A Scalable and Secure Genome Archiving and Communication System for the Clinical Enterprise

Rajeswari Swaminathan[1], Yungui Huang[1], Eric Yu[1], James Fitch[2], Ron Buckley[1]
Katherine Lintner[1], Peter White[2,3], Simon Lin[1,3]

1) Research Information Solutions and Innovation, The Research Institute at Nationwide Children's Hospital, Columbus, OH
2) Center for Microbial Pathogenesis, The Research Institute at Nationwide Children's Hospital, Columbus, OH
3) Department of Pediatrics, College of Medicine, The Ohio State University, Columbus, OH

NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.℠

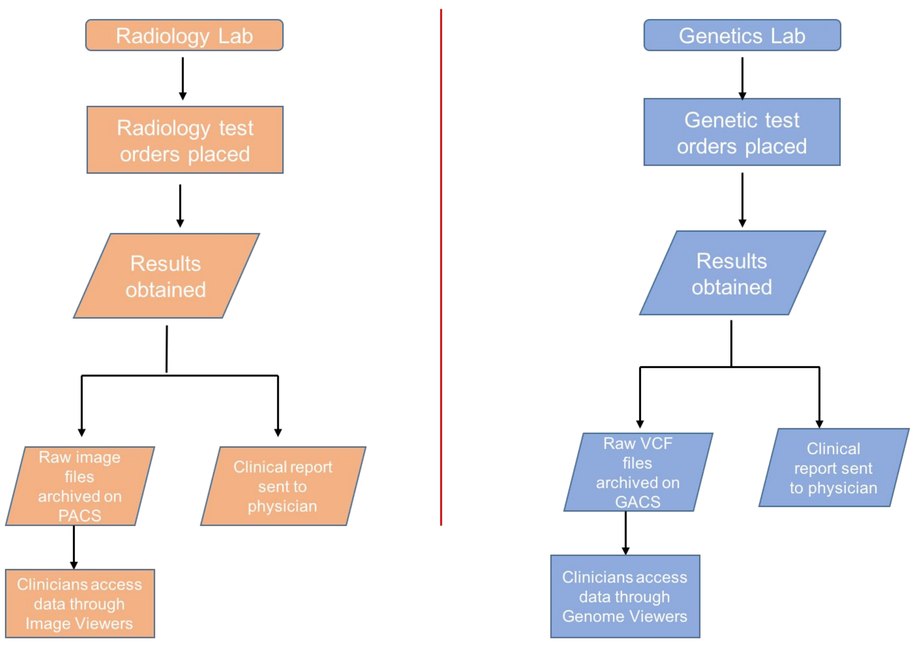
Research Information Solutions and Innovation Department

The Research Institute at Nationwide Children's Hospital

Abstract

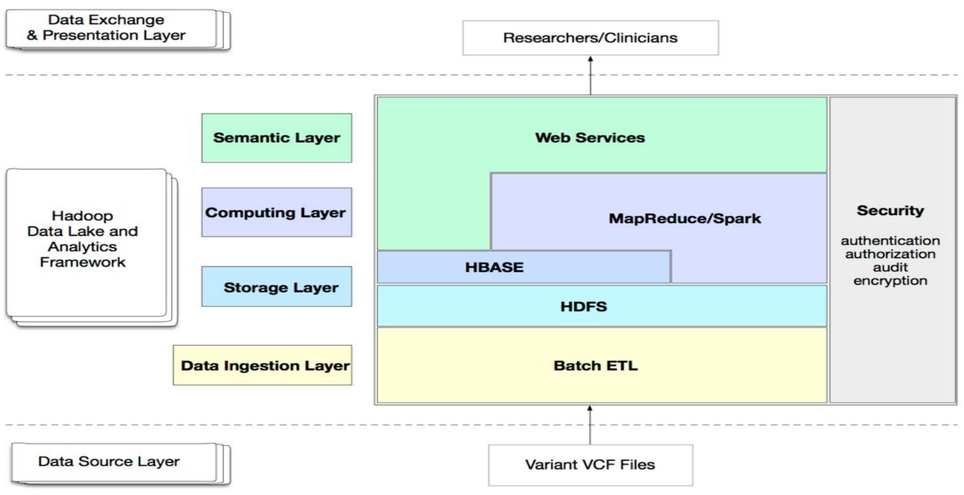
The plummeting costs and increased diagnostic yield of next generation sequencing technologies have resulted in its increased use for clinical diagnostics. The amount of sequencing data that humans as well as computer systems need to handle has risen tremendously. Genetic tests result in a raw data file, in VCF format, around 500-750 MB in size, as well as a text based clinical report, that summarizes information from within the VCF file. Being able to store, share and query this exponentially growing information is the biggest challenge that organizations are facing today. To overcome this, we are utilizing the strengths of the Hadoop framework, harnessing its distributed storage and processing capabilities, to build a Genome Archive and Communications system (GACS) within our institution. Since the primary use of GACS is for clinical genomics, data ingestion and query performance are key factors that need to be optimized. We also measure the performance of GACS by being able to store 1 million simulated whole genome VCF files as well as query variant level information from it, in near real time.

