Nile University

School of Communication and Information Technology (CIT)

Master of Science/Engineering MSCIT/ MCIT Program

Status Report 16

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version No.** | **Date** | **Description** | **Created By** | **e-mail** |
| 0.1 | 25-01-2019 | Status Report 16 | Ahmed Mohamed Abdel Rahman | [Robot209@gmail.com](mailto:Robot209@gmail.com) |
|  |  |  |  |  |

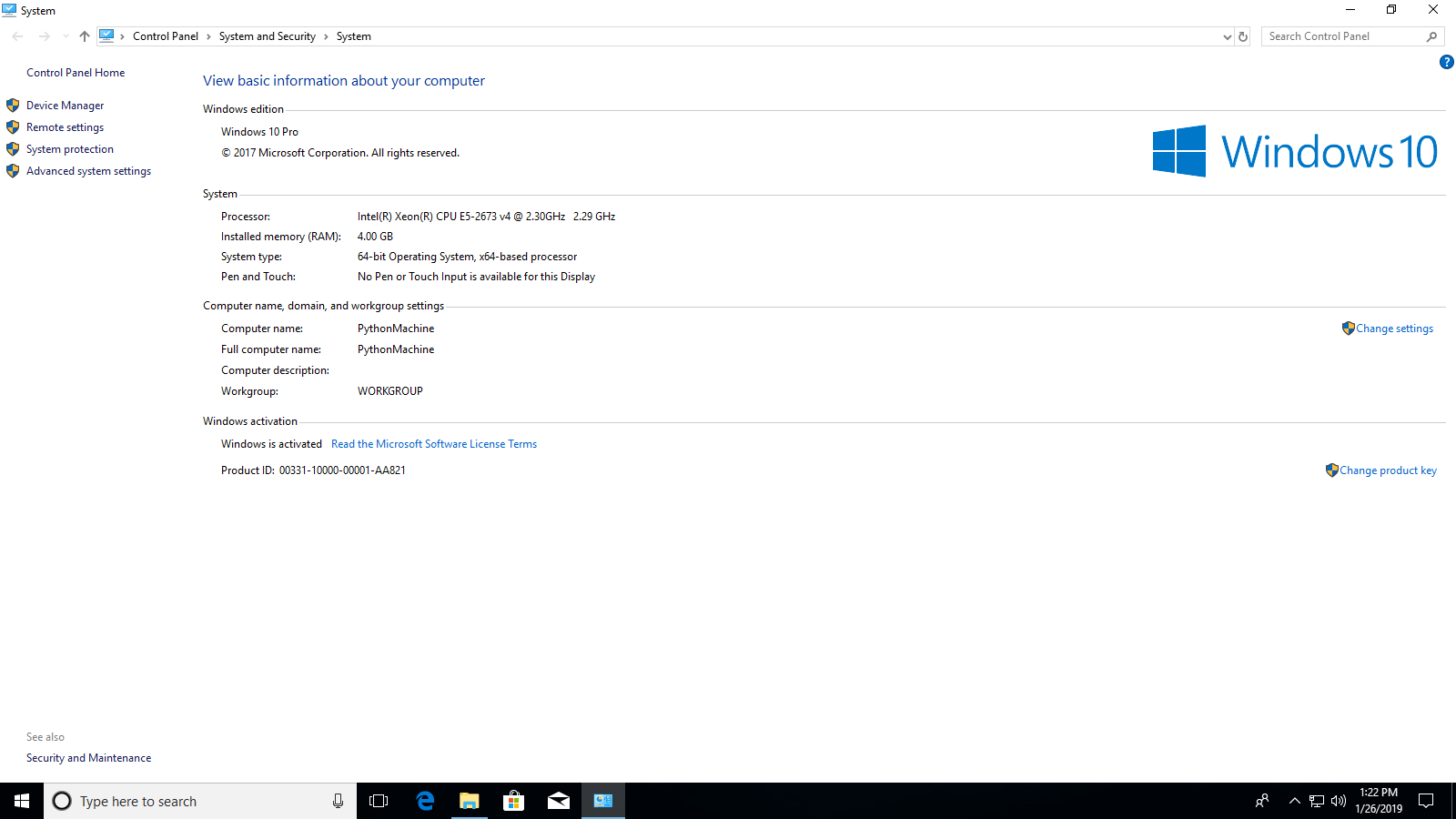
## Objective

* Compare between single node & Distributed LDA Classifier

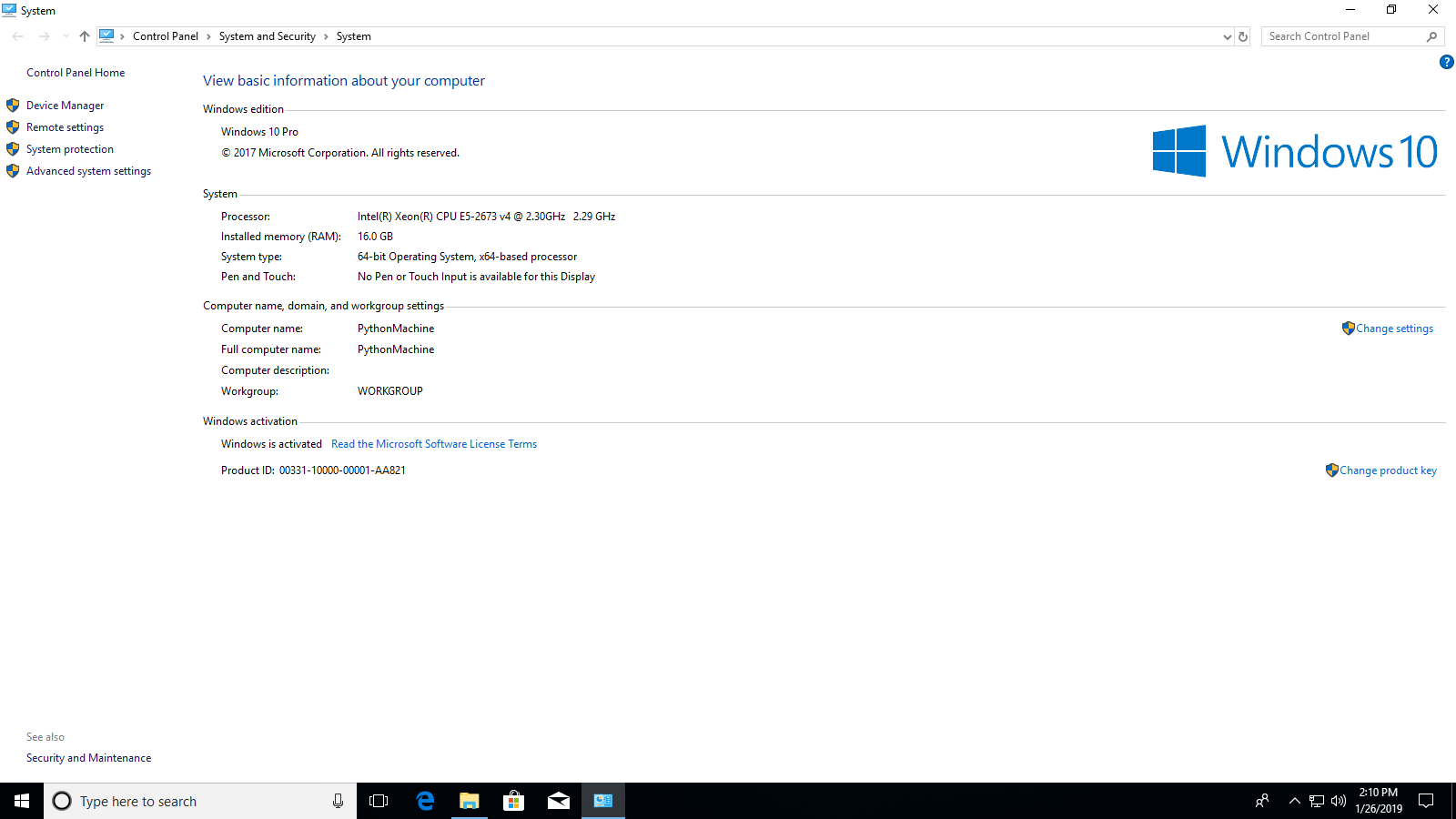
## The Environment Specification

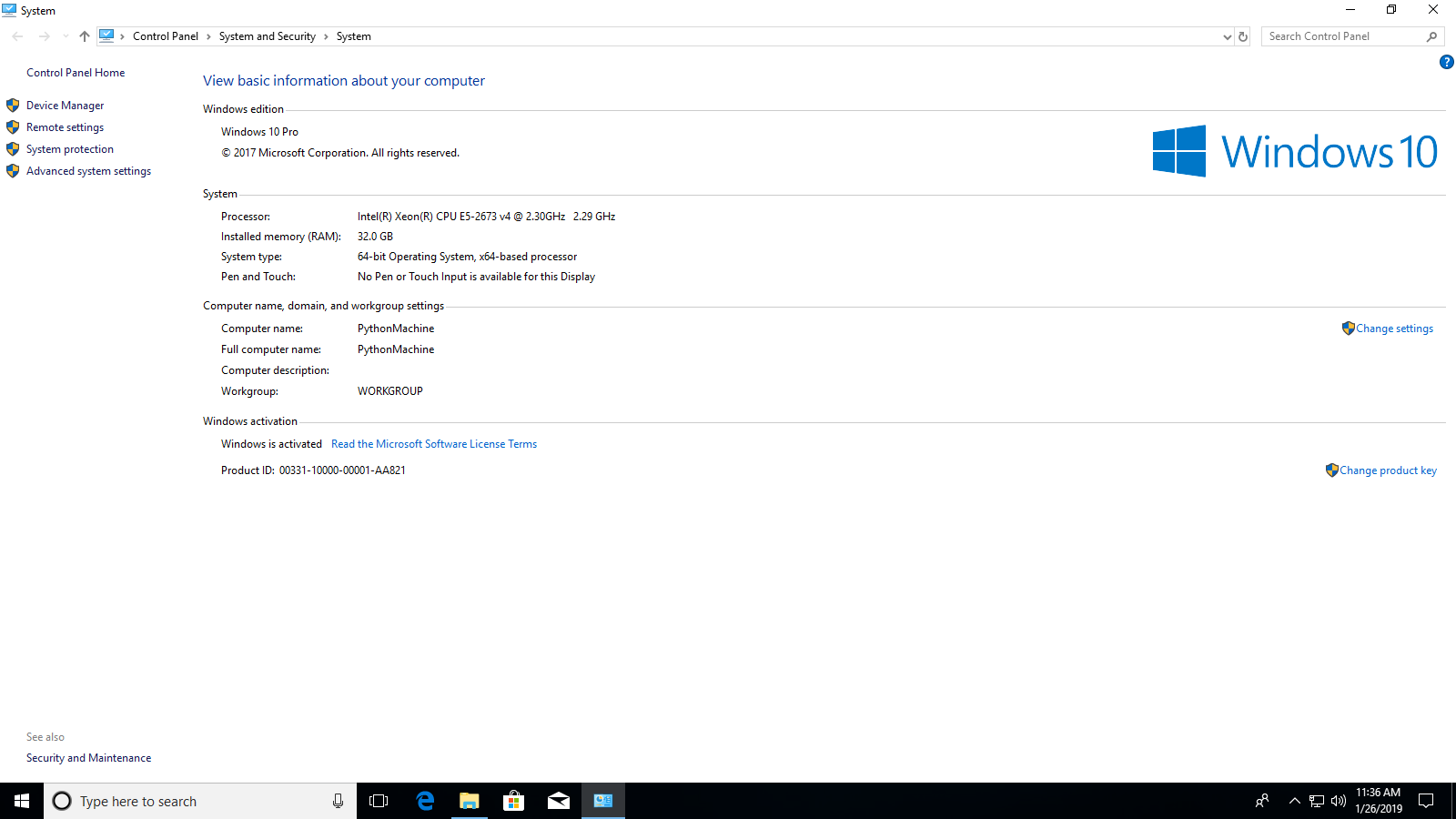
