

# Monitoring Spark

Proposed Architecture

Presenters :

Deepshi Garg [S4199456]

Pranav Vallala [S4210875]

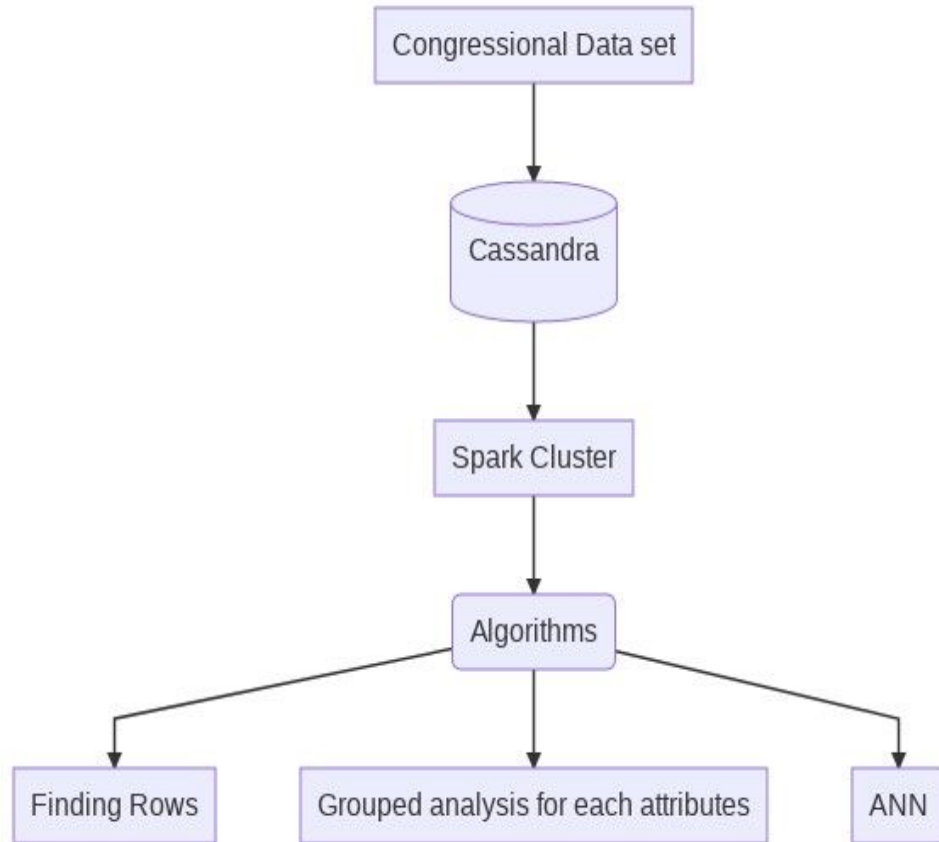
Vinayak Prasad [S4208110]

# Problem Statement

Your goal is to monitor the load on a Spark cluster and perform different types of profiling. Can you determine where bottlenecks are? How do different algorithms perform?

---

# Approach



- Deploy Apache Spark cluster over an out of box dataset
- Use Cassandra to store the data
- Performance of this cluster would be monitored over different algorithms to perform computations over this data

# Dataset

## Congressional Voting Results

Source :

Origin:

Congressional Quarterly Almanac, 98th Congress, 2nd session 1984, Volume XL: Congressional Quarterly Inc. Washington, D.C., 1985.

<https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records>

Citation:

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Attributes	Possible Outcomes
Class Name	(democrat/republican)
handicapped-infants	(Y/N)
water-project-cost-sharing	(Y/N)
adoption-of-the-budget-resolution	(Y/N)
physician-fee-freeze	(Y/N)
el-salvador-aid	(Y/N)
religious-groups-in-schools	(Y/N)
anti-satellite-test-ban	(Y/N)
aid-to-nicaraguan-contras	(Y/N)
mx-missile	(Y/N)
immigration	(Y/N)
synfuels-corporation-cutback	(Y/N)
education-spending	(Y/N)
superfund-right-to-sue	(Y/N)
crime	(Y/N)
duty-free-exports	(Y/N)
export-administration-act-south-africa	(Y/N)



Thank You!