Background



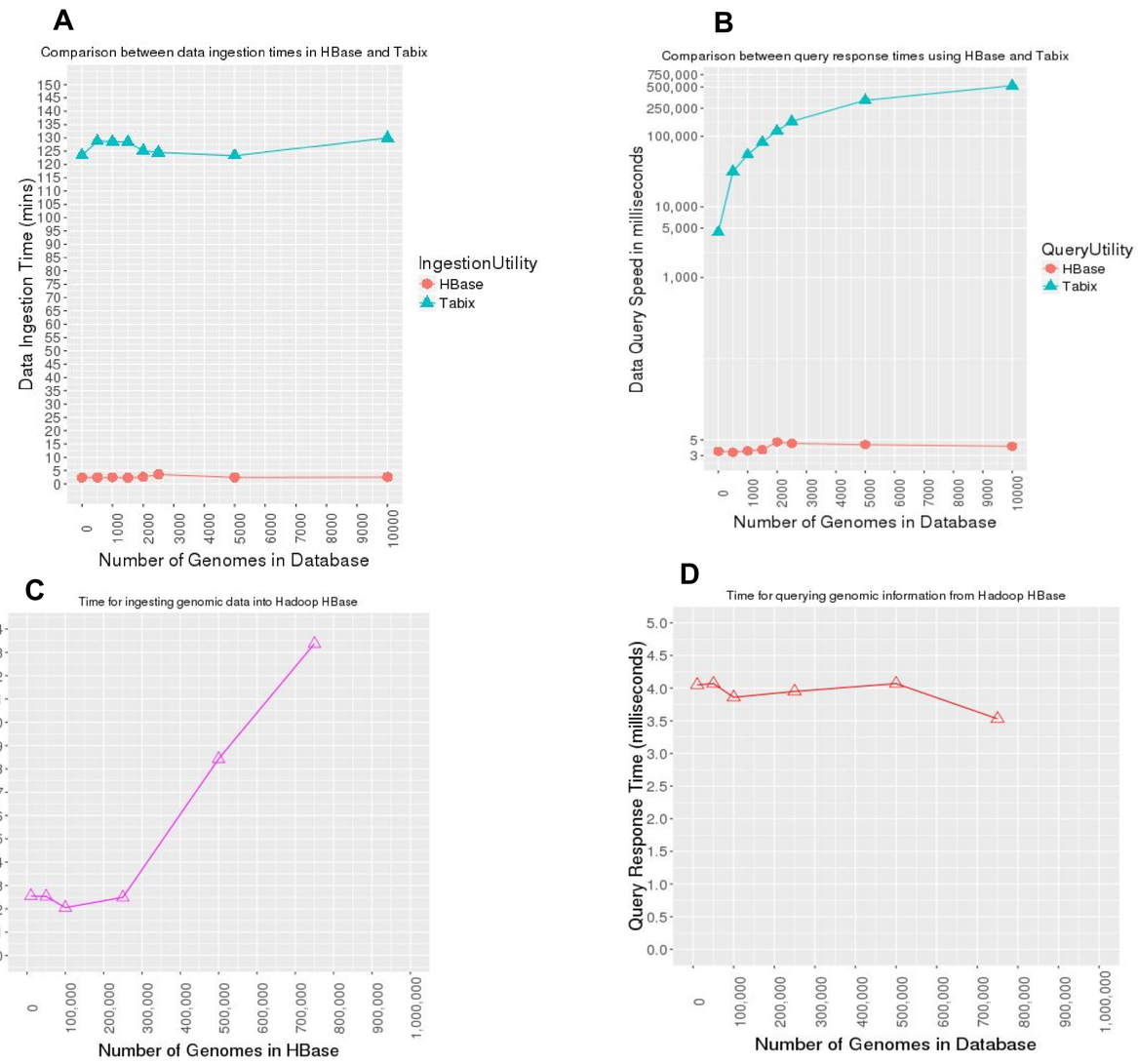
Similar to how PACS (Picture Archive and Communications System) is used to store radiology image data that can be accessed and shared across clinicians, we are putting together GACS (Genome Archive and Communications System), that can be used to store, share and query genomic information within our institution.