**Single Node**

* **Machine One (**B2s): 2 CPU, 4 GB RAM, 4 Data Disks and 1600 IOPS



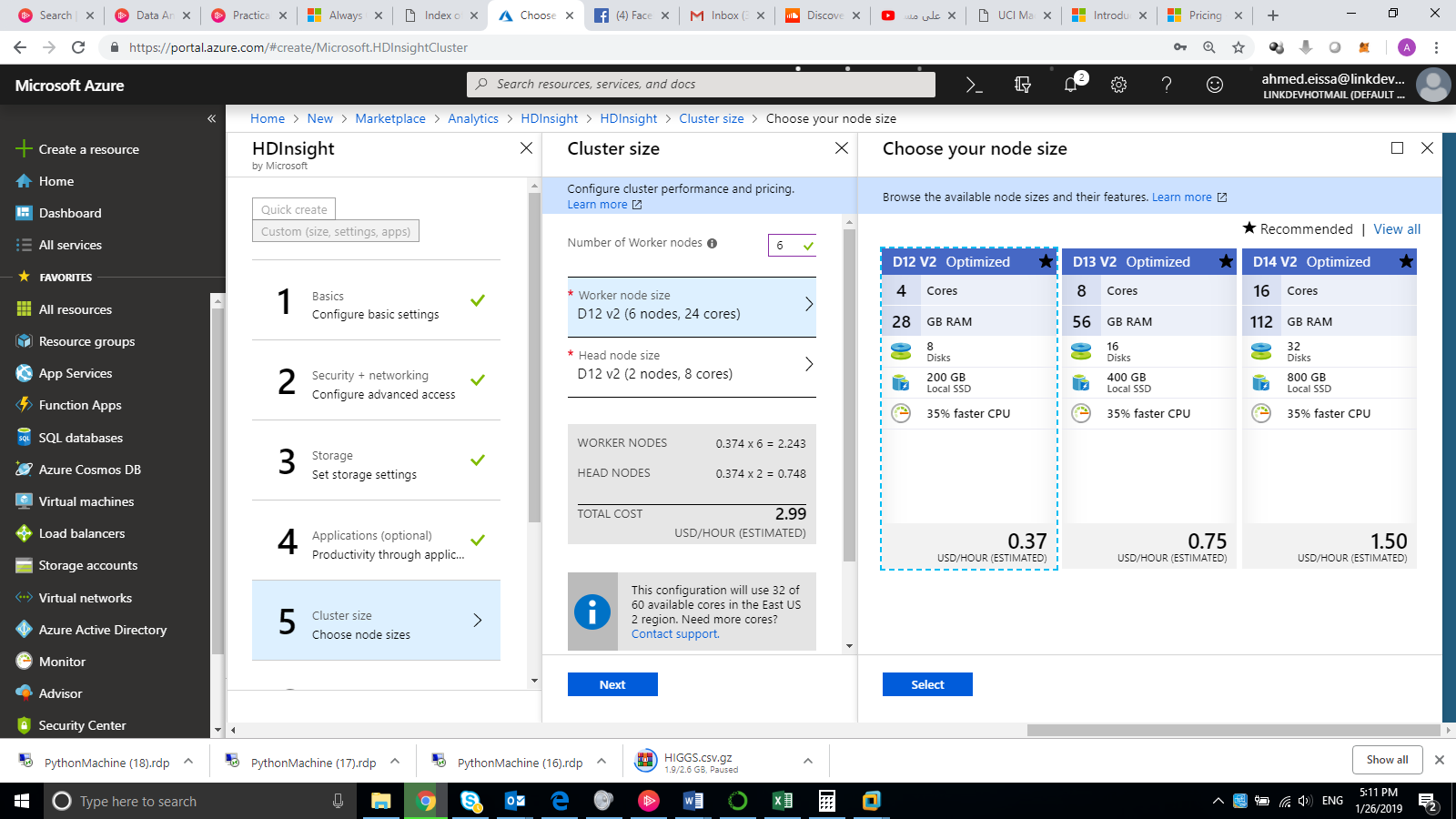
* **Machine Two (**B4ms): 4 CPU, 16 GB RAM, 8 Data Disks and 3600 IOPS



