Efficient Visualization of Medical Big Data Using R Shiny Application

Nikolai-Iraj Sanamrad, Giancarlo Garcia Deleon and Deep Gill (Advisor: Dr. Daehee Kim)
California State University, Stanislaus

Abstract

Approximately 20,000 patients of the Marshfield Clinic [1] shared their medical information in order to research the possible DNA sequences related to certain diseases.

This research can improve the healthcare quality and help predict certain risks, based on statistical data.

We are researching efficient methods of processing queries to the database and displaying both graphical and textual results in a R Shiny [2] Web Application.

We expect to have an efficient application, capable of displaying results as charts and text.

Motivation

The Research results are stored in a distributed file system using Apache Spark [3], which is connected to R using Spark-R. In order for a query to be processed, it has to pass through all these layers:

** node1 has a worker which is not shown for simplicity.
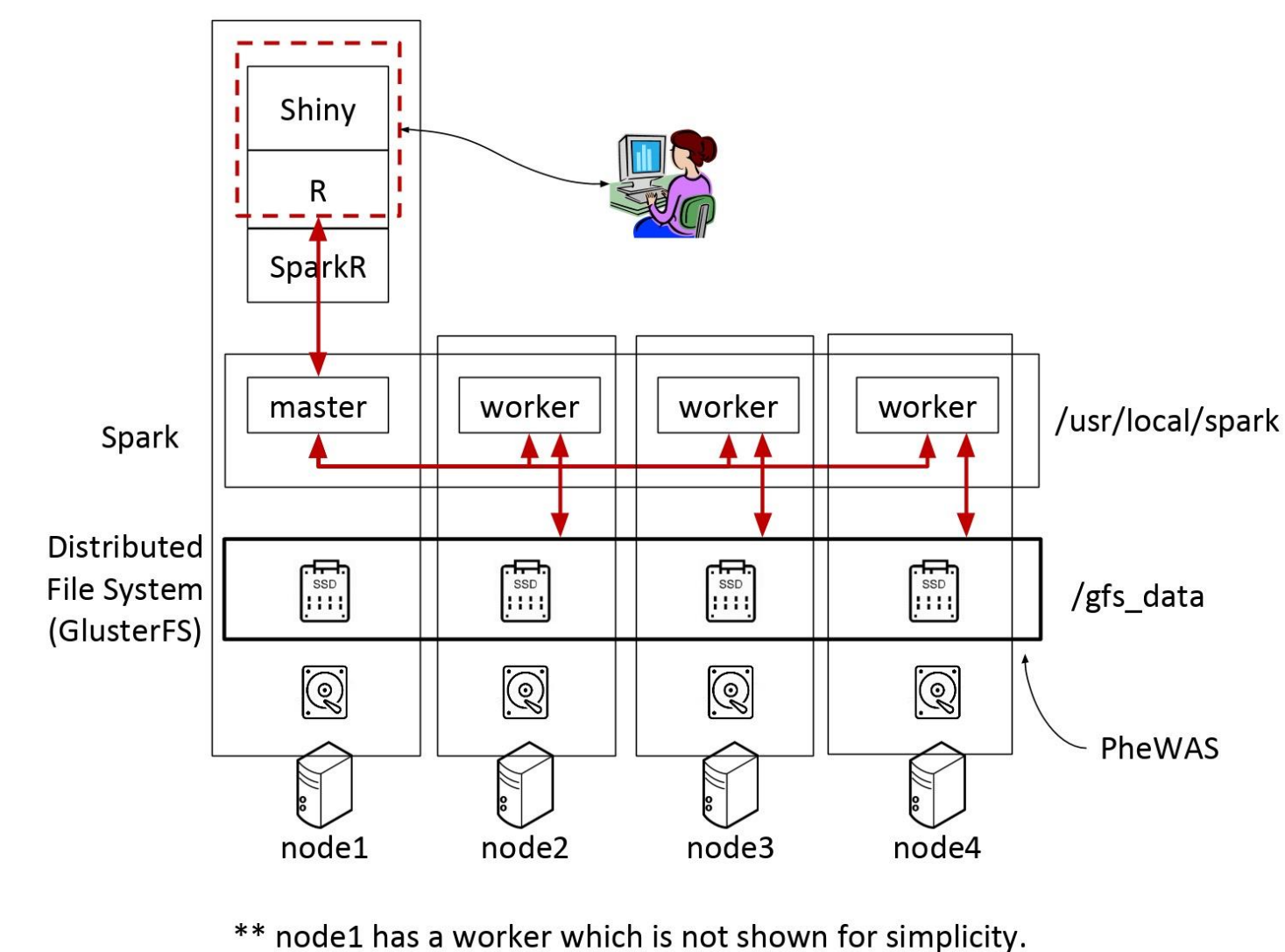
Related Work

The medical data has been analyzed in a Phenome-Wide Association Studies [4] and the results of that analysis are available for our project.

