Medical Big Data Analysis System to Discover Associations between Genetic Variants and Diseases

PRESENTED BY **JAIMIT JAMES**

DAEHEE KIM, STEPHANIE GAMBOA, VANESSA HERNANDEZ, **MARLEN MARTINEZ-LOPEZ**, SCOTT J. HEBBRING, JOHN MAYER & JAIME FOX

Accepted in IEEE International Conference on Communications (ICC) 2021
https://icc2021.ieee-icc.org/program/technical-symposia#S1569591512

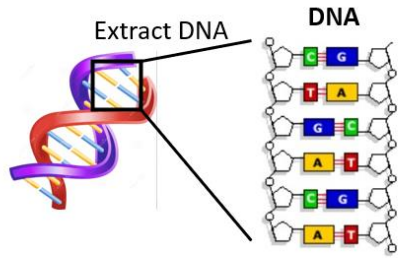
Background

- Health Record Data
  - Used to record data on patients
  - Biological measurements
    - Disease Diagnoses
    - Medical procedures
- Genetic Data
  - Retrieved from DNA in blood samples
- Marshfield Clinic Research Institute (MCRI)
- Genome-Wide Association Studies (GWASs)
  - Find genetic variants for certain diseases
  - Phenotype-to-genotype approach
- Phenome-Wide Association Studies (PheWASs)
  - Explore multiple diseases relevant to genetic variant
  - Genotype-to-phenotype approach

Name: Jane Doe
Medical History #: 111111
DOB: 01/01/1950
Weight: 150 lbs
Height: 5'5''
Address: 1000 N. Oak Street

Diagnosis & Procedure
(ICD9 codes):
250 = Diabetes
493.1 = Intrinsic Asthma
474.00 = Chronic Tonsillitis
28.2 = Tonsillectomy

