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

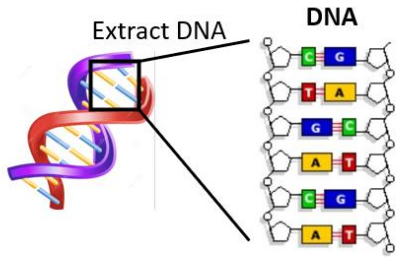
Background

- Health Record Data
  - Used to record data on patients
  - Biological measurements
    - Disease Diagnoses
    - Medical procedures
- Genetic Data
  - Retrieved from DNA in blood samples

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

Background- cont'd

- 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
- Electronic Health Records (EHRs) and DNA genotype are the main resources used to discover individual differences
- We designed and implemented a Medical Big Data analysis system that retrieves results from a GWAS-by-PheWAS dataset

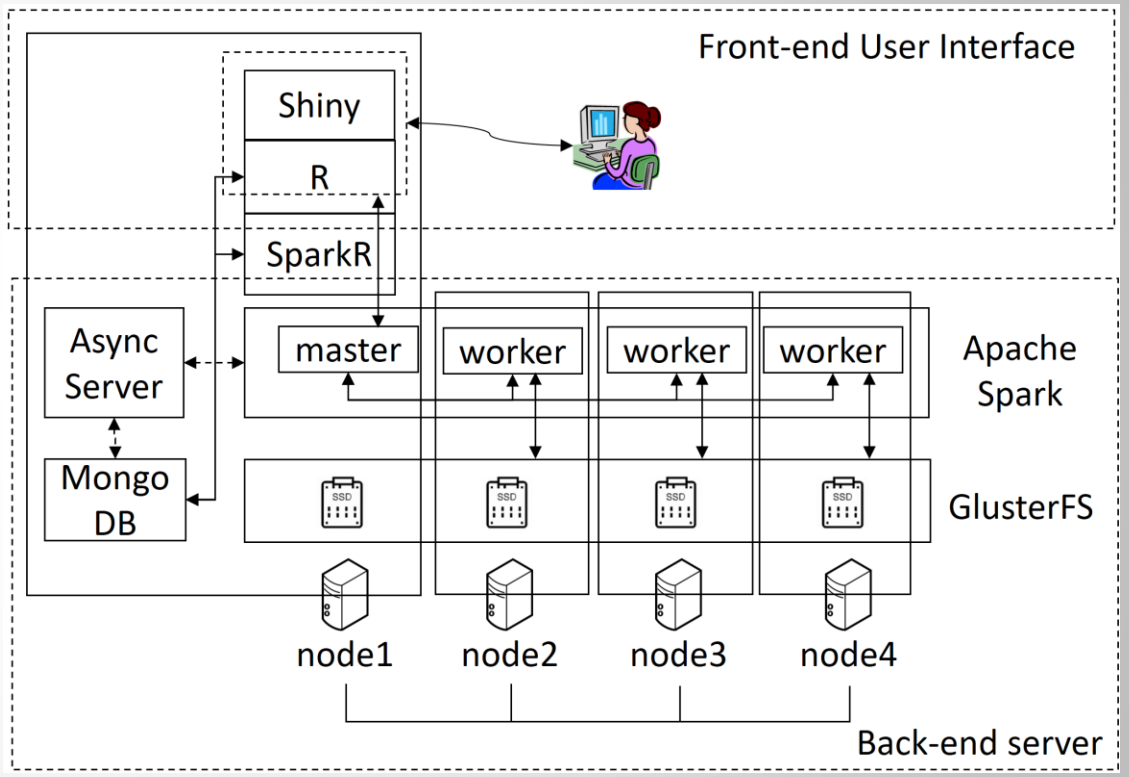
Dataset

- Data of biobank in Marshfield Clinic Research Institute (MCRI)
- Consists of genotype DNA and EHR of 20,000 patients
  - Age range 18 to 98.5
  - 57.2% were female.
  - PheWAS dataset searchable by RS ID or genetic position of SNP
  - GWAS dataset searchable by ICD-9 disease code or description

PheWAS Example	**22**,**29854579**,G,A,8613,0.19234,0.0065506,-,0.935493,dx903, **Type 1 (Juvenile Type) Diabetes Mellitus With Ketoacidosis Uncontrolled**, **250.13**
GWAS Example	**22**,17265124,**17265124**,A,C,exonic,XKR3,NA nonsynonymous SNV,XKR3:NM_175878:exon4:c:T765G:p.F255L,0.694489,0.6282, **rs5748623**, 1,T,0.0,B,0.0,B,0.001,N,1.000,P,-1.1,N,NA,,

