Big Data AtSNP Infrastructure

a case study for searching billions of records while providing significant cost savings over cloud providers

Christopher Harrison

Paper accepted to: The 4th IEEE International Workshop on High-Performance Big Data, Deep Learning, and Cloud Computing @The 32nd IEEE International Parallel and Distributed Processing Symposium (IPDPS 2018)

The atSNP story

- Hallway conversation
- 2TB of data on the web
- Another dataset to put online in the future
- Post-Doc will work with you
- Let me know what you need


WHAT COULD POSSIBLY
GO WRONG?

The data

- **Atsnp: Jaspar dataset 2TB (35.78TB)**
- **Encode dataset 21.2TB (360.37TB)**
- **Web accessible genomic data search and export in real-time**
- **Atsnp total uncompressed: ~396TB**
- **307 billion Single Nucleotide Polymorphisms (SNP) records**
- **Library of congress = 10TB Compressed**

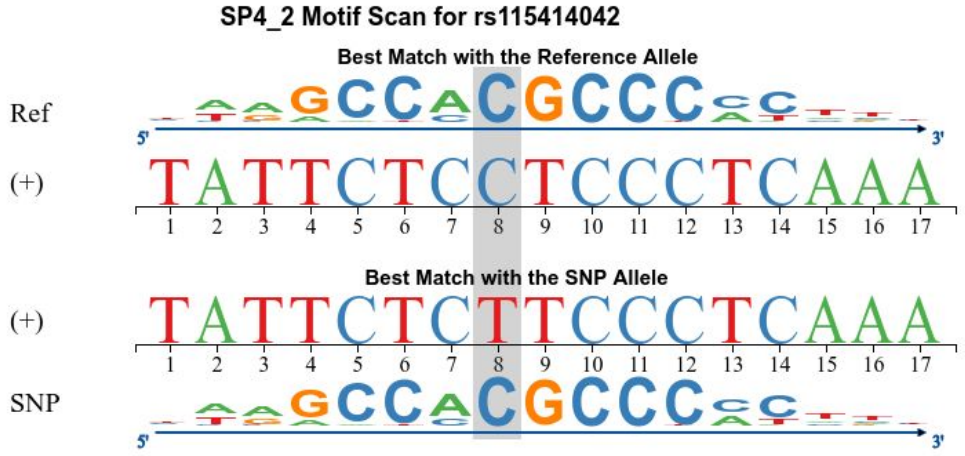
Image from LOC courtesy of: http://www.against-the-grain.com/2015/12/atg-newschannel-original-the-post-print-era-part-1-the-demise-of-library-binderies-2/

What is a SNP

- Technical: Single nucleotide polymorphisms
  - SNPs occur normally throughout a person's DNA
  - They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in each human genome
- Easy to understand:
  - A SNP is a part of an individuals DNA, small sigment
  - DNA is the library of what makes us
  - DNA is comprised of nucleotides:
    - Adenine (A)
    - thymine (T)
    - cytosine (C)
    - Guanine (G)

What is atSNP

- Software developed to evaluate SNP-Transcription factors-DNA interactions
- 115,500 CPU hours to compute SNP to Position Weight Matrix (Big Data)
  - Computed using HTCondor UW-CHTC and OSG
  - Wanted to make this compute power available to researchers without this amount of compute at hand
- Calculate p-values
- Determine SNP-PWM motif's
- Motif images for each of the 307 bill
  - Originally a PNG for each SNP-PWM
  - Would have consumed 3.7Petabytes