GACS Architecture and Design



The GACS architecture is organized into three major functional layers, data source layer, data source/analytics layer and data presentation layer. The data source layer consists of genomic information in various formats such as sequence data as FASTQ, alignment files as BAM, variant files as VCF. The data from the data source layer is first ingested into HDFS (Hadoop Distributed File System) and in the next step loaded into HBase for real time querying. Genomic APIs help present the data within HDFS and HBase through standard programmable interface and web interface.

Hadoop Performance Metrics



A: This figure tries to compare data ingestion times using HBase and Tabix utility. The x-axis represents the increasing size of the database (as number of genomes) and the y-axis denotes the time taken to ingest a batch of 100 files (in minutes) at different database sizes. Since tabix is a file-based utility, a database is nothing but a directory in the unix file system. For tabix, the ingestion time comprises of moving a given number of files into a folder, sort, compress and index them. The database sizes range from 0 to 10,000. The reason to limit at 10,000 is the increasing ingestion time with tabix. From the graph, we see that the ingestion times are fairly stable with increasing database sizes, for both HBase and Tabix, although Tabix consumes more time for each ingestion process.
B: This figure tries to compare query response times between HBase and Tabix, with increasing database sizes. The query executed is to extract all samples that contain the variant searched for. Since a system can have multiple users accessing the database at any given time, the query response time is the average time when having 10 parallel users accessing the data. The x-axis represents the database(in terms of number of genomes) and the y-axis represents the query response time in milliseconds.
C: One of the key points of GACS is its scalability to large data sets. This figure shows the scalability of the Hadoop framework to store close to 1 million whole genomes. The x-axis represents the increasing size of database, moving from 50,000, all the way to 750,000 and the y-axis represents the time taken to ingest a batch of 100 genomes at different database sizes. With a relational database, we typically notice a latency as the database sizes grows and that is where Hadoop comes to the rescue with its unprecedented scalability
D: This figure represents the time taken to query for a variant with increasing size of HBase. The x-axis represents the increasing database size (as number of genomes), going from 50,000 to 750,000 and the y-axis represents the query response time when executing 10 parallel queries.

Privacy and Clinical Data Security

- The Hadoop cluster is hosted on a secure private network within institutional firewalls
- Every user needs to be authenticated through Kerberos, implemented on the cluster
- Authorization is maintained through Ranger
- Data is encrypted, both at rest and in transit using AES encryption
- Have the system meet compliance requirements in order to host clinical genomics data

Discussion

- One of the first genomics systems with a capacity to work with millions whole genomes
- Overcomes some of the challenges associated with file-based or relational database systems
- It meets all requirements of security for storing clinical data in the event genomics data stats getting treated as PHI
- System is also suited to be in a "data lake" with clinical data, claims, etc.
- Given future needs, the system can be deployed and functional in the cloud too

References

- Heath, A.P., et al. (2014). Bionimbus: A cloud for managing, analyzing and sharing large genomics datasets. J Am Med Inform Assoc **21**(6): 969–975.
- O'Driscoll, A., et al. (2013). Big Data', Hadoop and cloud computing in genomics. Biomed Inform **46**(5): 774–781.
- Siegel, E., & Brown, A. (1994). Preliminary impacts of PACS technology on radiology department operations. *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 917–921.