Our project is expected to provide a more efficient solution with additional features.

Method

The dataset in storage is large, and there are multiple layers of different applications between the end-user interface, which is implemented in R Shiny, and the storage, which is distributed among multiple servers.

As shown in the Architecture Layer Diagram, the highest layer is the R Shiny Application layer, where the end-user Web Application is located. The R Shiny layer sends the user requests to the R Layer, which then sends it to the next layer – Apache Spark, through Spark-R. The Apache Spark then retrieves the necessary data from the dataset, which is located in a Distributed File System across multiple servers, and sends it to the upper layer.

The end result is then processed by R Shiny, a powerful package that generates simple, but functional and efficient Web Applications with results displayed both as tables and various charts. R Shiny has a large library of different chart tools that are easy to implement as below:
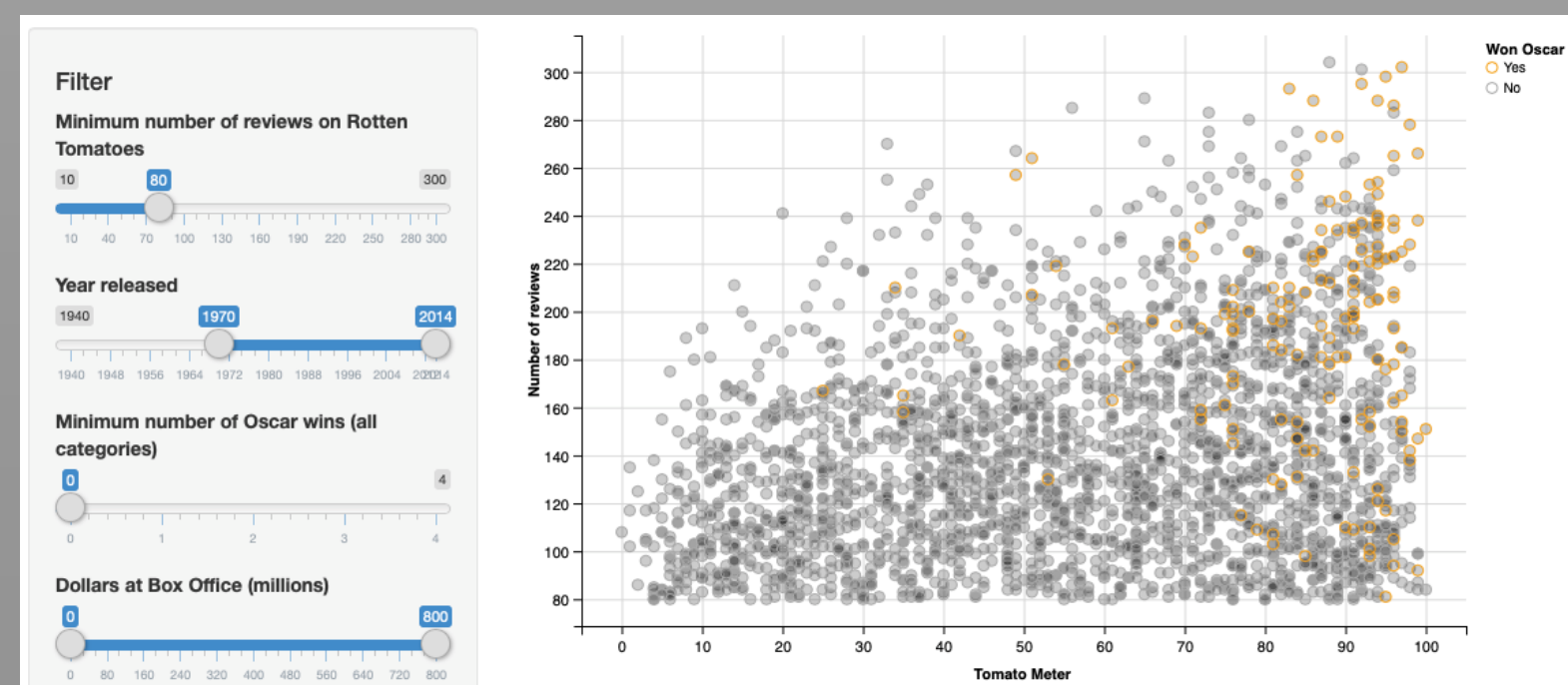

