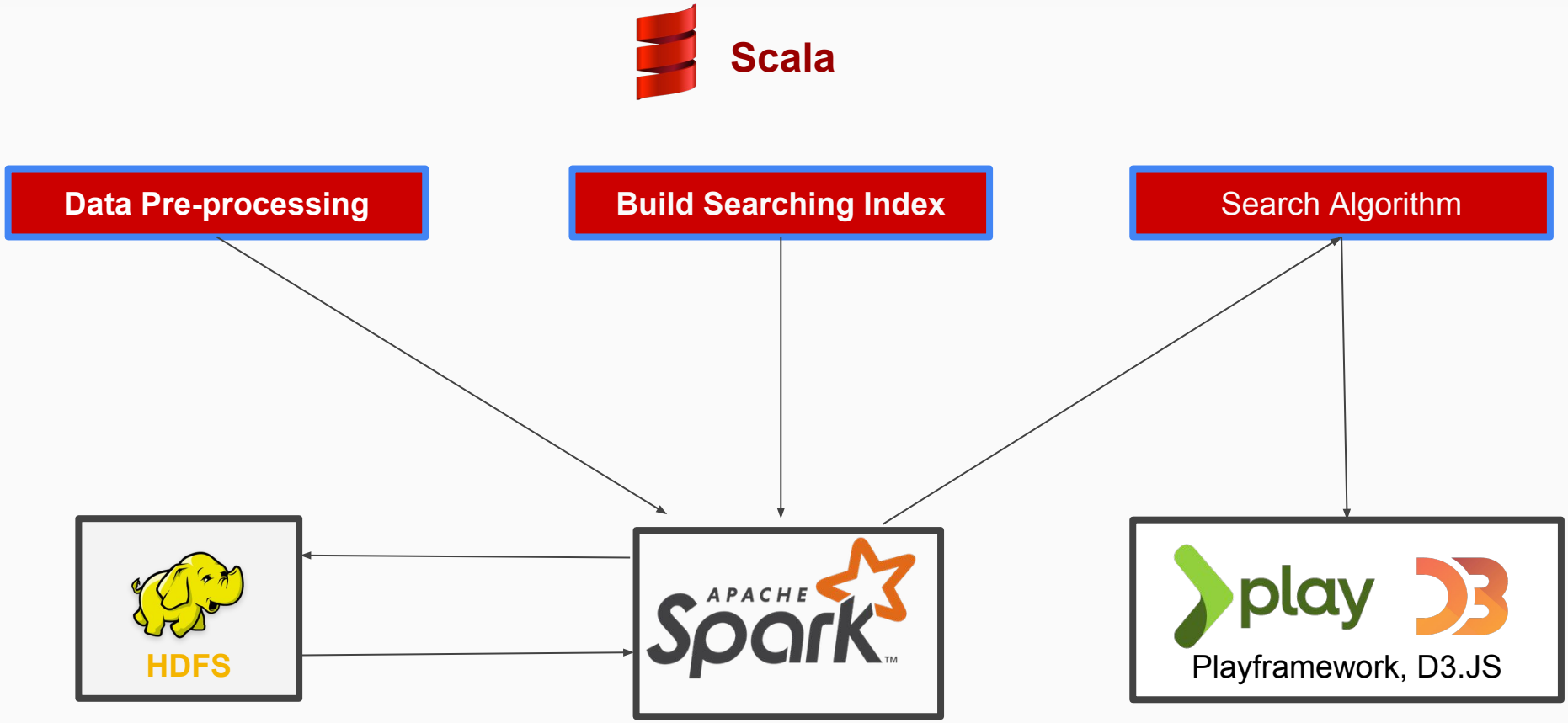
Goals of the project

- Research on the NCBI Taxonomy database structure and Implement efficient search method for sub-tree query;
- Implement Spark and GraphX(Data Visualization) to process the query and visualize the relationship of the nodes;
- Build search Index for querying the relationship of multiple nodes;
- Create web user interface for user by using D3.js(a JavaScript library for manipulating documents based on data).

Use Case


Search specific Species
Navigate to relevant sequence data
Search relationship of multiple species
User
View the database graphically

Methodology:



Methodology: Test Driven Development

Like we practiced in our assignment we will follow TDD methodology in the project development:

- Add a test
- Run all tests and see if the new one fails
- Write some code
- Run tests
- Refactor code
- Repeat

writes an (initially failing) automated test case that defines a desired improvement or new function, then produces the minimum amount of code to pass that test, and finally refactors the new code to acceptable standards.

Data Source: NCBI Taxonomy Database

Lineage: a group that can demonstrate their common descent from an apical ancestor or a direct line of descent from an ancestor

NCBI Taxonomy Database: A curated set of names and classifications for all of the organisms that are represented in GenBank.