* **Machine Three (**B8ms): 8 CPU, 32 GB RAM, 16 Data Disks and 4320 IOPS

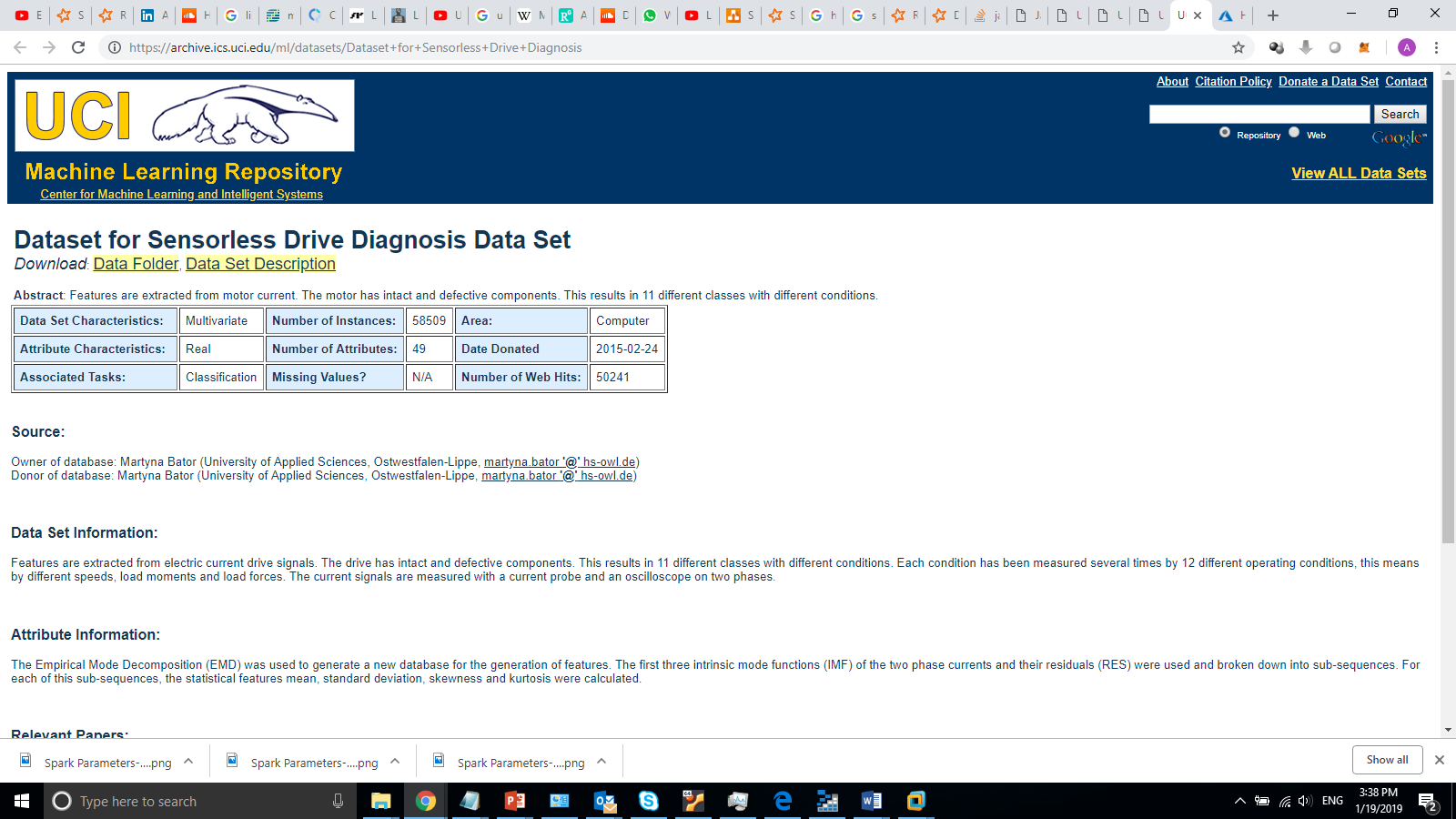
**Spark Cluster**

* **Hardware** 
  + Number of worker Nodes: 6
  + CPU: 4 core per worker node
  + Memory: 24 GB
  + Desk: 200 GB
* **Software**
  + OS: Linux
  + Spark Version: 2.3.0