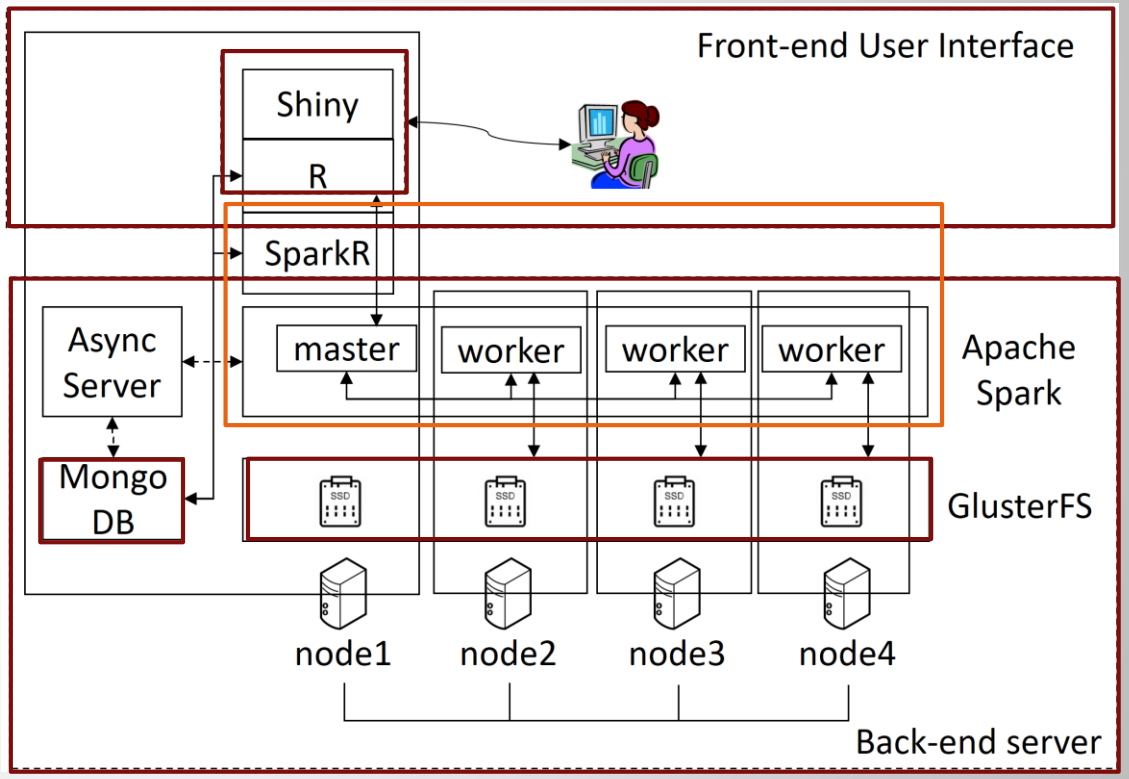
System Architecture

- Each node runs on
  - Dell PowerEdge R710
  - 2U rack sever (144GB)
  - 2 Intel Xeon 5660
- Each node has
  - 2 TB SSD
- 8 TB for Spark cluster
- Ubuntu 18.04
- Standalone cluster manager



Software Architecture

- Web Query System architecture
  - Front-end user interface
    - R Shiny
  - Back-end server
    - GlusterFS
    - Spark
    - MongoDB
    - Java daemon

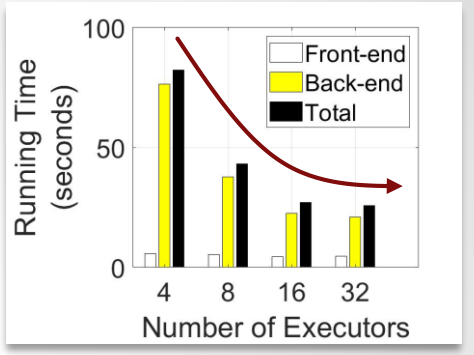


Evaluation Set Up

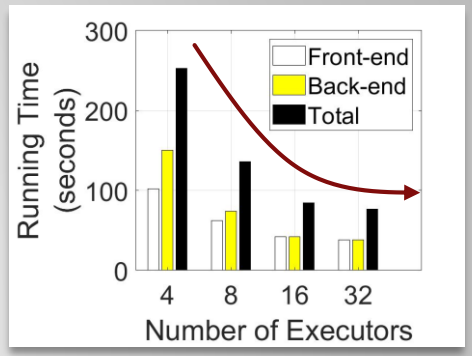
- SparkR (front-end), Spark-submit (back-end)
- Measured running time of:
  - Front-end & back-end operations
  - Averaged 5 times running the same request using ‘sar’ command
- Each executor: 2 CPU cores, 16 GB
- Varying the number of executors to 4, 8,16 and 32
- Equally distributed to four worker nodes
  - (e.g., 32 executors, each worker node runs 8 executors with 16 cores and 128 GB, resulting in 64 CPU cores and 512 GB in total for processing a user request)

Performance

- 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
  - Separating workloads between front-end and back-end is configurable
  - For all chromosomes, long time for front-end operation
  - Running time with 16 and 32 executors is similar, indicating the existence of upper bounds

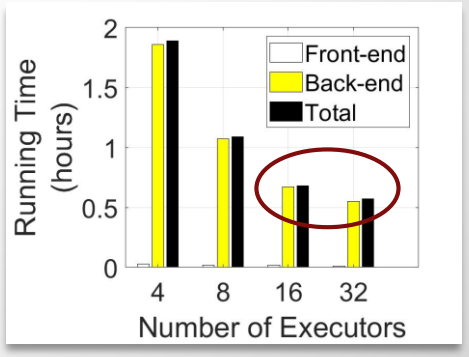


Chromosome 22

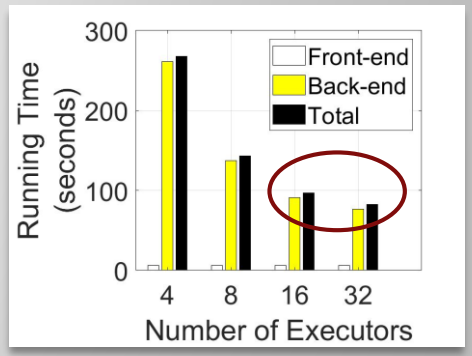


All chromosomes

Search disease



Chromosome 22



All chromosomes

Search genome

Future Work & Conclusion

- Medical big data analysis system is a prototype
  - Check the application design and system architecture
- To handle more data
  - Large scale Spark cluster
  - More worker nodes
  - MCRI biobank: 20 petabytes
- Future
  - Dynamic resource allocation
  - A hybrid system

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