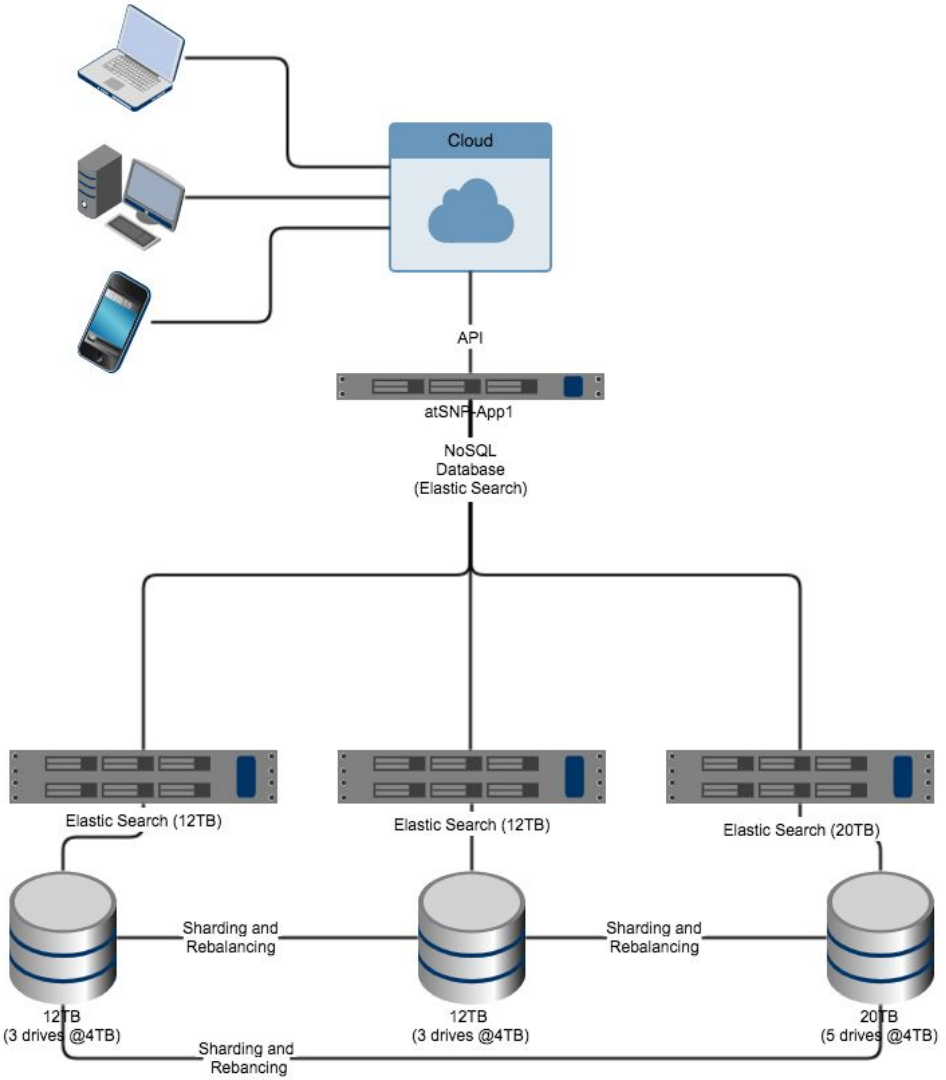
Constraints

- Cost
- Supportability (personal time, monitoring, domain knowledge)
- Speed to implementation
- Data center rackspace
- Query result times

Feasibility Candidates

- Objective: use a DB with a large usage and support base
- Cassandra
  - NoSQL known for quick access and search
- MySQL (or MariaDB)
  - Oldie and goodie
- Elasticsearch
  - Indexes log data
- Others
  - We needed quick turn around and widely supported platforms

Infrastructure for our initial feasibility testing



Cassandra

Pro’s

- Fast searches
- Fast imports (ETL) (14,664records/sec)
- Auto rebalancing on node failure

Con’s

- No range query support*
- No team domain expertise

* At evaluation time

MySQL (MariaDB)

Pro’s

- Team domain expertise
- Range query support

Con’s

- Slow ETL (ETL 1023records/sec)
- Partitioning of data across systems manually
- Auto rebalancing on node failure

