Treating Machine Learning Data as Source Code .

Subscribe to our newsletter!
mobilemonitoringsolutions.com/subscribe

Database engineers and administrators for years have been focusing on backing up and restoring data on a singular point in time basis for transaction based systems as well as data warehouses. With the popularity of Machine Learning mobilemonitoringsolutions.com/statistical-machine-learning/, Artificial Intelligence mobilemonitoringsolutions.com/deep-learning/ Algorithms, and unstructured data now enable organizations to make business decisions in seconds rather than hours or days. A new requirement and paradigm shift is emerging. The ability to manage various snapshots of raw data, training data, as well as data used to develop, test, benchmark, finalize, and validate a machine learning model.

Various factors could influence a model and the Data Scientists need access to raw data in the past as well as snapshots of the data as well as other forms such as aggregation of various data sources at various times in order to validate their models presents a massive challenge and paradigm shift in existing practices.

So what is raw data vs training data? What different states of data exist in a Machine Learning operations environment?

- Raw data is the data that exists in its most basic form. For example, Log Files would be considered raw data. It is important to save the state of this data as it is the source data for Machine Learning algorithms.

- Aggregation data is data that is summarized from the same raw data or combined with other raw data sources and/or Training data to form a whole new data model for use in a machine learning model.

- Training data is data that is used as inputs into a Machine Learning model. This could take the form of elements of raw data as well as data aggregations from various sources combined to create a training data set. Training data must be saved in order rebuild a model in its state at time of deployment as well as for compliance. The preservation of training data contains many of the same characteristics as source code control that software developers use to manage their code base. In addition, raw data may be transformed into aggregations, extractions, and combinations that result in data used for training Machine learning models. This results in the creation of data in various forms that looks much different than the raw source data. Various versions of that same data may also be created for Machine Learning model testing and creation.

Mobile Monitoring Solutions

Machine Learning models require intensive data inputs from various sources and forms to determine a model outcome. The model needs to be constantly tuned as external factors such as real world scenarios and human behavior could require a model adjustment in real-time.

For example, you may want to create a 360-degree view of your customers to include social media as well as web site visits and purchasing history. Data snapshots of purchases as well as web browsing activity would be helpful in determining a set of circumstances at a given point of time that the model can be executed against to determine model validity and detect anyanomalies in either the model or the data.

This presents a challenge for traditional database administrators who are used to transactional based systems that at most require point in time recovery of structured data. New Backup/Restore Use Cases for Machine Learning require a combination of structured and unstructured data sets as well as various versions and transformations that need to be backed up by snapshots taken at points of time. This in effect creates a “versioning” of data, data at various points in time and forms. Everything from Raw data, Data aggregated before being trained and training data sets, subsets of raw data and their associated aggregations and transformations from various data sources. As you can see, this can get complicated fast and traditional backup solutions capabilities are being outstripped by the day at many organizations.

Mongodb www.mongodb.com/ is an emerging NoSQL Json based database platform that provides an automation and backup tool called “Ops Manager”. This stand-alone utility has backup functionality that can take snapshots of data at various points in time allowing the storage/backup/recovery of structured and un-structured data on demand and on a prescheduled basis.

Mongodb allows for both processing of structured and unstructured data sets as well as aggregations of them. It allows for data from many sources and forms and has the ability to add data sources easily. Mongodb horizontal scaling (Sharding) allows for large data volumes that support fast data operations. In addition, Mongodb’s NoSQL capability supports access to data as it “was” both in raw data form as well as Machine Learning model input training data. Ops Manager combined with Mongodb database engine meets many Machine Learning and AI operation requirements and the ability to re-create data and model states of production Machine Learning algorithms.

Machine Learning and AI implementations will grow in the coming years and data engineers and administrators will need to adapt to these new requirements and adopt a “DataOps” operations and mentality. This is like the “DevOps” concept that infrastructure administratiors have adopted with tools such as the Cloud, Puppet, Chef, and Ansible. The ability to react quickly and automate aspects of data engineering to meet the needs of an AI implementation.

For more information on Machine Learning and AI consulting services, please email raul@mobilemonitoringsolutions.com