Geodistributed Analytics using Spark

Qifan Pu, Frank Austin Nothaft
{qifan, fnothaft}@berkeley.edu

Motivation

Many analytics workloads can benefit from geographically distributed processing:

1. **"Edge" Clusters:** It is becoming common to place small clusters near data sources (app users) to improve latency for interactive queries, but still need to process data at edge node with batch queries
2. **Very Large Datasets:** Datasets that are too large to process efficiently in a single datacenter need to be distributed
3. **Geodistributed Datasets:** Data sources may be geographically distributed (data acquisition instruments for scientific experiments)

Approach

Tech Specs:

- Built in Scala on top of Parquet and BDAS Spark
- Leverages new ADAM read/pileup/variant call format
- Scalability well past 30+ nodes; other pipelines are limited to 26 (1/chromosome)

Pipeline:

Design Principles:

- Use mapping quality/coverage as filtering heuristic
- Use assembly methods on high complexity regions
- Design is modular: easy to add new calling algorithms

Performance

Notes:

- Algorithm is currently disk bound due to shuffles: performance bug in pileup creation due to partitioning
- Plan to fix performance bug by doing interval-based rod conversion:
  - Lump reads by reference position group to maintain locality
  - Fewer objects created than reads $\rightarrow$ pileups $\rightarrow$ rods

% Reads in High Complexity Region

Performance Over Different Datasets

Applications

For calling SNPs on a single sample, we look at genome loci that show evidence of a SNP (at least one non-reference base). Genotype likelihoods are calculated by:

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} (m-g)\epsilon + g(1-\epsilon) \prod_{j=l+1}^{k} (m-g)(1-\epsilon) + g\epsilon$$

m = ploidy, g = genotype state, ϵ = likelihood of error,
l = bases matching reference, k = bases at locus

Genotyping is biased towards the reference. We compensate by the allele frequency and call a non-reference genotype if $g \in (1, 2)$ has the highest probability.

Future Work

For a few samples, one may look-up the MAF ϕ in a reference and compensate the the single sample likelihood

$$\hat{g} = \arg\max_{g} \mathcal{L}(g)\mathbf{P}(g|\phi)$$

When many samples are collected it can be desirable to compute a population MAF while performing genotype calling. For each SNP a, this is done via EM:

$$\phi_{a,t+1} = \frac{1}{M} \sum_{i=1}^{N} \frac{\sum_{g_i} g_i \mathcal{L}(g_i)\mathbf{P}(g_i|\phi_{a,t})}{\sum_{g_i} \mathcal{L}(g)\mathbf{P}(g|\phi_{a,t})}$$

$M = \sum_i m_i$ = total number of chromosomes N = number of individuals