- Aves (birds) *Click on organism name to get more information.*
  - Neognathae
    - Apodiformes
      - Apodidae (swifts)
      - Hemiprocnidae (tree swifts)
      - unclassified Apodiformes
    - Bucerotiformes
      - Bucerotidae
      - Bucorvidae
    - Caprimulgiformes
      - Aegothelidae
      - Batrachostomatidae
      - Caprimulgidae
      - Eurostopodidae
      - Nyctibiidae
      - Podargidae (frogmouths)
      - Steatornithidae
      - unclassified Caprimulgiformes
    - Charadriiformes (shorebirds)
      - Alcidae
      - Burhinidae
      - Charadriidae
      - Chionididae
      - Dromadidae
      - Glareolidae
      - Haematopodidae (oystercatchers)
      - Jacanidae (jacanas)
      - Laridae (gulls)
      - Pedionomidae
      - Recurvirostridae
      - Rostratulidae
      - Scolopacidae (snipes)
      - Stercorariidae
      - unclassified Charadriiformes

Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Inherited blast name: **primates**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
synonym: **humans**
common name: **man**
authority: **Homo sapiens Linnaeus, 1758**

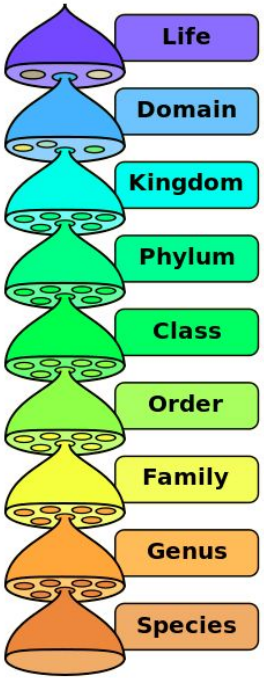
Lineage(full)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

Equus asinus

Taxonomy ID: 9793
Genbank common name: **ass**
Inherited blast name: **odd-toed ungulates**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
common name: **donkey**
common name: **domestic ass**
common name: **Somali wild ass**
common name: **African wild ass**
common name: **African ass**

Lineage(full)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Perissodactyla; Equidae; Equus; Asinus

