Related Work

Ryan Verdon

March 14, 2012

1 Introduction

None of the papers I found while researching my topic were completely related to my topic. The vast majority covered only small pieces of my problem. In the sections below I will go over related work.

2 Current state of biological databases

The papers below examine the current state of biological databases.

Database survey In 2003, a survey was done to examine what types of databases were running.[4] The survey found that of the 111 databases they sampled 40-44 were collections of flat files and 41-42 were implemented as relational databases.[4] Interestingly, the authors also noticed that the vast majority of databases had hypertext references to other databases.

Database list There is a journal that has been keeping a list of useful biological databases.[7] In the 2008 version there was over 1000 unique or interesting databases. Interestingly in 2001 there was only 281.

Biozon Biozon examined what biologists have to go through to do research.[3] The paper looked at how data stored in a database can be highly related to data stored in different databases. The paper also examined how multiple queries of different databases have to be done to complete research.

3 NoSQL databases

A major portion of my thesis will revolve around distributed databases. This section contains all of the distributed databases I am considering.

Cassandra Cassandra is an open-source, highly available, eventually consistent database modeled after Google's BigTable and Amazon's Dynamo.[6] Cassandra is a mix between key-value store and column-oriented database.

HBase HBase is an open-source, distributed, versioned, column-oriented database modeled after Google's Bigtable.[8] HBase ties in closely with Google's map-reduce framework. So close in fact that tables in HBase can be used as input and output of map-reduce jobs.

MongoDB MongoDB is a document-based NoSQL database created by 10gen. MongoDB supports ad hoc queries with nested fields, indexing, replication and map-reduce support..[2]

SimpleDB Amazon SimpleDB is a highly available, flexible, and scalable non-relational data store that is sold as a service.[1] Amazon charges for resources consumed in storing the data and serving requests. SimpleDB provides excellent flexibility allowing schema changes on the fly.

4 MapReduce

Data processing is a very big part of bioinformatics. The database I design will require an extremely fast implementation of BLAST (Basic Local Alignment Search Tool). To do that I researched the MapReduce framework.

MapReduce MapReduce is a framework Google created for processing large data sets in parallel.[5] The idea originally came from the notion of mapping a function over a list of data. This is commonly found in functional languages.

CloudBlast CloudBlast is an implementation of BLAST that uses the MapReduce framework in combination with the Hadoop file system.[9] The authors were able to get better performance than all of the other attempts at parallelizing BLAST.

References

[1] "Amazon simpledb," http://aws.amazon.com/simpledb/. [Online]. Available: http://aws.amazon.com/simpledb/

[2] "Mongodb." [Online]. Available: http://www.mongodb.org/

[3] A. Birkland, "BIOZON: a hub of heterogeneous biological data," *Nucleic Acids Research*, vol. 34, pp. D235–D242, Jan. 2006. [Online]. Available: http://nar.oxfordjournals.org/content/34/suppl_1/D235.abstract

[4] F. Bry and P. Krger, "A computational biology database digest," 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=607994

[5] J. Dean and S. Ghemawat, "MapReduce," *Communications of the ACM*, vol. 51, p. 107, Jan. 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1327492

[6] D. Featherston, "Cassandra: Principles and application." [Online]. Available: http://dfeatherston.com/cassandra-cs591-su10-fthrstn2.pdf

[7] M. Y. Galperin, "The molecular biology database collection: 2008 update," *Nucleic Acids Research*, vol. 36, no. Database issue, pp. D2–D4, Jan. 2008, PMID: 18025043 PMCID: 2238887.

[8] A. Khetrapal and V. Ganesh, "Hbase and hypertable for large scale distributed storage systems." [Online]. Available: http://www.uavindia.com/ankur/downloads/HypertableHBaseEval2.pdf

[9] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363 –1369, June 2009. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/25/11/1363.abstract