System Architecture Diagram
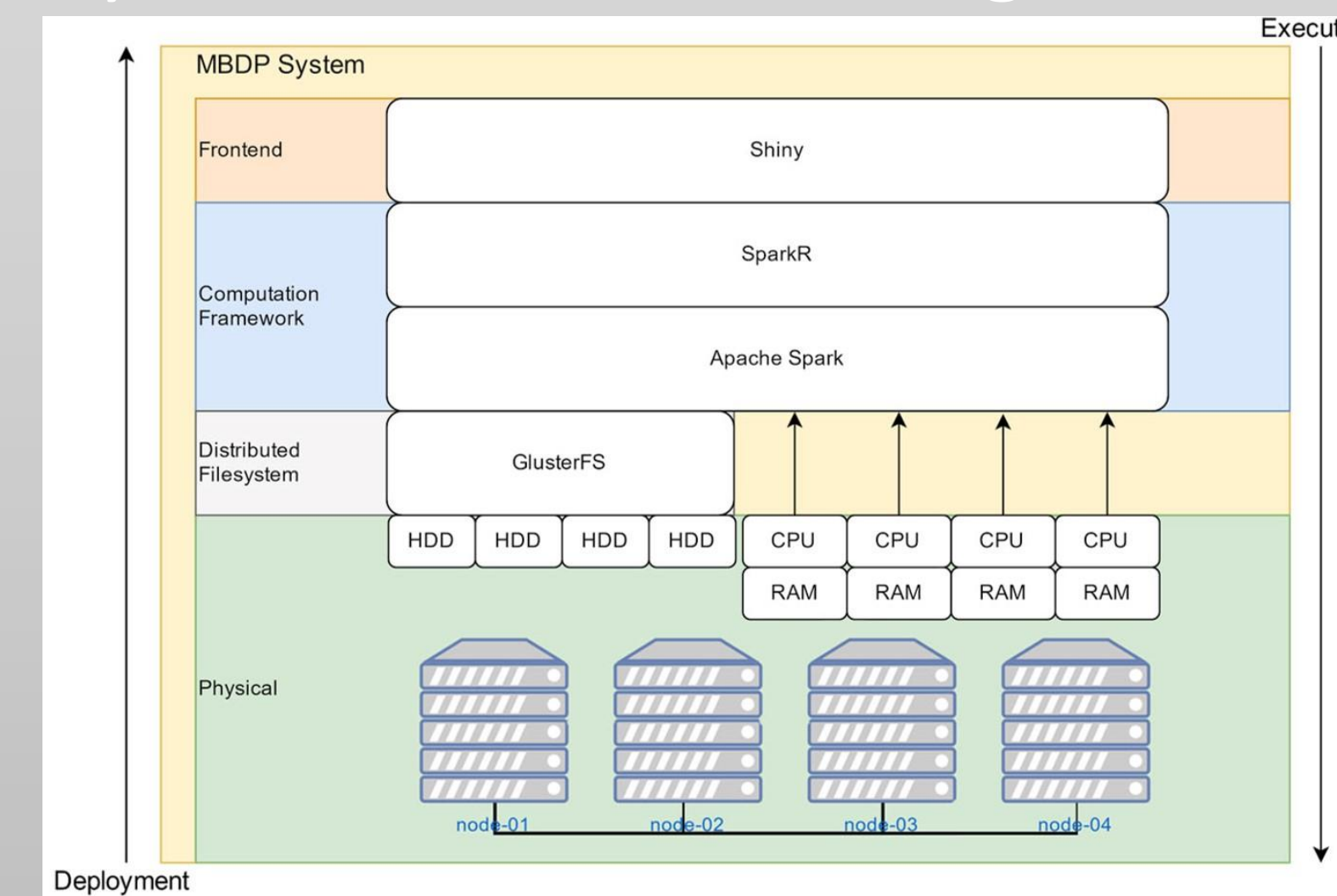
Dataset Structure

