High Performance Data Analytics with Heterogeneous Cancer Data in a Scale-Up Property Graph

Terry Antony[1], Sarah Bullard[1], Nicole Sabatelli[1], Shawn Ulmer[1], Doron Levy[1], Tianwen Chu[2], Frank D'Ippolito[2], Michael S. Atkins[2]

[1]University of Maryland, College Park, MD, [2]FedCentric Technologies, LLC., College Park, MD

Background

With advancements in high-throughput sequencing, genomic data from individuals around the world is rapidly being uploaded to a variety of databases, stored on cloud servers, and kept on localized computers. Genomic data for millions of individuals can be petabytes and even exabytes in size. Data this size is difficult to manage and genomic researchers need methods to easily and rapidly store, upload, query, and analyze it. Discoveries of patterns and relationships in data is crucial in understanding genetic diseases, such as cancer, but it is especially relevant in precision medicine, which utilizes the knowledge and insights contained in biomedical information to customize individual therapies and treatment options.

Introduction

Traditional relational databases are not optimized to perform specific, complex queries that Precision Medicine researchers are interested in. Structured query language (SQL) relational databases are not ideal to handle the volume or diversity of biomedical data that is available. Resource description framework (RDF) triples are a type of graph structure that comprise of subject and object nodes, and a predicate edge. While more flexible than relational databases, RDF triples tend to still be rigid and lead to database size increases.

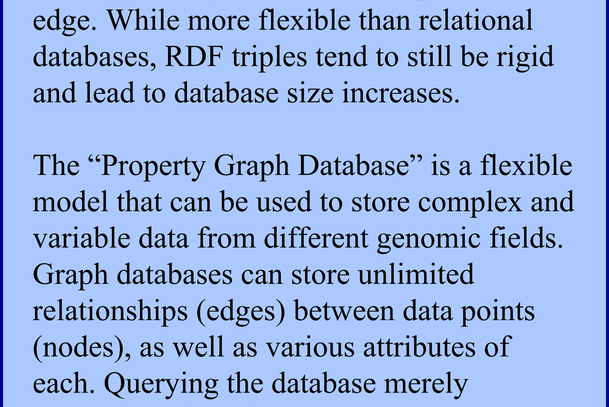
The "Property Graph Database" is a flexible model that can be used to store complex and variable data from different genomic fields. Graph databases can store unlimited relationships (edges) between data points (nodes), as well as various attributes of each. Querying the database merely accesses the list of relationship records, eliminating the need for a lengthy search and match computation. Graph databases do not require a pre-set schema and any amount of new data can be arbitrarily incorporated with ease.

To showcase the compatibility of a property graph database with biomedical data, we stored and analyzed nearly all the publicly available from The Cancer Genome Atlas (TCGA) data.

Methods and Materials

Hardware

Fig 1: Scale Up vs. Scale Out



Scale Up
- One large server
- Scales up by adding memory and CPUs to the same system
- Blades are linked with high speed internal bus

Scale Out
- Cluster of small servers
- Scales out by adding more physical servers
- Servers are linked through external network connection

Methods and Materials

Software Architecture - Apache Spark
- Big data processing framework
- Driver node manages parallel operations carried out by executor nodes
- Implements graphs through its GraphFrame data structure

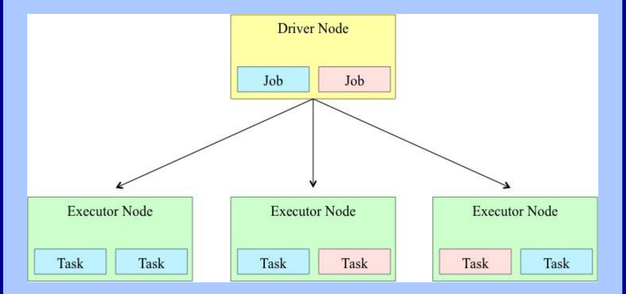


Fig 2: Apache Spark

Data
- Over 500GB of The Cancer Genome Atlas (TCGA) data from 33 Cancer Project for over 11,000 patients (Fig. 1)
- Clinical Information & VCF files* from NIH GDC
- Methylation, miRNA expression, & Copy Number Variation from ICGC
- Gene Expression from UCSC
- Gene names and additional information from Ensembl
- Probe ID from Illumina

