

# Predicting Readmission of Diabetic patients

KUMAR SATYAM<sup>1,\*</sup>, PIYUSH SHINDE<sup>1,\*\*</sup>, AND SRIKANTH RAMANAM<sup>1,\*\*\*</sup>

<sup>1</sup>School of Informatics and Computing, Bloomington, IN 47408, U.S.A.

\* Corresponding authors: [ksatyam@indiana.edu](mailto:ksatyam@indiana.edu)

\*\* Corresponding authors: [pshinde@iu.edu](mailto:pshinde@iu.edu)

\*\*\* Corresponding authors: [srikrama@iu.edu](mailto:srikrama@iu.edu)

project-000, March 25, 2017

We are trying to predict whether a diabetic patient will be readmitted to the hospital, using several features representing patient and hospital outcomes. We will use Hadoop/Spark distributed architecture on multiple clouds as the core infrastructure and machine learning classification algorithms for data analysis.

© 2017 <https://creativecommons.org/licenses/>. The authors verify that the text is not plagiarized.

**Keywords:** Hadoop, Spark, Ansible, Python

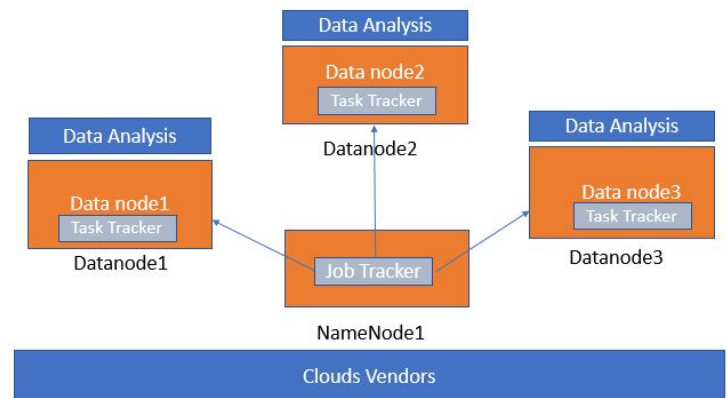
<https://github.com/cloudmesh/classes/blob/master/project/S17-IR-P004/report/report.pdf>

## CONTENTS

|   |                 |   |
|---|-----------------|---|
| 1 | Introduction    | 1 |
| 2 | Timeline        | 1 |
| 3 | Technologies    | 2 |
| 4 | Deployment      | 2 |
| 5 | Benchmarking    | 2 |
| 6 | Results         | 2 |
| 7 | Conclusion      | 2 |
| 8 | Acknowledgments | 2 |

## 1. INTRODUCTION

We will use Hadoop to split the dataset and transfer the data chunks to different data nodes. We will use Ansible to install pre-requisite softwares and push configurations on different machines. The data chunks would then be analyzed using machine learning techniques and the results would be aggregated predicting whether a patient would be readmitted or not. This information would help hospitals to be better prepared for readmitting patients.



**Fig. 1.** Deployment Architecture

## 2. TIMELINE

| Week | Target                                     |
|------|--|
| 1    | Finalizing Technologies, Data Cleansing    |
| 2    | Hadoop/Spark Deployment on Chameleon Cloud |
| 3    | Troubleshooting                            |
| 4    | Data Analysis                              |
| 5    | Deployment on other cloud using Ansible    |
| 6    | Benchmarking                               |
| 7    | Report Preparation                         |

### 3. TECHNOLOGIES

| <i>Technology</i>                                    | <i>Usage</i>   |
|--|--|
| <b>Hadoop</b> [1]/ <b>Spark</b> [2]                  | Distributed Data Storage                             |
| <b>Python</b> [3]/ <b>Java</b> [4]/ <b>Scala</b> [5] | Development  |
| <b>Ansible</b> [6]                                   | Application Deployment<br>& Configuration Management |
| TBD  | Benchmarking   |
| <b>LaTeX</b> [7]                                     | Document Preparation                                 |

### 4. DEPLOYMENT

We will deploy a master & multiple slave nodes in the Hadoop/Spark distributed cluster environment.

We will use **Ansible** as an automated application and configuration deployment tool. This will enable us to install softwares and push configurations simultaneously from master node to the respective target nodes.

### 5. BENCHMARKING

We will assess the performance of the Hadoop/Spark clusters deployed on different clouds. The parameters for benchmarking would be memory usage, storage size and IO throughput.

### 6. RESULTS

Results of data analysis and benchmarking will be showcased in this section.

### 7. CONCLUSION

Using the 130-US hospitals dataset [8] for years 1999-2008, we should be able to analyze factors pertaining to readmission of patients with diabetes.

### 8. ACKNOWLEDGMENTS

This project was a part of the Big Data Software and Projects (INFO-I524) course. We would like to thank Professor Gregor von Laszewski and the associate instructors for their help and support during the course.

### REFERENCES

- [1] "Welcome to Apache™ Hadoop®!" Web Page, accessed: 2017-03-12. [Online]. Available: <http://hadoop.apache.org/>
- [2] "Apache Spark: Lightning-fast cluster computing," Web Page, accessed: 2017-03-12. [Online]. Available: <http://spark.apache.org/>
- [3] "python," Web Page, accessed: 2017-03-12. [Online]. Available: <https://www.python.org/>
- [4] "java," Web Page, accessed: 2017-03-12. [Online]. Available: <https://www.java.com/en/>
- [5] "Scala," Web Page, accessed: 2017-03-12. [Online]. Available: <https://www.scala-lang.org/>
- [6] "ANSIBLE," Web Page, accessed: 2017-03-12. [Online]. Available: <https://www.ansible.com/>
- [7] "The LATEX Project," Web Page, accessed: 2017-03-12. [Online]. Available: <https://www.latex-project.org/>
- [8] "Diabetes 130-US hospitals for years 1999-2008 Data Set," Web Page, accessed: 2017-03-12. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#>