Big Data for Actionable Intelligence (BDAI)

SAND2019-9156C

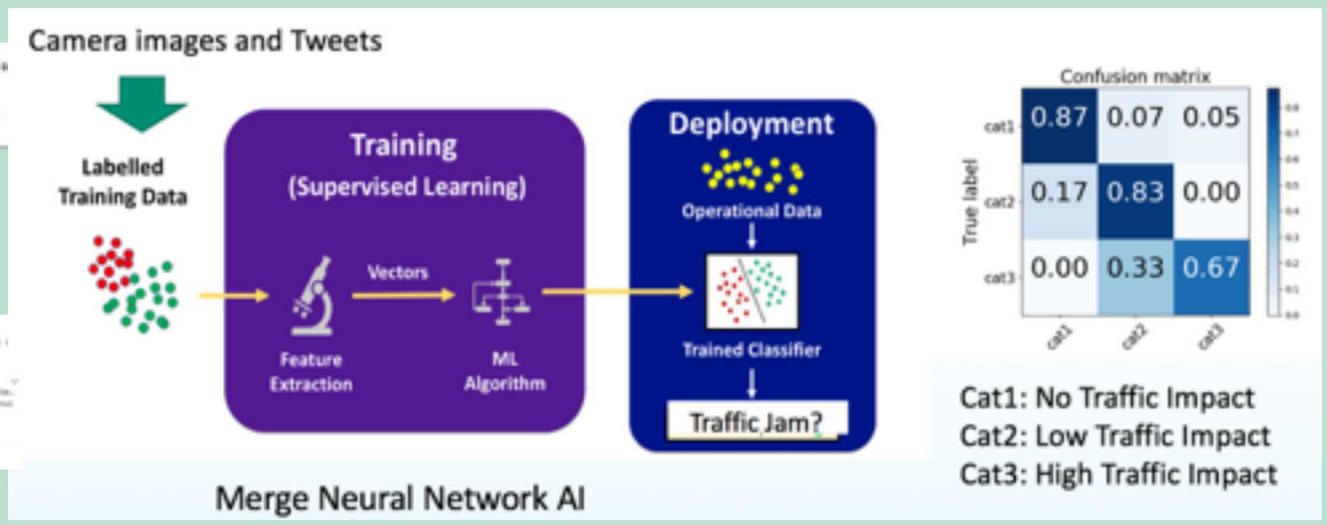
Motivation: Significant increase in amount of digital information

- "90% of the data in the world today has been created in the last two years alone, at 2.5 quintillion bytes a day" - IBM Marketing Cloud (Dec 2016)
- There are more data than humans can analyze

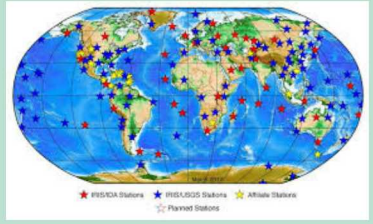
Goal: Perform an analytic assessment of technologies and algorithms around big geospatial data leveraging distributed analytical cluster platforms such as Apache Spark. Gain insight into how Spark and SparkDL run in different environments.

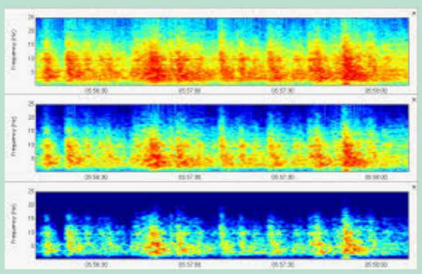
Two Exemplar Problems:

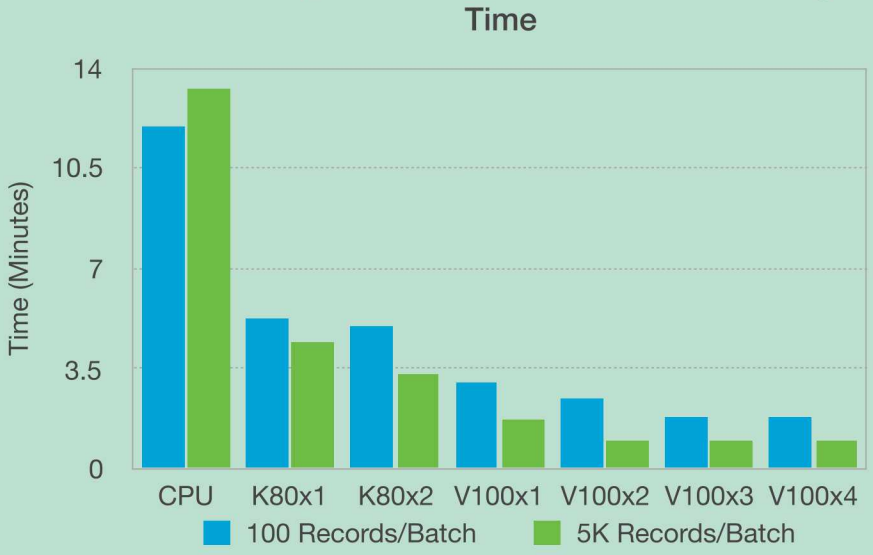
- Traffic Congestion Classifier
  - A data set containing images, sensor outputs, and metadata, with geospatial and temporal qualities, openly available Chicago traffic data
  - We can process about 250K events daily and have over 85 million events in our corpus

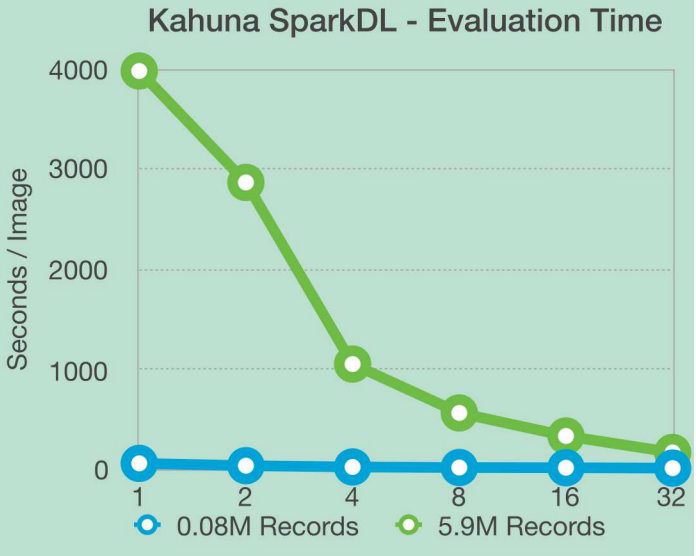


- Seismic Classifier
  - Streaming inference from current global seismic monitoring stations queried through IRIS
  - Apache Spark framework runs on Sandia-dedicated data analytics and ML hardware
    - HORTONWORKS (HWX) Data Platform running on Azure Stack (SNL Albuquerque)









Sandia Team Members:

Wes Brooks (6364), Forest Danford (6354), Tim Draelos (6362), Jenny Galasso (6364), Rudy Garcia (6332), Thushara Gunda (8825), Craig Hanna (6364), Jason Loyd (9368), Tian Ma (6321), Laura Patrizi (6354), Craig Ulmer (8753), Otto Venezuela (9368), Ben Ybarra (10779), Matthew Xie (8753)

Sandia National Laboratories is a multimillion laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525