Big Data Technologies for Storage and Querying of Large-Scale Medical Data in Clinical Genetic Research

Gittu George[1], Philip Appleby[1], Alex Doney [1], V. Mohan[2], Radha V [2]

[1]School of Medicine, University of Dundee, Dundee, United Kingdom

[2]Madras Diabetic Research Foundation (MDRF), Chennai, India

INTRODUCTION

SCOTLAND DATA:
- Clinical Data: 250K people
  - GoDARTS
  - GoSHARE
  - NHS Tayside and NHS File
- Genotype Data
  13,000 people X 40 Million SNPs

INDIAN DATA:
- Clinical Data: 400K people
- Genotype Data: 25K people

UK BIOBANK DATA:
- Clinical Data: 490K people
- Genotype Data:
  490K people and 93 Million SNPs

CHALLENGES

- Flat text files which are hard to break or naturally not indexed, makes it strenuous to manually separate and parallelize
- Lack of centralized storage for genotype and phenotype data
- Single node tools and programming methods that doesn't scale
- Difficulties in learning across multiple and deeper phenotypes & genotypes

SOLUTION

- Work in big data ecosystem
- Move to algorithms that can run in parallel
- Advanced file formats that are indexed and ready for parallelism
- Centralized storage for both phenotype and genotype data
- Provide best performance to the amount of hardware available

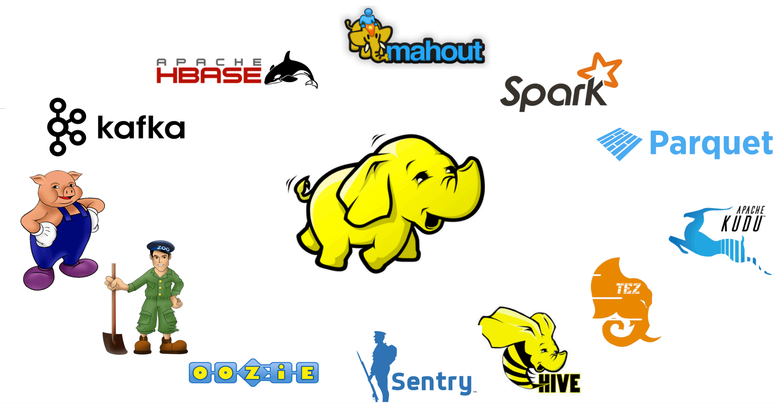
WHY HADOOP ?

- Open source framework for storing data and running applications in clusters.
- Horizontal scalability
- Failure is normal and expected
- Data Locality - Compute should move to the data

HADOOP ECOSYSTEM

HADOOP DISTRIBUTION FILE SYSTEM (HDFS):
- *File system for Hadoop framework*
- *Uses commodity hardware – low cost*
- *Optimized for MapReduce workloads - deliver data into the compute infrastructure at a huge data rate*
- *Support of highly efficient datatypes – Parquet , ORC*

MAPREDUCE:
- *Programming paradigm for Hadoop*
- *Consist of Mapper and Reducer*



Pig: An engine for executing data flows in parallel on Hadoop.

Hive: A Data Warehouse infrastructure for Hadoop

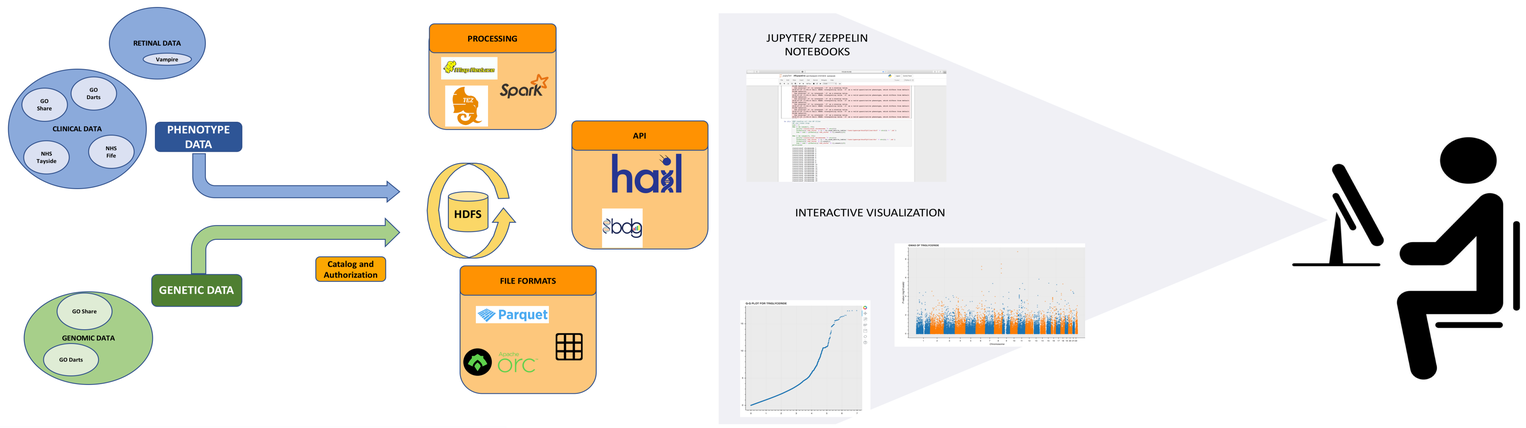
Oozie: Workflow scheduler system for Map Reduce jobs.

