CanDIG's Evolving Search Backend

CanDIG

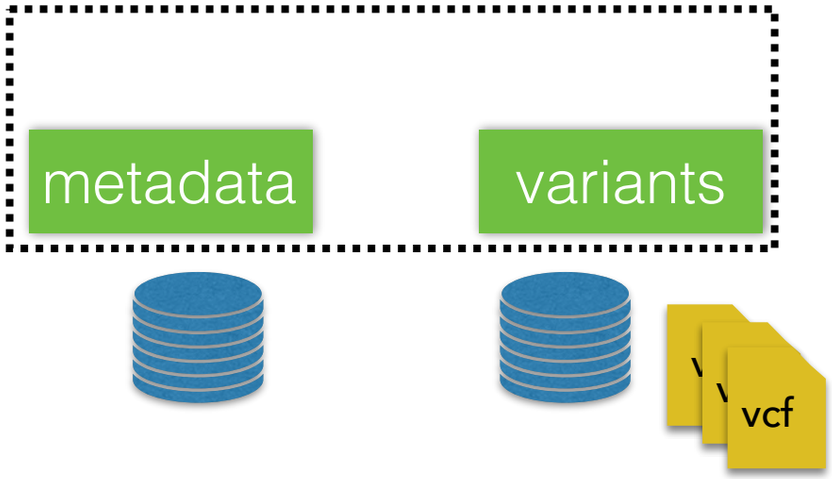
National Analysis of Distributed Private Genomic Data

CanDIG v1 Prototype

- Based on UCSC pre-GA4GH Connect “genomics API server”
- Used to validate federation approach for Canadian national projects

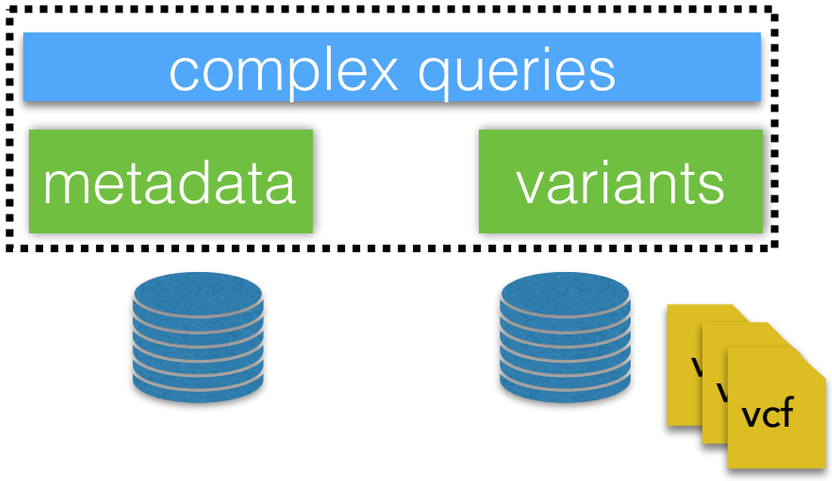
CanDIG v1

- Based on UCSC pre-GA4GH Connect “genomics API server”
- Prototype used to validate federation approach for Canadian national projects
  - PROFYLE
  - TF4CN
- Demonstrating feasibility



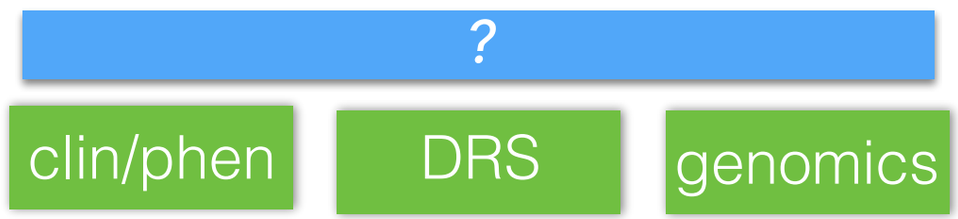
CanDIG v1

- Custom metadata schema, aimed for cancer projects
- Simple custom query language for querying across data types
- Adding RNA expression, imaging metadata, other services…
- Don’t want to keep manually updating query engine

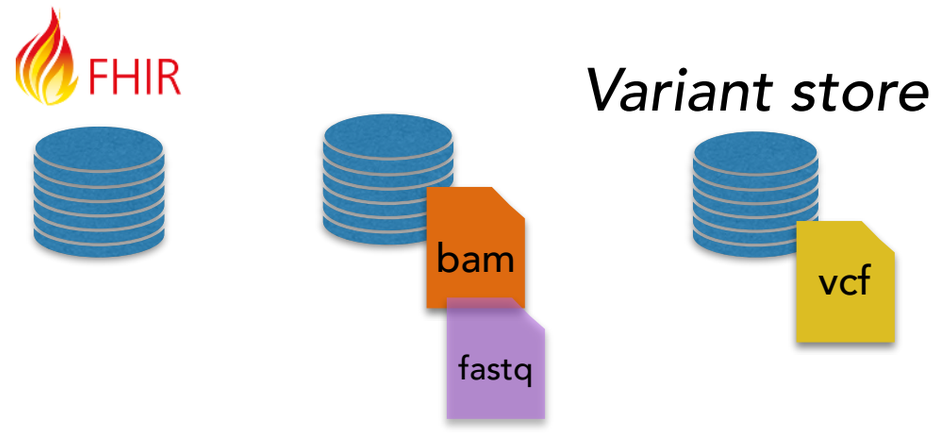


CanDIG v2 Prototype

- Adopting FHIR for clin/phen
- GA4GH DRS for data catalog & pipeline/ provenance info

