Amazon Writing Sample

What is the Most Inventive or Innovative Thing You've Done?

Russ Weeks
Sept. 27 2016

One of the most innovative and personally fulfilling projects that I've led is the design and implementation of a distributed database optimized for the storage and indexing of genomic data.

I am by no means a bioinformatician but for the purposes of this project it was sufficient to understand a few terms. Your **genome** is the total of your genetic material. It consists of a very long sequence of **alleles** represented by the characters G,T,A and C measured across a set of **chromosomes**. Some chromosomes are represented by numbers (1-22); some are represented by letters (X,Y).

Physicians and researchers are interested in the ways in which a person's genome differs from a reference human genome – these are called variant observations or **variants**. There are 4-5 million variants in your genome. The vast majority of these variants are meaningless – in fact, many are unique to you and have never been observed before. These variants are the noise in your genomic signal. Other variants are important or **clinically significant**. Some of these variants confer benefits: for example, the deletion of 32 alleles around position 46,000,000 on chromosome 3 may give immunity to the HIV virus. Unfortunately, many significant variants are **pathological**: they may lead directly to a disease like muscular dystrophy or cystic fibrosis, or they may put the carrier at increased risk of disease like Alzheimer's or obesity. Variant metadata such as clinical significance and allele frequency is available through a collection of public **reference datasets**.

Physicians and researchers study clinically significant variants from different perspectives: a physician wants to understand the variants in a patient's genome in order to plan the best course of treatment, and a researcher wants to understand what variants are present in an existing cohort of subjects. The technical challenge I faced was to design and build a genomic database that could satisfy both these access patterns while scaling horizontally to hundreds of thousands of patients.

The first design question I faced in this project was the selection of a distributed database, since the dataset was clearly too large for a centralized solution. The database needed to provide sub-second responses to narrowly-scoped queries such as "does this patient have a variant at chromosome Y, position 11134340" and interactive (1s - 10 min) responses to more broadly-scoped queries such as "show me all variants in the `FOXRED1` gene found in cancerous lung tissue". I needed to be able to integrate the database with the Spark distributed computing framework and I also needed to enforce strict access control rules due to the confidential nature of the data. The Apache Accumulo distributed key/value store was a great fit for these requirements. Accumulo is an open-source implementation of Google's BigTable data structure. It features a flexible and powerful distributed processing

stack which allowed me to prune results for many queries at the server-side, and it has a very mature and robust cell-level security model.

Having settled on a key/value store, the next design challenge was to determine an optimal key schema. It was clear that the keys would represent variant observations; it was also clear that a variant could be identified by the tuple of values $(patient\_id, chromosome, position, reference\_allele, alternate\_allele)$. What wasn't clear was the order of these components within the key. Conventional genomic databases, which are oriented towards the researcher's workflow, effectively put *patient_id* at the end of the key. This optimizes for research-oriented queries like "show me all patients with a known variant" but makes it nearly impossible to answer patient-oriented queries like "show me all clinically significant variants in this patient's `CCR5` gene". I can't disclose the solution we arrived at but through a combination of schema design and server-side processing we were able to satisfy both the researcher and physician query patterns.

Another challenge I faced was related to annotating variant observations. For example, the DBSNP dataset consists of 160M "known" variants that have been identified and catalogued by researchers around the world. When a new patient genome is ingested by the system, all 4-5M variants should be annotated with metadata from known public datasets. I implemented this denormalization because *(a)* our use of commodity hardware means that storage is relatively cheap, *(b)* genomic data is immutable, which means that the cost of processing once at write-time can be amortized across many reads, and *(c)* the alternative, which is an asymmetric join at read-time between two large datasets, is prohibitively expensive at this scale.

One especially frustrating aspect of the annotation process is that an inner join between a patient genome and a reference dataset produces very few results. Cross-referencing a full set of ~5M variant observations against the 160M catalogued observations in DBSNP may yield only ~1K annotations. Since the DBSNP database itself is so large that it must be distributed, every lookup involved an RPC and the vast majority of the lookups produced no match. I mitigated this problem and improved ingest performance by 28x by condensing the DBSNP dataset to a Bloom filter. The Bloom filter was small enough to be kept in RAM on all the worker nodes during the ingest workflow and avoided 99.99% of useless network RPCs.

This project was technically innovative due to the sheer volume of the data being processed as well as the variety of access patterns that we needed to support. More than that, it was meaningful to me because it was an opportunity to contribute to the important work that bioinformaticians, researchers and primary caregivers are doing all over the world to understand the effects of our genome on our physical and mental well-being.

There were many more interesting and challenging aspects of this project, for more information please check out my talk at this year's Accumulo Summit.