File/Data Type	Size (GB)	Donors	Website
Clinical	0.509	11,160	GDC
Methylation	461.22	8,467	ICGC
Gene Expression	3	9,775	UCSC
miRNA Expression	3.14	8,062	ICGC
Copy Number Variation	2.17	8,842	ICGC
VCF Files	39.91	10,429	GDC
Gene Names	1.52	N/A	Ensembl
Probes	0.197	N/A	Illumina
Total	511.67	N/A	N/A

Fig 3: File Information

The Graph
- All data types mapped to gene name and/or chromosome (Fig. 2)
- Includes over 20,000 genes and all chromosomes
- Over 57 million nodes and 3.5 billion edges

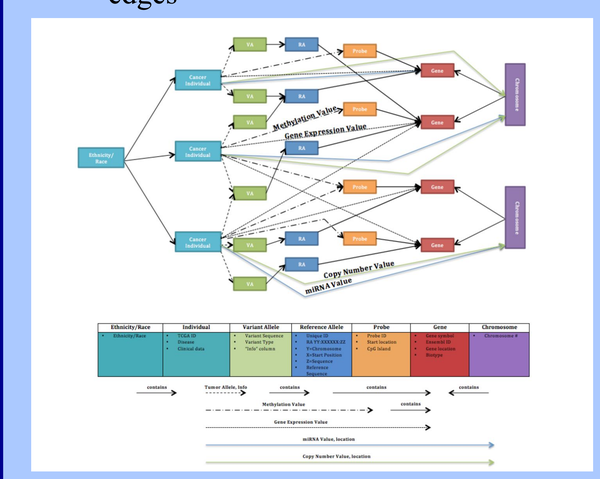


Fig 4: The Graph Model

*NIH Granted Access Data

Results

Ingestion
- Spark provides multithreaded parse and write functionality
- Ingestion and write to parquet takes 50 minutes using 192 cores
- Total size on disk is 3.4 GB compressed
- Total size in memory is 171 GB

Results

Simple Queries

Examples (Fig. 3)
- All information for one gene (IDH1)
- Percentage of patients that died due to GBM
- Most common gain/loss copy numbers for chromosomes

Performance Speeds (Fig. 3)
- Ability to derive subgraphs in seconds

	Query	Query Subgraph	Query Process	Time (sec)
1	How many gains (positive Copy Number Variations) occur within a location on a chromosome	Individual -> Chromosome	Chromosome = 1, Start Location < 1000000, Copy Number Variation > 0	0.54
2	All information for one gene - IDH1	Gene	Gene Name = IDH1	1.34
3	Percentage of patients that died from GBM	Individual	Disease = GBM, Life Status = Dead	1.88
4	Number of patients under the age of 30	Individual	Age < 30	2.82
5	All clinical data for one patient - TCGA-HT-A614	Individual	TCGA ID = TCGA-HT-A614	3.36
6	How many probes are associated with a certain gene - IDH1?	Probe -> Gene	Gene Name = IDH1	7.25
7	How many genes are on each chromosome?	Gene -> Chromosome	Group by Chromosome	12.88
8	Most common gain/loss copy numbers for chromosomes	Individual -> Chromosome	Copy Number Variation < 0, Copy Number Variation > 0, Group By Chromosome	15.57
9	Patient race and ethnicity background	Population -> Individual	Group By Race and Ethnicity	18.39

Fig 5: Simple Queries

Complex Queries

Examples (Fig. 4)
- Top 5 most commonly mutated genes & diseases
- All information for one gene for one patient
- Most commonly overexpressed genes

Performance Speeds (Fig. 4)
- Ability to scale up to 256 parallel processes
- Query times in seconds

	Query	Query Subgraph	Query Process	Time (sec)
1	Percentage of patients that have that a gene expression over a value (0.75) / under a value (-0.75) for IDH1	Individual -> Gene	Gene Name = IDH1, Gene Expression > 0.75, Gene Expression < -0.75	50.47
2	Gene that is most commonly highly expressed	Individual -> Gene	Group By Gene Expression With Average Aggregate	92.21
3	Top 5 most commonly mutated diseases	Individual -> Allele	Group By Individual, Group By Disease	112.3
4	Number of identical mutations across BRCA and all diseases	Individual -> Allele	Group By Mutation	125.63
5	Percentage of patients with an over/under expressed gene	Individual -> Gene	Max Individual Gene Expression > 0.75, Min Individual Gene Expression < -0.75	165.62
6	Top 5 most commonly mutated genes	Individual -> Allele -> Gene	Group By Gene Name	174.81
7	How often is IDH1 mutated across diseases and races	Population -> Individual -> Allele -> Gene	Gene Name = IDH1, Group By Disease, Group By Race	195.15
8	All information for one gene for one patient - TCGA-HT-A614, IDH1	Individual -> Probe -> Gene	TCGA ID = TCGA-HT-A614, Gene Name = IDH1	354.2