Prescriptions:
Antibiotics
Albuterol
Metformin



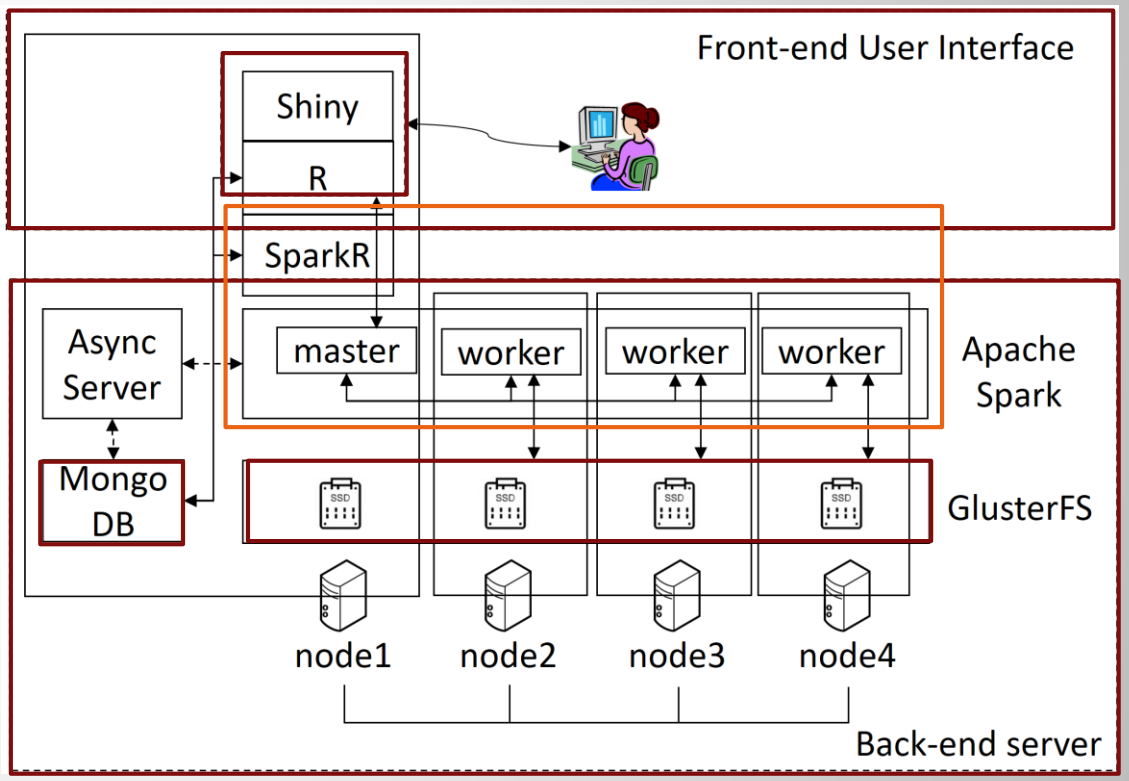
Person1: A–G–T–C–A–A–G
Without disease

SNP

Person2: A–G–T–A–A–A–G
With disease

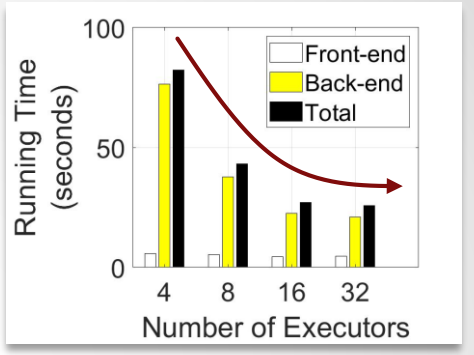
Architecture

- Web Query System architecture
  - Front-end user interface
    - R Shiny
  - Back-end server
    - GlusterFS, Spark, MongoDB, Java daemon
- Each node runs on
  - Dell PowerEdge R710
  - 2U rack sever (144GB)
  - 2 Intel Xeon 5660
- Each node has
  - 2 TB SSD

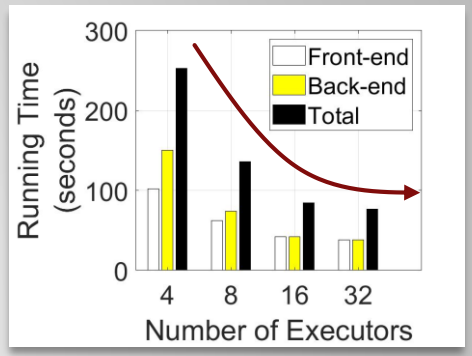


Evaluation

- SparkR (front-end), Spark-submit (back-end)
- Measured running time of:
  - Front-end & back-end operations
- Each executor: 2 CPU cores, 16 GB
- Varying the number of executors to 4, 8,16 and 32
- Running time for disease / genome data
  - Running time becomes faster with more executors on parallel processing
  - Running time of front-end is much less than back-end processing
  - Running time with 16 and 32 executors is similar, indicating the existence of upper bounds

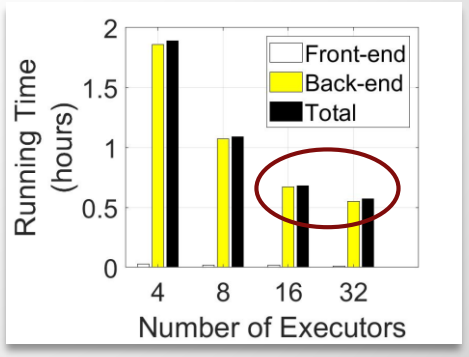


Chromosome 22

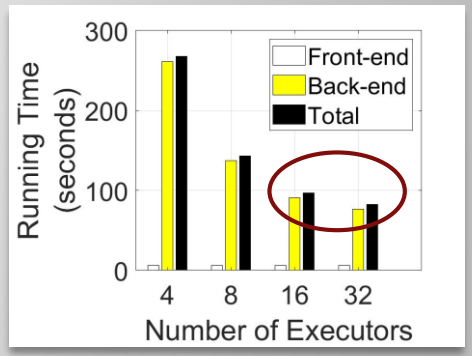


All chromosomes

Search disease



Chromosome 22



All chromosomes

Search genome

Thank you

California State University-Stanislaus

Jaimit James	jjames5@csustan.edu
Marlen Martinez-Lopez	mmartinezlopez@csustan.edu
Stephanie Gamboa	sgamoa@csustan.edu
Vanessa Hernandez	vhernandez27@csustan.edu
Advisor: Dr. Daehee Kim	dkim10@csustan.edu

Marshfield Clinic Research Institute

Dr. Scott J Hebbring	hebbring.scott@marshfieldresearch.org
John Mayer	mayer.john@marshfieldresearch.org

Prevention Genetics

Dr. Jaime Fox	jaime.fox@preventiongenetics.com