The PheWAS analysis previously performed on the raw data resulted in the dataset that we utilize for our project [5].

Column	Definition
Chrom	Chromosome number
Pos	location of variant on the chromosome
Ref_WT	"Reference" Allele
Alt_WT	"Alternative" Allele
N_Informative	# of individuals used in association test
AF	"Allele Frequency"
Stat	Statistics estimated from the model
Direction	"Direction of effect"
Pvalue	pValue calculated by Fisher exact test
Dx_code	"converted" ICD9 codes
Dx_desc	Description of ICD9
Dx	"actual" ICD9 code

Expected Results

The resulting Web Application is expected to be able to save query results and display them on-demand. The previous realization demonstrates some features:
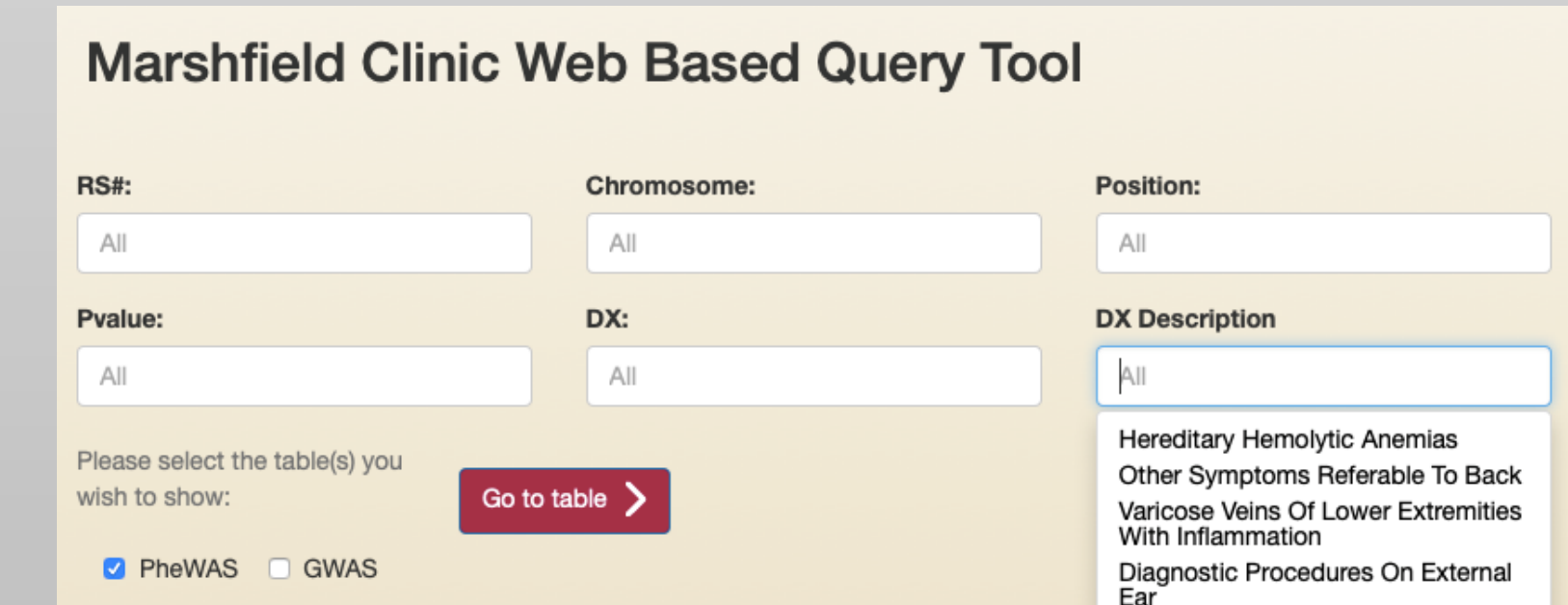