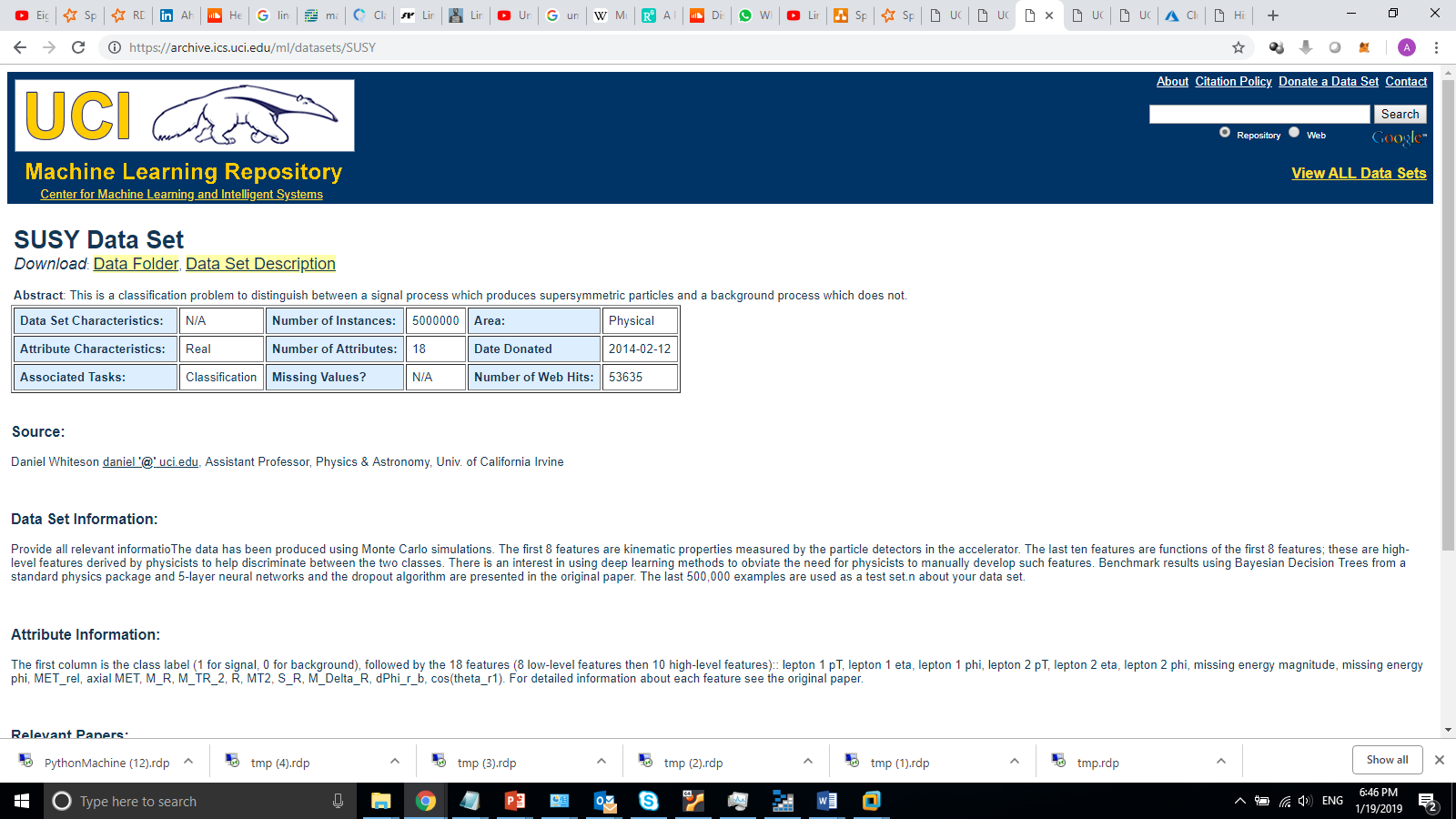
## The Experiment Datasets

* Sensor less Drive Diagnosis (25 MB)