←—hierarchical view

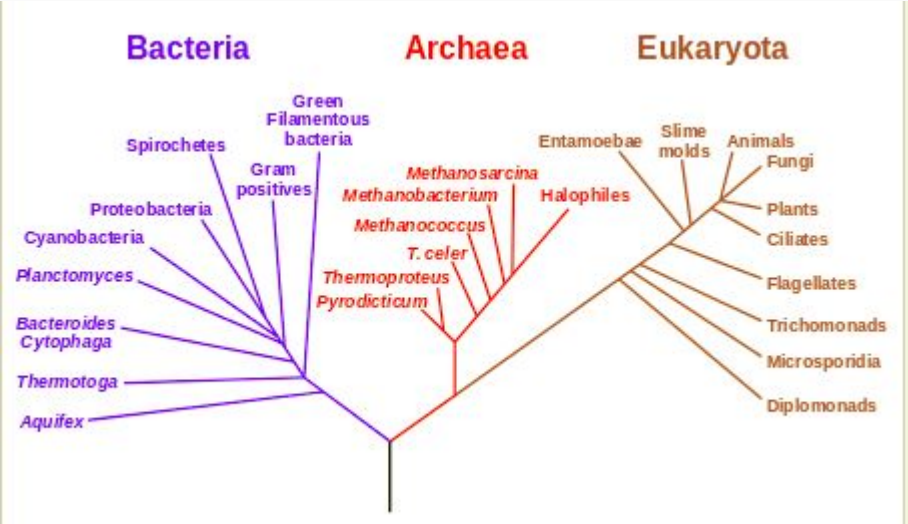


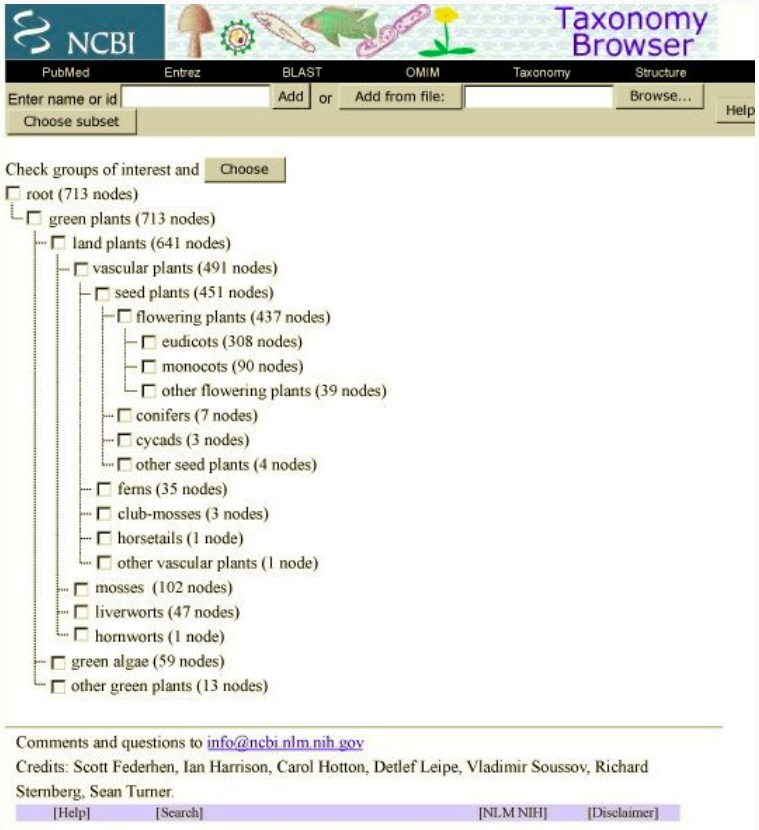
The hierarchy of biological classification's eight major taxonomic ranks. A genus contains one or more species.

Data source: Taxonomy structure - phylogenetic classification scheme

phylogenetic classification system names only clades — groups of organisms that are all descended from a common ancestor.

If two organisms (A and B) are listed more closely together in the taxonomy than either is to organism C, the assertion is that C diverged from the lineage leading to A+B earlier in evolutionary history, and that A and B share a common ancestor that is not in the direct line of evolutionary descent to species C.





Data Source:

Index of /pub/taxonomy

Name	Size	Date Modified
[parent directory]		
Ccode_dump.txt	48.8 kB	11/4/16, 9:31:00 AM
Cowner_dump.txt	1.4 MB	11/4/16, 9:31:00 AM
Icode_dump.txt	133 kB	11/4/16, 9:31:00 AM
accession2taxid/		10/30/16, 7:31:00 AM
coll_dump.txt	371 kB	11/4/16, 9:12:00 AM
gi_taxid.readme	1.3 kB	3/29/16, 12:00:00 AM
gi_taxid_nucl.dmp.gz	1.3 GB	10/31/16, 8:44:00 AM
gi_taxid_nucl.dmp.gz.md5	55 B	10/31/16, 8:44:00 AM
gi_taxid_nucl.zip	1.3 GB	10/31/16, 8:45:00 AM
gi_taxid_nucl.zip.md5	52 B	10/31/16, 8:45:00 AM
gi_taxid_nucl_diff.dmp.gz	2.9 MB	10/31/16, 8:44:00 AM
gi_taxid_nucl_diff.dmp.gz.md5	60 B	10/31/16, 8:44:00 AM
gi_taxid_nucl_diff.zip	2.9 MB	10/31/16, 8:45:00 AM
gi_taxid_nucl_diff.zip.md5	57 B	10/31/16, 8:45:00 AM
gi_taxid_prot.dmp.gz	813 MB	10/31/16, 8:44:00 AM
gi_taxid_prot.dmp.gz.md5	55 B	10/31/16, 8:44:00 AM
gi_taxid_prot.zip	799 MB	10/31/16, 8:45:00 AM
gi_taxid_prot.zip.md5	52 B	10/31/16, 8:45:00 AM
gi_taxid_prot_diff.dmp.gz	9.9 MB	10/31/16, 8:44:00 AM
gi_taxid_prot_diff.dmp.gz.md5	60 B	10/31/16, 8:44:00 AM
gi_taxid_prot_diff.zip	9.6 MB	10/31/16, 8:45:00 AM
gi_taxid_prot_diff.zip.md5	57 B	10/31/16, 8:45:00 AM
taxcat.tar.Z	7.7 MB	11/4/16, 6:37:00 PM
taxcat.tar.Z.md5	47 B	11/4/16, 6:37:00 PM
taxcat.tar.gz	5.5 MB	11/4/16, 6:37:00 PM
taxcat.tar.gz.md5	48 B	11/4/16, 6:37:00 PM
taxcat.zip	5.5 MB	11/4/16, 6:37:00 PM
taxcat.zip.md5	45 B	11/4/16, 6:37:00 PM

