unite for diabetes

Predict Onset of Diabetes

Big Data System Design - Proposal

John Carneiro - johnmcarneiro@gmail.com
Janakiram Sundaraneedi - janakiram_sundaraneedi@student.uml.edu
Jan Zaloudek - honzicekz@gmail.com

Agenda

unite for diabetes

- Introduction
- Motivations
- Proposed Approach
- Related Work
- Evaluation
- Timeline

Introduction

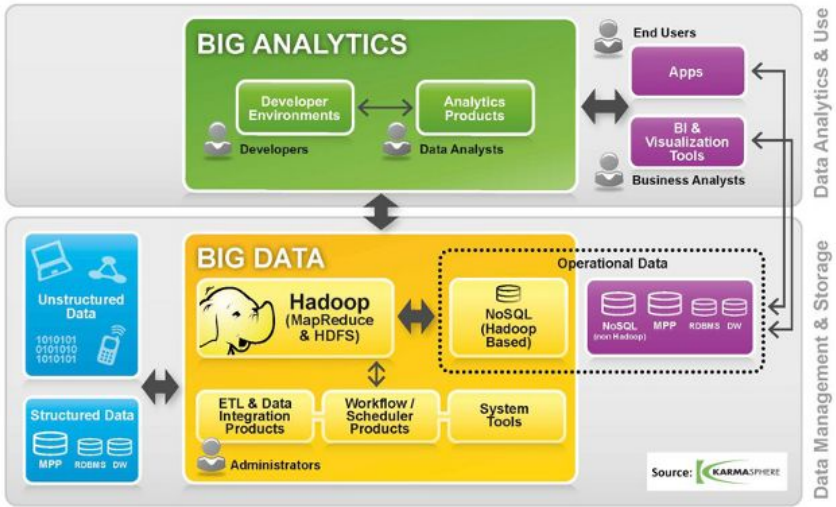
unite for diabetes

- The research/development question is to predict based on diagnostic measurements whether a patient has type 2 diabetes using the Kaggle based Pima Native American Diabetes Database originally from the National Institute of Diabetes and Digestive and Kidney Diseases.
- The CSV text dataset contains records of female patients of Pima Native American heritage in Arizona who historically have greater than 50% rate of type 2 diabetes. Data includes pregnancy count, glucose concentration, blood pressure, tricep skinfold thickness, insulin resistance, body mass index, diabetes family history factor, age, and diabetes positive.

Introduction (cont'd)

unite for diabetes

- The ultimate goal of your proposed system is to perform this prediction in an automated way using big data system design.
- Hadoop big data ecosystem (Hadoop FS, MongoDB, Spark, Machine Learning, Processing techniques such as MapReduce, Keras, TensorFlow)



Motivations

unite for diabetes

- This is interesting question for us due to a lot of our family members have diabetes
- It is important to develop a system to be able to predict the occurrence of diabetes if the patient’s lifestyle and eating habits are kept the same and possibly prevent the occurrence of diabetes if preventative steps are taken by patient
- 415 million people have diabetes worldwide
- 8.3% of the world adult population (equal parts men/women) have diabetes and is rising
- Diabetes doubles a person's risk of early death.
- 5 million deaths occur worldwide each year because of diabetes.
- The global economic cost of diabetes is estimated to be US $612 billion.

Proposed Approach

unite for diabetes

- Plan for working out the solutions to the question:
  - Research dataset
  - Survey knowledge experts (UML doctor, nursing staff, and people with the diabetes)
  - Improve via algorithms accuracy and precision level on missing values
  - Research ML algorithms, current solutions on kaggle.com
  - Use rapidminer.com, and Python ML libraries
  - Performance test ML algorithms
- Main features in your proposed system:
  - Hadoop based backend (HDFS, MongoDB (i.e. NoSQL DB), Python, Apache Web Server)
  - HTML5, JS, Cordova mobile development front end (desktop, iOS, Android)
    - Data entry and/or upload data set
    - Matplotlib, plotly, Google Chart, D3, etc.
    - Based on input data, one or more users can be told whether or not he/she has or is susceptible to get type 2 diabetes (data can be optionally added to learning dataset to improve algorithm accuracy)

Proposed Approach (cont'd)

unite for diabetes

- Implementation approach your proposed system:
  - Hadoop FS
  - MongoDB
  - Python, RapidMiner
  - HTML5 - JS
  - Spark, Keras, TensorFlow, Scikit Learn, MapReduce
  - Machine Learning Algorithm(s) (Classification, Regression, Clustering, Dimensionality reduction, Model selection, Preprocessing) - those found exhibiting best performance
    - Keras based neural network
    - Gaussian Naive Bayes
    - K-Nearest Neighbor Classifier
    - Decision Tree Classifier
    - Random Forest Classifier
    - Ada Boost Classifier
    - Gradient Boosting Classifier

Related Work

unite for diabetes

- The following are sources of existing research/development work has tried to answer the same or a similar question:
  - https://www.kaggle.com/uciml/pima-indians-diabetes-database (data set)
  - Using the ADAP learning algorithm to forecast the onset of diabetes mellitus
  - Genetic Studies of the Etiology of Type 2 Diabetes in Pima Native American
- A highly reliable and accurate automated prediction algorithm is still not available.
- Also, it would be self-evident that an accurate automated prediction algorithm should also work with men and persons not of Pima Native American ethnicity

Evaluation

unite for diabetes

- We will evaluate our solution as follows (to demonstrate that our solution/answer is good/reasonable):
  - Test ML algorithms locally with the stock dataset
  - Find the algorithm that performs the best in a reasonable amount of processing time at a small scale first
  - Scale up the dataset by several factors of magnitude (10x, 100x, 1000x, etc.)
  - Setup a Hadoop environment to handle the increase in size
  - Integrate Hadoop components
  - Test the scaled up components at the backend server side
  - Test the scaled up components at the client side
  - Compare all results with those available online in previous studies
  - Publish results on sites like kaggle.com to verify and respond to feedback

Timeline

unite for diabetes

- Week 1 (ending 2017-02-02) - Proposal presentation
- Week 2 (ending 2017-02-09) - Pre data set analysis and Github setup
- Week 3 (ending 2017-02-16) - Subject matter expert surveys
- Week 4 (ending 2017-02-23) - ML algorithm analysis
- Week 5 (ending 2017-03-02) - ML algorithm testing
- Week 6 (ending 2017-03-09) - ML algorithm testing
- Week 7 (ending 2017-03-16) - Scale up dataset
- Week 8 (ending 2017-03-23) - Hadoop ecosystem integration
- Week 9 (ending 2017-03-30) - Data plot development
- Week 10 (ending 2017-04-06) - Front end development
- Week 11 (ending 2017-04-13) - System performance testing
- Week 12 (ending 2017-04-20) - System performance testing
- Week 13 (ending 2017-04-27) - Final presentation