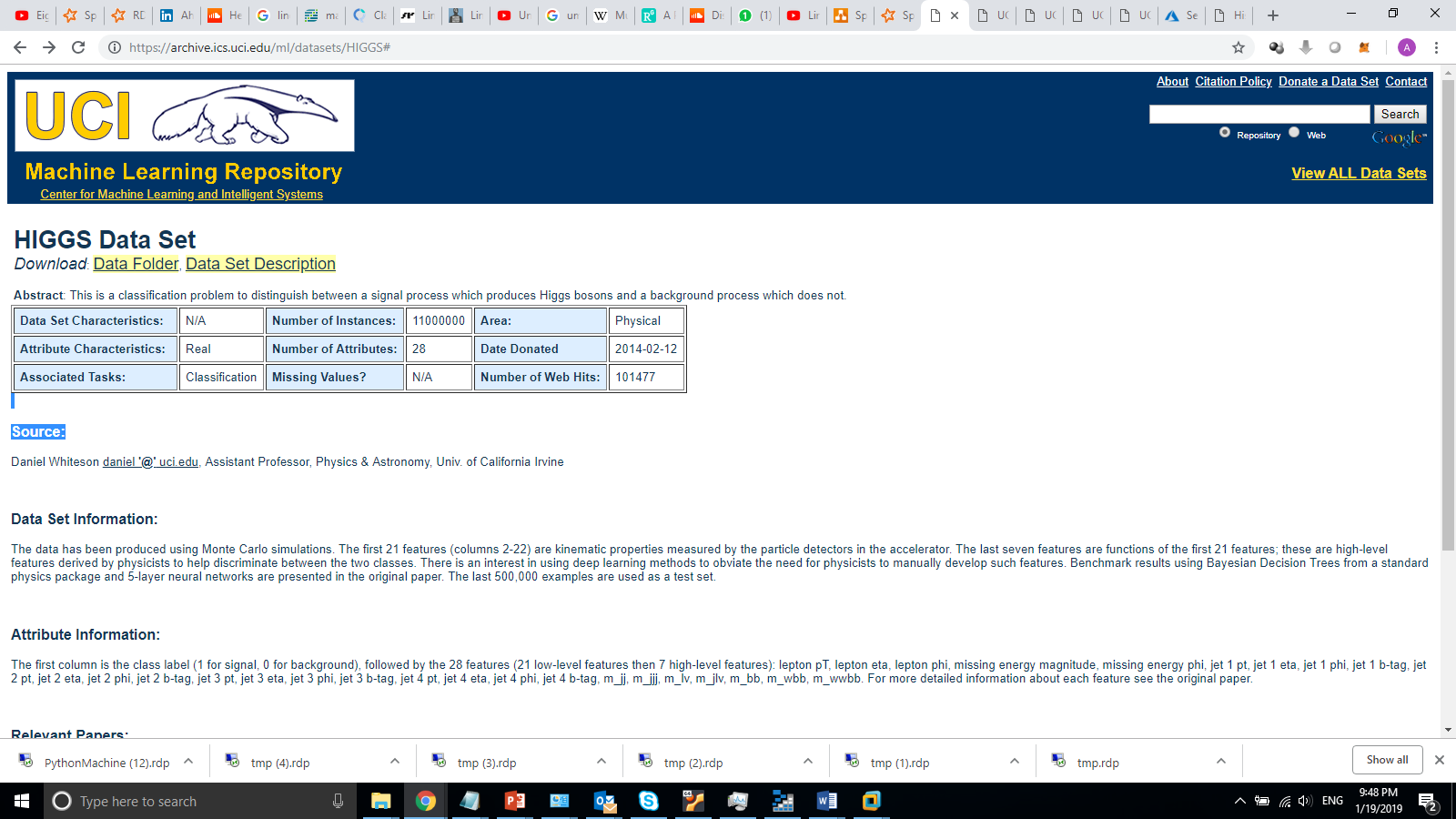
Ref: <https://archive.ics.uci.edu/ml/datasets/Dataset+for+Sensorless+Drive+Diagnosis>

* SUSY (2.3 GB)