Spark: Hadoop on steroids. Runs up faster in memory and even fast on disk than Hadoop.

HBase: A column-oriented database management system that runs on top of HDFS.

Kafka: Building real-time data pipelines and streaming apps on Hadoop.

GOAL

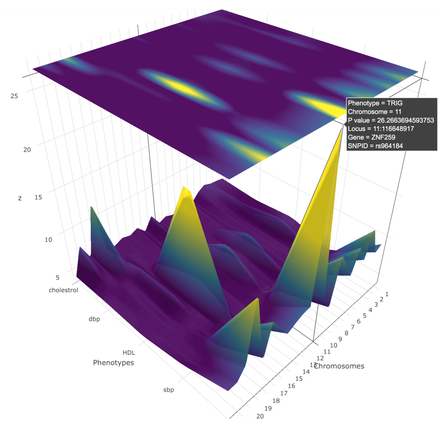


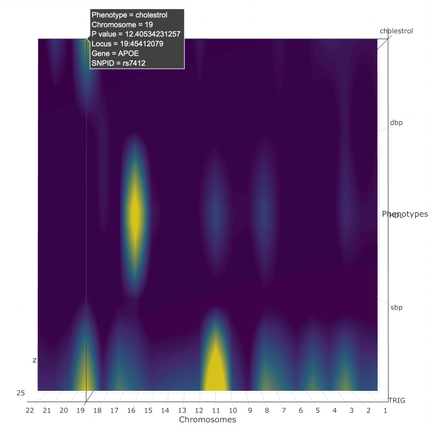
- Pipeline to load the data to HDFS
- Enrich the data in HDFS
- Merge the Phenotype and Genotype data into a single file for further analysis
- Create visualization UI to bridge the gap between the medical researchers and growing big data

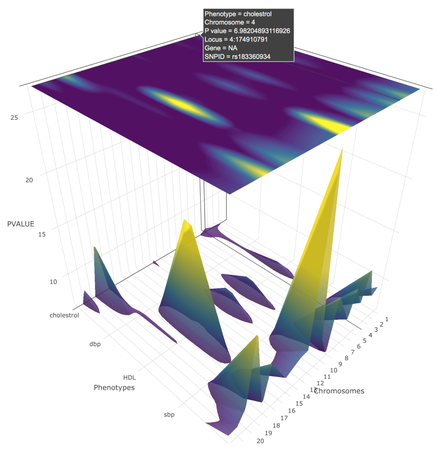
APPLICATIONS

FIGIWAS (Ongoing Work)

- Landscape visualization of Many Gene Variant and Many Disease
- Heatmap to visualize the variations among different phenotype on each chromosome
- Ability to zoom, pan, orbital and turntable rotation
- Provide a cut of the landscape for users to visualize the significant regions







FUTURE WORK

- Identify different questions to ask the data
- Expand the Hadoop cluster to become more powerful
- Showing genomic significance for each chromosome in FIGIWAS
- Dynamic UI for clinicians/geneticists to interact with the data.

ACKNOWLEDGEMENT

Special thanks to Shona Matthew the NIHR Global Health Research Unit Manager and Howard Rogers from HIC center UoD.

The research was commissioned by the National Institute for Health Research using Official Development Assistance (ODA) funding [INSPIRED 16/136/102].

Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.