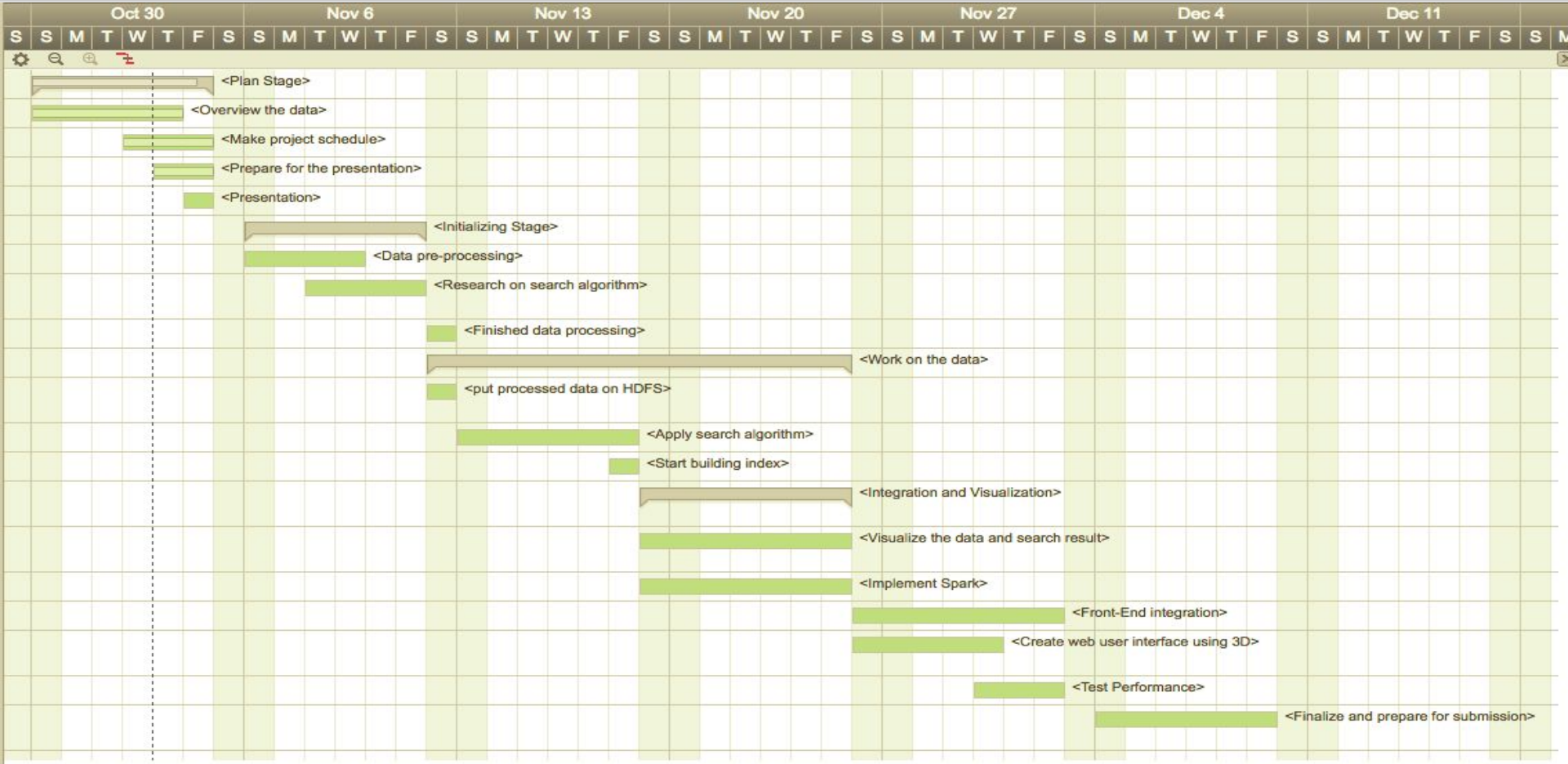
Taxonomy Nodes (all dates)

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	179	160	610	0	949
Bacteria	1.711	3.160	16.033	865	21.769
Eukaryota	22.385	76.016	355.197	25.942	479.540
Fungi	1.627	5.295	34.559	1.180	42.661
Metazoa	15.987	52.304	180.356	12.987	261.634
Viridiplantae	3.027	15.340	129.776	11.475	159.618
Viruses	686	540	2.174	0	3.400
All taxa	24.993	79.883	374.040	26.807	505.723

Dates: 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 all dates

Taxa: Customize Use Default

Milestones- Gantt Chart



Project Detail in Scala

We will be using Scala for…

- Data pre-processing
- Index building
- Search Algorithm implementation
- Web Service

Our working repository will be on Github to have version control
https://github.com/Houzl/CSYE7200FinalProject

Acceptance Criteria

- Have at least 2 search algorithms for sub-tree query.
- For querying the data, it should be quick for user to use. User should get the result in 2 secs.
- The relationship of the target node should be visualized clearly and accurately for its parent node, sibling nodes and children nodes from 505,723 nodes.

Thank you !