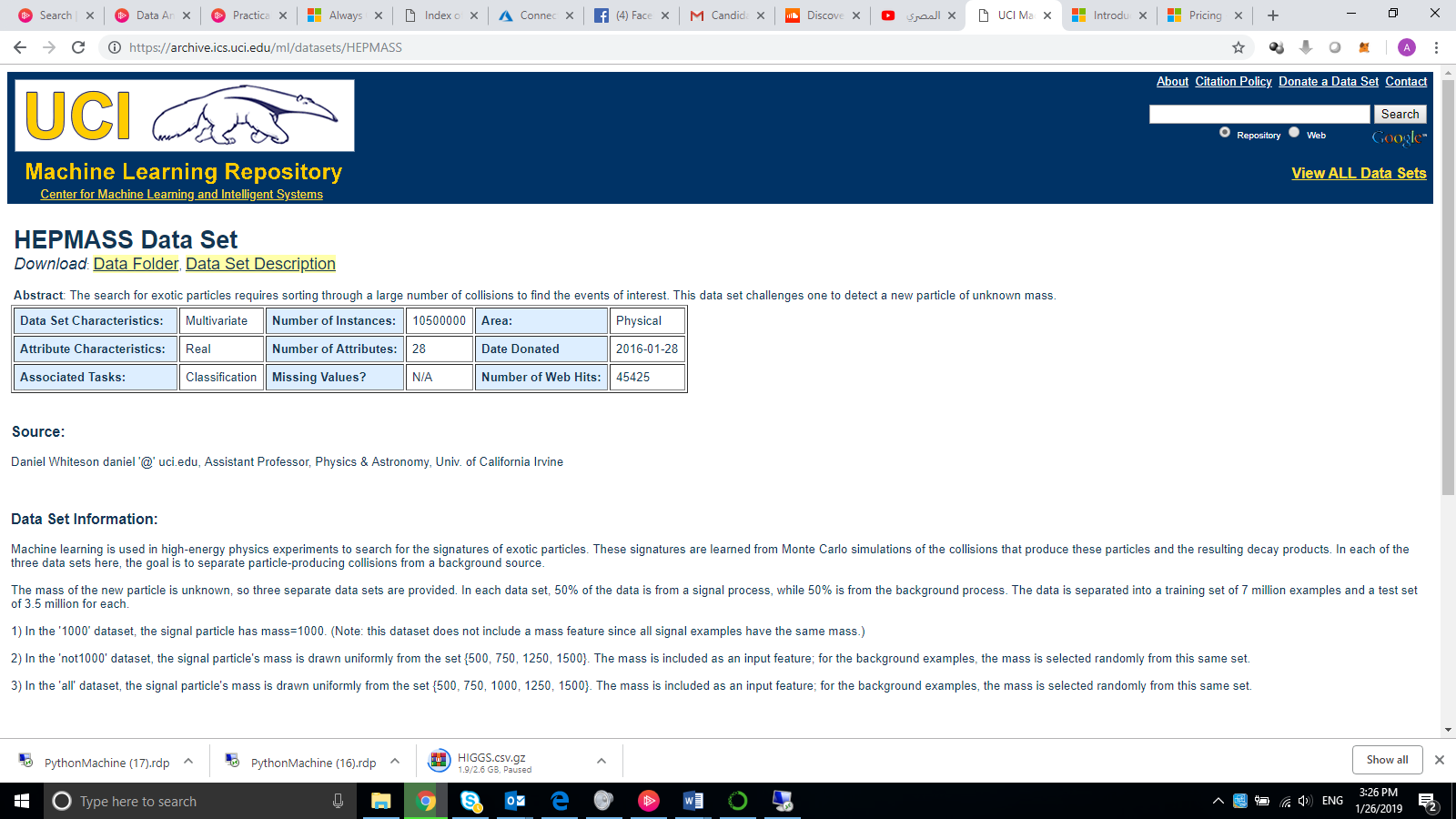
Ref: <https://archive.ics.uci.edu/ml/datasets/SUSY>

* HIGGS (7.8 GB)



Ref: <https://archive.ics.uci.edu/ml/datasets/HIGGS#>

* HEPMASS(7.5 GB)



Ref: <https://archive.ics.uci.edu/ml/datasets/HEPMASS>

## The Experiment Result

* Sensor less Drive Diagnosis (25 MB)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Machine One** | **Machine Two** | **Machine Three** | **Spark Cluster** |
| **Time** | 258 | 80 | 77 | 78 |
| **Accuracy** | 86.00% | 86.00% | 86.00% | 77.00% |

* SUSY

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Machine One** | **Machine Two** | **Machine Three** | **Spark Cluster** |
| **Time** | Memory Error | 393 | 392 | 90 |
| **Accuracy** | 76.00% | 76.00% | 76.00% |

* HIGGS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Machine One** | **Machine Two** | **Machine Three** | **Spark Cluster** |
| **Time** | Memory Error | Memory Error | 720 | 180 |
| **Accuracy** | 64.00% | 63.90% |

* HEPMASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Machine One** | **Machine Two** | **Machine Three** | **Spark Cluster** |
| **Time** | Memory Error | Memory Error | 737 | 240 |
| **Accuracy** | 84.00% | 83.80% |

## Conclusion

* In general, the algorithm can be distributed and learning process take less time when I scaled the environment horizontally (work on spark multiple worker nodes)
* Single machine will have a memory limitation during the calculation of Singular Value decomposition with big data (assuming I am able to read the data to the single machine memory)