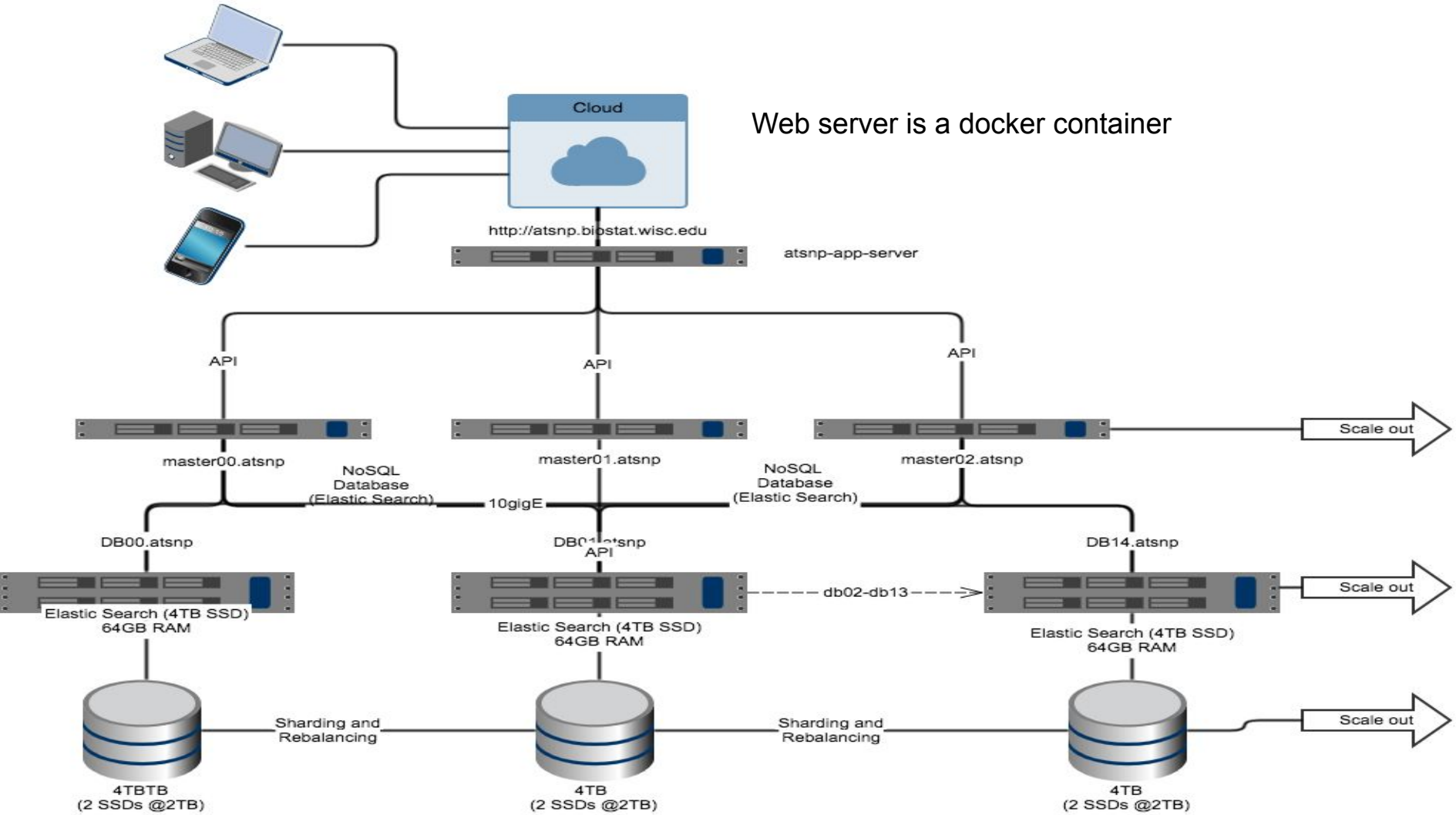
Elasticsearch

Pro’s

- Range queries
- Reasonable Load times (ETL- 11,944records/sec)
- Auto rebalancing on node failure

Con’s

- No domain expertise
- Data loading took longer than Cassandra


Cloud
Web server is a docker container
http://atsnp.biostat.wisc.edu
atsnp-app-server
API
API
API
master00.atsnp
master01.atsnp
master02.atsnp
Scale out
NoSQL
Database
(Elastic Search)
10gigE
NoSQL
Database
(Elastic Search)
DB00.atsnp
DB14.atsnp
API
db02-db13
Scale out
Elastic Search (4TB SSD)
64GB RAM
Elastic Search (4TB SSD)
64GB RAM
Elastic Search (4TB SSD)
64GB RAM
Sharding and
Rebalancing
Sharding and
Rebalancing
Scale out
4TBTB
(2 SSDs @2TB)
4TB
(2 SSDs @2TB)
4TB
(2 SSDs @2TB)

Results of final infrastructure

- Final results proved elasticsearch was a viable option for
  - loading
  - searching
  - and retrieving of data
- Scale-out infrastructure
  - Can add more nodes as data needs change/grow
  - Response time is critical for genomics data searches
  - Future improvements can be easily integrated
- Cost
  - Amazon, $0.135/GB/Month
  - Our final cost $0.039/GB/Month
  - 3.4x Cost Savings over Amazon

Key Contributions

- Feasibility testing is important for application infrastructure deployments
- Cloud providers are not always the lowest cost provider
- NoSQL databases are great for scalability and work for genomic data stores
- atSNP website:
  - http://atsnp.biostat.wisc.edu
- System engineers are rockstars

Acknowledgements

- NIH Big Data to Knowledge (BD2K) Initiative under Award Number U54 AI117924
- Center for Predictive Computational Phenotyping
- University of Wisconsin - Madison
  - School of Medicine and Public Health
    - Department of Biostatistics and Medical Informatics
- My Family


Thank You
Questions?
I know you do….
You in the blue shirt start, ask away