Results will be available as both text and charts:
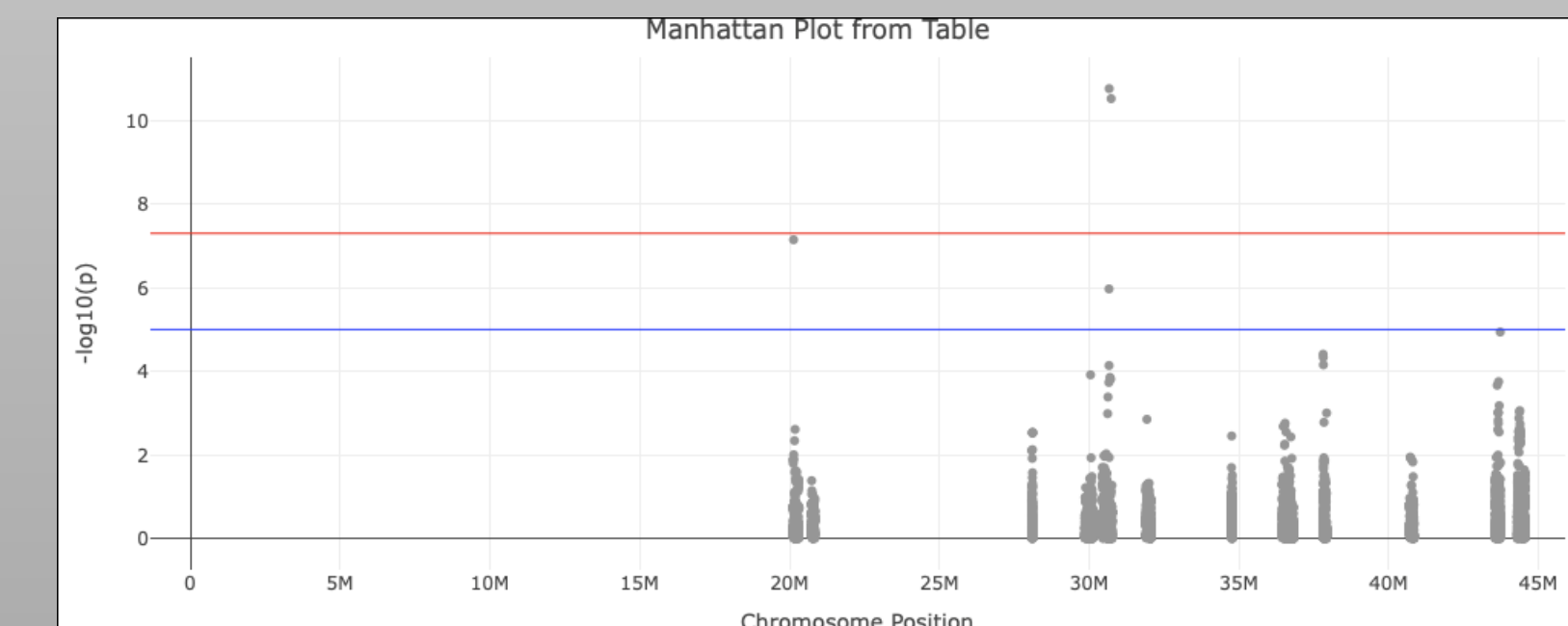
Conclusion

Implementing a productive Big Data Visualization system using Shiny is possible, but because of a large number of necessary layers required for the query processing, the solutions are not always efficient. Our project is aimed at optimizing the layers, to provide a more efficient performance with fewer delays.

References

[1] Marshfield Clinic Biobank. http://marshfieldclinic.org.
[2] R Shiny. http://shiny.rstudio.com.
[3] Apache Spark. http://spark.apache.org.
[4] Phenome Wide Association Studies (PheWAS). https://phewascatalog.org.
[5] Jamie Fox. Big Data: Applications in Health Care and Disease Research.