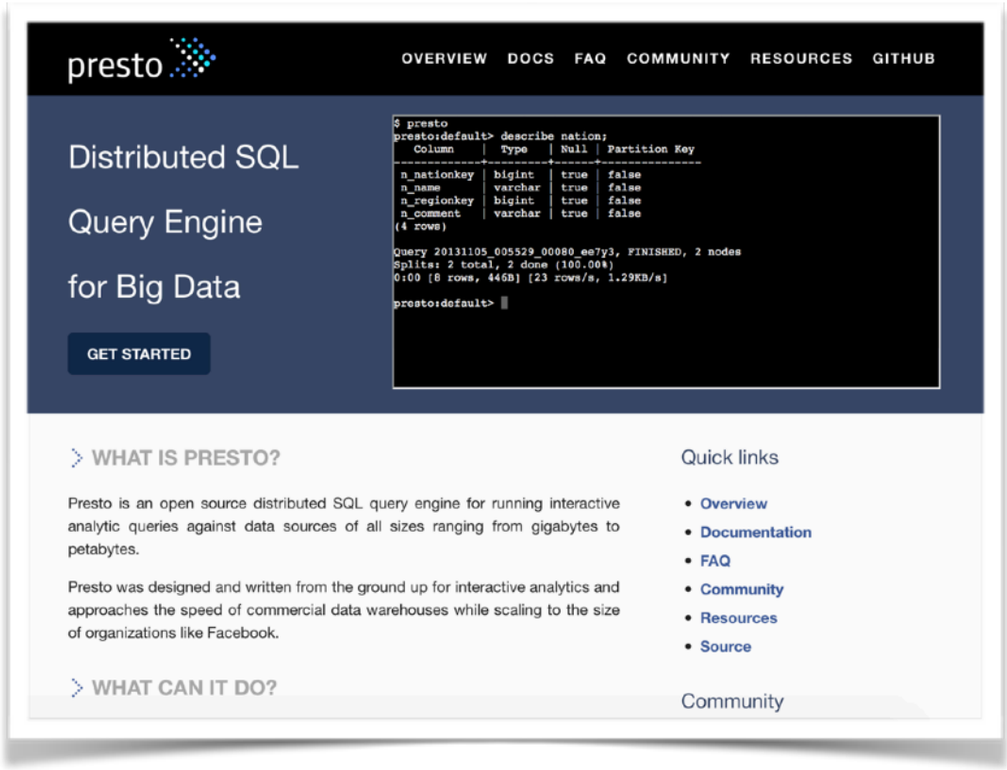


- Currently auditioning tool to stitch cross-service queries together
  - https://github.com/CanDIG/presto-graphql-testbed



PRESTO

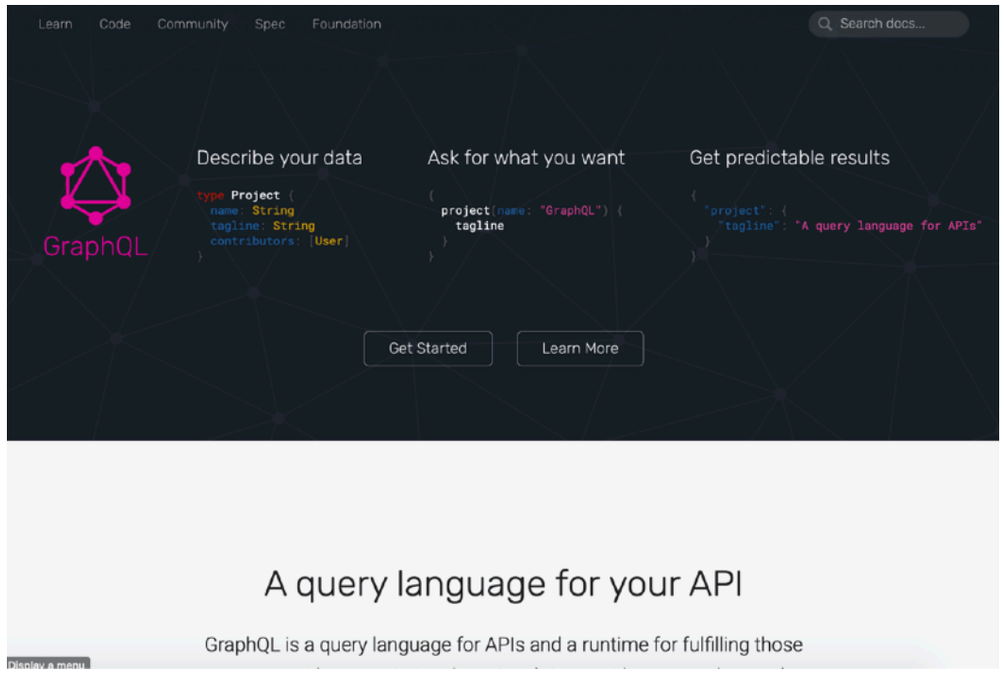
- SQL engine that can combine across multiple data sources
- Plugins for most databases
- Can configure API access
- Once connectors + schemas configured, can do regular SQL - full queries right away
- Like SQL, nested queries awkward
- Right now just connecting directly to underlying databases



https://prestodb.github.io

GraphQL

- Query language & tool
- Aimed particularly at graph data - complex connections, nested queries
- Have to set up schemas *and* any filters, logic within queries/ mutations
- Once done, arbitrarily nested queries quite straightforward
- Testing both directly to databases and across APIs



https://graphql.org