Fig 6: Complex Queries

StepMiner2D
- Bivariate algorithm determines thresholds for two data types at once
- Hypergeometric test to find association in data types
- Ran using methylation and gene expression data for 200 breast cancer patients (Fig. 5)
- Took ca. 16 minutes to run
- Very low p-value (high correlation)

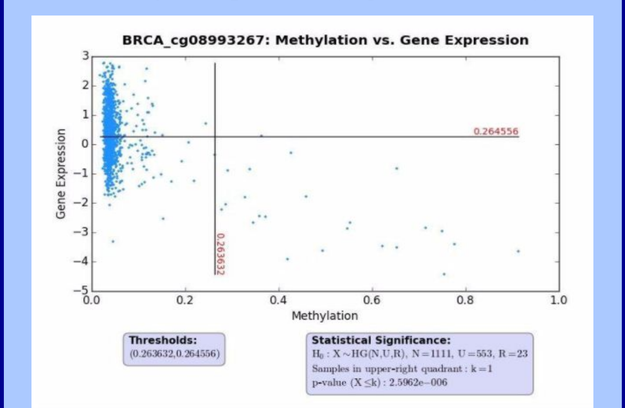


Fig 7: StepMiner2D Results

Clustering
- K-means clustering using Euclidean distance
- Clustered 2,275 patients by disease using 500 genes (Fig. 6)
- Over 99% accurately clustered
- Took ca. 1 minute to run

Results

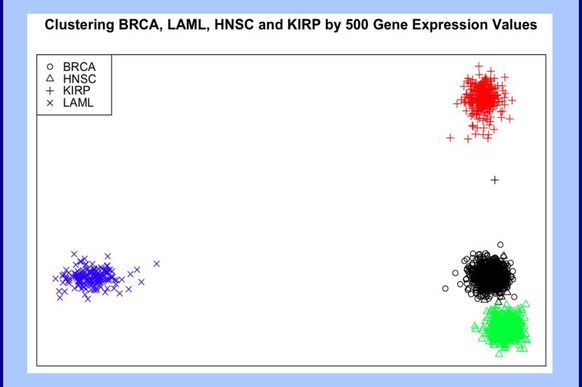


Fig 8: Clustering Results

Conclusions

In conclusion, FedCentric was able to successfully build a property graph database containing a variety of TCGA data, while storing the entire dataset in memory on a scale-up architecture to reduce latency and computational time. We were able to run simple and complex queries to find relationships, extract information, and cluster data types quickly.

This graph database gives oncology researchers the ability to make novel discoveries and find unique relationships between different types of data both quickly and seamlessly. These discoveries can lead to substantial advancements in Precision Medicine.

Future Work

With graph model's flexibility and capacity for intricacy, the team will use graph methodology for fusing different types of biomedical data. This methodology can be used to study cancer diagnosis, treatment, and survivorship (such as breast cancer, prostate cancer, and blood cancers).

Many researchers and doctors have a plethora of data generated daily that has the capacity to be stored and analyzed using graph technology. FedCentric can team with clinicians and experimentalists to study the applicability of graph technology in a variety of medical problems to find scientifically accurate relationships and connections deemed most useful in real-time.

References

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., Staudt, L. M. (2016) Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine 375:12, 1109-1112.

Goldman, M., Craft, B., Swatloski, T., Ellrott, K., Cline, M., Diekhans, M., Ma, S., Wilks, C., Stuart, J., Haussler, D., Zhu, J. The UCSC Cancer Genomics Browser: update 2013. Nucleic Acids Research 2012; doi: 10.1093/nar/gks1008.

Liu, Y., Ji, Y., & Qiu, P. (2013). Identification of thresholds for dichotomizing DNA methylation data. *EURASIP Journal on Bioinformatics and Systems Biology*, *2013*(1), 8. http://doi.org/10.1186/1687-4153-2013-8

Acknowledgements

The authors would like to acknowledge the International Cancer Genome Consortium for providing data found at https://icgc.org/.

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.